REPUBLIC OF TURKEY
ÇUKUROVA UNIVERSITY
INSTITUTE OF SOCIAL SCIENCES
DEPARTMENT OF ENGLISH LANGUAGE TEACHING

A CORPUS-BASED STUDY ON TURKISH EFL LEARNERS' WRITTEN ENGLISH: THE USE OF ADVERBIAL CONNECTORS BY TURKISH LEARNERS

M. Pınar BABANOĞLU

A PHD DISSERTATION

ADANA, 2012

REPUBLIC OF TURKEY
ÇUKUROVA UNIVERSITY
INSTITUTE OF SOCIAL SCIENCES
DEPARTMENT OF ENGLISH LANGUAGE TEACHING

A CORPUS-BASED STUDY ON TURKISH EFL LEARNERS' WRITTEN ENGLISH: THE USE OF ADVERBIAL CONNECTORS BY TURKISH LEARNERS

M. Pınar BABANOĞLU

Supervisor: Asst. Prof. Dr. Cem CAN

A PHD DISSERTATION

ADANA, 2012

To the Directorship of the Institute of Social Sciences, Çukurova University

We certify that this dissertation is satisfactory for the award of the Degree of Doctor of Philosophy.

Supervisor: Asst. Prof. Dr. Cem CAN

Member of Examining Committee: Prof. Dr. Hatice SOFU

Member of Examining Committee: Assoc. Prof. Dr. Ahmet DOĞANAY

Member of Examining Committee: Asst. Prof. Dr. Hülya YUMRU

Member of Examining Committee: Asst. Prof. Dr. Hasan BEDİR

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

...../...... /2012

Prof. Dr. Azmi YALÇIN
Director of the Institute

Note: The uncited usage of the reports, charts, figures, photographs in this thesis, whether original or quoted from other sources, is subject to the law of Works of Art and Thought No: 5846.

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

ÖZET

YABANCI DİL OLARAK İNGİLİZE ÖĞRENEN TÜRK ÖĞRENENLERİN YAZILI ANLATIMLARINDA DERLEME DAYALI BİR ÇALIŞMA: TÜRK ÖĞRENCİLER TARAFINDAN ZARF BAĞLAÇLARIN KULLANIMI

M. Pınar BABANOĞLU

Doktora Tezi, İngiliz Dili Eğitimi Anabilim Dalı
Danışman: Yrd. Doç. Dr. Cem CAN
Haziran 2012, 202 Sayfa

Aradil, uzun zamandır ikinci/yabancı dil edinimi araştırmalarının önemli bir konusu olmuştur. Birincil amaç ikinci dil edinimi ve süreci ile ilgili daha iyi tanımlamalar yapmaktır. Öğrenen dili ile ilgili yeni bir düşünce olan Bilgisayarlı Aradil Derlemi (Computer Learner Corpus), ikinci dil edinimi alanında tanımlayıcı ve dikkate değer deneysel bir öğrenen veri kaynağı sunmaktadır (Granger, 2004). Aradil derlemi, yabancı dil öğrenenlerce üretilen dilden oluşturulan bilgisayar kaynaklı veri tabanıdır (Leech, 1998). Bu aradil derlemi, İngilizce öğrenenlerin dilbilgisi, sözcük düzlemi ve bir kompozisyon yazarken karşlaştıkları zorlukları araştırmak için öğrenenlerin yazılı ürünlerinden oluşan güvenilir bir veri sağlamaktadır. Aradil hakkında daha iyi bir anlayışa sahip olmak için aradil derlemi yoluyla aradil ile aradil araştırması üzerine kurulu birçok derleme dayalı çalışma yapılmıştır (Altenber & Tapper, 1998; Granger & Tyson, 1998; Aijmer, 2002; Housen; 2002; Neff et al., 2003; Narita et. al., 2004). Bu çalışmada, ikinci dil olarak İngilizce öğrenen Türk öğrenenlerin İngilizce metinlerinde ki zarf bağlaç kullanımı araştırılmıştır. Zarf bağlaçlar, bu kullanımın eğer varsa olası bir anadil aktarımından ve farklı anadil artlanlarından gelen öğrenenler arasında ortak aradil özelliklerinin bulunup bulunmadığı açısından incelenmiştir. Çalışmada, zarf bağlaç kullanımında farklı öğrenenler arasında bazı ortak aradil özelliklerine rastlanmıştır. Ayrıca, Türk öğrenenlerin zarf bağlaç kullanmlarında anadil aktarımı adına bazı anadil etkileri bulunmuştur.

Anahtar Kelimeler: Bilgisayarlı Aradil Derlemi, Zarf Bağlaçlar, Uluslararası İngilizce Öğrenen Derlemi, Uluslararası Türk Öğrenenler Derlemi.

ABSTRACT

A CORPUS-BASED STUDY ON TURKISH EFL LEARNERS' WRITTEN ENGLISH: THE USE OF ADVERBIAL CONNECTORS BY TURKISH LEARNERS

M. Pınar BABANOĞLU

Ph.D. Dissertation, English Language Teaching Department
Supervisor: Asst. Prof. Dr. Cem CAN
June 2012, 202 Pages

Investigation of interlanguage has long been an important subject of second and foreign language acquisition research. The primary goal is to provide better descriptions for SLA and its process. Computer Learner Corpus (CLC), which is a new way of thinking about learner language (Granger, 2004), offers a source of learner data suggesting empirical base for a remarkable and descriptive contributions in the field of SLA. Learner corpus is the computer texture database formed by the language produced by foreign language learners (Leech, 1998). This interlanguage corpora provides a reliable data of learners written production in order to examine the learner grammar and lexis and the main difficulties experienced by learners of English when writing an essay. Many corpus-based studies have been conducted on interlanguage investigation through learner corpora (Altenber & Tapper, 1998; Granger & Tyson, 1998; Aijmer, 2002; Housen; 2002; Neff et al., 2003; Narita et. al., 2004) to gain insight for a better understanding of learner language. In the present study, the use of adverbial connector in L2 writings of Turkish adult learners has been investigated. Adverbial connectors have been examined whether, such usage is effected by a possible transfer from mother tongue and there is a common interlanguage properties among learners from different mother tongue backgrounds. In the study, some common interlanguage properties among different EFL learners have been identified in use of adverbial connectors. In addition, some features in the use of adverbial connectors by Turkish EFL learners have been found in respect of L1 transfer.

Keywords: Computer Learner Corpus (CLC), Adverbial Connectors, Turkish International Corpus of Learner English (TICLE), International Corpus of Learner English (ICLE).

ACKNOWLEDGEMENTS

This dissertation represents a milestone in my educational and professional career. During the process of preparation and completing of this project, I have received valuable support anf help from a number of people to whom I owe thanks and would like to acknowledge.

First of all, I would like to express my deepest gratitude to my supervisor, Professor Cem CAN who has provided tremendious support and mentorship througout my doctoral time . His guidiance not only has led to the complation of this study, but also he fostered my academic development. My most sincere thanks go to him.

I would like to present my great thanks to Professor Hatice Sofu who has provided valuable support for writing this dissertation with her profound knowledge. Her insightful comments and constructive suggestions on each draft have facilitated the creation of this dissertation. I would also like to extend my great thanks to Professor Ahmet Doğanay for his invaluable comments and suggestions especially in statistical considerations of the dissertation. I am also very greatful to him for his patience and kindness. I would also like to thank to Professor Hülya Yumru and Professor Hasan Bedir for accepting to be a comitee member.

My special thanks go to Professor Mike Scott who provided stimulating comments and important suggestions for the dissertations during my research visit at Aston University. I am fortunate for receiving his co-supervision and great advices.

I am also thankful to my close friend Eliz Can for her valuable friendship and for encouraging me during the process with her generous kindness.

Finally, my special thanks go to my precious family member who deserve my special thanks most. I eternally gratitude to my dear mom Fatma Babanoğlu for her constant love, patience and invaluable support. Without her precious parentship, I do not belive that I can go on in my life and I owe her so much. I am also greatful to my dad for his strong supports and always being there. I truly thank to my brother Onur Babanoğlu for his valuable helps as well as to my brother Cumhur Babanoğlu for being there as my family member.

TABLE OF CONTENTS

Pages

CHAPTER I
INTRODUCTION

CHAPTER II
REVIEW OF RELATED LITERATURE

CHAPTER III
METHODOLOGY

CHAPTER IV
RESULTS AND DISCUSSION

CHAPTER V
CONLUSION

LIST OF ABBREVIATIONS

AC	: Adverbial Connectors
BOE	: Bank of English
BNC	: British National Corpus
CA	: Contrastive Analysis
CEA	: Computer-aided Error Analysis
CIA	: Contrastive Interlanguage Analysis
CL	: Corpus Linguistics
DDL	: Data-Driven Learning
EFL	: English as a Foreign Language
ELT	: English Language Teaching
ESL	: English as a Second Language
ICE	: International Corpus of English
ICLE	: International Corpus of Learner English
JPICLE	: Japanese International Corpus of Learner English
KWIC	: Key Word in Context
L1	: First (native) Language
L2	: Second Language
LC	: Learner Corpus
LL	: Log-likelihood
LOCNESS	: Louvain Corpus of Native English Essays
NL	: Native Language
NNS	: Non-Native Speakers
NS	: Native Speakers
SLA	: Second Language Acquisition
SPICLE	: Spanish International Corpus of Learner English
TICLE	: Turkish International Corpus of Learner English
TNC	: Turkish National Corpus
TL	: Target Language
TUC	: Turkish University Corpus

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER I

INTRODUCTION

1.1. General Background

1.1.1. Corpus Linguistics and Corpus Research

Studies of language can be divided into two main areas as studies of structure and studies of use (Biber, Conrad & Rippen, 1998). Traditional linguistic analyses have generally focused on structure aiming to identify units and classes of language (e.g. morphemes, words, phrases and parts of language). On the other hand, studies of emphasizing language use investigate actual language use in naturally occurring language productions. In particular, how speakers exploit the resources of their language rather than looking at what is theoretically possible in a language. Corpus linguistics includes the system of all methods and principles of how to apply corpora in language studies. Therefore, corpus linguistics is not restricted to a particular aspect of language use and it can be employed to explore almost any area of linguistic research (McEnery, Xiao & Tono, 2006). Corpus linguistics is a methodological basis for pursuing linguistic research than a separate paradigm within linguistics (Leech, 1992), or, in other words, it is a methodology rather than an aspect of language requiring explanation or description (Stubbs, 1996).

Corpus linguistics is basically described as a study of language or a linguistic methodology based on samples of ‘real life’ language use (McEnery & Wilson, 1996). ‘Real life’ language use can be explained with natural language data which forms a corpus. A corpus can be described as a body of occurring language, any collection of more than one text. Corpus term is originated from Latin word that means ‘body’ (plural *corpora*). However, corpus in modern linguistics tends to have more specifications. In linguistics, a corpus is a collection of texts (or ‘body’ of language) stored in an electronic database (Baker et al., 2006). According to Sinclair (1996), a corpus can be defined as a collection of pieces of language that are selected and ordered according to explicit criteria in order to be used as a sample of the language (in McEnery, Xiao & Tono, 2006) . Or in Meyer's terminology, it is “a collection of texts or parts of texts upon which some general linguistic analysis can be conducted” (Meyer, 2002, p. xi). A corpus could comprise written texts such as in the *The Brown Corpus* or spoken language data as in *The London-Lund Corpus*, or both written and spoken forms of language as in *The Bank of English* (BOE) or *British National Corpus (BNC)*.

In corpus linguistic studies, major focus is empirical, based on what is observed in corpus. McEnery &Wilson (2001) claims that an empirical research can be carried out by using any written or spoken text and such individual texts from the basis of many kinds of linguistics analysis. Therefore, it can be accepted that a corpus-based study may serve an empirical research basis for linguistics.

As has been noted, the primary aim of corpus linguistics is to provide accurate explanations for qualitative and quantitative descriptions of language use based on representative samples of natural usage. Important part of this description is the information about the distribution and the frequency of different forms and functions under different linguistic conditions (Sigley, 2006). The corpus is ''the only reliable source of evidence for such features as frequency'' (McEnery & Wilson, 2001, p.12). Frequency is an aspect of language that plays a major role to understand what is possible and what is likely to occur in a particular language (Granger, 2002). Accordingly, the computer corpus methodology and corpus-based techniques provide a wide suitability for conducting quantitative research opportunities like quantitative comparisons of a wide range of linguistic features in corpora representing different varieties of languages in texts.

1.1.2. Corpus Approaches in Language Studies

Empirical corpus data can be contributed to different fields of language centered studies under corpus-based linguistic research for example heterogeneous fields like lexicography, grammar, speech, semantics, pragmatics and discourse analysis, sociolinguistics, stylistics and language teaching, or domains such as studies of language variation, dialect, register, style or diachronic studies.

Grammar studies have been the most frequent types which have used corpora because corpus research serves a representative for the grammar of a whole language variety and empirical data for testing hypotheses of grammar theories (McEnery & Wilson, 2001). Especially, description of grammar has underwent a dramatic change with the development and improvement o corpus linguistics techniques. For example, in *Comprehensive Grammar of the English Language*, Quirk et al (1985) provided descriptions of structures with occasional mentions of corpus analysis of use in. other grammar resource is *The Longman Grammar of Spoken and Written English* in which Biber et al. (1999) presents a corpus based descriptions of structures with specific corpus results. In addition, many corpus-based grammar studies (Aarts, 1991, McCarthy and Carter, 1995, Greenbaum, Breivik, 1999, Leech, 1999) have provided descriptions of functions of specific grammatical features within written and spoken data.

In lexicography, Biber (1993a) examined collocations using corpus data and Bauer (1993) studied morphology in their corpus-based works. For speech research, Wilson (1989) investigated intonation by Lancester/IBM Spoken English Corpus, and Altenberg (1990) conducted a phonological study using London-Lund corpus. In Semantics, Mindt (1991) demonstrates how corpus can be utilized in order to provide objective criteria for assigning meanings to linguistic terms (cited in McEnery & Wilson, 2001).

Corpus-based research comprises a wide variety of studies in the linguistics subfield as pragmatics and discourse analysis (Stenström, 1987; Myers, 1989), in sociolinguistics (Holmes, 1994), in stylistics (Wikberg, 1992). Also language variation has been studied with several corpora (Biber, 1987, 1988; Lee, 2005).

The most remarkable contribution of corpus linguists is in the areas of language studies as first/second language acquisition, language teaching and language pedagogy. Numerous research studies have been conducted on second language learning and language pedagogy related to corpus linguistics (Kennedy, 1987; Holmes; 1988; Gavioli, 1997 Leech, 1997; Reppen, 2001). In teaching/learning foreign language, the importance of corpus is related to its significance in empirical study of large databases of language. Corpus data gives opportunity to conduct studies with more data and variables, and to design new kinds of classroom activities for learners to analyze the target language (Conrad, 2005). In addition, corpus linguistics serves real life language data in text books or dictionaries as represented in Collins COBUILD project (Collins COBUILD English Grammar, 1990) which was drawn from Bank of English Corpus (Sinclair, 1987). In the classroom, Johns (1994) suggests data-driven learning (DDL) in which corpus techniques are used in the classroom by learners like researchers. In first language acquisition, CHILDES (MacWhinney, 1996) has been developed which contains a corpus of transcriptions of children and parents.

In the field of second language acquisition (SLA), researchers have began to develop 'learner corpora' which contains written or spoken texts of second language (L2) learners last two decades. By learner corpora, researchers are able to use corpora from second language learners to describe and explore the linguistic patterns of L2 learners rather than relying on information from case studies and single examples (Reppen, 2006). One of the larger learner corpora is ICLE (Granger, 1998, 2002, 2009) (will be discussed next chapters) and other one is Longman's Learner Corpus.

The development of learner corpora has enhanced the corpus linguistics on SLA and language pedagogy (Partington, 1998; Flowerdew, 1998; Ringbom, 1998; Conrad, 1999; Biber and Reppen, 2002; Granger et al., 2002; Meunier, 2002; Granger, 2003). Reppen (2006) states that ''as more second language corpora are developed, they will become powerful resources for cross-linguistic comparisons of different first language speakers producing different target languages'' (p.249).

1.2. Research Background

Altenberg & Tapper (1998) states that ''effective communication requires coherence and clarity'' (p.80). One way of achieving this goal is to signal logical semantic relations between units of discourse via connectors like 'however' (to indicate contrast) or 'therefore (to indicate result)'.

English adverbial connectors or 'Conjuncts' connect linguistic units such as sentences, paragraphs and even larger parts of texts (Quirk et al., 1985, pp.631-632). Connectors provide coherence by signaling logical and semantic relations between units of a discourse and they help reader/listener to relate units each other to make sense (Altenberg & Tapper, 1998). They can consist of either one single adverb like 'nevertheless' or a prepositional phrase like 'for example'.

In this particular study, the choice of adverbial connectors is based on the list of semantic conjuncts classification in Quirk et al. (1985). The conjunct function entails a conjunct specific set of semantic relations. They are connected with, but are frequently rather remote from, the adverbial relation that is assumed in the speaker-related clause to which they correspond (Quirk et al., 1985). Seven roles of conjunctives described by Quirk et al. (1985) are as *Listing*, *Summative*, *Appositive*, *Resultive*, *Inferential*, *Contrastive* and *Transitional*. Connectors like **firstly** and **first of all** are considered as Enumerative; and therefore, so, and thus are as Resultive adverbials.

In *The Longman Grammar of Spoken and Written English,* Biber et al., (1999) presents the following 'real life' examples of English adverbial connectors from corpora containing academic prose, news or conversation:

Example 1.

Enumerative

This new structure must accomplish two special purposes. **First**, as a part of overcoming the division of Europe there must be an opportunity to overcome through peace and freedom the division of Berlin and Germany. **Second**, the architecture should reflect that American's remains linked to Europe. (NEWS)

(Biber et al., 1999, p.875)

Example 2.

Resultive

This year's commitment we will not reach this year. **Therefore**, we'll go into deficit! (CONV)

(Biber et al., 1999, p.877)

Example 3.

Contrastive/Concessive

They were economically active; **yet**, as the workshops were closed down one after another, they had few places to go to be active. (ACAD)

(Biber et al., 1999, p.879)

In Turkish, Connectors generally explained as conjunctions and discourse connectives. Considered structures in the study referred to discourse connectives that can be used for purposes of forming a cohesive link between concepts expressed by group of sentences (Goksel & Karslake, 2005). In the study, Turkish equivalent structures of English adverbial connectors have been tried to find out considering the linguistic similarities and translation. Thus, a list of corresponding items of adverbials in Turkish was formed categorically. For example:

English	**Turkish**
Listing/Enumerative	Listeme Bağlaçları/Sayıma Ait Bağlaçlar

First, Firstly	Önce, İlk önce

When the structures are examined within the sentences, it can be seen that these connectors provide a similar base for connection both in English and Turkish. For instance:

Example 4.

English

We have to fight against racism. **<u>First</u>/ <u>Firstly</u>**, the mentalities must change.

Example 5.

Turkish

Irkçılığa karşı mücadele etmeliyiz. **<u>Önce</u>/ <u>İlk</u>** önce zihniyetler değişmeli.

In the present study, all Turkish equivalents of English adverbial connectors have been identified for the corpora analysis. Turkish counterparts of English adverbials were examined as native reference of subject matter focus of the current study.

1.3. Statement of the Problem

This study attempts to investigate the use of adverbial connectors of Turkish learners in their argumentative essays. The aim is to examine similarities and/or differences between native speakers of English and English as a foreign language learners from various mother tongue backrounds, whether there are common interlanguage properties across EFL learners and the possible transfer errors stemming from Turkish learners' interlanguage affected by their L1.

The reason of the selection of adverbial connectors as the linguistic elements to investigate in this study is their importance for the coherence and the cohesion of the texts that learners should be aware of. The correct use of connectors is important for two reasons: explicit signaling of connections and rhetorical purpose in terms of indications of attitude and emphases (McCarthy and Carter, 1994). Cook (1989) states that "language learners need to know both how and when to use them. Their presence or absence in discourse often contributes to style, and some conjunctions can sound very pompous when used inappropriately" (cited in Tanko, 2004, p.154). However, a number of studies have shown that the use of connectors is problematic for foreign language learners (Altenber & Tapper, 1998). One reason is that connectors are not always used and that they have to be used with discrimination. The other problem is that the use of connectors is sensitive to discourse type which might cause difficulty for learners. And the last issue is that connector usage may vary across languages and not all languages mark connectors explicitly as in English (Altenberg & Tapper, 1998). The problematic usage of connectors are often expressed as under-, over- and misuse by learners. Therefore, the present study aims to explore the tendency of these possible problems which Turkish learners of English might face when using connectors.

1.4. Purpose of the Study

This particular corpus-based study focuses on investigating adverbial connectors in Turkish learners' written English and to examine whether there is a transfer from their mother tongue, or traces of interlanguage properties. More specifically, in the present study, following targets were aimed: (1) to provide a comprehensive explanation of the use of the patterns of adverbial connectors with various forms and functions in Turkish learner corpus in comparison to native English corpus and other English learner corpora; (2) to describe and explain the distinctive and recurrent usage of adverbial connectors, particularly overuse/underuse of specific types; (3) to examine the similarities and differences between different interlanguages to see possible common interlanguage properties.

1.5. Research Questions

This study will try to find answers to the following research questions:

1. Which Adverbial Connectors does TICLE corpus contain and how can they be classified?
2. Do Turkish learners use English adverbial connectors as native speakers in a statistically similar way?
3. How is the Turkish EFL learners' use of adverbial connectors different from Spanish and Japanese EFL learners?
4. What are the sources of divergences in TICLE corpus?
   a) Are there any signals of L1 transfer?
   b) Are these divergences a property of interlanguage?

1.6. Limitations

The present study is limited to the size of the four corpora and the results of the study are limited to the analysis of them; TICLE as a learner corpus of Turkish EFL learners, LOCNESS as native English corpus, SPICLE as a learner corpus of Spanish EFL learners, JPICLE as corpus of Japanese EFL learners and TUC as a native Turkish reference corpus. In addition, the study has been carried out by limiting its scope to overuse and underuse of adverbial connectors by EFL learners; misuse of the structures have not been included in the analysis.

1.7. Operational Definitions

Adverbial Connectors: Single or multi-word units that signal connections between discourse segments and establish various discourse relations. They conjoin linguistic units such as sentences, paragraphs, or even larger discourse (Quirk et. al., 1985, p. 631-632)

Computer Learner Corpus (CLC): Electronic collection of authentic texts produced by foreign or second language learners. (McEnery and Wilson, 2001, p.177)

Concordance: 'A comprehensive listing of a given item in a corpus (most often a word or phrase), also showing its immediate context'' (McEnery & Wilson, 2001, p.197)

Contrastive Interlanguage Analysis (CIA): A method involves comparisons between native speakers and learners (L1 vs. L2), and between different learner groups (L2 vs. L2) (Granger, 1996).

Corpus: ''A corpus is a collection of texts (a 'body' of language) stored in an electronic database. Corpora are large bodies of machine readable texts containing thousands or millions of words'' (Baker, et. al., 2006, p.48).

Corpus Linguistics (CL): ‘‘A linguistics methodology which is founded on the use of electronic collections of naturally occurring texts, viz. corpora’’ (Granger, 2002, p.4).

English as a Foreign Language (EFL): Use or study of English by speakers of different native languages.

First Language (L1): Language that is acquired in early childhood, i.e., mother tongue, native language.

Interlanguage (IL): The language system of a second language learner at any stage in the process of second language acquisition (Gass & Selinker, 2001).

International Corpus of Learner English (ICLE): A learner corpus containing of argumentative essays written in English by learners from 16 different mother tongue backgrounds (Granger, 2009).

Japanese International Corpus of Learner English (JPICLE): A learner corpus containing the Japanese EFL learners’ written argumentative essays (Granger, 2009).

Key Word in Context (KWIC): A type of display of concordance in which the key (node) is centered and framed by the words occurring left and right of it (Baker et. al., 2006).

Learner Corpus (LC): A corpus containing written or spoken texts of second language (L2) learners.

Louvain Corpus of Native English Essays (LOCNESS): A native reference corpus containing British and Americn Students’ written essays (Granger, 2009).

Second Language Acquisition (SLA): The acquisition of a language after the native language has already become established in the individual (Ritchie & Bathia, 1996).

Second Language (L2): Any language that is acquired after the native language.

Spanish International Corpus of Learner English (SPICLE): A learner corpus containing the Spanish EFL learners' written argumentative essays.

Turkish International Corpus of Learner English (TICLE): A learner corpus containing the Turkish EFL learners' written argumentative essays (Granger, 2009).

Turkish University Corpus (TUC): A native reference corpus containing Turkish University Students' written essays.

1.8. Overview of the Thesis

The present dissertation consists of five chapters which organized as follows: introduction, review of related literature, methodology, results and discussion and conclusion.

In the first chapter, brief information about corpus field within general background is introduced as an introduction. Main focus of the study is initiated in research background, and then explained by purpose of the study and research questions. Finally, limitations, operational definitions and the summary of the first chapter sections are presented.

Second chapter presents a detailed historical and theoretical review of the literature which our study is based on. After a chronological historical background, key concepts and fundamentals of the corpus field are described in detail. As the major concern of the study, adverbial connectors are examined in four languages (English, Turkish, Japanese and Spanish) in order to support the linguistic background. Finally, the previous research on adverbial connectors related to interlanguage are explained as well as major issues of connectors in EFL.

Third chapter provides the methodological base of the present study. Information about the main methodology which the study was conducted on (CIA), the corpora used in the analysis (LOCNESS, TICLE, SPICLE, JPICLE and TUC), the software of used in data processing and the statistical method used in data analysis are given in detail.

Fourth chapter includes data analysis obtained from frequency and statistical processes. The results of the analysis are discussed by regarding the methodological and theoretical research background of the present study.

Fifth chapter presents the conclusion section in which the research questions of the study are discussed in relation to the results given in the previous chapter. This last chapter also proposes some the suggestions for future studies and implications for ELT.

1.9. Chapter Summary

In the first chapter, a brief introduction of corpus linguistics is given within a general outline of corpus linguistics and corpus approaches in language studies. In research background, Adverbial connectors are briefly illustrated in English and Turkish languages. Lastly, research questions, limitations and operational definitions of the study are presented.

CHAPTER II

REVIEW OF RELATED LITERATURE

2.0. Introduction

This chapter includes the historical and theoretical assumptions of corpus field. Firstly, the origins of corpus and corpus linguistics are reviewed in the perspective of their developmental progress. Then the fundamentals and key concepts in corpus linguistics are described as well as the related terminology and the research conducted in learner corpus and adverbial connectors.

2.1. Historical Overview of Corpus Linguistics

Despite the term 'Corpus Linguistics' first appeared in 1980s (Leech, 1992), corpus based language research has a longer and substantial history (McEnery, Xiao & Tono, 2006). Actually, corpus research dates back to thirteenth century when the first primitive samples of corpora in Bible concordances began to be used (Meyer, 2008). Then the inquiry evolved over time with an increasing trend. The considerable factor in the development of corpus is technology because the research has gained acceleration as the computer technology improved in time. The revival of corpus research has fallen into after the 1950s and then the progress continued through 1980s and present day. A chronological outline of corpus linguistics is presented below in Figure 1.:

Thirteenth - Nineteenth Centuries	Pre-electronic Period First samples: Biblical Concordances Biblical Concordances
⇩	
Nineteenth Century-1950s	Pre-electronic Period Paper-based data Studies on language acquisition/pedagogy Survey of English (SEU)
⇩	
1950s-1980s	Chomsky Revolution Machine-Readable Texts Early Modern Corpora: Brown Corpus / LOB Corpus
⇩	
1980s-Present	Advanced Computer Technology Bank of English by COBUILD British National Corpus (BNC)

Figure1. Timeline of corpus linguistics

The corpus trend started by Biblical concordances at thirteenth century continued with literary texts, grammar and dictionary compilations over eighteenth and nineteenth centuries. In the period between nineteenth century and 1950s, studies on language acquisition, language pedagogy, syntax and semantics have been the first examples based on the compilation of paper based data. Before 1950s, early corpus linguistics was affected by the descriptivism and comparative linguistics. The 1950s was the most significant decade for both corpus linguistics and linguistics itself as a field. Chomsky, on studies within the framework of Generative Linguistics, changed the direction of linguistics and caused a shift from empiricism to rationalism. Under transformational grammar, Chomsky (1957) argued in favor of competence rather performance for modeling the language therefore he criticized the nature of corpus as a source of evidence in linguistics inquiry. In 1960s, early machine readable corpora were formed and first computer-generated concordances with punched-card storage technique (Parrish, 1962) and then KWIC (Key Word in Context) (Hines, et.al, 1970) appeared. 1980 and onwards, considerable developments in computers and network technology led substantial improvements in corpus linguistics. Projects as COBUILD and BNC were the signs of the revival of corpus linguistics. Starvik (2007) claims that "While it is natural today to take 'corpus linguistics' to mean 'electronic corpus linguistics', we must not forget that there were language corpora BC, i.e. 'before computers'" (p.12).

2.1.1. Early Corpora and Pre-electronic Period: Thirteenth Century –1950s

Current corpus linguistics is associated with concordance lines and wordlists generated by modern computer software to analyze texts. Long before todays' computerized era, corpus history has a long pre-electronic period which had started at thirteenth century with biblical concordances, continued in eighteenth, nineteenth and mid-twentieth centuries with grammar and dictionary works and lasted until late 1950s with first electronic corpora samples.

The corpus phenomena had started at thirteenth century when Bible words manually indexed line by line and page by page. The aim was to simplify the arranging words in Bible in an alphabetical order with citations of where and in what passages they occurred (McCarthy and O'Keeffe, 2010). Kennedy (1998) states that Biblical concordances represent ''the first significant pieces of corpus-based research with linguistic associations...'' (in Meyer, 2008, p.19). Antony of Padua (1195-1231) is associated with the first known (anonymous) Bible concordance 'Concordantiae Morales' based on Vulgate (the fifth century Bible in Latin). Around the same dates, in 1230, Cardinal Hugo, by the help of 500 monks, formed word index of Bible in Latin. Many others followed the concordance of religious texts; Hebrew and English concordances in fifteenth century, and in eighteenth, Crudens' '*Complete Concordance to the Holy Scriptures'* in 1787 and Strong's '*Exhaustive Concordance of the Bible'* in 1890 (McCarthy and O'Keeffe, 2010). Cruden's was the most comprehensive one which consists of 2,370,000 words longer than Bible itself and took two years to write.

As a means of concordance in literary texts, Becket's 'A Concordance to Shakespeare' in 1787 can be considered as a source of former corpus work done manually in 18th century. In grammar, Lowth (1762) used corpus examples in his work 'Short Introduction to English Grammar'. This trend influendced many subsequent grammarians and linguists such as George Curme, Otto Jaspersen and Charles Fries in descriptively oriented grammar across nineteenth and early mid-twentieth centuries (Meyer, 2008). Jespersen's (1909-1949) '*A Modern English Grammar on Historical Principles*' based on examples from an extensive collection of texts consist of hundreds of books, essays and poems (Meyer, 2002). For instance, to show the widespread usage, Jespersen illustrates indefinite pronoun from extensive lists as:

Example 6.
God send **euery one their harts** desire. (Shakespeare, Much Do About Nothing, III, 4.60, 1623) (adapted from Meyer, 2008, p.5)

Example 7.
Each had **their** favorite. (Jane Austen, Mansfield Park, 1814)
(adapted from Meyer, 2008, p.5)

As a pioneer in the corpus linguistics field, Jespersen is one of the linguists who used the authentic corpus data recorded on slips of papers. In his autobiography, Jespersen (1938) points out that:

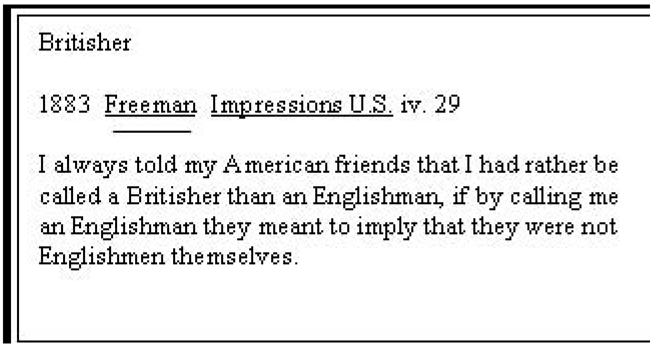
> I'm above all an observer. I quite simply cannot help making linguistic observations. In conversations at home and abroad, in railway compartments, when people passing in streets or roads, I am constantly noticing oddities of pronunciation, forms and sentence constructions- but more in my younger days than now when much of what was then striking is familiar to me… For these notes I have found in practical to use small slips of paper…It is impossible for me to put even a remotely accurate number on the quantity of slips I have had or still have: a lot of them have been printed in my books, particularly the four volumes of *Modern English Grammar*, but at least just as many were scrapped when the books were drafted, and I still have a considerable number of drawers filled with unused material. I think a total of 3-400.000, will hardly be an exaggeration.''
>
> (1938, cited in Starvik, 1992, p.7)

This old manual data formation of corpus-based study has a long honorable tradition in linguistics. Jespersen's methodology improved by Charles Carpenter Fries (1957, in Meyer, 2008) who provides grammatical descriptions in '*The Structure of English*':

> a large body of actual English speech observed and recorded in a university community in the North-Central part of the United States [...] The materials which furnished the linguistic evidence for the analysis and discussions of the book were primarily some fifty hours of mechanically recorded conversations on a great range of topics – conversations in which the participants were entirely unaware that their speech was being recorded. (Fries 1951, p. 3, cited in Starvik, 2007, p.13)

Fries is the first grammarian who used spoken texts besides written texts as a source of data for his grammar, and also frequency information taken from his corpus to discover the common and uncommon patterns (Meyer, 2008). Fries' empirical and behaviorist approaches in methodology was rejected during Chomsky Revolution in late 1950s.

Corpora have long tradition lexicography and dictionaries as well. In 1775 Johnson used illustrative quotations in '*Dictionary of English'*. This study influenced letter lexicographers who prepared the largest dictionary ever published, 'Oxford *English Dictionary*' (OED) in 1859, in which was included every word in English language from 1250 to 1858. The first edition of OED was published in 1928, it took fifty years to complete and it consists of nearly five million citations slips (Meyer, 2008). OED is the most famous example of 'corpus of slips on paper'. A citation slip from OED is presented in Figure 2 below:



Britisher

1883 Freeman Impressions U.S. iv. 29

I always told my American friends that I had rather be called a Britisher than an Englishman, if by calling me an Englishman they meant to imply that they were not Englishmen themselves.

Figure 2. A Citation slip from the OED (adapted from Meyer, 2008, p.9)

Dictionary of English and OED are the first dictionaries based upon pre-electronic corpora in eighteenth and twentieth centuries. However, the first most significant and influential pre-electronic corpus was the Survey of English Usage (SEU) corpus whose compilation began in 1959 by Randolph Quirk. SEU contains nearly 1 million words in written and spoken texts which were collected with paper slips.

Between eighteenth - mid-twentieth centuries, early corpora have examples can be found in a variety of linguistic fields. In language acquisition, the studies of child language (1876-1926) which carried out by parental diaries of child's locations can be considered as first pre-electronic corpus samples in the field. In spelling conventions, Käding (1897) used a large corpus of German (11 million of words) to collocate frequency distributions of letters and sequences of letters in German (McEnery and Wilson, 1996, p.2). On the other hand, Fries and Traver (1940) and Bongers (1947), Thorndike (1921) and Palmer (1933) used corpus in foreign language pedagogy. In comparative linguistics, Eaton's (1940) compared word frequencies in German, French, Italian and Dutch (in McEnery & Wilson, 1996).

In summary, early biblical and literal works provides a background for word searching and indexing. Leech (1992) claims that 1950s was the era that American structuralists as Fries, Harris and Hill and others are the forerunners of corpus research in terms of real data and data gathering. During 1950s, although there have been a lot of criticisms by generative linguists, essential advances in corpus linguistics were made.

2.1.2. Generative Grammar vs. Corpus Linguistics

Basic corpus methodology have proceeded in its classical route that formed by empiricism over years until 1950s. McEnery & Wilson (1996) point out that corpus methodology was widespread in linguistics in the early 20th century, however, after the late 1950s, corpus as a source of data underwent a period of unpopularity and rejection. In this period, Chomsky (1957, 1965) influenced the linguistics field with generative grammar inquiry and changed the direction of linguistics from empiricism to rationalism. Rationalist language theories based on the development of a theory in mind whereas empiricist approach relies on observation of naturally occurring data as in corpus methodology. According to Chomsky, linguists should model language competence (I-language) rather than performance (E-language). In addition, he invalidated the corpus as a source of evidence and suggested that corpus could never be a useful tool for the linguists, as the linguists must seek to model language rather than performance.

During the formative years in the 1950s, Chomsky himself were more active than others in developing transformational-generative theory. Therefore, an analysis of the relation between early generative linguistics and corpus linguistics is largely a study of the development of Chomsky's methodological practices, especially of how he used corpus observation methods, native speaker intuitions, and the linguist's own intuitions (Karlsson, 2008).

When Chomsky entered the field of linguistics, the immediate linguistic atmosphere he faced was that of North American structuralism, the era was in the dominance of American structuralists as Harris, Hill and Fries. Key concepts of language and linguistics were reliance on corpora as the starting point of linguistic analysis, emphasis on description rather than on theory formulation, inductivistic discovery procedures, classification of elements, separation of levels in the grammar, insistence on biuniqueness of phonemic transcriptions, physicalistic concept formation, and non-mentalism manifested especially as an aversion for semantics (Karlsson, 2008). When this approach was taken to its extremes, a grammar of a particular language was considered to be an inventory of elements (phonemes, morphemes, constructions, etc.), and linguistics was basically conceived as a classificatory type of scholarship.

On the other hand, in transformational-generative grammar, it is important to keep in mind that the following three types of phenomena are ontologically distinct: (i) language data in the form of sentences (utterances), (ii) the mentally represented competence of the native speaker-hearers' grammatical intuitions (tacit knowledge of the language), and (iii) the spatio-temporal performance processes underlying speaker-hearers' speaking and understanding. *Language data* (i) are accessible by observation, i.e. corpus work done for example by authors of comprehensive reference grammars, and elicitation, typically conducted by a field linguist working with an informant, both backed up by introspection in order to ensure that the language specimens so obtained are indeed grammatical. *Competence* (ii) is accessible by introspection, elicitation, experimental testing, and indirectly by observation of language data. *Performance* processes (iii) are accessible by observation of language data and by experimental testing, both guided by introspective consultation of competence.

Meyer (2008) states that corpus linguistics and generative grammar have had an uneasy relationship because they have different goals. According to Chomsky, there are three types of adequacy of that linguistic claims can meet; observational, descriptive and explanatory. The major conflict between a generative grammar and a corpus linguist reveals as; while the generative grammarian relies on explanatory adequacy (the highest level of adequacy, according to Chomsky); on the other hand, the corpus linguist aims for descriptive adequacy. Indeed the explanatory adequacy is arguable whether it can be achievable through corpus analysis. In generative grammar, the highest level of adequacy is explanatory adequacy, which is achieved when the description or a theory reaches both descriptive adequacy and abstract principles of *Universal Grammar* (UG). Since generative grammar has placed so much emphasis on UG, explanatory adequacy has always been the priority. Therefore, as for descriptive adequacy, there has never been so much emphasis in generative grammar on data based on representative of the language and also language variation (Meyer, 2008).

Chomsky believed that main task of a linguist should be the definition of a model of linguistics competence so that it is hard for corpus linguistics to achieve this goal since it relies on performance data (McEnery and Wilson, 1996). Competence is our tacit knowledge, i.e. internalized knowledge of a language; on the other hand, performance is external evidence of language competence and its usage on particular occasions. Chomsky argued that it is our competence rather than performance that a linguist should model and its competence which both explains and characterizes a speaker's knowledge of language. Performance is a poor mirror of competence and may be influenced by factors other than competence as short-time memory or drinking. Therefore, Chomsky states that since corpus data is a collection of externalized utterances and it is performance data, it must be a poor modeling of linguistic competence. However, Leech (1992) argues that this characterization is overstated: the distinction between competence and performance is not as great as it is often emphasized, ''since the latter is the product of former'' (1992, p.108). In addition, a corpus can be used as a basis for any theoretical issue and indeed it serves excellent source for verifying the falsibility, completeness, simplicity, strength, and objectivity of any linguistic hypothesis (Leech, 1992).

Another issue about corpus itself was insufficient in explaining the infinity of natural of a language. The number of sentences in a natural language is potentially infinite and some of the rules are recursive. Recursion expresses the repeating which describes the infinity of sentences. For example, following phrase structure rules include recursion:

Phrase Structure Rules:

S → NP VP
NP → AT N
NP → AT N PP
PP → Prep NP
VP → V JP
JP → J

Figure 3. Phrase structure rules of recursion (McEnery a& Wilson, 1996, p.14).

In this set of rules above, the second NP rule and the sole PP rule refer to one another. That is, there could be an infinite number of prepositional phrases enclosing an infinite number of noun phrases within a sentence according to these rules. These rules may give infinite sentences as in the following example:

Example 8.

The dog of the man (one recursion) was old.

S → NP VP
NP → AT N PP
PP → Prep NP
VP → V JP
JP → J

The dog of the man from the house (two recursions) was old.

S → NP VP
NP → AT N PP
PP → Prep NP
NP → AT N PP
PP → Prep NP
NP → AT N
VP → V JP
JP → J

The dog of ….(infinitely many recursions) was old.

S → NP VP
NP → AT N PP (infinitely many recursions start here)
PP → Prep NP
NP → AT N
VP → V JP
JP → J

(adapted from McEnery & Wilson, 1996, p. 8)

Here, there is certain circularity in the phrase structure of English. This recursive nature of phrase structure rules shows that sentences in a natural language are infinite. Accordingly, the argument is that the corpus could never describe this syntactic competence since it is a performance data. Language is non-enumerable and corpora itself is in complete in nature so that no finite corpus can represent a language, i.e., corpora are 'skewed' (McEnery and Wilson, 1996). Chomsky (1959) argues that:

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others are impolite. The corpus, if natural, will be so wildly skewed that the description [based upon it] would be no more than a mere list.
>
> (1959, cited in McEnery & Wilson, 1996, p.8)

That is to say, corpora are partial in two senses; first, they are incomplete; they will contain some but not all valid sentences in a language. Second, they are partial since they are 'skewed'; with frequency of a feature in the language being a determiner of inclusion (McEnery and Wilson, 1996). Chomsky himself stated the sentence ***I live in New York*** is fundamentally more likely than ***I live in Dayton Ohio*** to show the fact that there are more people likely to say the former than the letter. As a matter of fact, this partiality was seen a major failing of early corpus linguistics.

Chomsky underlines the power of introspection by saying that 'if you sit and think for a few minutes, you're just flooded with relevant data'. To illustrate this idea, Chomsky can be seen in the following exchange:

Chomsky: The verb perform cannot be used with mass word objects: one can *perform* a task, but one cannot *perform* a labour.

Hatcher: How do you know, if you don't use a corpus and have not studied the verb *perform*?

Chomsky: How do I know? Because I am a native speaker of the English language.

(Hill, 1962, p.29, cited in McEnery & Wilson, 2001, p.11)

According to McEnery & Wilson (2001), Chomsky was wrong. For a check in a corpus, *perform* magic occurs once and *performing* magic occurs three times in BNC corpus. Therefore, native speaker intuition merely allowed Chomsky to be wrong with an air of absolute certainty. Still, it must be conceded that intuition can save time in searching a corpus. Chomsky saw the linguist, or native speaker of a language, as the sole explicandum of linguistics and introspective judgments helps to distinguish ungrammatical utterances or ambiguous structures. A corpus may not contain ungrammatical sentences like **He shines Tony books*, and indeed there may be evidence suggests that it is grammatical. The construction '*He shines..'* followed by a proper name does not occur in BNC corpus, whereas constructions like '*He gives Keith the stare…'*, or '*he pushes Andy down..'* do occur in BNC. Then, it is the native speaker who can differentiate the grammaticality of a sentence not the corpus itself. McEnery & Wilson (2001) concludes that language is non-finite and corpus is finite, the problem is real and intuition must be considered (pp.12-13).

In summary, since the generative grammarian and the corpus linguist have very different views of what constitutes an adequate linguistic description, it is clear that these two groups of linguists have had difficult time in communicating and valuing each other's work (Meyer, 2008). Fillmore (1992) satirizes this situation as; when the corpus linguist asks the theoretician (or 'armchair linguist') 'Why should I think that what you tell me is *true*?, the generative grammarian replies as 'Why should I think that what you tell me is *interesting*?' About corpus linguist and 'armchair' linguists, Fillmore (1992) states that:

> Armchair linguistics does not have a good name in some linguistic circles. A caricature of the armchair linguist is something like this. He sits in a deep soft armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting ''Wow, what a neat fact!'', grabs his pencil, and writes something down. Then he paces around for hours in the excitement of having some still closer knowing to what language is really like. (There isn't anybody exactly like this, but there are some approximations.
>
> Corpus linguistics does not have a good name in some linguistics circles. A caricature of the corpus linguists is something like this. He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy in determining the relative

> frequencies of the eleven parts of speech as first word of a sentence versus the second word of a sentence''
>
> (Fillmore, 1992, p.35)

The point is that the primary concern of corpus linguists is an accurate description of language whereas the major importance of generative grammarian is a theoretical discussion of language that advances our knowledge of UG. Fillmore (1992) concludes that two linguists need each other, indeed, wherever possible, they should exist in the same body.

Different critics have been made by others which centers around the problem of data processing in corpus linguistics. Abercrombie (1965) claimed that corpus-based approach as being composed of 'pseudo-techniques' (McEnery and Wilson, 1996). For example, as in Käding (1897), searching 11 million word corpus only with eyes is very impractical, too slow and time consuming, indeed too expensive and error prone. McEnery & Wilson (1996) state that whatever Chomsky's criticisms, Abercrombie's criticisms were very real and undoubtedly correct. The impact of criticism on corpus analysis as being time consuming, more expensive, less accurate and less feasible had levelled at early corpus linguistics in the 1950s and continued until the faster computers became available.

During the 1950s, a series of criticisms were made of corpus-based approach to language study. According to McEnery & Wilson (1996), some of these criticisms were right, some were half-right and some were having proved themselves in time to wrong or irrelevant. The first important point is that these criticism were not necessary fatal ones although they were widely perceived and the second is that some linguists carried on studies balancing between the use of corpus and the use of intuition.

2.1.3. Early Modern Corpora: 1950s - 1980s

Since the thirteenth century to 1950s, corpus linguistics have had a long historical background passed with a constant developments in corpus methodology and also criticisms in the linguistics field. Although corpus linguistics underwent a period of unpopularity during 1950s, the major developments encounters these times around late 1950s and early 1960s when the first generation of computer based corpus studies began to be used. The real breakthrough in corpus linguistics came with the access to machine readable texts which could be stored, transported, and analyzed electronically in the early 1960s. The first modern corpus of the English language is the Brown Corpus (the Brown University Standard Corpus of Present-day American English) was built in 1961. With important developments in technology, computers equipped with more processing power, massive data storage and exploitation of massive corpora with relatively low costs. Tasks which were done by manually with human capacity or required enormous such as frequency lists and concordances could now be done easily with improved computers. In 1970s and 1980s, there has been an explosion in the quantity and the variety of texts prepared for analysis by computer. Since then, the number and the size of corpora and corpus based studies have considerably increased during 1980s onwards.

After Quirk initiated the important pre-computational Survey of English Usage (SEU) in 1959, Francis and Kucˇera has begun the compiling process of Brown corpus for American English in 1961 (Johansson, 2008). Following Brown Corpus, its British counterpart London-Oslo Bergen (LOB) corpus was constructed by Geoffrey Leech in 1974 and London-Lund corpus by Jan Starvik in 1975. Kennedy (1998) characterizes these three corpora as the '**first-generation corpora'**. Johansson (2008) points out that ''Although they are not the only early computer corpora compiled for language research, they are the ones which have been influential in the development of English corpus linguistics, and they have no doubt also stimulated corpus studies more generally'' (p.35). Next two decades after 1960s, many considerable corpora followed Brown Corpus such as CHILDES, ICE, BNC and COBUILD. The availability of computerized corpus and availability of institutional and private computing opportunities provided the revival of corpus linguistics. The growth of corpus linguistics can be seen in the number of corpus-based studies between 1965 and 1991 in Table 1.:

Table 1

The Revival of Corpus Linguistics (Johansson, 1991, p.312, adapted from McEnery & Wilson, 1996, p. 18)

Date	**Studies**
To 1965	10
1966-1970	20
1971-1975	30
1976-1980	80
1981-1985	160
1986-1991	320

Initial corpus works generally restricted to English language, therefore the development in corpus studies were mostly carried out in English language.

2.1.3.1. First Computer-based Corpus: Brown Corpus and Beyond

Brown corpus can be considered as the pioneer in the history of corpus linguistics being the first computer based corpus. One of the compilers, W. Nelson Francis, gives a vivid account in his paper titled 'Problems of assembling and computerizing large corpora' (Francis 1979). When planning the Brown Corpus with Henry Kucˇera, they convened a conference of 'corpus-wise scholars' at Brown University, including Randolph Quirk (complier of SEU). Francis continues:

> This group decided the size of the corpus (1,000,000 words), the number of texts (500, of 2,000 words each), the universe (material in English, by American writers, first printed in 36 I. Origin and history of corpus linguistics _ corpus linguistics vis-a`-vis other disciplines the United States in the calendar year 1961), the subdivisions (15 genres, 9 of 'informative prose' and 6 of 'imaginative prose') and by a fascinating process of individual vote and average consensus, how many samples from each genre (ranging from 6 in science fiction to 80 in learned and scientific).
>
> (Francis 1979, p.117, cited in Johansson, 2008, p.p.35-36)

In the beginning, the aim of compiling Brown corpus was to create ‘‘a standard sample of present-day English for use with digital computers’’ (Starvik, 1991, p. 7). In 1967, Kucˇera and Francis published their classic work *Computational Analysis of Present-Day American English*, which provided basic statistics on what is known today simply as the Brown Corpus. The Brown Corpus was a carefully compiled selection of current American English, totaling about a million words drawn from a wide variety of sources. Kucˇera and Francis subjected it to a variety of computational analyses, from which they compiled a rich and variegated opus, combining elements of linguistics, psychology, statistics, and sociology (http://en.wikipedia.org/wiki/Brown_Corpus).

The Brown corpus consists of 500 samples, distributed across 15 genres in rough proportion to the amount published in 1961 in each of those genres. All works sampled were published in 1961; as far as could be determined they were first published then, and were written by native speakers of American English. Each sample began at a random sentence-boundary in the article or other unit chosen, and continued up to the first sentence boundary after 2,000 words. In a very few cases miscounts led to samples being just under 2,000 words. The original data entry was done on upper-case only keypunch machines; capitals were indicated by a preceding asterisk, and various special items such as formulae also had special codes. The composition of Brown corpus and its British counterpart LOB corpus is presented in Table 2:

Table 2

The Composition of Brown Corpus and LOB Corpus (adapted from Johansson, 2008, p. 36)

Text categories	**Number of texts in each category**	
	Brown	**LOB**
A Press: reportage	44	44
B Press: editorial	27	27
C Press: reviews	17	17
D Religion	17	17
E Skills, trades, and hobbies	36	38
F Popular lore	48	44
G Belles lettres, biography, essays	75	77
H Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)	30	30
J Learned and scientific writings	80	80
K General fiction	29	29
L Mystery and detective fiction	24	24
M Science fiction	6	6
N Adventure and western fiction	29	29
P Romance and love story	29	29
R Humour	9	9
Total	**500**	**500**

The building of the Brown Corpus is remarkable considering the unsupportive environment among leading linguists at the time. The process told by W. Nelson Francis in 1979: has often been quoted (Francis 1979, 110):

> In 1962, when I was in the early stages of collecting the Brown Standard Corpus of American English, I met Professor Robert Lees at a linguistic conference. In response to his query about my current interests, I said that I had a grant from the U.S. Office of Education to compile a million-word

> corpus of present-day American English for computer use. He looked at me in amazement and asked, ‘Why in the world are you doing that?’ I said something about finding out the true facts about English grammar. I have never forgotten his reply: ‘That is a complete waste of your time and the government’ s money. You are a native speaker 38 I. Origin and history of corpus linguistics _ corpus linguistics vis-a`-vis other disciplines of English; in ten minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text.
>
> (Francis 1979, 110, cited in Johansson, 2008, p.37-38)

The Brown Corpus has been significant in a number of ways. Firstly, it established a pattern for the use of electronic corpora in linguistics, at a time when corpora were negatively regarded by many linguists in the United States and elsewhere. In addition, it was significant in the care which was taken to systematically sample texts for the corpus and provide detailed documentation in the accompanying manual (Francis and Kucˇera 1964, 1979 cited in Johansson, 2008). Johansson (2008) adds “But the world-wide importance of the Brown Corpus stems from the generosity and foresight shown by the compilers in making the corpus available to researchers all over the world” (p.38).

In the beginning of 1970s, Geoffrey Leech at the University of Lancaster initiated to collect the British counterpart of the Brown Corpus (Leech/Leonard 1974, cited in Johansson, 2008). After the majority of work had been done at Lancaster, the project was taken over and completed in Norway, in cooperation between the University of Oslo and the Norwegian Computing Centre for the Humanities at Bergen. Therefore, the name of the corpus is *L*ancaster-*O*slo/*B*ergen Corpus. The LOB Corpus matches its American counterpart as closely as possible; see the detailed documentation on sources, sampling, and coding in the accompanying manual (Johansson/Leech/Goodluck 1978, cited in Johansson, 2008, p.38). Despite the technical advances that being echieved in the decade since the Brown Corpus was first produced, compiling the LOB Corpus was not an easy task. One difficult problem, which threatened to stop the whole project, was the copyright issue. This led indirectly to the beginning of the *I*nternational *C*omputer *A*rchive of *M*odern *E*nglish (ICAME) (http://icame.uib.no/).

In February 1977, a group of people, including Jostein Hauge, director of the Norwegian Computing Centre for the Humanities, W. Nelson Francis, Geoffrey Leech, Jan Svartvik, and Stig Johansson, met in Oslo to discuss the copyright issue as well as corpus work in general. The outcome of the meeting was a document announcing the beginning of ICAME as following:

> The undersigned, meeting in Oslo in February 1977, have informally established the nucleus of an International Computer Archive of Modern English (ICAME). The primary purposes of the organization will be:
> (1) collecting and distributing information on English language material available for computer processing;
> (2) collecting and distributing information on linguistic research completed or in progress on the material;
> (3) compiling an archive of corpuses to be located at the University of Bergen, from where copies could be obtained at cost.
> One of the main aims in establishing the organization is to make possible and encourage the coordination of research effort and avoid duplication of research.
>
> (ICAME Journal 20, 101 f., cited in Johansson, 2008, p.38)

The document announcing the establishment of ICAME was circulated to scholars active in the field, and it was used to support applications for permission to include texts in the LOB Corpus.

Fourteen years after Brown corpus, in 1975, Jan Starvik started to construct to The Survey of Spoken English (SSE) project at Lund University, which was then called London-Lund Corpus (LLC). The primary goal was to computerize the spoken corpus material collected and transcribed in SEU (which was first compiled by Quirk in 1959 in University College London) and make it available in machine-readable form (Johnsson, 2008). The process included editing and checking the corpus, which at the time consisted of 87 texts, each of 5,000 words. As described by Starvik and Quirk (1980), the study depended on the reductions of very detailed prosodic and paralinguistic transcriptions:

> […] the basic prosodic distinctions (tone units, nuclei, boosters, onsets, and stresses) have been retained in the SSE version. Other features, including tempo (allegro, clipped, drawl, etc.), loudness (piano, forte, etc.), modifications in voice quality (pitch range, rhythmicality, and tension), voice qualifiers (whisper, creak, etc.), and voice qualifications (laugh, sob, etc.) have been omitted. The reasons for reducing the number of features were partly practical and technical, partly linguistic. While we do not want to minimise the importance of paralinguistic features, it is clear that they are less central than the basic distinctions (such as tone units, types of tone, place of nucleus) for most grammatical studies of spoken English."
>
> (Svartvik/Quirk 1980, 14, cited in Johnsson, 2008, p.39)

The London-Lund Corpus was the most important source for the computer-based study of spoken English. Because of the difficulties of handling spoken material (to do with recording, transcription, prosodic coding, etc.), spoken corpora have not been so popular, and the imbalance in the availability of spoken and written material in machine-readable form is likely to remain for the predictable future (Johnsson, 2008).

In the early 1980s, whilst the conditions became possible for larger-scale corpus development, there were still formidable practical issues to be overcome. John Sinclair (1982) pointed out some inadequacies of Brown corpus:

> […]the limitation on continuous text is 2,000 words, and so any study of largish text patterns is likely to be inappropriate. Its vocabulary is controlled only indirectly via the genre classification, so any study of the patterning of infrequent words is doomed […]
>
> (1982, p.2, cited in Johnsson, 2008, p.41)

The problem was about both the short text samples and the limited size of the early text corpora. Sinclair (1982) suggested the improvement in hardware and software development offered opportunities for compiling bigger corpora and "these developments mean that everything which has ever been printed, or will ever be, is within the reach of the determined researcher" (1982, p.3, cited in Johnsson, 2008, p.41). These were the premises for the Birmingham Collection of Texts, which formed the basis for the innovative COBUILD project in lexical computing and the Collins

COBUILD English Language Dictionary led by Sinclair (1987) at University of Birmingham. Bank of English (BoE) is the name of COBUILD corpus in which 'monitor corpus' concept firstly used for the production of the COBULD dictionary (1987). In monitor corpus, there is no limit on the length of the texts, as Sinclair claims "Sampling can be done to order on gigantic, slowly changing stores of text, and detailed evidence of language evolution can be held efficiently." (1982, p.4). Accordingly, BoE consists of 500 million words of British/American written and spoken texts and it is still running since it was started to be compiled at 1980.

Another landmark in mid 1980s is the corpus-based comprehensive grammar study '*A Comprehensive Grammar of the English Language*' composed by Quirk, Greenbaum, Leech and Starvik in 1985. In many sections of these grammars, discussions of grammatical constructions were informed by analyses of the London Corpus (within SEU).

During the period of 1980s, the term '**corpus linguistics'** first appeared in ICAME conference held in Nijmegen in 1983; Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research (Aarts/Meijs 1984, cited in Johansson, 2008).

The period between 1960s-1980s passed by learning how to build and maintain corpora of up to a million words; no material was available in electronic form, so everything had to be transliterated on a keyboard (Bonelli and Sinclair, 2006). Toward the end of the period of 1980s, as the main developments on computer technology came, the interest in corpus-based language research increased.

2.1.4. Contemporary Corpora: 1990s-Present Day

Besides the developments in English corpus linguistics, we should to provide a picture of the main trends more generally. At the beginning of the 1970s, corpora were few and small, corpus use was limited and cumbersome, and the users were restricted to a dedicated few outside the mainstream of linguistics at the time. Hence, computer corpus studies rarely went beyond indexes, concordances, and quantitative lexical studies. Twenty years later a fast increasing number of users had easy access to vast amounts of machine-readable text (different types of written and spoken material, modern and historical texts, general and specialized corpora, machine-readable dictionaries), new analysis tools were developed (concordancers, taggers, text analysis software), and the uses expanded to encompass a wide range of linguistically sophisticated studies in syntax, lexis, discourse, language variation and change (Johnsson, 2008).

An important shift in the theoretical perspective this time has been caused by corpus linguistics from rationalism which was dominant until 1970s to empiricism which was revived in the early 1990s. When the technology of corpus analysis became actually usable, empiricism as a methodology has proved its value. Observability of phenomena, verifiability of theories and frequency information which cannot be obtained through introspection provided an upgrade for corpus linguistic methodology.

The study of acquisition and development is crucially depend on transcriptions of interaction between and among children and parents (or caregivers) in natural settings. Over the last three decades many important and rich language acquisition data were recorded and transcribed for particular purposes (Kennedy, 1998). It is accepted that the corpus that consists of child language acquisition data is a useful tool to find out more about first language acquisition. In the early 1990s, important corpora and software projects for language acquisition research were improved. The Child Language Data Exchange (CHILDES) (McWhinney and Snow, 1990) database consists of transcriptions and media data collected from conversations between young children of different ages and their parents, playmates and caretakers (in Kennedy, 1998; Lu, 2010). These data, which were gathered from 500 children, were contributed by researchers from many different countries, following the same data collection and transcription standards.

CHILDES database is large and contains 20 million words. Each file in the database contains a transcript of a conversation and includes a header encoding information on the target child or children (e.g. age, native language, whether the child is normal in terms of language development, etc.), other participants, the location and situation of the conversation, the activities going on during the conversation, and the researchers and coders collecting and transcribing the data. The conversation is transcribed in a one utterance- per-line format, with the producer of each utterance clearly marked in a prefix. Each utterance is followed by another line consisting of a morphological analysis of the utterance. Any physical actions accompanying the utterance are also provided in a separate line. The Computerized Language Analysis (CLAN) software (McWhinney, 2000) is a bundle of computational tools designed to automatically analyze data transcribed in the CHILDES format (in Lu, 2010).

Between 1991 and 1995, according to Kennedy (1998), the most ambitious corpus compilation project yet attempted was undoubtedly the British National Corpus (BNC). The project was launched with a wide collaboration between major academic, commercial publishing and publicly funded institutions. With the financial support of British government paying the half of costs, he project was established to create a corpus of about 100 million words of contemporary spoken and written British English. The written part of the BNC (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part (10%) consists of orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different age, region and social classes in a demographically balanced way) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-talks (http://www.natcorp.ox.ac.uk/corpus/index.xml).

Another large-scale corpus project for the comparative study of English worldwide is the International Corpus of English (ICE) began to be launched in the early 1990s by Sydney Greenbaum, director of SEU at University College of London. The primary aim is collecting material for comparative studies of English worldwide. Twenty-four research teams around the world are preparing electronic corpora of their own national or regional variety of English. Each ICE corpus consists of one million words of spoken and written English produced after 1989. For most participating countries, the ICE project is stimulating the first systematic investigation of the national varieties (http://ice-corpora.net/ice/index.htm).

There are recent corpus projects such as METU Turkish Corpus and Turkish National Corpus (TNC). METU Turkish Corpus was compiled by Say et. Al. (2004) at Middle East Technical University, consisting of 2 million words of written texts gathered from books, journals and newspapers from the period of post-1990. TNC was led by Aksan & Aksan (2009) and was compiled at Mersin University and completed at 2011. TNC approximately includes 50 million words collected from written texts and spoken transcriptions covering the period between 1990 and 2008.

In summary, Sinclair and Bonelli (2006) remind that “Corpora have come a long way from the time that they were rejected by applied linguists on the grounds that their evidence could not be trusted by learners.” (p. 217). Up to present, corpus linguistics demonstrated the utility of the use of corpora in language studies. The construction and the use of large-scale corpora provided the wide-spread recognition of the validity of the corpus as a tool in the analysis of language. Today, corpora are proving their worth by the results such as authoritative grammar, language processing tools, better dictionaries, and new methods for constructing thesauri (McEnery and Wilson, 1996).

2.1.5. Future of Corpora

Sinclair and Bonelli (2006) state that one feature of the working environment for linguists, will not go away, no matter what happens to the corpus. That is, linguists now work in a data-rich environment, and even if they make only minimal use of the resources now commonly available, they are able to make statements with greater authority than before, and with greater generality. Moreover, many researchers are already seeking ways of understanding how to interpret search results from the Web, how to improve the quality of the information obtained, how to replicate results, or how to perform a similar task to replication in an ever-moving torrent of text. This kind of progress may lead to a new and more modern idea of a corpus – not the sample corpus from which almost all our present corpora are derived, but something closer to the notion of a monitor corpus, letting the text flow over a finely-tuned set of filters, continuously sampling, updating records, and maintaining a stable set of descriptive tools for users, rather than a stable description.

A significant factor which may have impact the future of corpora is the ‘World Wide Web’. Hundt, Nesselhauf and Biewer (2007) point out that “We will, in future, have to make use of the web as one additional resource to complement the evidence we can extract from our carefully compiled ‘standard’ corpora.” (p.4). Using web as a source of corpus compilation, corpus building and corpus analysis is a rising trends in the field of corpus linguistics. A recent example of web based corpus work is an online web tool named ‘WebCorp’ (http://www.webcorp.org.uk/live) which provides access to world wide web as a corpus consisting of large collection of texts for linguistics data search. WebCorp operates as a search engine; a word or a phrase is entered, then WebCorp takes the list of URLs and extract the all concordance lines of the particular item from each pages. Although WebCorp works as an ordinary search engine as Google or Bing, it differs in some respects. WebCorp is designed to retrieve linguistics data from web within concordance lines showing the context in which the searched item occurs. There are studies which focused on WebCorp in order to describe it or to use it for a linguistic analysis (Schmied, 2006; Kehoe, 2006).

In sum, as the computer technology is a field that growing fast, new technologies are becoming available for corpus linguists. In addition, McEnery and Wilson (2001) claims that there is a stunning amount of potential for the exploitation of multimedia technology to improve the representation, manipulation and retrieval of corpus data and it cannot be too long before somebody takes this challenge up and develops a truly multimedia corpus. In brief, McEnery and Wilson (2001) states that "Corpus linguistics is constantly developing" (p.175).

2.2. The Scope of Corpus Linguistics

In the language sciences, a corpus is a body of written or transcribed speech which can serve a basis for linguistics analysis and description. (Kennedy, 1998). Over the last three decades, the compilation and the analysis of corpora stored in computerized databases have led to a new enterprise as 'corpus Linguistics'. In terms of research on language, corpus linguistics is a source of evidence for improving descriptions of the structure and the use of languages, and for various applications, including natural language processing by machine or how to learn or teach a language. As a research activity, Leech (1992) states that the focus of a study corpus linguistics is on performance rather than competence, and on observation of language in use leading to theory rather than vice versa.

Corpus linguistics primarily is concerned with the description and of the nature, structure and use of language and with particular interests such as language acquisition, variation and change. Nevertheless, corpus linguistics has developed a tendency within linguistics sometimes focusing on lexis and lexical grammar rather than pure linguistics. This may be the result of methodological aspects of corpus such as concordance (Kennedy, 1998).

Many early studies in corpus linguistics relied on simply counting the occurrence of linguistic items. For example, some lexical studies compare the frequency of particular words and some grammatical studies counted the frequency of nouns, verbs and adjectives. However, according to Biber et al. (1998) a carefully exploited representative corpus can provide much additional information about language use.

Biber et al. (1998) characterizes four essential properties of corpus-based analysis:

- It is empirical, analyzing the actual patterns of use in natural texts;
- It utilizes a large and principled collection of natural texts, known as a 'corpus' , as the basis for analysis;
- It makes extensive use of computers for analysis, using both automatic and interactive techniques;
- It depends on both quantitative and qualitative analytical techniques.

(Biber, et al., 1998, p.4)

At the present day of corpus linguistics, some researchers tend to focus on corpus compiling, others on methodology for text analysis and processing, and still others on corpus-based linguistic descriptions and the applications of such descriptions.

2.2.1. Corpus Linguistics: Theory or a Methodology?

There have been arguments on the corpus linguistics whether it is a methodology or an independent branch of linguistics. McEnery and Wilson (1996) claims that corpus linguistics is not a branch of linguistics in the same sense as syntax, semantics, sociolinguistics which generally concentrate on describing/explaining some aspects of language. "In contrast, corpus linguistics is a methodology rather than an aspect of language requiring explanation or description" (McEnery and Wilson, 1996, p.2).

Tognini-Bonelli (2001) claims that corpus linguistics goes well beyond this methodological role so far and it has become an independent discipline. McEnery, Xiao and Tonio (2006) agree that corpus linguistics is a real domain research and has become a new research enterprise and a philosophical approach of linguistics theory. On the other hand, they maintain the idea that corpus linguistics is indeed a methodology than an independent branch in the same sense as phonetics, semantics, syntax or pragmatics. Different from these linguistics areas, corpus linguistics is not restricted to an aspect of a particular language, better, it can be employed to almost any areas of linguistics research. For instance, syntax can be examined using a corpus-based or non-corpus-based approaches (McEnery, Xiao and Tonio , 2006).

2.2.2. Corpus-based vs. Corpus Driven Approaches

As has been emphasized in the previous section, corpus linguistics is essentially a methodology, or sets of methodologies rather than a theory of language descriptions. According to Hunston (2006) corpus linguistics seems to be 'theory neutral' although the practice of doing corpus linguistics is never neutral, "..as each practitioner defines what is meant by a 'feature' and what frequencies should be observed, in line with a theoretical approach to what matters in language" (p.244). Approaches by which the use of a corpus that essentially rely on the existence of categories derived from non-corpus investigations of language are sometimes referred to as "corpus based" (Tognini-Bonelli, 2001). The corpus-based approach is a method that uses an underlying corpus as an inventory of language data. From this repository, appropriate material is extracted to support intuitive knowledge, to verify expectations, to allow linguistic phenomena to be quantified, and to find proof for existing theories or to retrieve illustrative samples. It is a method where the corpus is interrogated and data is used to confirm linguistic pre-set explanations and assumptions. It acts, therefore, as an additional supporting material.

On the other hand, the corpus-driven approach is a methodology whereby the corpus serves as an empirical basis from which lexicographers extract their data and detect linguistic phenomena without prior assumptions and expectations (Tognini-Bonelli 2001). Any conclusions or claims are made exclusively on the basis of corpus observations. Sinclair (1991, 2004) argues that the kind of patterning observable in a corpus (and nowhere else) require descriptions of a markedly different kind from those commonly available. Both the descriptions and the theories that they in turn inspire are, in Tognini-Bonelli's (2001) terms, ''corpus driven.'' Some of the challenges of corpus-driven theories are:

- Lexis and grammar are not distinct, and grammar is not an abstract system underlying language.
- Choice of any kind is heavily restricted by choice of lexis
- Meaning is not atomistic, residing in words, but prosodic, belonging to variable units of meaning and always located in texts.

(Hunston, 2006, p.244)

Due to the corpus-driven approach, it is accepted that the notion of pattern grammar focuses on the way that different lexical items behave differently in terms of how they are complemented. Grammatical generalizations about complementation cannot be made without describing the individual lexical behavior. Sinclair (1991, 2004) suggests that meaning is not only expressed by the examined (node) word, but also neighbouring, co-selected words so that a lexical item consists of several words and their relationships, that is, words typically occur with specific collocations in specific grammatical configurations.

The point is that corpus-driven approach focuses on lexical item as the primary object of the study and put the lexis in the heart of the description of the language. Another point is noted by McEnery, Xiao and Tonio (2006) as corpus-driven approach is not so different from corpus-based approach;" while latter allegedly insulates theory from data or standardizes data to fit the theory, the former filters the data via apparently, though there is no guarantee that corpus is not explored selectively avoid inconvenient evidence (p.9).

2.3. Aspects of Corpora

For revisiting the notion of corpus, Francis who is the first complier of the first ever machine-readible corpus, Brown corpus, defines it as "..a collection of texts assumed to be representative of a given language, or other subset of a language, to be used for linguistic analysis" (1964, cited in Bartsch, 2004, p..118). As emphasized by Francis (1964), a corpus should represent the language as it exists; it should combine different sources or kinds of languages as text types, genres, and domains, medium, written or spoken. In contrast to a simple body of text, corpus is described as fine-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration. Leech (1992) explains corpus by emphasizing representativeness:

> It should be added that computer corpora are rarely haphazard collections of textual material: They are generally assembled with particular purposes in mind, and are often assembled to be (informally speaking) representative of some language or text type. (Leech, 1992, p.116)

In modern linguistics McEnery and Wilson (1996) suggests four main properties that a corpus should meet: (1) machine- readable, (2) finite size, (3) sampling and representativeness, (4) a standard reference. That is, a corpus typically implies a finite body of texts, sampled to be maximally representative of a particular variety of language, and which can be stored and manipulated using a computer.

Machine-readable form is essential requirement for modern corpora. When compared to past corpora in 'book form', machine readable corpora have considerable advantages: they can be searched and analyzed quickly and indeed they can easily be enriched with extra information (McEnry and Wilson, 1996).

A corpus consists of a finite number of words which is determined at the beginning of a corpus-building project. The term "corpus" also implies a body of text of finite size, for example, 1,000,000 words as Brown Corpus contains. However, some corpora, which are called 'monitor corpus', are open-ended, i.e. texts are constantly being added and it is getting bigger in time like COBUILD project.

Representativeness, as noted above, is an essential feature in a corpus since it distinguishes a corpus from an archive (i.e. a random collection of texts). A corpus is designed to represent a language or a language variety whereas archive is not (McEnery, Xiao & Tono, 2006). Moreover, the issue of sample is unavoidable, whether the sample is representative of the language or language variety is under consideration. Sampling and balance are the factors that ensure representativeness so that as the key concept of corpus, representativeness should be explained related to issues of sampling and balance. Biber (1993) defines representativeness as it " refers to the extent to which a sample includes the full range of variability in a population" (1993, cited in McEnery, Xiao &Tono, 2006, p.13). A corpus is essentially a *sample* of a language or language variety and the sampling is entailed in the compilation of virtually corpus of a living language. From this point of view, representativeness of most corpora is determined by the range of genres included, i.e. balance, and how the text chunks for each genre are selected, namely sampling.

Balance simply means that how a corpus is balanced, that is, the range categories included in a corpus. As with representativeness, the balance of a corpus is specified by its intended uses. For instance, BNC contains both spoken and written data, so it can be accepted as a balanced corpus. A balanced corpus covers a wide range of proportionally sampled text categories which are supposed to be representative of the language or language variety under consideration.

Sampling is closely associated with corpus representativeness and balance. The corpus is typically a sample of a much larger population and a sample is assumed to be representative for that general population. In other words, a sample which is maximally representative of the language or language variety provides an accurate picture of the tendencies of that entire language, including proportions (McEnery & Wilson, 1996). As for sampling examples, in Brown and LOB corpora, the target population for each corpus was first grouped into fifteen categories such as news reportage, academic prose and different types of fiction; and then samples were drawn from each text category (McEnery, Xiao & Tono, 2006).

It is accepted that a corpus constitutes a *standard reference* for the language variety that it represents, so it is available for other researchers like the Brown Corpus, the LOB corpus and the London-Lund corpus. One advantage is that widely available corpus provides a yardstick by which successive studies can be measured. Also, a standard corpus also means that a continuous base of data is being used, therefore, 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. (McEnery & Wilson, 1996).

2.3.1. Processing and Annotation of Corpus

Corpus-based approaches to language studies provided new dimensions to linguistics descriptions and to various applications by automatic analysis of text. The identification, counting and sorting of words, collocations, grammatical structures that existing in a corpus can be carried out quickly and accurately on. During the two decades, various tools have been improved for corpus processing such as software and automatic text encoding procedures for corpus annotation.

A corpus can be in two forms: unannotated; as it exists in raw state of plain text, and annotated; enhanced with additional linguistics information. Annotation is defined by Leech (2004) as the practice of adding linguistic information to a corpus. A corpus should be annotated to be useful to potential users (Meyer, 2002). Certain kinds of linguistic annotation, which involve the attachment of special codes to words in corpus in order to indicate particular features, are generally known as 'tagging' rather than annotation (McEnery & Wilson, 1996). The code which are assigned to the words are known as 'tags'. A tag usually consists of a code, which can be attached to a phoneme, morpheme, word, phrase or longer stretch of text in a number of ways, for example, using Standard Generalized Mark-up Language (SGML) which is a way of encoding electronic texts created in 1980s (Baker, et al.,2006). There are various types of tagging such as part of speech (POS), lemmatization, parsing, problem-oriented tagging, semantic tagging.

2.3.1.1. Part-of- Speech Tagging

The most basic type of linguistic corpus annotation is part-of-speech (POS) tagging which is also known as grammatical tagging or morph-syntactic tagging. It is a type of annotation or tagging by which grammatical categories are assigned to words (or in some cases morphemes or phrases), usually by an automatic tagger although human post-editing may take place as a final stage. As automatic taggers, various POS tagging software like Constituent Likelihood Automatic Word-tagging System (CLAWS) or TAGGIT is utilized. The aim of POS is to assign to each lexical unit in the text a code indicating its part of speech, for example singular common noun, comparative adjective, past participle (McEnery & Wilson, 1996). An example of POS tagging from LOB corpus is presented in Figure 4:

hospitality_NN is_BEZ an_AT excellent_JJ virtue_NN ,_, but_CC
not_XNOT when_WRB the_ATI guests_NNS have_HV to_TO sleep_VB
in_IN rows_NNS in_IN the_ATI cellar_NN !_!

Figure 4. An example of POS tagging from LOB corpus (adapted from http://ucrel.lancs.ac.uk/annotation.html).

In this example in Figure 4, POS codes are attached to words using underscore (_) character. This tagging analysis was made by CLAWS and following codes are used:

NN	singular common noun (boy, pencil,..)
BEZ	(is)
AT	singular article (a, an)
JJ	general adjective (happy, red,..)
CC	co-ordinating conjunction (and, or, but, so, then, yet, only, for)
XNOT	(not, n't)

WRB	wh-adverb (where, when, how, why, whenever, wherever, however ...)
ATI	article (the, ze, no)
NNS	plural common noun (pencils, skeletons, days, weeks ...)
HV	(have)
TO	infinitival TO
VB	base form of lexical verb
IN	preposition (after, by, of, for, since ...)

(adapted from http://ucrel.lancs.ac.uk/annotation.html)

This tag-set above has been used in CLAW software which was developed by Lancaster University at 1980s. CLAWS has been used in many corpora as BNC and LOB for POS annotations. CLAWS has consistently achieved 96-97% accuracy (the precise degree of accuracy varying according to the type of text). Judged in terms of major categories, the system has an error-rate of only 1.5%, with 3.3% ambiguities unresolved within the BNC (http://ucrel.lancs.ac.uk/annotation.html).

POS annotation is one of the first type of annotation to be performed on corpora and it is the most commonly used one. One reason for this is that POS tagging is a task which can be carried out by a computer to a high degree of accuracy without manual intervention, since the correct POS for any given word is highly predictable from surrounding context like common word suffixes and their possible parts of speech (McEnery and Wilson, 1996).

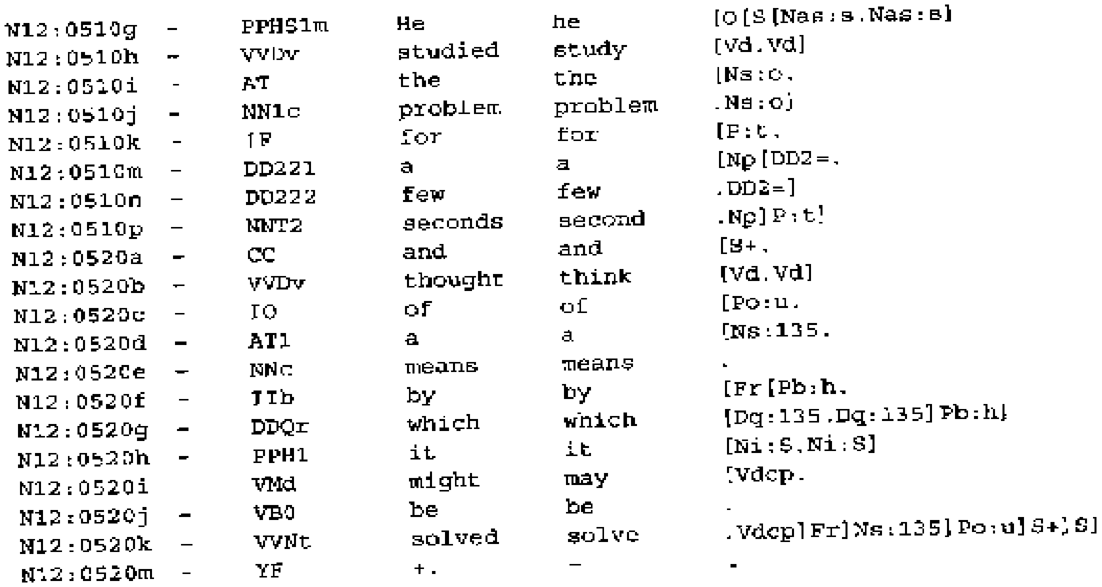
2.3.1.2. Lemmatisation

Lemmatisation is another form of automatic annotation that is closely allied to the identification of parts-of- speech and involves the reduction of the words in a corpus to their respective lexemes. Lemmatisation allows the researcher to extract and examine all the variants of a particular lexeme without having to input all the possible variants, and to produce frequency and distribution information for the lexeme (Baker, Hardie, and McEnery, 2006).

Lemmatization is a process of classifying together all the identical or related forms of a word under a common headword, as in dictionary making of the various morphological inflections or derivations of a word are listed under a single entry (Kennedy, 1998). For example, *go*, *gone*, *going*, *goes, went* are classed under the headword *go*; *better* is counted under *good*; and *broke* is classified under *break*. In list in the second column of words have been lemmatized:

He	he
studied	study
the	the
problem	problem
for	for
a	a
few	few
seconds	second
and	and
thought	think
of	of
a	a
means	means
by	by
which	which
it	it
might	might
be	be
solved	solved

This list above can be seen as a lemmatized intersection from SUSANNE corpus in Figure 5:



```
N12:0510g   -   PPHS1m   He        he        [O[S[Nas:s.Nas:s]
N12:0510h   -   VVDv     studied   study     [Vd.Vd]
N12:0510i   -   AT       the       the       [Ns:o.
N12:0510j   -   NN1c     problem   problem   .Ns:o]
N12:0510k   -   IF       for       for       [P:t.
N12:0510m   -   DD221    a         a         [Np[DD2=.
N12:0510n   -   DD222    few       few       .DD2=]
N12:0510p   -   NNT2     seconds   second    .Np]P:t]
N12:0520a   -   CC       and       and       [S+.
N12:0520b   -   VVDv     thought   think     [Vd.Vd]
N12:0520c   -   IO       of        of        [Po:u.
N12:0520d   -   AT1      a         a         [Ns:135.
N12:0520e   -   NNc      means     means     .
N12:0520f   -   JIb      by        by        [Fr[Pb:h.
N12:0520g   -   DDQr     which     which     [Dq:135.Dq:135]Pb:h]
N12:0520h   -   PPH1     it        it        [Ni:S.Ni:S]
N12:0520i       VMd      might     may       [Vdcp.
N12:0520j   -   VB0      be        be        .
N12:0520k   -   VVNt     solved    solve     .Vdcp]Fr]Ns:135]Po:u]S+]S]
N12:0520m   -   YF       +.        -         .
```

Figure 5. Example of lemmatization from SUSANNE corpus (adapted from McEnery and Wilson, 1996, p. 43)

In this lemmatization example from SUSANNE, the format is as follows:

N12:0510h	**VVDv**	**studied**	**study**
Corpus file	POS tag code	actual word	head word (lemma)

As can be seen, every word in corpus is on a separate line. Lemmatization can be useful process for certain purposes, it is considered as a very useful technique especially for lexicography.

2.3.1.3. Parsing

Parsing is the procedure of identifying morphosyntactic categories in a text, then bringing these categories into higher level syntactic relations with one another. It briefly refers to annotating with syntactic structures using ‘treebanks’ for parsing. A parsing example from BNC corpus, the sentence *Claduia sat on a stool* is illustrated by a tree diagram in Figure 6.:

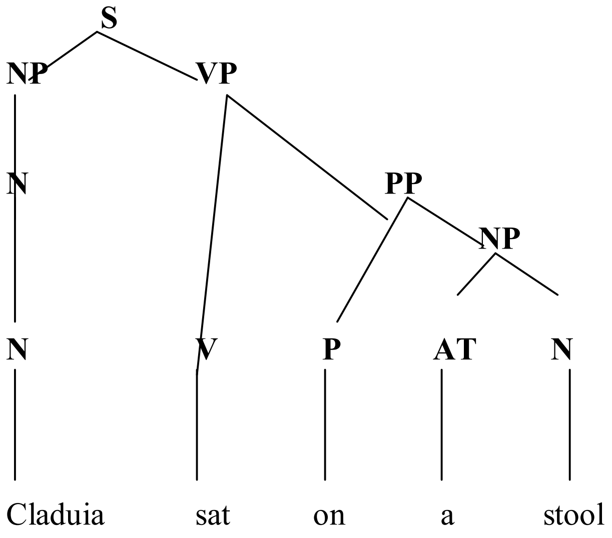
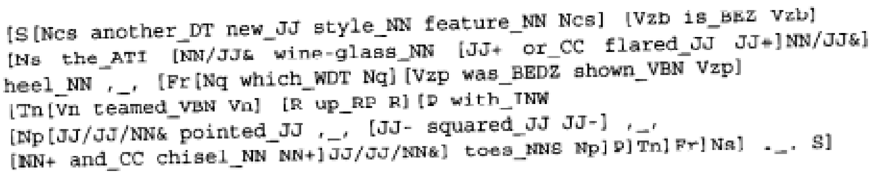


Figure 6. Parsing example from BNC corpus (adapted from McEnery and Wilson, 1996, p. 43).

The parsed sentence above can be displayed like this:

[S [NP Claduia_NP1 NP] [VP sat_ VVD[PP on_II[NP a_AT1 stool_NN1 NP]PP]VP]
The constituents are indicated by opening and closing brackets annotated with the parsed type, using the same abbreviations in the tree bank, morphosyntactic information is attached to the words with underscore characters in the form POS tags. Parsing may be applied in two forms as '**full parsing**' in which all structures in the sentence given as detailed as possible; and '**skeleton parsing**' which is a less detailed approach which tends to provide less distinguished set of syntactic constituent types. Parsed corpora examples include the Lancaster–Leeds Treebank, the Penn Treebank, the Gothenburg Corpus and the CHRISTINE Corpus (Baker, Hardie and McEnery, 2006). In figure 7. An example of full parsing from Lancester-Leeds Treebank is presented:



```
[S[Ncs another_DT new_JJ style_NN feature_NN Ncs] [Vzb is_BEZ Vzb]
[Ns the_ATI [NN/JJ& wine-glass_NN [JJ+ or_CC flared_JJ JJ+]NN/JJ&]
heel_NN ,_, [Fr[Nq which_WDT Nq][Vzp was_BEDZ shown_VBN Vzp]
[Tn[Vn teamed_VBN Vn] [R up_RP R][P with_INW
[Np[JJ/JJ/NN& pointed_JJ ,_, [JJ- squared_JJ JJ-] ,_,
[NN+ and_CC chisel_NN NN+]JJ/JJ/NN&] toes_NNS Np]P]Tn]Fr]Ns] ._. S]
```

Figure 7. An example of full parsing from the Lancester-Leeds Treebank (adapted from McEnery and Wilson, 1996, p.45)

2.3.1.4. Semantic Annotation

Semantic annotation is the next step after grammatical annotations. An example of semantic annotation is shown in Figure 8:

PPIS1	I	Z8
VV0	like	E2+
AT1	a	Z5
JJ	particular	A4.2+
NN1	shade	O4.3
IO	of	Z5
NN1	lipstick	B4

Figure 8. An example of semantic annotation (adapted from http://ucrel.lancs.ac.uk/annotation.html)

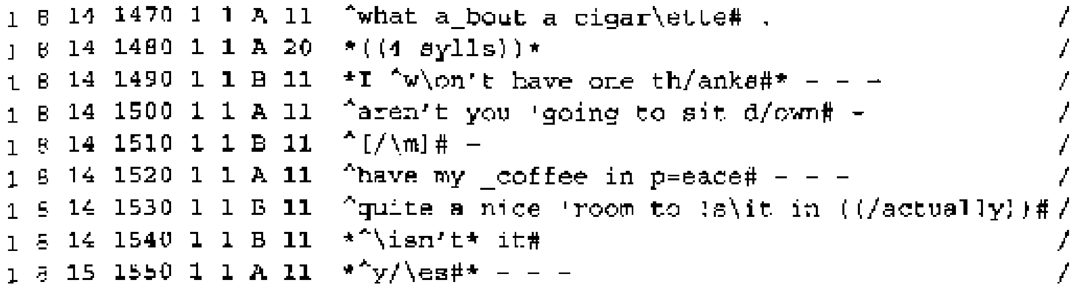
In this example, the text is read downwards in which grammatical tags on the left, and the semantic tags on the right. The semantic tags are composed of:

- an upper case letter indicating general discourse field
- a digit indicating a first subdivision of the field
- (optionally) a decimal point followed by a further digit to indicate a finer subdivision
- (optionally) one or more `pluses' or `minuses' to indicate a positive or negative position on a semantic scale

For instance, A4.2+ indicates a word in the category `general and abstract words' (A), the subcategory `classification' (A4), the sub-subcategory `particular and general' (A4.2), and `particular' as opposed to `general' (A4.2+). Similarly, E2+ belongs to the category `emotional states, actions, events and processes' (E), subcategory `liking and disliking' (E2), and refers to `liking' rather than `disliking' (E2+). The semantic annotation is designed to apply to open-class or `content' words. Words which belong to closed classes, as well as proper nouns, are marked by a tag with an initial Z, and set aside from the statistical analysis (http://ucrel.lancs.ac.uk/annotation.html).

2.3.1.5. Prosodic Annotation

Prosodic annotation aims to indicate patterns of intonation, stress and pauses in speech. The spoken parts of SEU which was collected in 1960s were prosodically annotated and later encoded on computer as London-Lund corpus. A recent prosodic annotation was employed in Lancester/IBM Spoken English corpus. An example of prosodic annotation from London-Lund Corpus is given in Figure 9.:



```
1 8 14 1470 1 1 A 11  ^what a_bout a cigar\ette# .                /
1 8 14 1480 1 1 A 20  *((4 sylls))*                               /
1 8 14 1490 1 1 B 11  *I ^w\on't have one th/anks#* - - -         /
1 8 14 1500 1 1 A 11  ^aren't you 'going to sit d/own# -          /
1 8 14 1510 1 1 B 11  ^[/\m]# -                                   /
1 8 14 1520 1 1 A 11  ^have my _coffee in p=eace# - - -           /
1 8 14 1530 1 1 B 11  ^quite a nice 'room to !s\it in ((/actually))#/
1 8 14 1540 1 1 B 11  *^\isn't* it#                               /
1 8 15 1550 1 1 A 11  *^y/\es#* - - -                             /
```

Figure 9. An example of prosodic annotation from London-lund corpus (adapted from McEnery and Wilson, 1996, p. 55).

In prosodic annotation in Figure 9, the codes below are used to annotate:

#	end of tone group
^	onset
/	rising nuclear tone
\	falling nuclear tone
/\	rise-fall nuclear tone
_	level nuclear tone
[]	enclose partial words and phonetic symbols
‘	normal stress
!	booster: higher pitch than preceding prominent syllable
=	booster: continuance
(())	unclear
**	simultaneous speech
-	pause of one stress unite

(adapted from McEnery & Wilson, 1996, p. 55)

The major difficulty of prosodic annotation is that it is considerably more impressionistic than other linguistic levels in corpus annotation. Prosodic transcription is task which requires manual involvement of highly skilled phoneticians. Unlike POS tagging, it cannot be delegated to the computer (McEnery and Wilson, 1996). For instance, the prosodic annotation of the Lancaster/IBM Spoken English Corpus (SEC) was carried out by two phoneticians (Gerry Knowles and Briony Williams). A set of 14 special characters was used to represent prosodic features. Stressed syllables were marked with a symbol indicating the direction of the pitch movement. Syllables which were felt to be stressed but with no independent pitch movement were marked with a circle (or bullet in the printed version). Unstressed syllables, whose pitches are predictable from the tone marks of surrounding accented syllables, were left unmarked (http://ucrel.lancs.ac.uk/annotation.html#acamrit).

2.3.2. Corpus Analysis

A number of procedures are used to search a corpus, to recover information, or to organize, categorize or display the facts languages which are under investigation. The most basic format used in displaying information about linguistic elements in a corpus is obtained by the agency of listing and counting (Kennedy, 1998). The lists are generated and processed by software and analyzed in different kinds ranging from simple wordlists to more sophisticated analyses such as classic concordance formats. Hunston (2002) states that “a corpus does not contain new information about language, but the software offers us a new perspective on the familiar” (cited in Evison, 2010, p. 122) and in order to gain this new perspective, the first analytical steps generally involve two related processes: the production of frequency lists (either in rank order, or sorted alphabetically) and the generation of concordances.

There is an increasing trend of software available to carry out such processes, from established commercial software such as WordSmith Tools (Scott 1999) (which has been utilized in this present study), Monoconc Pro (Barlow, 2000) and Word Sketch Engine (Kilgarriff et al. 2004). Via these software, frequency lists and concordance are built on the very basic foundation in which electronic texts collections can be searched easily and rapidly. These two basic corpus analysis techniques themselves serve both qualitative and quantitative insights in terms of linguistics analysis. McEnery and Wilson claims that qualitative analysis can provide richness and prediction whereas quantitative analysis can provide statistically reliable and generalizable results. Hence, both qualitative and quantitative analyses have something to contribute to corpus study since it supplements qualitative analyses with quantitative data and serves quantification with many sophisticated statistical techniques.

2.3.2.1. Concordance

A concordance is a list of all of the occurrences of a particular search term in a corpus, presented within the context in which they occur – usually a few words to the left and right of the search term. It is also referred to as key word in context (KWIC). A search item is often a single word although many concordance programs allow users to search on multiword phrases, words containing wildcards, tags or combinations of words and tags. Concordances can usually be sorted alphabetically on either the search term itself or to x places left or right of the search term, allowing linguistic patterns to be more easily observed by humans (Baker, Hardie and McEnery, 2006). KWIC presentation, as it is known, has a number of uses. Even the small amount of context is usually enough to show what the word or phrase means, what phrases it often occurs in, and/or the discourse function that it has. In Figure 10, the concordance of the word ‘witnessed’ is presented which sorted one place to the right (the sorted token is given in bold):

1	y told Tom Jones that he had never before	witnessed	**a** Cabinet scene like it.” All who were
2	the early decades of the twentieth century	witnessed	**an** increase in the power of medical m
3	uld be drawn up carefully and signed and	witnessed	**in** a particular way. If you write it
4	The first attitude has been	witnessed	**in** the 1930s and during our more rece
5	nk had recovered from the breakdown we	witnessed	**in** late 1986 and, despite the months al
6	fought essentially on national issues and it	witnessed	**the** return not only of a reforming Libe
7	he last year of Ayliffe’s Presidency	witnessed	t**he** fulfilment of one of the BDDA’s ea
8	eneration after the coming of Cyrus which	witnessed	**the** most brilliant speculations of the “
9	dirt, gloom and misery as I never before	witnessed	“. Queen Victoria had the curtains of h
10	ood that this small Year Niner has been “	witnessed	“ to and moves on to his next victim.

Figure 10. A sample concordance of ‘witnessed’ (adapted from Baker, Hardie and McEnery, 2006, p. 43).

Even though it is a small section of a concordance, it is possible to interpret some linguistic explanation about ‘witnessed’. For instance, the verb ‘witnessed’ tends to precede an article or the preposition ‘in’. The concordance also shows the different meaning of ‘witnessed’ from a legal usage in line 3 (*signed and witnessed*) to a meaning to do with noting a considerable event as in line 8 (*witnessed the most brilliant speculations*). In addition, the phrase *never before witnessed* occurs twice in the concordance as in lines 1 and 9, and this may suggest that ‘witnessed’ is often used to denote a remarkable or unusual event (Baker, Hardie & McEnery, 2006).

2.3.2.2. Frequency

The concept of frequency supplies much of the analytical work that is carried out within corpus linguistics. According to McEnery & Wilson (2001), the most straightforward approach to work with quantitative data is to simply classify items according to a particular scheme and perform an arithmetical count of number of items (or tokens) within the text which belong to each classification (or type) within the scheme. In order to do a simple frequency count, for example, we set up a classification scheme including four major parts of speech as noun, verb, adjective and adverb. Every time we meet a word in the corpus which belongs to one of these categories- a token of one the types-we would simply add 1 to the count for corresponding category type (McEnery and Wilson, 2001).

Frequencies can be given as raw data, for example, there are 58,860 occurrences of the word *man* in the British National Corpus (BNC). Or (more often) they can be presented as percentages or proportions – *man* occurs 602.91 times per million words in the BNC – allowing comparisons between corpora of different sizes to be made (Baker, Hardie and McEnery, 2006). Frequency analyses also allow comparisons to be made between different words in a corpus, for example *man* (602.91 per million) tends to occur more frequently than *woman* (225.43 per million), suggesting that *man* is the marked or ‘prototype’ term. Another one, *homosexual* (8.41 per million) occurs more than *heterosexual* (3.86 per million), which in this case is due to the term *homosexual* being marked because homosexuality has been considered problematical and non-standard by society in the past (Baker, Hardie and McEnery, 2006).

When a frequency list for a particular corpus is generated, the software searches every item in that corpus in order to establish how many tokens there are in total (at the simplest level a token and a word can be considered to be the same thing) and how many different types constitute this total. The software such as WordSmith Tools (Scott, 1999) then outputs the final counts as a frequency list, which can be displayed in rank order of frequency or in alphabetical order. Table 3. displays a rank order frequency list:

Table 3

A Sample of Rank Order Frequency List (adapted from Evison 2010, p.124)

N	**Token**	**Freq.**	**%**
1	the	203	4.76
2	I	129	3.02
3	a	116	2.72
4	and	109	2.55
5	it	89	2.09
6	to	86	2.02
7	think	81	1.9
8	of	80	1.87
9	you	78	1.83
10	yeah	76	1.78

In Table 3, the rank order (N), raw frequency (actual number of occurrences) of each token and percentage of tokens in the whole corpus that each frequency count represent are presented.

Hunston (2006) points out that "Information about frequency is not very informative unless it is comparative." (p.235), and frequency is generally used to compare one corpus with another and, by implication, to compare two languages, varieties of a language, or text types. Table 4. presents a compare of rank order of items in two corpora.

Table 4

A Sample of Comparison of Rank Frequency (adapted from Evison, 2010, p. 126)

N	BNC	TTFN
1	I	the
2	you	and
3	it	of
4	the	I
5	and	a
6	a	to
7	to	that
8	that	you
9	yeah	in
10	oh	it

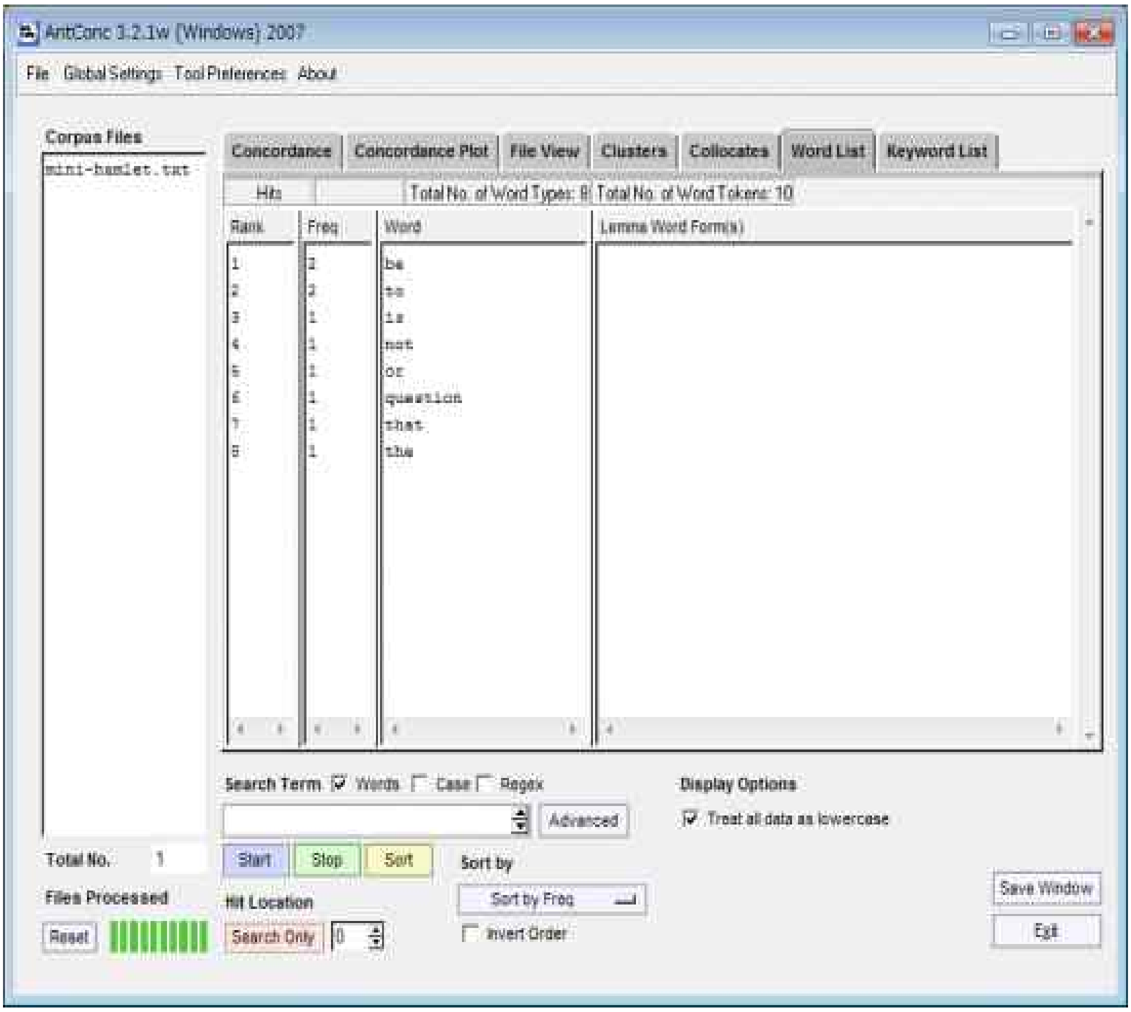
In Table 4., ten most frequent items in 50,000 words of conversation extracted from the BNC, and the top ten items from TTFN corpus (TESOL Talk from Nottingham) of 54,000 words of podcast talk. The data from BNC are considered as intimate conversations and from TTFN as academic conversations. The difference between two corpora can be seen in first and second personal pronouns *I* and *You*: they occur in higher ranks in intimate conversations whereas they placed lower down for more academic conversations (Evison, 2010).

2.3.2.3. Wordlists

A word list is a list, generally arranged either alphabetically or in frequency order, of all the words in a given corpus with information about the number of times that word occurs in the corpus. The simplest word lists interpret 'word' as simply a string of letters; so, for example, the number of occurrences of run is given without distinction between the noun and the verb, and the occurrences of runs, running, and ran are given separately. Word lists are usually formed through software like 'Antconc' concordance program which offers variety of corpus analysis tools such as KWIC concordancing, key word (Anthony, 2007).

Word lists give the frequencies of each word (or token) in the corpus. Words are most usually ordered alphabetically, or in terms of frequency, either with a raw frequency count and/or the percentage that the word contributes towards the whole text. In addition, word lists can be lemmatised or annotated with part-of-speech or semantic information (including probabilities – for example, the word *house* occurs as a noun about 99 per cent of the time and as a verb 1 per cent of the time) (Baker, Hardie and McEnery, 2006).

A basic step to analyze a corpus is to make a word list. It is usually arranged from highest to lowest **frequency** of *types*. A **type** is a unique form of a word. A "word" is defined as running letters separated by space or punctuation. For example in the sentence "*To be or not to be; that is the question.*" has 8 types (*to*, *be*, *or*, *not*, *that*, *is*, *the* and *question*). Below in Figure 11, the example of word list of these types made by AntConc is presented:



Corpus Files
mini-hamlet.txt

Hits
Total No. of Word Types: 8
Total No. of Word Tokens: 10

Rank	Freq	Word	Lemma Word Form(s)
1	2	be	
2	2	to	
3	1	is	
4	1	not	
5	1	or	
6	1	question	
7	1	that	
8	1	the	

Search Term ☑ Words □ Case □ Regex

Display Options
☑ Treat all data as lowercase

Total No. 1
Files Processed

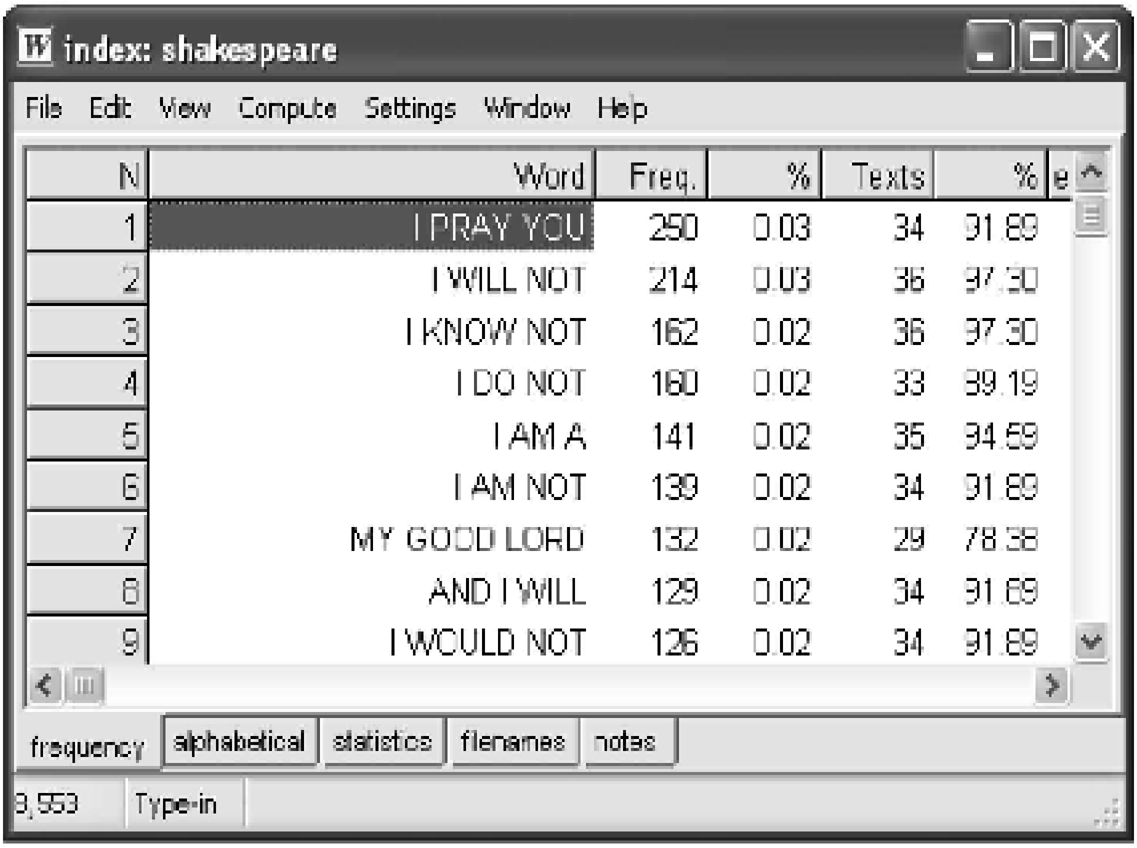
Sort by
Sort by Freq
□ Invert Order

Hit Location
0

Figure 11. A screenshot of a word list from antconc software (adapted from http://wmtang.org/corpus-linguistics/).

As shown in Figure 11., the types “to” and “be” have frequencies of 2 (namely, they occurred twice in our example). And every word is counted then there are 10 **tokens**.

Another example of word list based on three word cluster from Sheakespeare’s play which extracted by WordSmith Tools software is presented in Figure 12:



N	Word	Freq.	%	Texts	%	e
1	I PRAY YOU	250	0.03	34	91.89	
2	I WILL NOT	214	0.03	36	97.30	
3	I KNOW NOT	162	0.02	36	97.30	
4	I DO NOT	160	0.02	33	89.19	
5	I AM A	141	0.02	35	94.59	
6	I AM NOT	139	0.02	34	91.89	
7	MY GOOD LORD	132	0.02	29	78.38	
8	AND I WILL	129	0.02	34	91.89	
9	I WOULD NOT	126	0.02	34	91.89	

Figure 12. A screenshot of a wordlist sample by WordSmith Tools Software (adapted from Scott, 2010, p. 148)

It can be seen that the cluster ‘I pray you’ is not only a frequent structure but it is also widespread since it occurs in thirty-four (out of thirty-seven) plays (Scott, 2010).

More sophisticated words lists distinguish between, for example, the noun and verb occurrences of run and give summary figures for a whole lemma, such as for run, runs, running, ran, all occurring as verbs (Leech et al., 2001, cited in Hanston, 2006). Much more difficult, and indeed not publicly available, are word lists that distinguish between senses (e.g., between run meaning ‘move in fast motion’ and run meaning ‘manage an event or organization’) (Hanston 2006).

2.3.2.4. Keyword and Keyness

Keyword is the word which appears in a text or corpus statistically significantly more frequently than would be expected by chance when compared to a corpus which is larger or of equal size. Keywords can be calculated automatically in WordSmith Tools (Scott, 1999) software program. In order to compare two wordlist for deriving keywords, usually 'log-likelihood' or 'chi-squared' tests are used as statistical analysis. Commonly found keywords include (1) proper nouns; (2) grammatical words that are often indicators of a particular stylistic profile; (3) lexical words that give an indication of the 'aboutness' of a text (Baker, Hardie and McEnery, 2006). "Keywords are often taken to be markers of the "aboutness" and the style of a text (Scott & Tribble 2006, cited in Bondi, 2010, p.10): what we want to investigate here is what structures of textually keywords point to and how far they are also influenced by the position of the writer, in the context of text production (Bondi, 2010).

Key words are not necessarily the most frequent words in a corpus, but they are those words which are identified by statistical comparison of a 'target' corpus with another, larger corpus, which is referred to as the 'reference' or 'benchmark' corpus. This identification is involves the automatic comparison of word lists using the 'Keyword' program within WordSmith Tools (Scott, 1999) software. This program identifies key words on a mechanical basis by comparing patterns of frequency. A word is said to be "key" if:

a) it occurs in the text at least as many times as the user has specified as a Minimum Frequency

b) its frequency in the text when compared with its frequency in a reference corpus is such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a 'p value' specified by the user. (http://www.lexically.net/downloads/version5/HTML/index.html?keyness_definition.htm)

A keyword list includes items that are either significantly frequent (positive key words) or infrequent (negative key words), and is a useful starting point for many corpus linguistic analyses. A word which is *positively* key occurs *more* often than would be expected by chance in comparison with the reference corpus. A word which is *negatively* key occurs *less* often than would be expected by chance in comparison with the reference corpus. Table 5., illustrates the compare of keyness in specific items from two corpora:

Table 5

Positive Keywords in Sociology and History Texts (Adapted from Evsion, 2010, p. 127).

N	Key word	Freq.	%	RC Freq.	RC%	Keyness	P Value
1	social	372	0.5	229	0.02	1,269.90	0.000
2	p	394	0.53	294	0.03	1,258.83	0.000
3	class	259	0.35	159	0.01	884.6	0.000
4	society	222	0.3	179	0.02	688.54	0.000
5	women	263	0.35	341	0.03	658.91	0.000
6	power	209	0.28	269	0.03	525.51	0.000
7	archer	87	0.12	1		465.55	0.000
8	of	3,408	4.59	33,798	3.15	410.23	0.000
9	ibid	67	0.09	0		366.84	0.000
10	that	1,110	1.5	8,555	0.8	333.6	0.000

Note: RC = Reference Corpus.

Source: Data Extracted from BNC Sampler and BAWE Corpus.

In Table 5., the highest keyness values of four of the nouns (class, society, women, power) and the only adjective (social) reflect typical sociological or historical topics of the essays gathered from the British Academic Written English (BAWE) corpus of student writing is compared with a larger, more general corpus made up of five million words of written English (the reference corpus) extracted from the BNC Sampler. At first glance, however, the reason for the high keyness value of archer (a person who fires an arrow) is not apparent. In fact, this is a case where the analyst is likely to examine the item in context (by a concordance search for archer) in order to find some kind of explanation for its relatively high frequency. In this example, the examination shows that the key item is in fact ‘Archer’, a very commonly cited reference in a number of the essays. The statistical significance of the two abbreviations in the essay corpus (‘p’ and ‘ibid’) is also related to referencing: the convention of writing p for page number, and ibid to indicate reference to a previously cited work. The last two items in Table 5 are the grammatical items ‘of’ and ‘that’ which have strong associations with academic writing: the former because it is a constituent in postmodified noun phrases (e.g. the end **of** the Cold War) and the latter because of its multi-functionality – not only does that function as a subordinator, it also follows reporting verbs, often as part of it patterns such as it is reported 'that' (Evison, 2010, p.128).

Bondi (2010) claims that keywords are not necessarily a key to culture, however they can facilitate understanding of the main point of a text by constituting chains of repetition in text. Whether referring to words that are key to the intepretation of a text or key to the interpretation of a culture, the study of keywords has become central in corpus linguistics, especially through the development of techniques for the analysis of the meaning of words in context. "In a quantitative perspective, keywords are those whose frequency (or infrequency) in a text or corpus is statistically significant, when compared to the standards set by a reference corpus" (Scott 1997; Baker 2004; Scott & Tribble 2006, cited in Bondi, 2010, p.3).

2.3.3. Types of Corpora

Various kinds of machine-readable corpora can be used for linguistic analysis. Corpora can differ in a number of ways according to the purpose for which they were compiled, their representativeness, organization and format.

2.3.3.1. General Corpora

Some corpora are assembled simply to make available text base for unspecified linguistic research. Such corpora called *general corpora* which consist of a body of texts that linguists can analyse to seek answers to particular questions about different structures of the language as vocabulary, grammar or discourse. SEU corpus was an early example of a general corpus which has been used especially for grammar research. A general corpus is designed to be balanced, by including texts from different genres and domains of use as spoken or written, private or public, academic or general (Kennedy, 1998). It contains written texts such as newspapers and magazine articles, academic prose, works of fiction and non-fiction; as well as spoken transcripts from informal conversations to business meetings and government proceedings. General corpora or 'balanced' corpora sometimes refer to as 'core' corpora, which can be utilized as a basis for comparative studies.

The general corpora are the broadest type of corpus which is very large, approximately more than 10 million words. It contains the certain varieties of a particular language so that the findings from it may be 'generalized'. General corpora or so called 'generalized' corpora aims to represent a whole language as possible. Therefore, some national corpora such as British National Corpus (BNC) and American National Corpus (ANC) can be considered as generalized corpora.

A recently built general corpus project is the Turkish National Corpus (TNC) which was compiled by Aksan & Aksan at Mersin University. The project was started at 2009 and run until 2011; it approximately includes 50 million words collected in the period between 1990 and 2008. TNC consist of 95% written texts gathered from bestseller and prize-winner books, periodicals, newspapers and magazines and 5% spoken samples elicited from natural conversations and contexts-governed transcriptions.

TNC is considered as a mixed corpus since it contains both spoken and written texts, at the same time; it is accepted as a general corpus being not restricted to a particular genre or field. Indeed, TNC can be seen as a synchronic corpus features as it includes the imaginative and informative texts representing contemporary use Turkish language in the late twentieth century. Aksan and Aksan (2009) states that TNC aims to represent Turkish language in the most comprehensive and balanced way in order to provide relevant information for various types of research purposes (http://tnc.org.tr).

2.3.3.2. Specialized Corpora

Corpora designed for a particular research project are called *specialized corpora*. They are usually smaller in scale than general language corpora precisely because of their narrower focus. One important specialized area is that of academic English, and quite a few corpora have been created to serve the needs of practitioners of English for Academic Purposes (EAP). MICASE (the Michigan Corpus of Academic Spoken English; 1.8 million words) is a corpus of spoken English transcribed from about 190 hours of recordings of various speech events in a North American university (Simpson et al. 2003, cited in Lee, 2010, p.114). Topic of specialized corpora can be variable such as child language development (Carterette and Jones, 1974, in Kennedy, 1998) or the English used in petroleum geology exploration, drilling and refining (Zhu, 1989, in Kennedy, 1998).

2.3.3.3. Written and Spoken Corpora

Written corpus only contains texts that have been produced or published in written format. Written corpora may include traditional books, novels, textbooks, newspapers, magazines or unpublished letters and diaries. It can also include written texts that are produced electronically; for example, emails, bulletin board contributions and websites. The criteria for what exactly constitutes a written text can have grey areas as prepared speeches or television/film/radio scripts are probably better considered as written-to-be spoken texts. Written corpora generally tend to contain a higher number of conjunctions and prepositions than spoken data, suggesting longer, more complex sentences (Baker et.al., 2006). Brown corpus is both the first electronic corpus ever known and also an example of written corpus entirely consisting of various kinds of written texts. Also Hong Kong University of Science and Technology (HKUST) corpus also contains only samples from texts books in computer science.

Spoken corpus consists entirely of transcribed speech and can be gathered from a range of sources: spontaneous informal conversations, radio phone-ins, meetings, debates, classroom situations, meetings, etc. Spoken corpora can present problems to traditional taggers due to repetitions, false starts, hesitations, vocalisations and interruptions that occur in spontaneous speech. London-Lund Corpus is the first electronic spoken corpora followed by other spoken corpora as The Lancester/IBM Spoken English. When compared to written corpora, spoken corpora tend to have a higher proportion of pronouns (particularly first and second person) and discourse markers (Baker et. al, 2006). However, Biber (1988) showed that some spoken and written genres are considerably similar to each other (for example personal letters and face-to-face conversations) in terms of frequencies of certain linguistic phenomena.

2.3.3.4. Synchronic and Diachronic Corpora

Synchronic corpus is “A corpus in which all of the texts have been collected from roughly the same time period, allowing a ‘snapshot’ of language use at a particular point in time” (Baker et.al., 2006, p. 153). Therefore, it can be used for comparing language varieties. Examples of synchronic corpora are: ICE which was specifically designed for the synchronic study of world Englishes; next is Longman Spoken American Corpus which can be used to compare regional dialects in the USA; and

Linguistic Variations in Chinese Speech Communities (LIVAC) corpus is another synchronic corpus. METU Turkish Corpus (Say et. al., 2004) is an example of synchronic corpus as it contains Turkish samples of written texts collected from newspapers, articles and books that published after 1990. (http://www.ii.metu.edu.tr/tr/research_group/metu-turkish-corpus-project)

On the other hand, d*iachronic corpus* is a corpus which is carefully built in order to be representative of a language or language variety over a particular period of time, so that it is possible for researchers to track linguistic changes within it. Diachronic Corpus of Present-day Spoken English (DCPSE) which was constructed at the Survey of English Usage, University College London by a team led by Bas Aarts. The corpus includes spoken corpus data drawn from both the London–Lund Corpus and the spoken section of the British International Corpus of English (ICE) corpus in order to develop a diachronic corpus of relatively contemporary spoken English covering a period of a quarter of a century or so from the 1960s and early 1990s. (http://www.ucl.ac.uk/english-usage/diachronic).

2.3.3.5. Historical Corpora

Historical corpus is one which is intentionally created to represent and investigate past stages of a language and/or to study language change. In all other respects, the defining characteristics of a corpus apply: it is a finite electronic collection of texts or parts of texts by various authors which is based on well-defined and linguistically relevant sampling criteria and aims for some degree of representativeness. A historical corpus concerns periods before the present-day language, which may be taken to end roughly thirty to forty years (one generation) before the present: in other words, any corpus compiled in or around 2000 that goes back beyond ca. 1960/1970 can be called historical (Claridge, 2008).

There are three main collections of historical English that cover a wide span of time and genres: the diachronic part of the Helsinki Corpus of English, ARCHER (A Representative Corpus of Historical English Registers), and COHA (Corpus of Historical American English). The Helsinki Corpus (1.6 million words) covers the period from around 750 to 1700, and thus spans Old English (413,300 words), Middle English (608,600 words) and early modern (British) English (551,000 words). ARCHER is a multi-genre corpus (currently 1.8 million words) covering the early modern English period right up to the present (1650–1990) for both British and American English. It is divided into fifty-year blocks to facilitate comparisons (though not all periods are available for American English). ARCHER is, at the time of writing, undergoing correction, expansion and tagging. The corpus is not publicly available, but the several universities involved in the project are willing to host visits by interested scholars. COHA's aim is to create a 300-million-word corpus of historical American English covering the early 1800s to the present time, and is 'balanced' in each decade for the genres of fiction, popular magazines, newspapers and academic prose (Lee, 2010).

2.3.3.6. Multilingual Corpora

A *parallel corpus* consists of two or more corpora that have been sampled in the same way from different languages. The prototypical parallel corpus consists of the same documents in a number of languages, which is a set of texts and their translations. Since official documents (technical manuals, government information leaflets, parliamentary proceedings etc.) are frequently translated, these types of text are often found in parallel corpora. The Corpus Resources and Terminology Extraction (CRATER) corpus consisting of French, English and Spanish texts from telecommunication domain is an example of this type of corpus (Baker, Hardie and McEnery, 2006).

Translation corpora, which is a sub-type of parallel corpus contains 'original texts and their translations into one or more other languages. Another type of parallel corpus, called *comparable corpus*, consists of different texts in each language: it is merely the sampling method that is the same. For example, the corpus might contain 100,000 words of fiction published in a given timeframe for each language. Johnsson (2007) explains that parallel corpus is reserved for *bidirectional translation corpora*, a combination of translation corpora and comparable corpora that use the same framework (i.e. comparable originals in at least two languages plus their translations into the other language(s) (cited in Lee, 2010, p. 119). In parallel corpora, the two components are aligned on a paragraph-to-paragraph or sentence-to-sentence basis. The English–Norwegian Parallel Corpus (ENPC) and English– Swedish Parallel Corpus (ESPC) are good examples of a parallel bidirectional corpus: each corpus has four related components, allowing for various types of comparison to be carried out (Lee, 2010).

On the other hand, Aijmer (2008) states that parallel corpus under the term *multilingual corpus* as a sub-type with comparable corpora as presented in Figure 13:

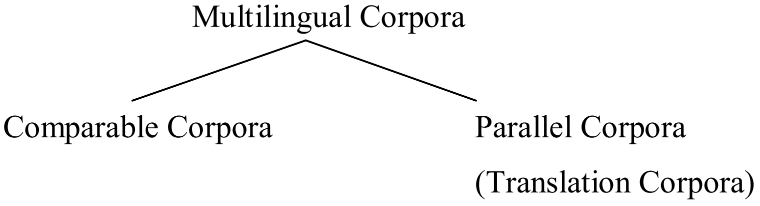


Figure 13. Types of multilingual corpora (adapted from Aijmer, 2008, p.276)

Aijmer defines multilingual corpora as:

> There are two types of multilingual corpora. A fundamental distinction is that between parallel and comparable corpus. Parallel corpora consist of a source text and its translation into one or more languages. They can be further characterised in terms of the direction of the translation. If a corpus consists of e. g. English texts translated into Swedish, it is unidirectional; on the other hand, if this corpus also contains translations into English, it is bidirectional. Very often parallel corpora are aligned, either by sentence or by word. A comparable corpus on the other hand does not contain translations but consists of texts from different languages which are similar or comparable with regard to a number of parameters such as text type, formality, subject-matter, time span, etc.
>
> (2008, p.276)

As examples for multilingual corpora, the Canadian Hansard Corpus which is a parallel French-English texts proceedings from Canadian parliament (750,000 words) and The Aarhus of Contract Law corpus including three sub-corpora of Danish, English and French texts from the area of contract law(1 million words).

2.3.3.7. Monitor Corpora

Meyer (2002) defines monitor corpus, as “ a large corpus that is not static and fixed but that is constantly being updated to reflect the fact that new words and meanings are always being added to English” (p.15). Monitor corpus, or ‘dynamic corpus’ is continually growing over time and constantly (annually, monthly or even daily) supplemented with fresh materials and keeps increasing in size (McEnery, Xiao, Tono, 2006). The Bank of English (BoE) is the best known example of monitor corpus.

Actually, the monitor corpus forms the philosophy of the Collins COBUILD Project at Birmingham University in England, which has produced a number of dictionaries based on two monitor corpora: the Birmingham Corpus and the Bank of English Corpus. The Birmingham Corpus was created in the 1980s (cf. Renouf 1987 and Sinclair 1987), and while its size was considered large at the time (20 million words), it would now be considered fairly small, particularly for the study of lexical items. Because of this, the Birmingham Corpus has been superseded by the Bank of English Corpus, which was recently totaled 650 million words (Meyer, 2004).

2.3.3.8. Learner Corpora

Corpus-based language studies conducted over the last two or so decades have led to much better descriptions of many of the different registers like informal conversation, formal speech, journalese, academic writing, sports reporting and dialects of native English as British English vs American English; male vs. female language and others. On the other hand, investigations of non-native varieties have been a relatively recent prologue because it was not until the late 1980s and early 1990s that academics and publishers started collecting corpora of non-native English, which have come to be referred to as *learner corpora* (Granger, 2002).

Computer learner corpora (CLC) are “electronic collections of authentic Foreign/Second language textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardized and homogeneous way and documented as to their origin and provenance” (Granger, 2002, p. 7). In other words, learner corpora are foreign language learners’ computerized representations of their L2 performance or output, usually written. Granger (2008, p.337) points out that:

> Analyzing learner language is a key component of second and foreign language education research and serves two main purposes: it helps researchers better understand the process of second language acquisition (SLA) and the factors that influence it, and it is a useful source of data for practitioners who are keen to design teaching and learning tools that target learners' attested difficulties.
>
> (2008, p.337)

CLC research enables to investigate learner language through particularly designed corpora that may give insights in respect of both SLA and FLT purposes. Granger (2004b) characterizes CLC research as "a new research enterprise, a new way of thinking about ***learner*** language, which is challenging some of our most-deeply rooted ideas about ***learner*** language." (p.123).

The major purpose of compiling a learner corpus is to gather natural interlanguage data for describing and analysing learner language (Granger, 1998). Leech (1998) also points out that a large-scale and well-assembled database of learners' language should prove to be a very useful resource to SLA researchers and educators. Therefore, modern learner corpora are often connected with interlanguage analysis. A number of learner corpora worldwide have been established during the past two decades, such as the International Corpus of Learner English (ICLE)) (see section 2.3.3.8.1.) and Longman Learners Corpus (LLC).

Granger (2008) points out that learner corpora fall into two major categories as commercial learner corpora, initiated by major publishing companies, and academic learner corpora, which are compiled in educational institutions. Since there are more academic than commercial corpora, commercial corpora tend to be much larger and have a wider range of mother tongue backgrounds. For instance, in English language, there are two major commercial learner corpora as the Longman Learners' Corpus and the Cambridge Learner Corpus, both of which contain over 10 million words and represent numerous mother tongue backgrounds. On the other hand, academic corpora, come in all shapes and sizes and usually cover learners from only one mother tongue background, for example, the International Corpus of Learner English (ICLE) is the most notable exception in this respect (Granger, 2008). As for spoken interlanguage corpus, Louvain International Database of Spoken English Interlanguage (LINDSEI) consists of 1.1 million words of transcripts from speech by speakers from eleven different L1 backgrounds.

2.3.3.8.1. International Corpus of Learner English (ICLE)

The ICLE project is the probably best known learner corpora which contain approximately three million words of essays written by foreign language learners of English from sixteen different mother tongues. In addition to allowing the comparison of the l2 wrings of learners from different L1 backgrounds, the corpus can be used in combination with the Louvain Corpus of Native English (LOCNESS) to compare native and learner English (as applied in this present study). Can (2010) states that ICLE not only provides a large-scale compare opportunity for SLA investigations, but it also serves a reliable database for interlanguage error analysis.

Turkish sub-corpus of ICLE, namely TICLE, was compiled by Can & Kilimci (2009) and included to the second version of ICLE in 2009. In the compilation process of TICLE, previous ICLE version (2002) sub-corpora design criterion have been considered in respect of learner sub-corpus of ICLE (see chapter three for a detailed description of ICLE corpora design).

2.3.3.8.2. Learner Corpora Analysis

Linguistic exploitation of learner corpora usually involves one of the following two methodological approaches: Contrastive Interlanguage Analysis (CIA) and Computer-aided Error Analysis (CEA). The first method is contrastive, and consists in carrying out quantitative and qualitative comparisons between native (NS) and non-native (NNS) data or between different varieties of non-native data. The second focuses on errors in interlanguage and uses computer tools to tag, retrieve and analyze them (Granger, 2002). Granger et al. (2009) claims that the differences between non-native (NNS) and native (NS) varieties are twofold:

- L2 varieties have a much higher error density than L1 varieties;
- the two varieties display major differences in frequency patterns, with some linguistics entities being overused and other underused. (p.40)

Granger et al. (2009) suggests CIA and CEA methods to undercover these differences.

2.3.3.8.2.1. Contrastive Interlanguage Analysis (CIA)

The contrastive interlanguage analysis (CIA) methodology involves comparing learner data with native speaker data (L2 vs. L1) or comparing different types of learner data (L2 vs. L2) (Granger 1996). That is, CIA compares varieties of one and the same language: either NS and NNS varieties (L1 vs. L2) or different NNS varieties (L2 vs. L2). Figure 14 illustrates CIA:

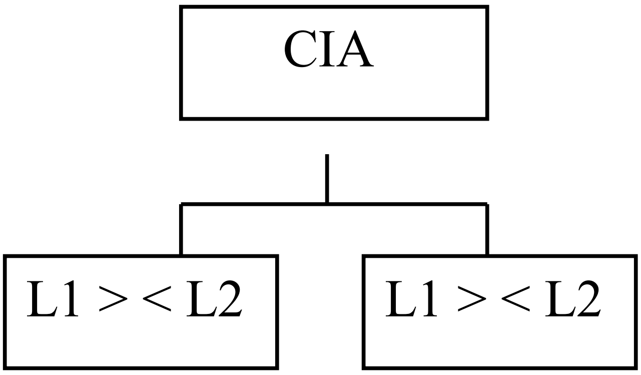
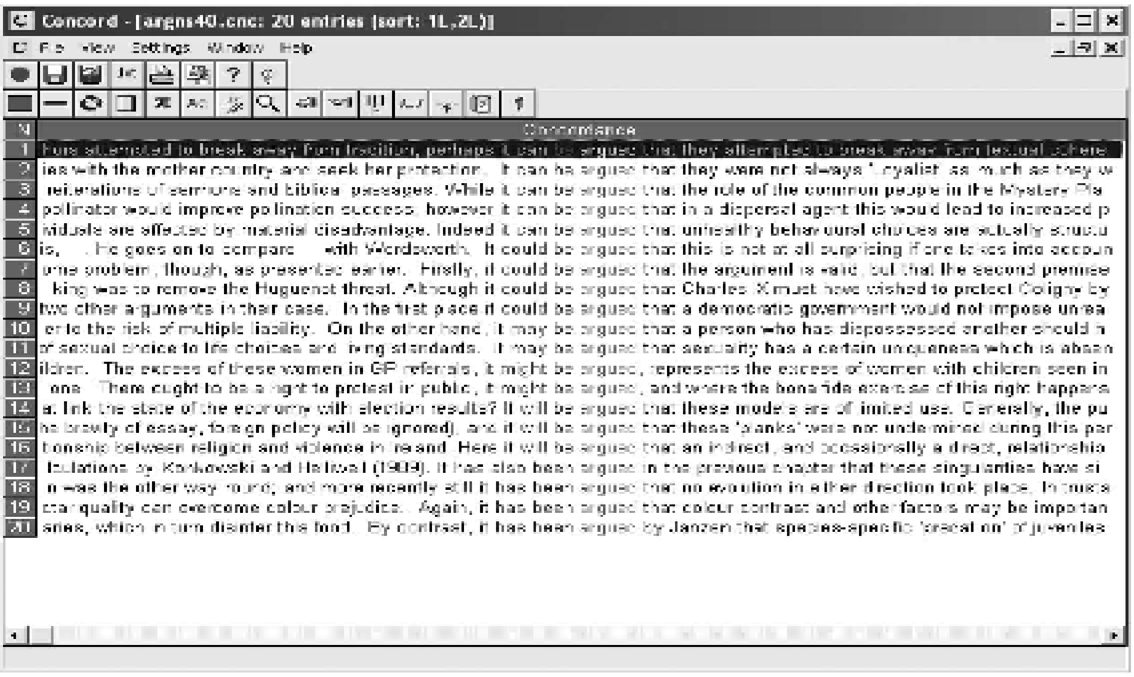


Figure 14. Contrastive interlanguage analysis (adapted from Granger et al., 2009, p.40).

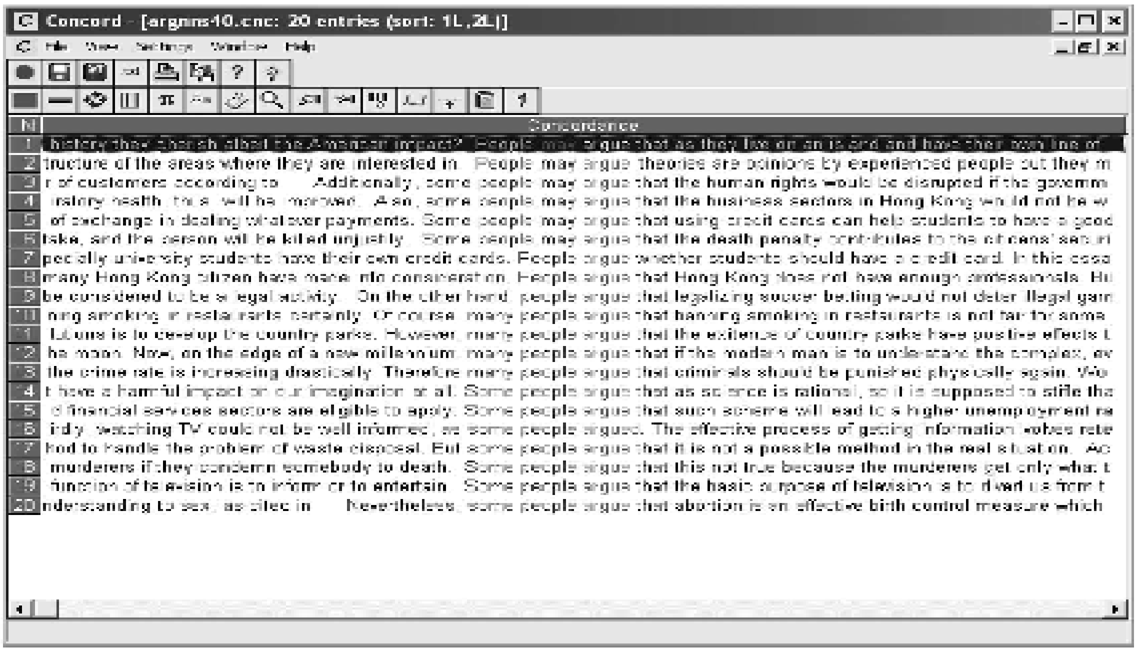
A lot of learner corpus studies to date have used this approach to investigate a wide range of topics, some of which _ like high frequency vocabulary, modals, connectors and phraseological units _ have received particular attention (Granger, 2002). Most of these CIA studies are based on unannotated learner corpora, some of them have used POS-tagged corpora to compare the frequency of grammatical categories or sequences of grammatical categories in native and learner corpora. Granger (2008) states that these studies have made it possible to bring out the words, phrases, grammatical items and syntactic structures that are over- or underused by learners and which therefore contribute to the foreign-soundingness of perhaps otherwise error-free advanced interlanguage.

A typical CIA investigation focuses on one particular linguistics phenomenon (e.g. modals, verbs) and has the automatic extraction of all occurences of that phenomenon from NS and NNS corpus through a software such as WordSmith Tools (Scott, 1999). Once the occurrences have been retrieved the results can be counted and sorted in various ways to allow significant patterns to emerge. For example, in Figure 15 and Figure 16, the typical patterns of the use of the verb 'argue' has been compared in a large academic corpus of English and a corpus of EFL learner writing:



N	Concordance
1	tors attempted to break away from tradition, perhaps it can be argued that they attempted to break away from textual cohere
2	ies with the mother country and seek her protection. It can be argued that they were not always 'Loyalist' as much as they w
3	reiterations of sermons and biblical passages. While it can be argued that the role of the common people in the Mystery Pla
4	pollinator would improve pollination success, however it can be argued that in a dispersal agent this would lead to increased p
5	viduals are affected by material disadvantage. Indeed it can be argued that unhealthy behavioural choices are actually structu
6	is, . He goes on to compare with Wordsworth. It could be argued that this is not at all surprising if one takes into accoun
7	one problem, though, as presented earlier. Firstly, it could be argued that the argument is valid, but that the second premise
8	king was to remove the Huguenot threat. Although it could be argued that Charles IX must have wished to protect Coligny by
9	two other arguments in their case. In the first place it could be argued that a democratic government would not impose unrea
10	ere to the risk of multiple liability. On the other hand, it may be argued that a person who has dispossessed another should h
11	of sexual choice to life choices and living standards. It may be argued that sexuality has a certain uniqueness which is absen
12	ildren. The excess of these women in GP referrals, it might be argued, represents the excess of women with children seen in
13	one. There ought to be a right to protest in public, it might be argued, and where the bona fide exercise of this right happens
14	at link the state of the economy with election results? It will be argued that these models are of limited use. Generally, the pu
15	he brevity of essay, foreign policy will be ignored), and it will be argued that these 'planks' were not undermined during this per
16	tionship between religion and violence in Ireland. Here it will be argued that an indirect, and occasionally a direct, relationship
17	lculations by Konkowski and Helliwell (1993). It has also been argued in the previous chapter that these singularities have si
18	n was the other way round; and more recently still it has been argued that no evolution in either direction took place. In contra
19	star quality can overcome colour prejudice. Again, it has been argued that colour contrast and other factors may be importan
20	aries, which in turn disinter this food. By contrast, it has been argued by Janzen that species-specific 'predation' of juveniles

Figure 15. A screenshot of the Verb '*argue*': Concordance lines from a native corpus of academic writing in wordsmith tools (adapted from Granger, 2008, p. 268).



N	Concordance
1	history they cherish about the American impact? People may argue that as they live on an island and have their own line of
2	tructure of the areas where they are interested in. People may argue theories are opinions by experienced people but they m
3	r of customers according to . Additionally, some people may argue that the human rights would be disrupted if the governm
4	ratory health, thus will be improved. Also, some people may argue that the business sectors in Hong Kong would not be w
5	of exchange in dealing whatever payments. Some people may argue that using credit cards can help students to have a good
6	take, and the person will be killed unjustly. Some people may argue that the death penalty contributes to the citizens' securi
7	pecially university students have their own credit cards. People argue whether students should have a credit card. In this essa
8	many Hong Kong citizen have made into consideration. People argue that Hong Kong does not have enough professionals. Hu
9	be considered to be a legal activity. On the other hand, people argue that legalizing soccer betting would not deter illegal gam
10	ning smoking in restaurants certainly. Of course, many people argue that banning smoking in restaurants is not fair for some
11	ious is to develop the country parks. However, many people argue that the existence of country parks have positive effects t
12	he moon. Now, on the edge of a new millennium, many people argue that if the modern man is to understand the complex, ev
13	the crime rate is increasing drastically. Therefore many people argue that criminals should be punished physically again. Wo
14	t have a harmful impact on our imagination at all. Some people argue that as science is rational, so it is supposed to stifle the
15	d financial services sectors are eligible to apply. Some people argue that such scheme will lead to a higher unemployment ra
16	idly, watching TV could not be well informed, as some people argued. The effective process of getting information solves rele
17	hod to handle the problem of waste disposal. But some people argue that it is not a possible method in the real situation. Ac
18	murderers if they condemn somebody to death. Some people argue that this not true because the murderers get only what t
19	function of television is to inform or to entertain. Some people argue that the basic purpose of television is to divert us from t
20	nderstanding to sex, as cited in . Nevertheless, some people argue that abortion is an effective birth control measure which

Figure 16. A screenshot of the Verb '*argue':* Concordance lines from a learner corpus of academic writing in wordsmith tools (adapted from Granger, 2008, p. 268).

A useful function of WordSmith Tools is the KeyWords function which compares all the words in two corpus such as a NS corpus list and NNS corpus list. Then it reports all the those which appear significantly more often in one than in the other (Granger et al., 2009).

2.3.3.8.2.2. Computer-aided Error Analysis

Computer-aided error analysis (CEA) basically involves analyzing learner errors on the basis of learner corpora in which error tags and possible corrections have been inserted with the help of a purpose-built editing tool. CEA origins dates back to 1970s, the era of Error Analysis which is the classic method of analyzing learner errors. EA was criticized due to the data use: the material explored was broadly associated with lists of errors gathered from elicited practices where little attention was paid to task, learner variables and also heterogeneous and non-natural data (Ellis, 1994, in Granger, et. al., 2009). However, CEA differs from previous EA studies in some major respects, not least of which is the fact that errors are not isolated from the texts in which they originated, as was the case in traditional Error Analysis. Rather, studied are usually based on the correct use and over- and underuse.

CEA usually based on one of the following two methods. The first simply consists in selecting an error-prone linguistic item (word, phrase, word category, syntactic structure) and scanning the corpus to retrieve all instances of misuse of the item with the help of standard text retrieval software tools. The second method is more time-consuming but also much more powerful in that it may lead the analyst to discover learner difficulties of which he was not aware. The method consists in devising a standardized system of error tags and tagging all the errors in a learner corpus or, at least, all errors in a particular category (for instance, verb complementation or modals). The error tagging process can be greatly helped by the use of an error editor and, more importantly, once the work has been done and researchers are in possession of a fully error-tagged corpus, the range of possible applications that can be derived from it is absolutely huge.

This approach has led to a much more limited number of publications than CIA, partly due to the difficulty of error annotation and the investment of time it involves, but also to the unpopularity of error analysis within the SLA community and a more general rejection of the notion of error in SLA and FLT. Recent years, however, have seen a revival of interest in error analysis, especially in pedagogical lexicography and language testing (Granger, 2008).

Granger (2002) claims that EA often causes negative reactions: it is felt to be retrograde, a return to the old days when errors were considered to be an entirely negative aspect of learner language. However, analyzing learner errors is not a negative enterprise: on the contrary, it is a key aspect of the process which takes us towards understanding interlanguage development and one which must be considered essential within a pedagogical framework. Moreover, teachers and materials designers need to have much more information about what learners can be expected to have acquired by what stage if they are to provide the most useful input to the learners, and analyzing errors is a valuable source of information (Granger, 2002).

2.3.3.8.3. Studies on Learner Corpora

In the field of corpus linguistics, learner corpora investigations have special considerations differing from other corpus research areas because of their relevance to language learning theory and practice (Granger, 2004b). Learner corpora are often used for studies that compare learners' language with that of native speakers. These studies can guide researchers for the areas of difficulties for learners or sources of errors made in L2 production. On the other hand, studies of learner corpora also often include comparisons of different learner groups from different L1 backgrounds. Therefore, common interlanguage patterns or L1 effects on L2 can be examined with such comparisons. That is to say, in order to identify which grammatical structures, lexis or discourse items are underused or overused by students in academic writing, findings from a learner or non-native speaker corpus are either usually compared with a parallel corpus of native speaker writing or sometimes with a larger reference corpus of expert writing.

To date, several learner corpus studies have been carried out in various topics in respect of learner corpora. The analytical and methodological issues in learner corpora were considered by Leech (1998), Granger (1998, 2002). Methods of analysis in learner corpora have also regarded by researches as Meunier (1998), Mönnink (2000) and Rooy & Schäfer (2003) (in Granger, 2002). However, the majority of learner corpus work are usually based on contrastive approach as in CIA methodology (see section 2.3.3.8.2.1.) that enables to compare non-native speakers and native speakers (L2 vs. L1) or non-native speakers and other non-native speakers (L2 vs. L2). For instance, Ringbom (1998), Altenberg (2002) studied vocabulary; Aijmer (2002); McEnery and Kifle (2002) investigated modals; De Cock (1998), Lorenz (1998) dealt with intensifiers, Granger (1998) and Nesseshauf (2005) looked at prefabs and collocations over learner corpora in a contrastive perspective. Connectors (see chapter three) were also examined by mostly comparing with native and different non-native speakers (Milton and Tsang 1993; Field & Yip, 1993; Granger and Tyson 1996; Altenberg and Tapper 1998; L. Flowerdew 1998b). In addition, grammatical categories or sequences were regarded by Granger and Ryson (1998) and Tono (2000) and stance adverbials by Can (2012). Some of the important learner corpus studies based on comparison are displayed in Table 6.;

Table 6

Studies on Learner Corpora

Study	**Method**	**Corpora**		**Result**
		NNS	**NS**	
Expressions of doubt and certainty (Hyland & Milton, 1997)	L2vs.L1	HKUST Cantonese L1	British L1 (Cambridge (exam scripts)	-difficulty in certainty and doubt in NNS
Vocabulary (Ringbom, 1998)	L2vs.L1	ICLE French, Spanish German, Dutch Finnish, Fin-Swedish Swedish	LOCNESS	-Transfer errors in NNS -Underuse in Finnish L1
Direct Questions (Virtanen, 1998)	L2vs.L1	ICLE French Spanish German Dutch Fin-Swedish	LOCNESS	-Lower Freq. in NS -Higher freq. in Fin-Swedish L1
Adverbial Connectors (Altenberg & Tapper, 1998)	(L2vs.L1) (L2vs.L2)	ICLE Swedish French	LOCNESS	-Overall underuse in NNS -Transfer traces

Table 6 (continued)

Demostratives (Petch-Tyson, 2000)	 (L2vs.L1)	ICLE French Finnish Dutch Swedish	LOCNESS	-Underuse in most of NNS
Morphemes (Tono, 2000) (Replication of Dulay & Burt, 1974)	Case Study	LLC Japanese	-	- later acquisition of plural –**s** -earlier acquisition of possessive –**s** (than previous study)
Modality (Aijmer, 2003)	 (L2vs.L1) (L2vs.L2)	ICLE Swedish Swedish French German	LOCNESS	-Overall overuse in NNS -speech-like patterns in Swedish L1
Stance Adverbials (Can, 2012)	 (L1 vs. L2)	 ICLE Turkish	 LOCNESS	 -less variety of stance adverbials with more Frequency against NS

NS=Native Speakers ; NNS= Native Speakers
ICLE= International Corpus of Learner English
LLC= Longman Learner Corpus
HKUST= Hong Kong University of Science and Technology Learner Corpus
LOCNESS= Louvain Corpus of Native English Essays

There are few works in learner corpora studied by CEA (see section 2.3.3.8.1.2.) method focusing on L2 errors. For example, lexical errors were studied by Chi Man-lai et al. (1994), Källkvist (1995) and Lenko-Szymanska (2003) and tense errors by Granger (1999) (in Granger, 2002).

Flowerdew (2001) points out that although a significant amount of research has been carried out in recent years on learner corpora, not many of the findings have been applied directly to pedagogy and remained at the level of implications. Flowerdew also suggests that the findings from comparative studies of learner and expert corpora could be applied to materials design as has been the case in the compilation of dictionaries for non-native speakers. For instance, many of the entries in the *Longman Dictionary of Common Errors* (2nd ed.) (Turton and Heaton 1996) are based on a comparison of usage in the *BNC* and the Longman Learner Corpus (in Flowerdew, 2001, p.366).

To conclude, in learner corpus research, "a great deal of spadework has to be done before the research results can be harvested" (Leech, 1998, p. xvii). Moreover, Granger (2004) states that many efforts have been made by researcher who spared a great amount time of building and analyzing learner corpora. However, there are needs for wider wider range of learner corpora (in particular, ESP, speech and longitudinal data) with more elaborate processing (POS-tagging and error-tagging). Thus, results need to be interpreted in the light of current SLA theory and incorporated in syllabus and materials design. Because, CLC have the potential the gap between SLA and ELT. Although ELT community joined the corpus inquiry more quickly and enthusiastically than SLA, there are signs that the value of CLC data has begun to recognize by SLA researchers. Because, the size and representativeness of CLC data can help them validate their hypotheses and indeed formulate new ones (Granger, 2004).

2.4. Corpus-based Research in Language Learning and Teaching

Over the past two decades, corpora and corpus evidence have not only been used in linguistic research but also in the teaching and learning of languages. Applications based on corpus investigation are found in a number of different areas, for example lexicography, translation, stylistics, grammar, gender studies, forensic linguistics, computational linguistics and, equally importantly, in language learning and teaching (Tognini-Bonelli 2001). In the early 1990s, there has been an increasing interest in applying the fining of corpus-based research to language pedagogy-teaching and learning of languages. When looking at the literature, large number of works which covers the issues related to corpora in language pedagogy have been produced (Mindt, 1996; Witchmann et al., 1997; Leech, 1997; Ketteman and Marko, 2000; Aston 2001; Hunston, 2002; Granger et al. 2002; Sinclair, 2004a; Aston at al., 2004; Nesselhauf, 2005; Scott and Tribble, 2006). In addition, there is now a wide range of fully corpus-based reference works (such as dictionaries and grammars) available to learners and teachers, and a number of dedicated researchers and teachers have made concrete suggestions on how concordances and corpus-derived exercises could be used in the language teaching classroom, thus significantly "[e]nriching the learning environment" (Aston 1997, cited in Römer, 2008, p. 112). Corpus approach made it possible to conduct studies with more data and more variables than was previously feasible, and to design new kinds of classroom activities that actively engage learners in the analysis of language. The contribution of corpus linguistics to foreign language teaching is related to the importance that it provides an empirical study of large databases of language (Conrad, 2005). Chambers (2005) claims that:

> The advances in the direct access to corpora by language teachers and learners have created the need to research into a number of pedagogic issues, including 'the types of corpora to be consulted, large or small, general or domain-specific, tagged or untagged'; the kinds of learning strategies to benefit from direct corpus consultation; and the means by which direct access to corpora can be integrated into the language learning context.
>
> (2005, cited in Cheng, 2010, p. 319)

In respect of Second Language Acquisition (SLA) and Foreign Language Teaching (FLT), Granger (2002) states that previous mainstream language teaching approaches great efforts have been made to improve the description of the target language. There has been an increased interest in learner variables, such as motivation, learning styles, needs, attitudes, etc., and our understanding of both the target language and the learner has contributed to the development of more efficient language learning tasks, syllabuses and curricula. Mark (1998) illustrates the mainstream interests in language teaching as in Figure 17.:

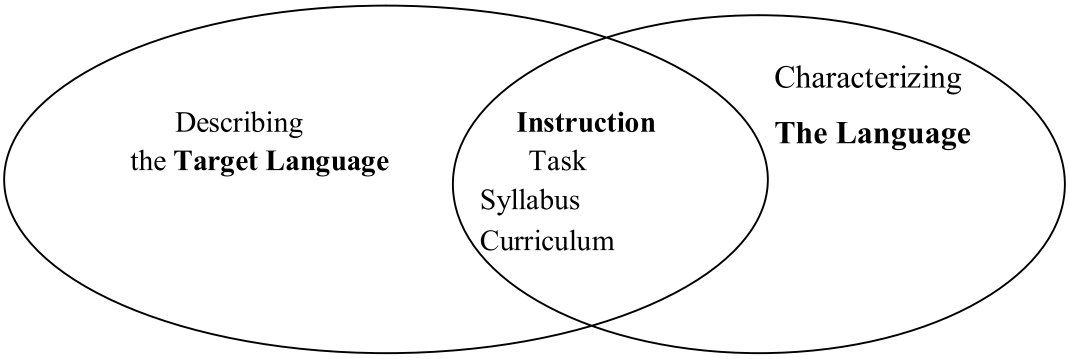


Figure 17. The concern of mainstream language teaching (Mark, 1998, adapted from Granger, 2002, p. 6)

The missing component was the 'learner output' which Mark (1984) emphasizes "it simply goes against common sense to base instruction on limited learner data and to ignore, in all aspects of pedagogy from task to curriculum level, knowledge of learner language"(cited in Granger, 2002, p.6). Granger (2002) emphasizes that the gradual attention of SLA and FLT communities is now turning towards learner corpora and the types of descriptions and insights they have the potential to provide. Figure 18. presents the focus on learner output (Mark, 1998):

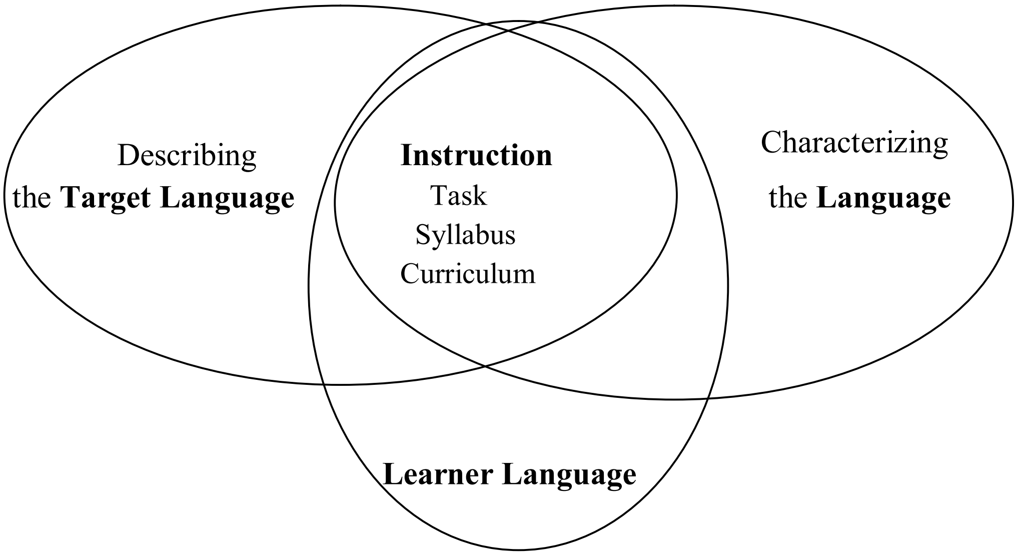


Figure 18. The focus on learner output (Mark, 1998, adapted from Granger, 2002, p. 7)

Granger (2002) then concludes as "it is to be hoped that learner corpora will contribute to rehabilitating learner output by providing researchers with substantial sources of tightly controlled computerized data which can be analyzed at a range of levels using increasingly powerful linguistic software tools" (2002, p.7).

It is assumed that corpus linguistics and language teaching are very closely related fields (Römer 2008). This relationship is very dynamic so that two fields greatly influence each other. As presented in Figure 19 below, while language profits from the resources, methods, and insights provided by corpus linguistics, it also provides important impulses that are taken up in corpus linguistic research.

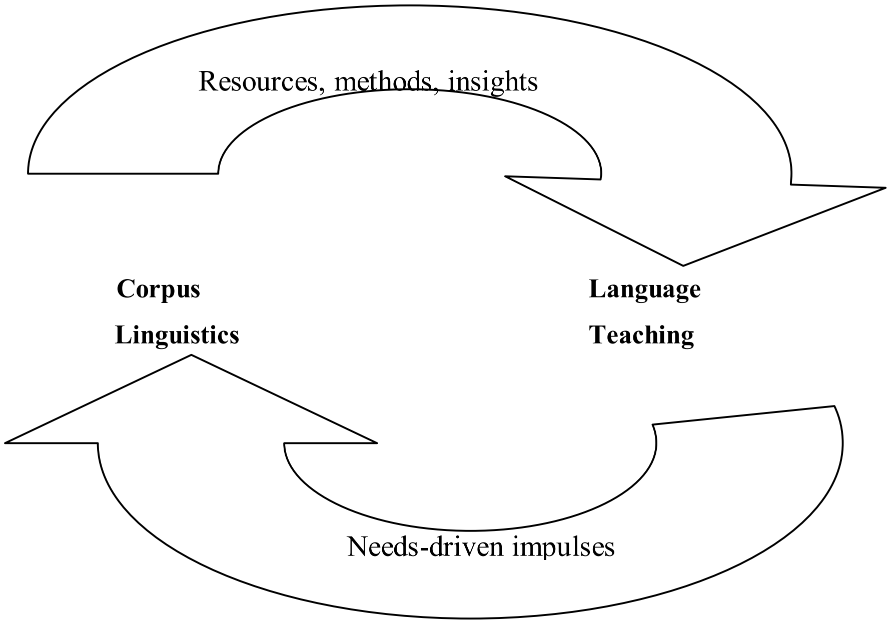


Figure 19. The relationship between corpus linguistics and language teaching (adapted from Römer, 2008, p. 113).

The relationship between corpus linguistic and language teaching is illustrated by Römer (2008) as in Figure 20. The requirements of language teaching have a pact on research projects in corpus linguistics and on the development of suitable resources and tools (Römer, 2008).

2.4.1. The Applications of Corpora in Language Learning and Teaching

The types of applications of corpora includes both the use of corpus tools such as text collections and software and corpus methods like various corpus analytic techniques. The classification of these corpus applications leads a distinction of the concepts as 'direct' and 'indirect' applications. On the other hand, Fligelstone (1993) describes three aims of corpus-based linguistics in teaching: teaching about (the principles and theory behind the use of corpora), teaching to exploit (the practical, methodological aspects of corpus-based work), and exploiting to teach (using corpora to derive or drive teaching materials) (cited in Cheng, 2010, p. 319). In another perspective, Leech (1997) notes that there is a convergence between corpora and language teaching. That convergence has three focuses: the direct use of corpora in teaching (teaching about, teaching to exploit and exploiting to teach), the indirect use of corpora in teaching (reference, publishing, material development and language testing).

That means, 'indirectly' corpora can help with decisions about what to teach and when to teach it, but that they can also be accessed 'directly' by learners and teachers in the language teaching classroom, and so "assist in the teaching process" (Fligelstone 1993, 98, cited in Römer, 2008, p. 114), thus affecting how something is taught and learnt. The third and the less central component is 'further teaching oriented corpus development (LSP, L1 developmental corpora, L2 learner corpora). The direct and the indirect applications of corpus linguistics on language learning and teaching is illustrated in Figure 20:

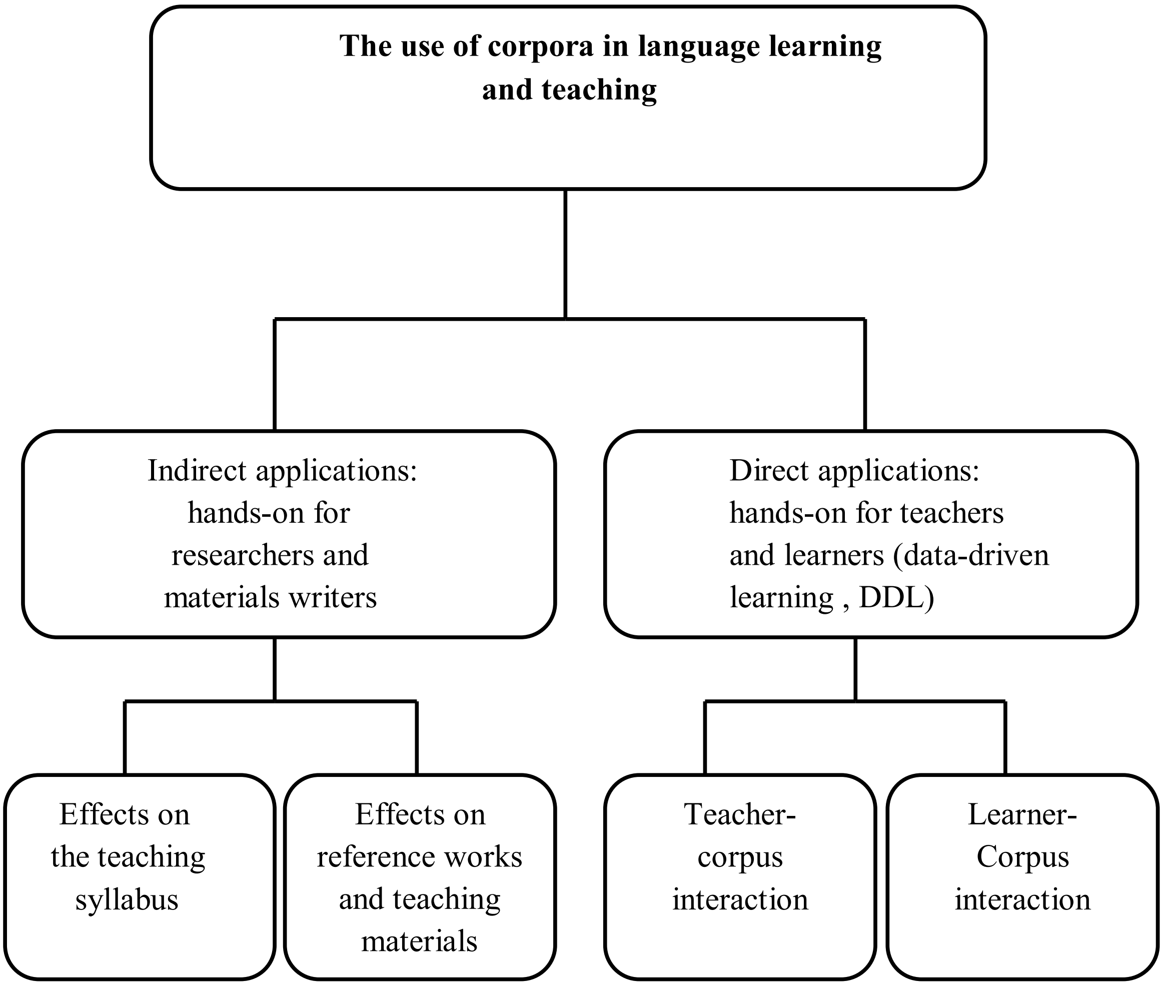

The use of corpora in language learning and teaching
Indirect applications: hands-on for researchers and materials writers
Direct applications: hands-on for teachers and learners (data-driven learning , DDL)
Effects on the teaching syllabus
Effects on reference works and teaching materials
Teacher-corpus interaction
Learner-Corpus interaction

Figure 20. Applications of corpora in language teaching (adapted from Römer, 2008, p 113)

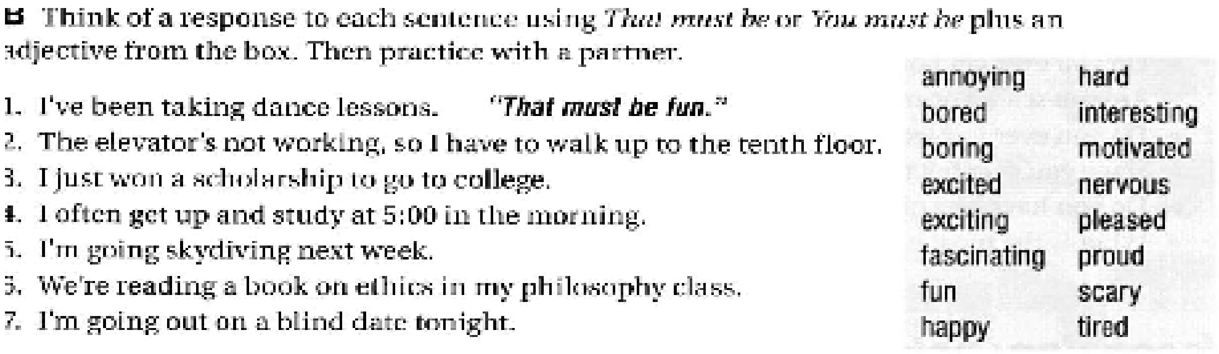
As presented in Figure 20, indirect and direct applications of corpora serve a wide range of opportunity for language learning and teaching. In terms of indirect applications, corpora can provide invaluable resource for teaching language syllabus that oriented communicative competence with real life communication situations that learner may encounter. For example, corpus-driven designed Collins COBUILD English course project (Sinclair, 1987) is the most prominent resource for language teaching syllabi including ‘corpus-driven lexical syllabus in which ‘the commonest words and phrases in English and their meanings’ (Römer, 2008). ‘Touchstone’ series (McCarthy et al. 2005) is another example of teaching materials based on corpus evidence and demonstrates how everything from syllabi to textbook examples can be informed entirely by corpus data (in Cheng, 2010). In Touchstone, various kinds of examples gathered from Cambridge International Corpus (CIC) are used as authentic samples. McCarten (2010) explains that “By observing in a corpus how people choose their language according to the situation they are in and the people they are with, the course-book writer can select appropriate, realistic and typical contexts in which to present and practise grammatical structures or vocabulary” (p. 421). An example from grammar is the use of ‘must’ in conversation, which is used frequently to express speculation rather than obligation and is usually found in responses and reactions to what the speaker hears or sees. A spoken section from CIC is presented in Example 9.:

Example 9 (Woman talking about business travel to a stranger)

Speaker 1: They put me up in hotels and things.
Speaker 2: Oh that’s nice. That’s always fun.
Speaker 1: It’s not too bad.
Speaker 2: Yeah. It **must** be tough though. Moving back and forth a lot.
(Cambridge International Corpus, North American English conversation)

(adapted from McCarten, 2010, p.421)

This real life example that contains modal structure can be practiced in examples such as in Figure 21.:



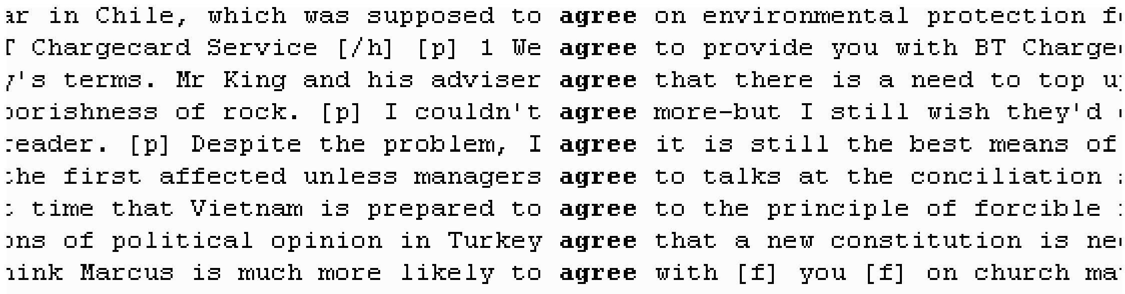
B Think of a response to each sentence using *That must be* or *You must be* plus an adjective from the box. Then practice with a partner.

1. I've been taking dance lessons. *"That must be fun."*
2. The elevator's not working, so I have to walk up to the tenth floor.
3. I just won a scholarship to go to college.
4. I often get up and study at 5:00 in the morning.
5. I'm going skydiving next week.
6. We're reading a book on ethics in my philosophy class.
7. I'm going out on a blind date tonight.

annoying	hard
bored	interesting
boring	motivated
excited	nervous
exciting	pleased
fascinating	proud
fun	scary
happy	tired

Figure 21. An extract from touchstone student book 3 (McCarthy, McCarten and Sandiford, 2006a, cited in McCarten, 2010, p.421)

In ELT teaching, COBUILD project also a major achievement as reference words and teaching materials since it has grammar series, guides, concordance usage and samplers which are based on 'real English and compiled with the needs of the language learners in mind (Sinclair, et.al., 2001). COBUILD offers teachers and learners more reliable information about English rather than traditional reference grammars and older non-corpus works. For instance, if learners or teachers want to see how the word 'agree' is used, they can look at the examples on BoE and Collins corpus as shown in Figure 22.:



```
ar in Chile, which was supposed to agree on environmental protection f
[ Chargecard Service [/h] [p] 1 We agree to provide you with BT Charge
y's terms. Mr King and his adviser agree that there is a need to top u
oorishness of rock. [p] I couldn't agree more-but I still wish they'd
reader. [p] Despite the problem, I agree it is still the best means of
the first affected unless managers agree to talks at the conciliation
: time that Vietnam is prepared to agree to the principle of forcible
ons of political opinion in Turkey agree that a new constitution is ne
hink Marcus is much more likely to agree with [f] you [f] on church ma
```

Figure 22. An extract of '*agree*' from collins corpus (adapted from http://www.mycobuild.com/about-collins-corpus.aspx)

Various usages of verb 'agree' can be seen on concordance which extracted from BoE such as 'agree on something', 'agree with somebody', or 'agree to something'. All COBUILD dictionaries are based on the information on the Bank of English (BoE) and the Collins corpus. Since the corpus is so large, lots of examples of how people really use the words can be examined. The data shows how words are used; what they mean; which words are used together; and how often words are used. The examples chosen from BoE help to show how the words are really used and to demonstrate, typical grammatical patterns, typical vocabulary and typical context of a particular word.

Two main advantages that COBUILD and other corpus-based reference works such as Longman, Macmillian, OUP and CUP (Römer, 2008) that they incorporate corpus-derived findings on frequency distribution and register variation, and that they contain genuine rather than invented examples. Another entire corpus-based resource is the student version 'The Longman Grammar of Spoken and Written English' (Biber, et at., 1999) contains authentic language examples. Also 'Cambridge Grammar of English' (Carter and McCarthy, 2006) is a starting point which has English as it is spoken and written today (in Cheng, 2010).

2.4.2. Data-driven Learning

Direct applications of corpora in learning and teaching concerned with the the direct access of corpora by teacher and learner oriented, so these applications are more teacher and learner oriented. Accordingly, Johns (1991) suggested a concept "confront the learner as directly as possible with the data, and to make the learner a linguistic researcher" (cited in Römer, 2008, p. 118). That is, when corpus techniques are used in the classroom and corpus data is analyzed by learners, then learners become language researchers. In this way, Johns (1991) describes this as data-driven learning (DDL). DDL method is based on learner- centered activities with the teacher as the facilitator. The method has usually been a reference in ELT and English language corpora; however, it can be applied in teaching other languages. The introduction of DDL method has begun in 1991 and developed over 1990s onwards. Johns (1991) initially used a concordance program MicroConrad as a tool in his work of English for specific purposes to non-native speakers. Then it was realized that a concordancing was much more effective way of studying the use of common prepositions, finding that an exercise such as underlining the headword colligating with the preposition on ('depending on', 'on demand') was more helpful than a gap-filling exercise involving filling in the prepositions (Chambers, 2010).

DDL method enables learners and teachers to have access to real life examples of particular structures through concordance programs within software such as WordSmith Tools (Scott, 1999) and MonoConc (Barlow, 2000). For instance, a teacher or a learner who wants to practice examples of the phrasal verb **end up,** can easily access to the following examples from COBUILD as presented in Example 10.:

Example 10.
If you drink with other people who regularly buy rounds for each other, it's easy to **end up** drinking more than you want.
if you forget to spray it with simazine every March you **end up** with a lot of extra weeding.
It's true you do get stared at in clubs, but you know, I am fat, I do live in the real world, and I don't want to **end up** some kind of fat separatist.
was a very tough little man, a very hard little man who knew what he wanted, where he was going and where he was going to **end up**.
As a result, the child may **end up** in a distress-provoking, or even physically dangerous, situation.
Many politicians **end up** simply hating the press.
We're gonna **end up** living in a broom cupboard.
the kids **end up** you know homeless and uneducated at sixteen.
Tony Galluci visited Italy for the first time and almost **ended up** in the army.
Those who have tried to be honest have **ended up** at the bottom of the ladder.
(adapted from Chambers, 2010, p. 347)

Although the above examples involve only English, multilingual resources are also available. Within DDL, utilizing concordance is a useful way of natural contextual learning and indeed researchers have recently also highlighted its use and usefulness for error correction in foreign or second language writing (Bernardini 2004; Chambers 2005; Gaskell/Cobb 2004; Gray 2005) (cited in Römer, 2008, p.120). These studies demonstrate that corpora suggest a reliable complement for existing reference works and that they may provide information which a dictionary or grammar book may not provide (Römer, 2008). Gilquin and Granger (2010) point out number of advantages of DDL in pedagogy. Firstly, it brings authenticity into the classroom by corpora so that learners can be exposed to authentic language and they identify the authentic instances of a particular linguistic item. Secondly, DDL has a corrective function by which learners compare their written productions by native writers' or they can examine common learner errors in a learner corpus. Indeed, learners can find the support they need to correct their own interlanguage features (misuse, overuse and underuse) and thus they can improve their L2 writing. In addition, DDL approach has an advantage of including discovery element which provides motivation and fun in language learning. As language researchers, learners are encouraged to observe corpus data, make hypotheses and formulate rules in order to gain insights int language (inductive approach) and to check the validity of rules from their actual grammar or textbooks (deductive approach) (Gilquin and Granger, 2010). O'sullivan (2007) suggests that, with DDL learners can acquire various learning skills such as predicting, observing, noticing, thinking, reasoning, analysing, interpreting, reflecting, exploring, making inferences (inductively or deductively), focusing, guessing, comparing, differentiating, theorising, hypothesising, and verifying' (2007, cited in Gilquin and Granger, 2010, p. 359-360). These skills can be used to explore language; however, since they are general cognitive skills, they may also be transferred to other fields of study.

2.5. Adverbial Connectors

2.5.1. Adverbial Connectors in English

The choice of adverbial connectors will be based on the list of semantic conjuncts in Quirk et al. (1985). Adverbials, or 'Conjuncts' (Quirk, et.al., 1985) have the function of conjoining independent units and they have a detached and superordinate role (Quirk et at., 1985). English Adverbial connectors are realised by a variety of syntactic forms. For instance:

- Single adverbs and adverb phrases: **First**, **Second**, **Then**, **So**;
- Adverb phrases: **More precisely**, **More accurately**;
- Prepositional phrases: **In addition**, **by the way**, **for example**;
- Finite clauses: **That is**, **That is to say;**
- Non-finite clauses: **To conclude**.

(Biber et. al., 1999)

The conjunct function entails a conjunct specific set of semantic relations. They are connected with, but are frequently rather remote from, the adverbial relation that is assumed in the speaker-related clause to which they correspond (Quirk et al. ,1985). Seven conjuntive roles of Quirk et. al. (1985) are presented as follows:

1. **Listing :** (i) Enumerative (e.g. First, Second, Finally)
   (ii) Additive: (iii) Equative (e.g. in the same way, likewise)
   (iv) Reinforcing (e.g. moreover, further)

2. **Summative:** (e.g. In sum, Altogether)

3. **Appositive :** (e.g. For example, Namely)

4. **Resultive :** (e.g. As a result, Therefore)

5. **Inferential :** (e.g. In that case, Otherwise)

6. **Contrastive:** (i) Reformulatory (e.g. more precisely, rather)
   (ii) Replacive (e.g. better, again)
   (iii) Antithetic (e.g. by contrast, instead)
   (iv) Concessive (e.g. in any case, however, yet)

7. **Transitional:** (i) Discoursal (e.g. by the way, incidentally)
   (ii) Temporal (e.g. in the meantime, meanwhile)

(Quirk et al. ,1985, p.634)

1. Listing Adverbial Connectors

The Listing adverbial connectors are used to give orientation to a list. This category consists of two sub-types as Enumerative and Additive. Additive is divided to two inner sub-type as Equative and Reinforcing. Enumerative adverbials show order, connote relative priority and endows the list with an integral structure as beginning and end:

Example 11.

First, the economy must be recovered, and **second (ly)**…

Equative indicates an item has a similar force to a preceding one:

Example 12.

Twins go to the same school. **Equally**, they go to the same sport centre.

Reinforcing assesses an item as adding greater weight to a preceding one.

Example 13.

He is a talented musician, **in addition** he is a teacher.

Listing Adverbial Connectors consist of 50 types within 3 sub-types. There are 24 types in enumerative and 6 types in Equative, and 20 in reinforcing. In total, 50 listing adverbial connectors have been searched over corpora. **Then** have role both in enumerative and reinforcing, this means **then enumerative** is a one type and **then reinforcing** is another listing connector (see section 2.5.5.1). Table 7. presents Listing adverbial connector types:

Table 7

Types in Listing Adverbial Connectors

1. LISTING		
1.1 Enumerative	**1.2 Additive**	
First, Second, Third, One, Two, Three Firstly, Secondly, Thirdly, In the first place, In the second place, First of all, Second of all, On one hand…on the other hand(*), For one thing…for another thing, For a start, to begin with, to start with, Next, Then(*), To conclude(*), Finally, Lastly, Last of all	**a. Equative** Correspondingly, Equally, Likewise, Similarly, In the same way By the same token	**b. Reinforcing** Again(*), Also, Further, Furthermore, More, Moreover, In particular, Then(*), Too, What is more, In addition, Above all, On top of it all, To top it all, To cap it all, Particularly, In fact (*), Indeed, Actually, As a matter of fact
24	**6**	**20**
Total: 50 Types		

(*) Connectors which can act in different categories

2. Summative Adverbial Connectors

In summative category, adverbials precede an item that looked in relation to all previous items, and they introduce an item that embraces the preceding ones:

Example 14.

She washed the dishes, cooked the dinner and cleaned the bedrooms. She took children to the park after shopping. **In sum**, it was another a busy day at home.

In Summative category, there are 15 types of adverbial connectors without a sub-category as above in Table 3. Many multi-functional connectors like **then**, **therefore**, **thus** and t**o conclude** have one of their role in this group.

Table 8

Types of Summative Adverbial Connectors

2. SUMMATIVE
Altogether, Overall, Then(*), Therefore(*), Thus(*), (all) in all, In conclusion, In sum, To conclude(*), To sum up, To summarize, In summary, In brief, In short, To be brief
Total: 15 Types

(*) Connectors which can act in different categories

3. Appositive Adverbial Connectors

Appositive conjuncts precede an item which related to previous items and express the content of preceding item.

Example 15.

The baby has some problems**, that is**, she doesn't sleep much and always cries.

Appositive category contains 10 types in which there are multi-functional connectors and some very common connectors like **for example** and **I mean**. Table 9. represents the Appositive connectors.

Table 9

Types of Appositive Adverbial Connectors

3. APPOSITIVE
Namely, Thus(*), In other words(*), For example, For instance, That is, That is to say, Specifically, To illustrate, I mean
Total: 10 Types

(*) Connectors which can act in different categories

4. Resultive Adverbial Connectors

Resultive adverbials express the result:

Example 16.

The weather was so rainy yesterday, **so** I got cold.

Resultive adverbials consist of 16 connector types in which some of them can act in other categories such as **therefore**, **of course**, **now**, t**hus.** Table 10. represents Resultive averbials:

Table 10

Types of Resultive Adverbial Connectors

4. RESULTIVE
Accordingly, Consequently, Hence, Now(*), So, Therefore(*), Thus(*), As a consequence, In consequence, As a result, Of course (*), Somehow, Due to (this), In order to do (this), Because of (this), For this purpose
Total: 16 Types

(*) Connectors which can act in different categories

5. Inferential Adverbial Connectors

Inferential connectors are used indicate a conclusion based on logic and supposition:

Example 17.

You have to come to dinner tonight. **Otherwise**, mom will be sad.

Inferential adverbial category has 6 connector and some of them are multi-functional adverbial connectors, namely, Else, Then, In other words. Inferential adverbial connectors are displayed in Table 11.:

Table 11

Types of Inferential Adverbial Connectors

5. INFERENTIAL
Else, Otherwise, Then(*), In other words(*), In that case, Or else
Total: 6 Types

(*) Connectors which can act in different categories

6. Contrastive Adverbial Connectors

Contrastive adverbials present either contrastive words or contrastive matter in relation to what has proceeded. **Concessive** is used where one unit is seen as unexpected in the light of the other. **Reformulatory** provides a different formulation whereas **Replacive** helps to replace an item with a more important one, and **Antithetic** introduces direct antithesis when an item contrasts the preceding one.

Example 18.

She was on a diet since last month. **Still,** she hasn't lost weight yet. (Concessive)

Example 19.

He wasn't at the party. **In other words**, he hasn't been invited.(Reformulatory)

Example 20.

You can send the documents with e-mail. **Alternatively**, you can give them to a courier. (Replacive)

Example 21.

I was waiting for my friend, **instead** my mom came. (Antithetic)

Table 12

Types of Contrastive Adverbial Connectors

6. CONTRASTIVE			
6.1. Reformulatory	**6.2. Replacive**	**6.3. Antithetic**	**6.4. Concessive**
Better, Rather, More accurately, More precisely, Alternatively, In other words(*)	Again(*), Alternatively, Rather, Better, Worse, On the other hand (*)	Contrariwise, Conversely, Instead, Oppositely, Then(*), On the contrary, In contrast, By contrast, By way of contrast, In comparison, By way of comparison, On one hand…on the other hand(*)	Anyhow,Anyway, Anyways, Besides, Else (*),However, Nevertheless, onetheless, Notwithstanding, Only, 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 (*), Admittedly, All the same, Of course (*), Still and all, That said, In fact (*), Even so
7	**6**	**13**	**30**
Total: 56 Types			

(*) Connectors which can act in different categories

Table 12. presents all contrastive adverbial connector types within their four sub-categories. The Contrastive adverbial group consists of 56 adverbial connectors types and 30 of them belong to Concessive type. Some concessive adverbial connectors can take part in different sub-types as well. Connectors like **again, on the other hand**, **then**, **on one hand ….on the other hand**, **else**, **of course**, **in fact** can function in different categories.

7. Transitional Adverbials

Transitional that helps to shift attention to another topic, (**Discoursal**) or to a temporally related event (**Temporal**).

Example 22.

I'm going to start a new project tomorrow, **by the way,** where is my notebook? (Discoursal)

Example 23.

I waited them all evening. **Eventually**, they came at midnight. (Temporal)

Transitional which comprise of two sub-types as **Discoursal** and **Temporal**. There are 22 connector types in total as shown in Table 13.:

Table 13

Types in Transitional Adverbial Connectors

7.TRANSITIONAL	
7.1. Discoursal	**7.2. Temporal**
Incidentally, Now (*), By the way, By the by, As for, As to, With regard to, With respect to, As regards, Regarding, As far as ..x...concerned	Meantime, Meanwhile, In the meantime, ın the meanwhile, Originally, Subsequently, Eventually, At first, Afterwards, Later, Then (*)
11	**11**
Total: 22 Types	

(*) Connectors which can act in different categories

In the present study, all connector types in each category and sub-category (see App. A) which are described below were analysed as the major linguistic item for the investigation.

2.5.2. Adverbial Connectors in Turkish

In Turkish, there are three types of to coordinate sentences: 1. by simply stringing the coordinated sentences one after another, without using any coordination marking (juxtaposition of two or more constituents); 2. By attaching the coordination post-clitics (such as **-da**) or subordinating suffixes (such as **-(y)ıp**); 3. using the unbound conjunction markers, (or conjunctions and connectives) (Kornfilt, 1997; Göksel and

Kerslake, 2005). Third type of coordination, which is the major concern of the study, can be considered the equivalence structures of English adverbial connectors.

"Conjunctions are expressions such as **ve** 'and', *fakat* 'but', and **ya da** 'or', which join two or more items that have the same syntactic function" (Göksel & Karslake, 2005, p. 440). These structures can be phrases, subordinate clauses or sentences. On the other hand, the conjoining function of **discourse connectives,** is minimally to join two sentences. For instance, discourse connectives such as ***aksine*** 'on the contrary', **üstelik** 'moreover' and **sonuç olarak** 'as a result' can be used for purposes of forming a cohesive link between concepts expressed by entire groups of sentences. Another difference between the two classes is that while a conjunction always joins two (or more) linguistic items, this is not always the case with discourse connectives, which can sometimes be used on their own if the context presents a situation (e.g. a recent experience shared by speaker and hearer) to which a cohesive link can be made (Göksel & Karslake, 2005).

The structure of Turkish connectors (equivalent of English adverbial connectors) is mostly consisting of adverbs. However, other type of structures that includes more than one structure to indicate conjunction is possible. Various forms which constitute the structures of connector are as follows:

- Single Adverb: **Önce** (First), **Böylece** (Therefore), **Yani** (Namely, That is);
- Adverb combinations: **Daha sonra (**Afterwards)
- Noun: **Dahası** (Moreover, Furthermore);
- Finite clause: **Demek istediğim** (I mean);
- Particle: **Mesela** (For example);
- Adjective: **Şüphesiz** (Of course);
- Adjective + Noun: **Bu sırada** (In the meatime);
- Adjective compound : Öyleyse (Thus); Öyle (adj.) + (y)se :compound;
- Affix+Adjactive: **-e/-a dair** (with regard to);-e/-a(prefix dative case) + dair (adj.).

(http://www.tdk.gov.tr/index.php?option=com_gts&view=gts)

In the study, Turkish equivalent structures of English adverbial connectors have been tried to find out considering the linguistic similarities and translation. Thus, a list of corresponding items of adverbials in Turkish was formed categorically.

1. Listing

In listing group, as in English, connectives are used to list of fact indicated in previous sentences. Three sub-categories of Listing adverbials connectors in English, Enumerative, Equative and Reinforcing have similar structures in Turkish. Table 14. presents some of Turkish equivalents of Listing Adverbial connectors in English:

Table 14

Turkish Equivalents of Listing Adverbial Connectors in English

Listing Adverbial Connectors		**Turkish Equivalents**
Enumerative	First, Finally	Önce, İlk Önce, Son olarak
Equative	Likewise, Similarly	Aynı şekilde, Bunun gibi
Reinforcing	Moreover, Also	Ayrıca, Hatta

Example 24.

Şahinde hanım 'Ne biliyorsun' demedi, **hatta** bunu düşünmedi bile. (Reinforcing)

Lady Sahinde did not say 'What do you know', **moreover,** she did not even think of it.

(extracted from TNC)

As shown in example 24., connector 'hatta' has a similar meaning as 'moreover' in English adding importance to preceding one.

2. Summative

In Turkish, Summative adverbial connectors are introduced as represented in Table 15.:

Table 15

Turkish Equivalents of Summative Adverbial Connectors in English

Summative Adverbial Connectors	**Turkish Equivalents**
In conclusion, In summary	Sonuç olarak, Özetlemek gerekirse
Then, Therefore	Öyleyse, En nihayetinde

Example 25.

Le Corbusier ne düşünürdü acaba? **Sonuç olarak** Le Corbusier de, utopist mimarlar gibi, başkaları adına karar veren bir uzman..

What would Le Corbuiser think? **In conclusion,** Le Corbuiser too, like other utopist architechts, is an expert who decide on behalf of others.

(extracted from Metu Corpus)

The sentence in example 25. is the last sentence of a paragraph in an article which was retrieved from Metu Corpus. As having a similar meaning as its English equivalent, 'in conclusion' which summing up the previous items.

3. Appositive

Appositive adverbial connectors are used in Turkish as shown in Table 16.:

Table 16

Turkish Equivalents of Appositive Adverbial Connectors in English

Appositive Adverbial Connectors	**Turkish Equivalents**
That is, Namely, I mean	Yani, Şöyle ki
For example, For instance	Mesela, Örneğin

Example 26.

Yeni açılan yerlerde kısıtlama yok. **Mesela** Galata köprüsü'nde ve bazı parklarda ki tesislerde sınırlama yok.

There are no constraints at newly opened places. **For instance**, there is no limitation at Galata bridge and facilities at some parks.

(extracted from Metu Corpus)

In the example above, '**mesela**' is used to express the content of preceding item (like **for instance**).

4. Resultive

Resultive adverbial connectors are represented in Turkish by such structures as in Table 17.:

Table 17

Turkish Equivalents of Resultive Adverbial Connectors in English

Resultive Adverbial Connectors	**Turkish Equivalents**
Therefore, So, Thus	Böylece, Bu yüzden, Dolayısı ile
Of course	Ebette ki, Tabi ki

In Turkish, connectors as **böylece** as in the example establish cause-effect relationship in a similar way (as **therefore**) in English.

Example 27.

Toplumlar çeşitli üretim tarzlarından geçerler. **Böylece** önemli değişimlere uğrarlar.

Societies experience various production styles. **Therefore**, they undergo important changes.

(extracted from Metu Corpus)

5. Inferential

In Turkish, inferential adverbial connectors are introduced as represented in Table 18. :

Table 18

Turkish Equivalents of Inferential Adverbial Connectors in English

Inferential Adverbial Connectors	**Turkish Equivalents**
Otherwise, Else	Aksi takdirde, Yoksa
In that case	O zaman, O halde

Example 28.

Bu soruna hemen bir çözüm bulmalıyız. **Aksi takdirde**, durum kötüleşir.

We must find a solution for this problem immediately. **Otherwise**, the condition gets worse.

In Turkish, connectors like **aksi halde** or **yoksa** are used to indicate a logical conclusion and supposition as **otherwise** in English form.

6. Contrastive

Contrastive adverbial connectors include four sub-categories as Reformulatory, Replacive, Antithetic and Concessive. Similar structures of contrastive adverbial connectors are displayed in Table 19.:

Table 19

Turkish Equivalents of Contrastive Adverbial Connectors in English

Contrastive Adverbial Connectors		**Turkish Equivalents**
Reformulatory	Better, More accurately	Daha doğrusu
Replacive	Rather, Worse	Daha beteri
Antithetic	On the contrary, Conversely	(Tam) Tersine, (Tam) Aksine
Concessive	However, Yet, Nevertheless	Fakat, Buna ragmen, Yine de

Example 29.

Yesterday the weather was very good. **However / Yet**, it’s very bad today.

Dün hava çok güzeldi. **Fakat/ Ama** bugün çok kötü.

In the example above, connectors in the sense of Concession type are presented. Both **fakat** and **ama** can be used to mark contrasts or clear differences with preceding idea.

Example 30.

Kararsız seçmen istikrarı en fazla arayandır. Yani belirsizliğe oy atmaz. **Tam tersine** güvene yönelir.

Indecisive voter is the one who is looking for stability most. That is, s/he does not vote uncertainty. **On the contrary**, s/he gravitates to confidence.

In example 30., **tam tersine** is used to indicate a clear antithesis for the preceding sentence in a similar way of **on the contrary** as an Antithetic type of connector.

7. Transitional

There are two sub-categories within English Transitional adverbial connector category; Discoursal and Temporal. Turkish connector types which act like Transitional adverbial connectors in English are presented in Table 20. :

Table 20

Turkish Equivalents of Summative Adverbial Connectors in English

Transitional Adverbial Connectors		**Turkish Equivalents**
Discoursal	By the way, As regards	Bu arada, (-) Hakkında
Temporal	Meanwhile, Afterwards	Bu sırada, Daha sonraları

Example 31.

Ama mağara yaşantısının sonu hep mutlu bitmez. **Bu arada** bir şeyi vurgulamam gerektiğine inanıyorum..

However, the cave life does not always a happy ending. **By the way**, I believe that I have to emphasize one thing…

(extracted from Metu Corpus)

In the example above, **bu arada** provides a transition from a topic to another similar to **by the way** in English.

In summary, all Turkish equivalents of English adverbials have been determined and gathered into a list (see Appendix B). Since the aim of TUC analysis is to investigate the general usage of adverbial connectors in Turkish, it is necessary to go to in a parallel line with TICLE analysis in which the aim is to identify all English adverbial connectors in L2 writing. Therefore, to examination of connectors in Turkish are important to give explanations for connector usage in writing in general.

2.5.3. Adverbial Connectors in Japanese

In Japanese, adverbial connectors are normally used to conjoin sentences. They connect sentences in a variety of meaning. The categorization of connectors in Japanese is as follows:

- Addition: besides, moreover
- Consequence: therefore, consequently
- Immediate consequence: just then, thereupon
- Contrast: however, on the contrary
- Qualification: though, however
- Reason: because, the reason for
- Sequence: first X.. then Y, thereafter
- Choice: or
- Alternative: on the one hand, on the other hand
- Paraphrasing: in other words, for example
- Change of topic/ Coming to the point: well, by the way

(Kaiser et al, 2001)

Since the classification of adverbial connectors in Japanese includes a number of different structures than in English, adverbial connectors which are considered common were handled in the present study.

- **Addition:** The connector types used instead of **besides** and **moreover** are;

soskite, **mata**, **shikamo**, **sono ue**, **sore ni**, **sara ni, oyobi**.

Example 32.

蛤の殻は一つとして同じ模様のものがない。【また】、もとの片割れでなければ噛み合わせが合わない。

Hamaguri no kara wa hitotsu to shite onaji moyō nomono ga nai. **Mata** moto no kataware de nakereba kamiawase ga awanai.

There is not one shell of the cherrystone clam that the same pattern. **Moreover**, t he shell fits only its original counterpart

(Kaiser et al, 2001, p. 73)

- **Consequence:** the connector types used in the sense of **consequently** and **therefore** are: **da kara**, **sore do**, **soko de**, **shitagatte**, **sono tame**.

Example 33.

大学の公開講座は回数が少ない上に、担当教員も毎回変わる講座が多い。【そこで】、通常の講義のように16回とおしで、上級レベルの講座を開いた。

Daigaku no kōkai kōza wa kaisū ga sukunai ue tantō kyōin mo maikai kawaru kōza ga ōi. **Soko de** tsūjō no kōgi no yō ni jūrokkai tōshi de jōkyū reberu no kōza o hiraita.

Univesity courses for the general public are short and often have different lectures each time. **Therefore**, we have established an advanced-level cpurse that runs continously for 16 classes, just like a regular course.

(Kaiser et al, 2001, p. 74)

- Contrast: to indicate the meaning of **however, but**, and **on the contrary**, the following connectors are used: **shikashi**, **keredomo**, **da ga**, **datte**, **sore demo**, **demo**, **tokoro ga**, **to wa ie**.

Example 34.

豪華なシャンデリアもなければ赤い絨毯が敷かれたエントランスホールもない。【けれども】、私には、この簡素な場所が東京でもっとも贅沢な劇の場であるように思える。

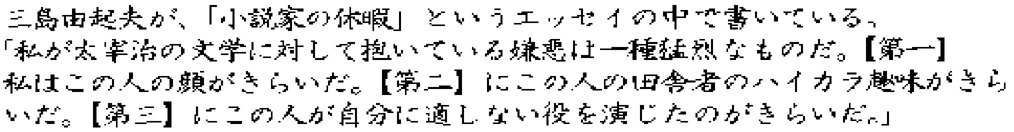
Gōka na shanderia mo nakereba akai ga shikareta entolansu hōru mo nai. **Keredomo** watashi ni wa kono kanso na basho ga tōkyō de mottomo zeitaku na geki ba de aru yō ni omoeru.

There is no luxurous chandelier, nr an entrance hall with red carpet. **However**, for me this simple place feels like the most luxurous spot for (satging) plays.

(Kaiser et al, 2001, p. 75)

- **Sequence**: In the sense of **first X**..**then Y** and **firstly**/ **secondly/ thirdly**, the following connectors are used: **mazu**, **hajime ni**, **soro kara**, **daiichi/ni/san ni.**

Example 35.


三島由起夫が、「小説家の休暇」というエッセイの中で書いている。
「私が太宰治の文学に対して抱いている嫌悪は一種猛烈なものだ。【第一】私はこの人の顔がきらいだ。【第二】にこの人の田舎者のハイカラ趣味がきらいだ。【第三】にこの人が自分に適しない役を演じたのがきらいだ。」

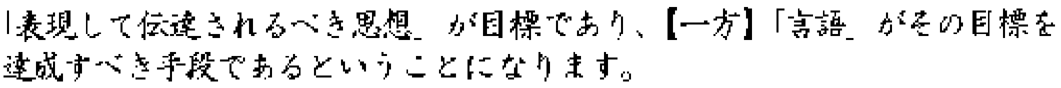
Mishima Yukio ga shōsetsuka no kyūka essē no naka kaite iru. Wtashi ga Dazai Osamu no bungaku ni taishite idaite iru ken'o wa isshu mōretsu na mono da. **Daiichi** watashi wa kono hito no kao ga kirai da. **Daini ni** kono hito no inakamono no haikara shumi ga kirai da. **Daisan ni** kono ga jibun ni tekishinai yaku o enjita no ga kirai da.

Mishima Yukio writes in an essay titled 'The Novelist's Vacation': 'The aversion I have to Dazai Osamu's works is quite strong. **Firstly**, I dislike his face. **Secondly** I dislike his country-bumpkin sense of stylishness. **Thirdly**, I dislike the fact that he played a part for which he was unsuited.'

(Kaiser et al, 2001, p. 79)

- **Alternative**: connectors as **ippō, ippō** and **tahō** are used instead of **on the one hand** and **on the other hand**.

Example 36.


「表現して伝達されるべき思想」が目標であり、【一方】「言語」がその目標を達成すべき手段であるということになります。

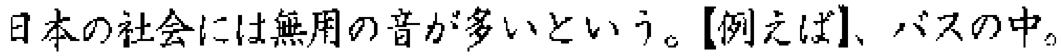
Hyōgen shite dentatsu sarerubeki shisō ga mokuhyō de ari **ippō** gengo ga sono mokuhyō o tassei subeki shudan de aru to iu koto ni narimasu.

The goal is 'An idea that needs to be expressed and communicated' but **on the other hand** 'language' is the means to achieve that goal.

(Kaiser et al, 2001, p. 80)3

- Paraphrasing: In the sense of **for example**, **that is**, **in short** and **in other words,** following equivalents are used: **tsumari**, **sunawachi**, **yōsuru ni**, **tatoeba**, **iwaba**.

Example 37.



日本の社会には無用の音が多いという。【例えば】、バスの中。

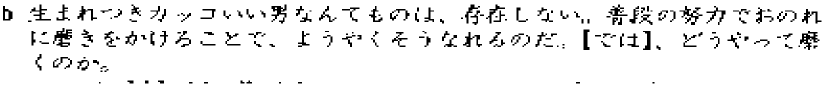
Nihon no shakai ni wa muyō no oto ga ōi to iu. **Tatoeba** no naka

He says that Japanese society there are many unnecessary sounds. **For instance**, inside a bus.

(Kaiser et al, 2001, p. 80)

- **Change of topic/ Coming to the point:** Instead of **by the way, well** and **ok/well then**, the following connectors are used: **sate** , **tokoro de**, **de wa**, **ja**.

Example 38.



b 生まれつきカッコいい男なんてものは、存在しない。普段の努力でおのれに磨きをかけることで、ようやくそうなれるのだ。【では】、どうやって磨くのか。

Umaretsuki kakko ii otoko nante mono wa sonzai shinai. Fudan no doryoku de onore ni migaki o kakeru koto de yōyaku sō nareru no da. **De wa** dō yatte migaku no ka.

There's nı such tings as an elegant man by birth. By making constant efforts to polish oneself one finally gets there. **OK then-** how does one do the polishing?

(Kaiser et al, 2001, p. 80)

As can be seen, English adverbial connectors have Japanese equivalents which are used in similar aims. The major difference between two languages in terms of adverbial connectors is their classification.

2.5.4. Adverbial Connectors in Spanish

Spanish connectors are usually used as conjunctions in adverb forms. Conjunctions in Spanish are A word which links other words or *phrases*, e.g. **y** 'and', **o** 'or', **pero** 'but'. Subordinating conjunctions introduce a subordinate *clause*, e.g. **que** 'that', **cuando** 'when', **aunque** 'although' (Bradley and Mackenzie, 2004, p.318).

Similarly to English adverbial connector classifications, Spanish grammarians often classify conjunctions referring to **conjunciones aditivas** (additive conjunctions such as "and" or y), **conjunciones adversativas** (contrastive conjunctions such as "but" or pero and "nevertheless" or sin embargo), **conjunciones causales** (causal conjunctions such as "because" or **porque**), and **conjunciones temporales** (temporal conjunctions such as "then" or **entonces**) (http://www.spanishbooster.com/SpanishConjunctions.htm).

- **Additive Conjunctions (Conjunciones Aditivas)**

Conjunction for addition in Spanish is usually indicated by ‘y’ as in the example below:

Example 39.

Compré musica rusa **y** turca.

I bought Russian **and** Turkish music.

Other additive conjuntions in Spanish like **ademias/es pas** for furthermore/moreover (http://www.gvsu.edu/mll/swc/index)

Example 40.

Además, es menos dañino para el medio ambiente.

Furthermore, it is environmentally friendly

- **Contrastive Conjunctions (Conjunciones Adversativas)**

Contrastive conjunctions indicate opposition among the elements that they join. Some conjunctives referring opposition in Spanish are as follows:

Pero / but

Sin embargo / nevertheless, however

Mas / however

Antes bien / on the contrary

Example 41.

Quería un helado, **mas** no tenía dinero.

I wanted an ice cream, **however** I did not have enough money.

- **Causal Conjunctions (Conjunciones Causales)**

In Spanish, casual conjunctions always subordinate one sentence to another. The most common are:

Porque ⁄ because

Por lo tanto/ therefore

Puesto que ⁄ although, since, as long as

Example 42

Ella no comprendía, **por lo tanto** se fué.

She did not understand, **therefore** she left.

- **Temporal Conjunctions (Conjunciones Temporales)**

As in English, temporal conjunctions mark temporarily related events. Some of temporal conjunctives are as follows:

Mientras/ meanwhile

Pues/then

A proposito/ by the way

Example 43.

Quieres dinero?, **pues** trabaja!.

Do you want money? **Then** work!

Spanish conjunctions include a wide variety of items usually consisting of adverbs as in English adverbials. The explanations for conjunctions below are given to support the L1 forms of connector types in Spanish grammar.

2.5.5. Corpus-based Studies on Adverbial Connectors

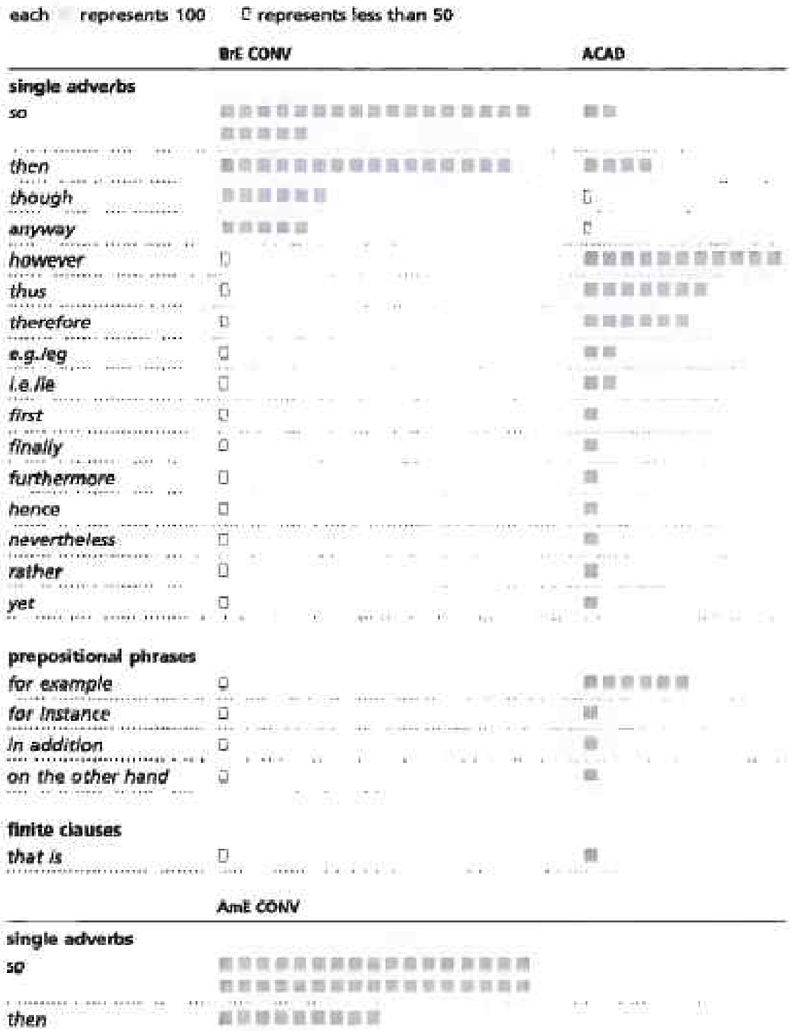
Adverbial connectors in L2 have been an important concern in the literature. Considerable amount of studies have been carried out on connectors, especially focusing on connector usage by EFL and ESL learners in writing. Most of the previous studies of connectors usage were based on learner corpora (Milton and Tsang, 1993; Granger and Tyson, 1996; Altenber &Tapper, 1998; Bolton, et. al., 2002; Narita, et. al., 2004; Tanko, 2004; Fei, 2006; Chen, 2006; Bikeliene, 2008), on the other hand, some non-corpus-based initial studies can be considered as the pioneer of the investigations of adverbial connectors in L2 (Connor, 1984; Crew, 1990; Field & Yip, 1992). This section explains previous studies on adverbial connectors majorly related to the area of EFL learning.

The text cohesion concept was first initiated and by Halliday and Hasan (1976) who introduced and worked categorisations of cohesive devices in sentence coherence. Afterwards, a number of studies were carried out on cohesive devises in learner writing which are now accepted as inconclusive since they did not utilized the computer technology like Connor's (1984) study in which only six ESL learner essays have been analyzed in terms of cohesion (in Granger and Tyson, 1996). In 1990s, the researchers began to study conjunctive adverbials using learner corpora. Besides previous L2 studies with learner corpora, a fundamental investigation on adverbial connectors in daily English have been worked by Biber et. al. (1999) by reference corpus in English.

In corpus-based framework Longman Grammar of Spoken and Written English (LGSWE), Biber et. al. (1999) provides descriptions of actual use of grammatical features in different varieties of English through LGSWE corpus (approx. 40 million words) which contains written texts of fiction, conversation, language or academic prose. Biber et.al. (1999) investigated adverbial connectors in by LGSWE corpus in order to identify the most common and higher frequency ones. In the study, adverbial connectors were examined in two genres as spoken and written corpora, namely through British English and American English conversations and academic prose. Table 4. represents the high frequency adverbial connectors in Biber et. al. (1999)'s study. In Table 21., BrE CONV stands for British English conversations, AmE CONV for American English conversations and ACAD for academic prose.

Table 21

Most Common Linking Adverbials in Conversation and Academic Prose in LGSWE Corpus: Occurrences per Million Words (adapted from Biber et. al., 1999, p.887).



According to Table 21., two adverbials are common in conversation in English language as **so** and **then** within result/inference adverbials category. Next two common adverbials in conversation are **tough** and **anyway** in contrast/concession class. AmE conversation differs from BrE in having a higher frequency of **so** and a lower frequency of **then.** These four connectors play important role in the development of a conversation, for instance, so is generally used in narrative accounts, it moves the story along and makes clear how an event follows from another (Biber et. al., 1999). In respect of genre difference, adverbial connectors are more frequent in academic prose than in conversation. While conversation has four common items (**so**, **then**, **though**, **anyway**), academic prose has several moderately common items in different groups (**then**, **therefore**, **thus**, **hence** in result/inference; **however**, **rather**, **yet**, **nevertheless**, **on the other hand** in contrast/concession group).

Several studies focus on the analysis of usage patterns of logical connectors in ESL or EFL academic writing to obtain empirical evidence to support the contradictory claims as ESL/EFL learners tend to overuse logical connectors in their English essay writing. Because the majority of previous studies have shown that the use of adverbial connectors seems to be problematic for learners because much of the studies reported misuse-overuse or underuse of connectors by learners. In table 22, the studies on adverbial connectors in L2 field are presented in a chronological order.

Table 22

Previous Studies on Adverbial Connectors in L2

Year	**Reference**	**Learner Group**	**Results**
1990	Crew	Chinese ESL Learners	-Misuse/overuse of connectors
1992	Field& Yip	Chinese ESL Learners	-Overall overuse of connectors
1993	Milton & Tsang	Chinese ESL Learners	-Overall overuse of connectors
1996	Granger & Petch-Tyson	French EFL Learners	-Over/underuse in individual connectors
1998	Altenberg & Tapper	Swedish EFL Learners	-Overall Underuse -L1 Transfer traces
2002	Bolton et. al	Chinese ESL Learners	-Overall overuse in both NS and NNS groups
2004	Narita, Sato & Sugiura	Japanese EFL Learners	-Overall overuse-over/underuse of individual connectors
2006	Tanko	Hungarian EFL Learners	-Over/underuse of individual connectors
2006	Chen	Taiwanese EFL Learners	-Overuse in individual connectors -Misuse of connectors
2006	Fei	Chinese EFL Learners	-Overuse of individual connectors
2008	Bikeliene	Lithuanian EFL Learners	-Underuse in Resultive Connectors -No significant over- underuse of individual connectors
2010	Heino	Swedish EFL Learners	-Overall Underuse -Over/underuse of individual connectors
2011	Can	Turkish EFL Learners	-Over/underuse of individual connectors

NS=Native Speakers, NNS=Non-native speakers, L1=First (native) language,
EFL=English as a Foreign Language
ESL=English as a Second language

First two initial studies (Crew, 1990; Field & Yip, 1992) were non-corpus based investigations on the use of connectives in academic writing of Hong Kong Chinese students. Crew (1990) conducted his study to examine misuse and overuse of logical connectors in writings of ESL students at Hong Kong University. Crew found frequent misuse of connectors like 'on the contrary' and overuse of others. Crew (1992) states that overuse is a way of 'disguising of a poor writing' and concludes as:

> Over-use at best clutters up the text unnecessarily and at worst causes the thread of the argument to zigzag about, as each connective points it in a different direction. Non-use is always preferable to misuse because all readers, native-speaker or non-native-speaker, can mentally construe logical links in the argument if they are not explicit, whereas misuse always causes comprehensive problems and may be so impenetrable as to defy normal decoding.
>
> (1990, p.324)

Field and Yip (1992) studied 'internal conjunctive cohesion' in ESL writing of senior/high school students in Hong Kong. They compare the use of connectors with cohesive devices in learners' and native group' (students from Autralia) essays. Field and Yip (1992) again suggests that L2 writers from Hong Kong tend to overuse such devices.

Milton and Tsang (1993) conducted one of the first using a corpus gathered from English learners in Hong Kong (HKUST corpus). They included 25 connectors in the analysis and founded that 20 of them were overused by ESL learners, contributing an overall patterns of overuse (Shea, 2009). The results of this study then questioned by the authors because of not comparing the learner results with native speakers' and selecting a limited number of connectors.

One of the cornerstone in adverbial connector investigation related to learner corpora was carried out by Granger &Tyson (1996), using two sub-corpora of ICLE (French and German) as learner data and LOCNESS as native English corpus. They compared French EFL learners' use of conjunctive adverbials with native speakers of English and other German EFL learners and hypothesised an overuse in general usage. However, Granger & Tyson (1996) found no overall overuse; instead they suggested that some patterns were the result of L1 conventions and translation equivalents. They concluded that "'heightened awareness of the semantic, stylistic and syntactic properties of connectors will lead students to think more carefully about the ideas these connectors are linking' (Granger & Tyson 1996, p.26).

A similar study was carried out by Altenberg and Tapper (1998) examining the use of adverbial connectors of Swedish EFL learners. They used Swedish sub-corpus of ICLE and compared it with LOCNESS. The overall results revealed that Swedish learners underuse the connectors but they showed evidence of overuse and underuse in individual connectors. They also compared their results with Granger and Tyson's (1996) study and found certain similarities and differences between the two learner groups, therefore they suggested that the learners' connector usage might not be entirely influenced by their mother tongue.

Narita et. al (2004) used Japanese sub-corpus of ICLE and LOCNESS for L1 reference to investigate connector usage. They found significant overuse in the use of connectors and they also indicated that, parallel to Altenberg and Tapper (1998) and Tanko (2004), some connectors were used more by the learners while others were less often.

In the same way, other studies reported overuse of adverbial connectors by learners such as in Taiwnese (Chen, 2006) and Chinese learners (Fei, 2006). Can (2011) found overuse and underuse of connectors in Turkish EFL learners essays. Also overuse and underuse in certain individual connectors is a common result in some studies (Tanko, 2004; Bikeliene, 2008, Heino 2010).

In sum, according to the above mentioned studies, although there are differences between the frequencies and the particular limitations in investigations, the overuse, underuse and misuse of adverbial connectors by L2 learners seems to be general tendencies. These issues then require discussing the problematic nature of adverbial connectors and possible solutions.

2.5.6. Difficulty of Connectors in EFL

Altenberg & Tapper (1998) states that connectors contribute to a better understanding of a spoken or written discourse, indeed, when properly used, connectors have a positive effect on the clarity and the comprehensibility of a text. However, the majority of previous studies have shown that the use of adverbial connectors seem to be problematic for learners because much of the studies revealed misuse-overuse or underuse of connectors by learners in their L2 writing. Crew (1990) states that:

> Logical connectives should be seen as higher-level discourse units which organize chunks of text in relation to the direction of the argument. If the links are misused, the argument as a whole, not merely the sentence containing the connective, becomes difficult to process and may even appear illogical. (p.316)

There are several reasons for connectors being difficult for language learners. Halliday and Hasan (1976) notes that conjunctive elements are not easily classifiable, they establish relation between meanings rather than grammatical units. That is, they provide a semantic relation on 'how' elements are connected instead of simply marking 'which' elements are connected. Moreover, the spaces of connectors in linguistic units can vary from clauses to paragraphs and even longer discourse (Quirk et. al. 1985; Hatch, 1992;). Therefore, learners first need to familiarize individual connectors, then the type of units they normally occur, finally the distance they can span between units.

Hatch (1992) and McCarthy (1991) (cited in Tanko, 2004, p. 160) point out another characteristic feature of connectors is that there is no one-to-one correspondence of connectors and their functions. For example, the one word conjunct **then** can be found in Halliday and Hasan's (1976) categorisation in three subcategories: in the **sequential** group within the **temporal** category; and within both the **simple** and **conditional** groups in the **causal** category (cited in Tanko, 2004, p.160). Similarly, in Quirk et al.'s (1985) classification of adverbial connectors **then** is entered twice under the listing category (**enumerativ**e and **reinforcing**), as well as under the **summative**, **inferential**, and **contrastive** (**antithetic**) categories (p. 634). These facts may cause confusion for learners in retrieving the proper connector in the proper unit.

Another problem is that the use of connectors is sensitive to register and discourse type (Altenberg and Tapper, 1998). For instance, the connectors used in conversation are highly differing from the ones used in expository prose. Therefore, learning to use connectors appropriately is to learning to produce different types of discourse. That is, connector usage is depends on the development of learners' communicative skills and how language is thought. Altenberg and Tapper (1998) also adds that one more problem for foreign language learners is that the use of connectors tends to vary from one language and culture to another, thus “Languages do not provide identical sets of connectors, and some cultures do not seem to require overt marking of textual relations to the same extent as others”(p. 80)

Lastly, the variety of connectors and the process of the acquisition of them is another problem for language learners. It is difficult for learners to memorize the given lists of conjunctives like other regular lexical items. The range and types of connectors given in textbooks or writing books are varies to a large extent and the selection of connectors are not supplied by empirical evidence (Biber et. al., 1999). Thus, learners get both exhaustive lists of connectors in an impractical way and randomly selected connectors rather than the learning most frequently used ones which may help them in building cohesive links in writing (Tanko, 2004). In addition, textbooks do not explain the different structural forms of connectors consisting of one adverb, a phrase or a clause which may be difficult for learners to identify.

All these factors mentioned above are the potential issues for language learners, particularly for EFL learners in using connectors. According to the studies, these difficulties are generally expressed as misuse, over or under use of connectors in writing as the common point reached by several previous studied explained in the previous section. Accordingly, the research provides opportunity to examine more effective pedagogical approaches in teaching/learning in adverbial connectors. The researchers (Zamel, 1983; Crew, 1990, in Tanko, 2004) suggest that textbooks should provide contextual lists of adverbial connectors for learners which are easy to understand sematic relations within context. Another approach (Granger and Tyson, 1996, Tanko, 2004) indicates that learners need to increase their knowledge on different registers and learn how to use adverbial connectors.

2.6. Chapter Summary

This chapter includes three parts: first, historical overview of corpus linguistics is presented; second, related literature of corpus linguistics is explained in general terms; and lastly, descriptions and related background of adverbial connectors are presented as the main topic of the present study. Explanations in general and specific literature related to current study are focused in order to be comprehensive in respect of corpus-based research.

PART III

METHODOLOGY

3.0. Introduction

This chapter describes the design and the data analysis procedure of the particular study. In the first part, Contrastive Interlanguage Analysis (CIA) is discussed in terms of a specific methodology of data processing procedure for a corpora analysis. Next, the selection and the description of corpora are explained in detail. As learner corpora, ICLE is described with a specification of its Turkish and Spanish sub-corpora which were utilized in this study. In addition, as for native corpora, LOCNESS for native English and TUC Corpus for native Turkish are emphasized. Finally, special attention is given to defining the research process that makes it possible to see the specific structural situations built on adverbial connectors in their entirety.

3.1. Methodology for Corpora Analysis and Research Design

Granger (2002) suggests possible methodological approaches to Computer Learner Corpus analysis and the main method is Contrastive Interlanguage Analysis (CIA). Unlike classic contrastive approaches, CIA compares different of one and the same language and involves the fallowing two types of comparison:

1. Comparison of learner and one or more native speaker reference corpora (L2 vs. L1) and
2. Comparison of different varieties of learner language (L2 vs. L2)

According to Granger (2002), L2 vs. L1 type of comparison helps to uncover the distinguishing features of learner language. In the same way, L2 vs. L2 comparison makes possible to assess the degree of generalizability of interlanguage features across learner populations and language situations.

In the study, both descriptive and quantitative type of research design was conducted in order to see the specific structures built on adverbial connectors in corpora. Therefore, four types of corpora were scanned in terms of eliciting information for the purpose of the study – to see how adverbial connectors in English performed by Turkish adult learners of English and other L2 learners and whether there are native language transfer signals.

The data analysis procedure followed four phases for each corpus:

1. The Analysis of TICLE (Turkish Sub-corpus of ICLE*v*2): TICLE corpus was searched in terms of Turkish interlanguage for the identification of adverbial connector usage. The identified connector structures were evaluated for their type and frequency.
2. The Analysis of Spanish Sub-corpus of ICLE*v*2 : Spanish sub-corpus of ICLE*v*2 was examined in terms of a different L2 interlanguage for the adverbial connector as their type and frequency.
3. The Analysis of Japanese sub-corpus of ICLE*v*2 : Japanese sub-corpus of ICLE*v*2 was examined in terms of a different L2 interlanguage for the adverbial connector as their type and frequency.
4. The Analysis of LOCNESS: LOCNESS was processed as a native English reference corpus for adverbial connector structures. The structures were analyzed for their frequency and type.
5. The Analysis of TUC: TUC corpus was searched as native Turkish reference in terms of adverbial connector use in Turkish language. Identified structures in Turkish were analyzed for their frequency and type.

After the analysis data, four corpora were compared in order to explain the research questions of the study:

- L2 vs. L2: Turkish and Spanish sub-corpus of ICLE corpus were compared to identify the adverbial connector usage to understand whether there were similarities or differences between interlanguages. (Research Questions 1 and 2)
- L2 vs. L1: TICLE corpus was compared with LOCNESS to see whether there were similarities or differences between an L1 and L2. (Research Question 3)

- L2 vs. L1: TICLE and TUC were compared in order to see any L1 transfer signal. (Research Question 4)

3.2. Instruments

In the study, the basis for data collection utilized with five main corpora. Three learner corpora from ICLE*v*2, namely TICLE, JPICLE and SPICLE were selected as learner corpora. As for L1 reference corpora; LOCNESS for English L1 and TUC for Turkish L1 were used in the study. WordSmith Tools software and ICLE*v*2 software have been used in order to analysis of five corpora. Log-likelihood calculation was used as the statistical analysis method for the analyzed data.

3.2.1. Learner Corpora: ICLE*v*2

Granger (2004b) states that learner language is highly variable and it is influenced by a wide variety of linguistic, situational and psycholinguistic factors, and failure to control these factors greatly limits the reliability of findings in learner language research. Therefore, the strict design criteria which should govern all corpus building make corpora a potentially very attractive type of resource for SLA research. Atkins et al. (1992) list 29 variables to be considered by corpus builders. While many of these variables are also relevant for learner corpus building, the specific nature of learner language needs for the interaction of L2-specific variables (Granger, 2004b).

3.2.1.1. The Design of ICLEv2

In the design of ICLE, Granger et al. (2009) decided to adhere the corpus design criteria of Atkins et al., (1992, in Granger et. al., 2009) as possible. In addition, because of the heterogeneous nature of learner data, rigorous data collection procedures which were emphasized by important SLA specialists such as Ellis (1994) were taken into consideration. Ellis (1994) criticizes EA studies in respect of learner data collection and lists some of the factors that can bring about variation in learner output and notes that “unfortunately, many EA studies have not paid sufficient attention to these factors, with the result that they are difficult to interpret and almost impossible to replicate.”(p.49). Gass and Selinker (2001) point out the same issue in relation to cross-sectional SLA studies: “there is often no detailed information about the learners themselves and the linguistic environment in which production was elicited.” (p.33).

The requirements of ICLE which were set at the beginning were as following:

- learners: young adults (university undergraduates); advanced proficiency level; learners of English as a Foreign Learners (EFL) rather than as a Second Language (ESL);
- language: academic writing (mainly argumentative); 200,000 words per corpus.

(Granger et al., 2009, p. 3)

It was decided to include several variables that may influence learner productions. Total of variables were gathered via a learner profile questionnaire, which all learners were requested to fill, and afterwards added in the ICLE data base where they can be used as search criteria. Figure 23 displays the all task and learner variables which were considered in ICLE*v*2 design:

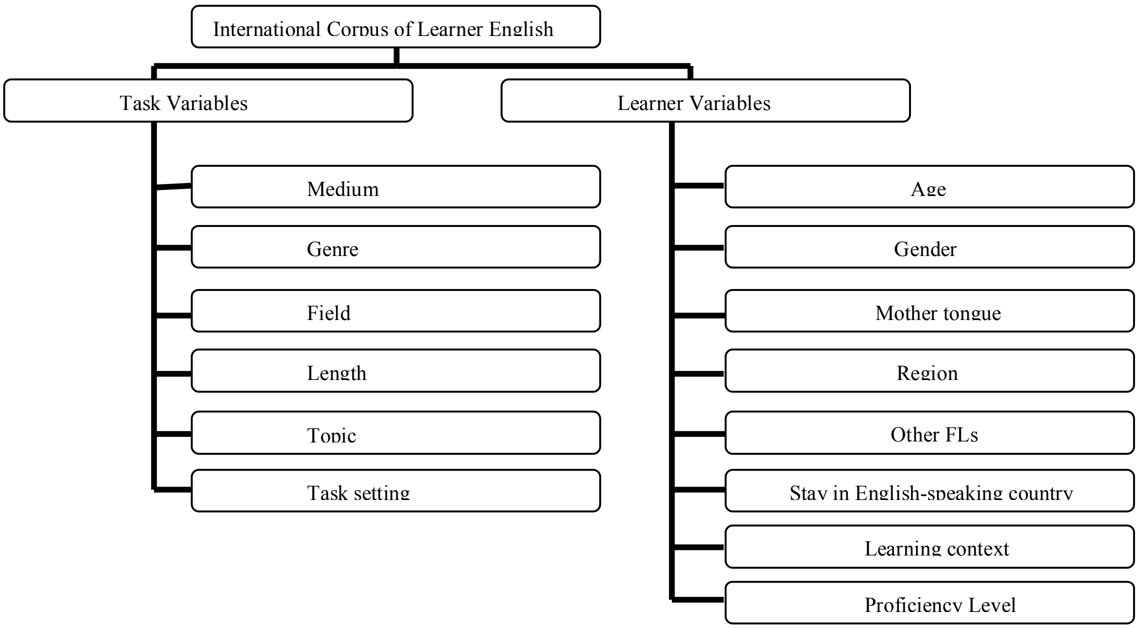


Figure 23. ICLE*v*2 Task and learner variables (adapted from Granger et al., 2009, p. 4).

As resented in Figure 23., ICLE consists of two main variable groups as task and learner variables. Task variables include six components as medium, genre, field, length, topic, task setting. On the other hand, learner variables contains eight major criteria as age, gender, mother tongue, region, other foreign languages, stay in English-speaking country, leaning context and proficiency level.

3.2.1.1.1. Task Variables

The ICLE project aims to collect learner productions that shared a large number of task variables in respect of medium (writing), genre (academic essay), field (general English rather than English for specific purposes) and length (between 500 and 1,000 words. The choice of topic and task settings which requires the arrangement timing, exam conditions and use of reference tools were left to the national coordinators by ICLE team (Granger et al., 2009). All these variables are recorded in database and can later be searched to compile homogeneous corpora.

The majority (91%) of ICLE*v*2 texts consist of argumentative essays which allow for discourse-oriented (cohesion, coherence, argumentative patterns, etc.) as well as lexical and grammatical exploration. Table 23. shows the proportion of argumentative essays in ICLE corpus:

Table 23

Proportion of Argumentative Essays in ICLEv2 (adapted from Granger et al,2009, p.5)

NATIONAL SUBCORPUS	**Argumentative**
BULGARIAN	**100%**
CHINESE	**100%**
CZECH	**81%**
DUTCH	**96%**
FINNISH	**92%**
FRENCH	**85%**
GERMAN	**97%**
ITALIAN	**34[1]%**
JAPANESE	**100%**
NORWEGIAN	**98%**
POLISH	**99%**
RUSSIAN	**100%**
SPANISH	**79%**
SWEDISH	**85%**
TURKISH	**100%**
TSWANA	**100%**
ICLE*v*2	**91%**

The possibility of difficulty in collecting this type of material, national coordinators were given the opportunity to include up to 25% of literary essays (typically literature exam papers). As presented in Table 1, the proportion of argumentative essays ranged from 79% (Spanish corpus) to 100% (Bulgarian, Chinese, Japanese, Russian, Turkish, and Tswana corpora) (Granger et al., 2009).

All the essays are unabridged and have an average length of 617 words. Table 24 shows the average length of all sub-corpora:

Table 24

Average Essay Length in ICLEv2 (adapted from Granger et al., 2009)

NATIONAL SUBCORPUS	**Average Length**
BULGARIAN	663 words
CHINESE	500 words
CZECH	830 words
DUTCH	893 words
FINNISH	704 words
FRENCH	654 words
GERMAN	526 words
ITALIAN	572 words
JAPANESE	542 words
NORWEGIAN	668 words
POLISH	641 words
RUSSIAN	832 words
SPANISH	789 words
SWEDISH	564 words
TURKISH	713 words
TSWANA	384 words
ICLE*v*2	**617 words**

The average length of essays in each sub-corpora is changing. For example the average essay length of Tswana is 384 words whereas Finnish is 704 words.

The essays in ICLE*v*2 contain a wide range of topics. Top ten most popular topic and lists the sub-corpora that have the highest proportion of e each of them are presented in Table 25.:

Table 25

Top Ten Essay Topics in ICLEv2 (adapted from Granger et al., 2009, p. 6-7)

Essay Topic	**Number of essays**	**Country of origin**
Some people say that in our modern world, dominated by science, technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?	491	29% Bulgarian
Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very value.	249	22% Turkish
Poverty is the cause of the HIV/AIDS epidemic in Africa	243	100% Tswana
Marx once said that religion was the opium of the masses. If he was alive at the end of 20th century, he would replace religion with television.	237	19% Russian
The prison is outdated. No civilized country should punish Its criminals; it should rehabilitate them.	176	32% Tswana
Discuss the advantages and disadvantages of banning smoking in restaurants.	156	100% Chinese
Discuss the advantages and disadvantages of using credit cards.	149	100% Chinese
Feminists have done more harm to the cause of women than good.	139	100% Russian
In the words of the old son "Money is the root of the evil" .	133	22% Russian
In his novel "Animal Farm", George Orwell wrote "All men are equal: but some are more than others". How true is this today?	127	39% Bulgarian

Some topics recur in the corpus since many national coordinators used the list of suggested topics provided by the ICLE*v*2 coordinator team in Louvain (Granger et al., 2009).

Task settings in ICLE depend on three conditions: whether the task was timed or untimed, whether it was part of an exam or not, and whether students were allowed to use reference tools to complete the task. Table 4 displays the proportion of task conditions in ICLE:

Table 26

The Proportion of Task Conditions in ICLEv2 (adapted from Granger et al., p.7)

Timed	Untimed	Written under exam conditions	Not written under exam conditions	With the use of reference tools	Without the use of reference tools
38%	62%	39%	61%	48%	52%

Timing and exam conditions are clearly linked; a timed essay is usually part of an examination and an untimed essay is usually written at home. The majority of the ICLE*v*2 essays are untimed (62%), not written under exam conditions (61%) and nearly half of the essays have been written with the help of reference tools (Granger et al., 2009)

3.2.1.1.2. Learner Variables

In respect of learner variables, six of eight variables are clear-cut features as age, gender, mother tongue background, region, knowledge of other foreign languages and time spent in an English-speaking country. Two remained variables as learning context and proficiency level are much fuzzier (Granger et al., 2009).

In respect of age, the essays in ICLE*v*2 were gathered from undergraduate university students so that they are usually in their twenties. Yet there are average age differences among national sub-corpora. For instance, the average age of Turkish learners is higher than Japanese learners. The average age of each national sub-corpora are displayed in Table 27.

Gender distribution in ICLE*v*2 is not completely balanced between male and female learners among national sub-corpora. Some corpora are more female dominated than other such as in Italian (92%). Table 6 also present the gender proportions in ICLE*v*2:

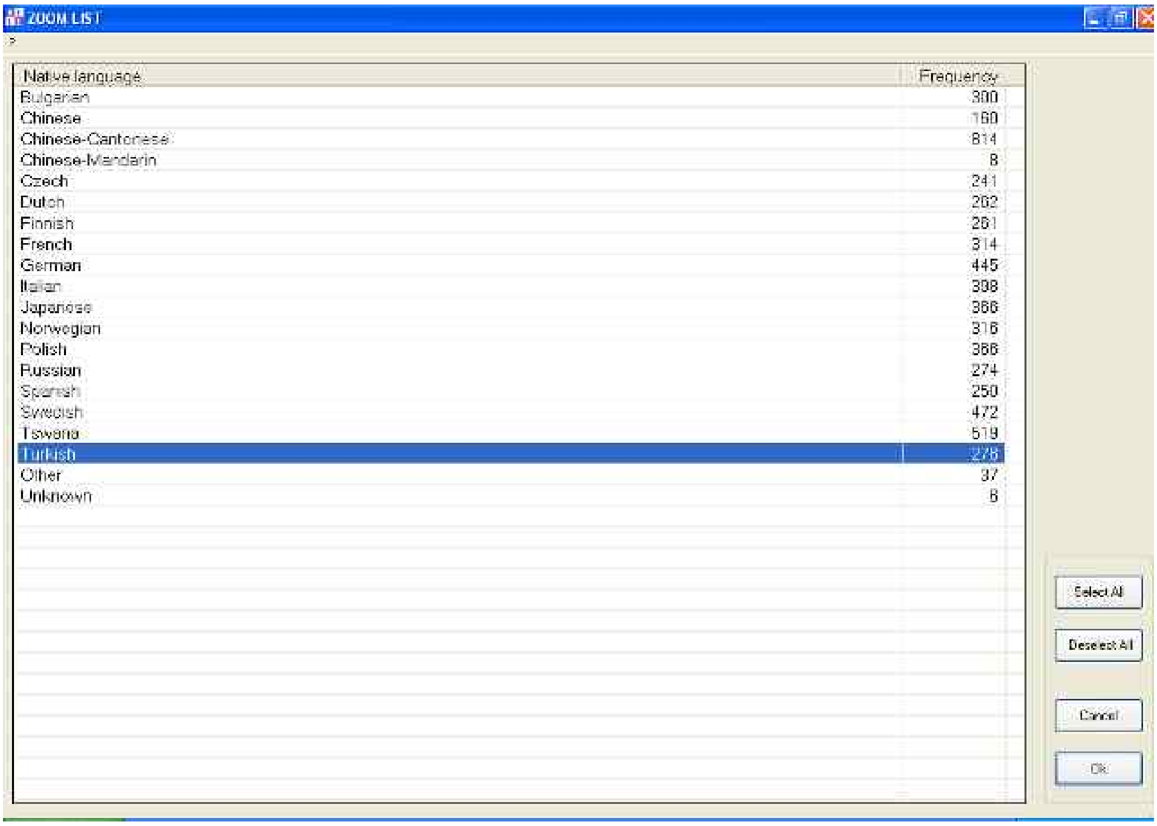
Table 27

Age and Gender Distribution in ICLEv2 (Granger et al., 2009)

National Sub-corpus	**Average Age**	**Learners' Gender**	
		Female %	Male %
Bulgarian	20.55	83%	17%
Chinese	20.49	64%	36%
Czech	22.07	72%	28%
Dutch	20.75	73%	27%
Finnish	22.73	85%	15%
French	21.70	88%	12%
German	23.39	78%	22%
Italian	24.59	92%	8%
Japanese	20.06	73%	27%
Norwegian	23.94	74%	26%
Polish	23.39	80%	20%
Russian	21.19	84%	16%
Spanish	21.72	86%	14%
Swedish	27.73	77%	23%
Turkish	22.08	81%	19%
Tswana	22.47	60%	40%
ICLE*v*2	**22.30**	**76%**	**24%**

As can be seen in Table 27, the total age average of ICLE*v*2 participants is 22.30. on the other hand, 76% of participants in ICLE*v*2 are females and 24% are males.

In ICLE, 16 different native languages are represented by learners. These mother tongue backgrounds are Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish and Tswana. In Figure 24, the screenshot of ICLEv2 software which displays the all native language types and their frequency list is presented:



Native language	Frequency
Bulgarian	300
Chinese	160
Chinese-Cantonese	814
Chinese-Mandarin	8
Czech	241
Dutch	262
Finnish	261
French	314
German	445
Italian	398
Japanese	366
Norwegian	316
Polish	366
Russian	274
Spanish	250
Swedish	472
Tswana	519
Turkish	278
Other	37
Unknown	6

Figure 24. A Screenshot of Learners' Native Language list in ICLEv2 Software (Granger et al., 2009).

In addition to their mother tongue, any other languages that learners speak at home were recorded as well. These languages are listed in decreasing order of use as fist, second or third 'language at home' (Granger et al., 2009, p. 9).

The region variable covers the learners' country of origin. This factor is relevant especially for languages which are spoken in more than one country such as:

- Chinese: Mainland China and Hong Kong;
- Dutch: Belgium and Netherlands
- German: Germany, Austria and Switzerland
- Swedish: Sweden and Finland

(Granger et al., 2009, p.9)

Knowledge of other languages is useful to indicate as the learners' L2 may not be influenced by mother tongue, but also by their knowledge of foreign languages other than English. In ICLEv2 The first foreign language of learners after English is German (32%) and then French (27%) (Granger et al., 2009).

Time spent in an English-speaking country is another learner criterion. Large proportion (45%) of learners reported no stay in an English-speaking country, while 23% reported a stay of three months or more and 19% a stay of less than 3 months (13% is unknown) (Granger et al. 2009).

Learning context is a variable which is described as a 'fuzzy' variable by Granger et. al. (2009). All the learners in ICLE corpus have learned English in a non-English-speaking country so that English is a foreign language rather than second language for them. Granger et al. (2009) states that the line between EFL and ESL can be extremely fuzzy, because the status of exposure to English is changing as being limited in some countries or extensive in some others. However, a certain fact is that learners in ICLE corpus have learned English primarily in classroom setting (Granger, et. al., 2009).

Proficiency is the essential factor to create a generalized picture of EFL learners. The most of ICLE subjects are undergraduate university students (usually in their third or fourth year) and the level of proficiency ranges from higher intermediate to advanced. By ICLE team, a random sample of 20 essays from 16 sub-corpora were submitted to a professional rater who was asked to rate them on the basis of Common European Framework of Reference for Languages (CEF) descriptors of writing. Table 28 presents these CEF results:

Table 28

CEF Results – 20 Essays per Sub-corpus

Mother Tongue	**B2 (and lower)**	**C1**	**C2**	**Total**
Bulgarian	2	16	2	20
Chinese	19	1	0	20
Czech	11	9	0	20
Dutch	1	11	8	20
Finnish	3	8	9	20
French	3	11	6	20
German	1	12	7	20
Italian	10	9	1	20
Japanese	18	2	0	20
Norwegian	8	7	5	20
Polish	1	12	7	20
Russian	3	15	2	20
Spanish	12	8	0	20
Swedish	0	14	6	20
Tswana	18	0	2	20
Turkish	16	4	0	20
Total	**126**	**139**	**55**	**320**

According to CEF results, 60 % of sample essays were rated as advanced (C1 or C2). The proportion is much higher in some sub-corpora, reaches 100% (e.g. Swedish) but it can be low as 10% or less in others. Granger et. al. (2009) point out that

"Although these results need to be firmed up on the basis of more rigorous assessment methods, they are clear indication that some of ICLEv2 sub-corpora are rather in the higher intermediate range while others clearly qualify as advanced" (p.11).

3.3.1.1.3. General Breakdown of ICLEv2

In total, ICLE*v*2 comprises 6,085 essays and 3,753,030 words within sixteen national sub-corpora. Table 29 shows distribution of essays/words per corpus:

Table 29

Distribution of Essays/words per Sub-corpus (adapted from Granger et al., 2009, p. 25)

NATIONAL SUBCORPUS	**Number of Essays**	**Number of Words**
BULGARIAN	302	200,194
CHINESE	982	490,617
CZECH	243	201,687
DUTCH	263	234,723
FINNISH	390	274,628
FRENCH	347	226,922
GERMAN	437	229,698
ITALIAN	392	224,222
JAPANESE	366	198,241
NORWEGIAN	317	211,725
POLISH	365	233,920
RUSSIAN	276	229,584
SPANISH	251	198,131
SWEDISH	355	200,033
TURKISH	280	199,532
TSWANA	519	199,173
ICLE*v*2	**6,085**	**3,753,030**

Each sub-corpus is divided into several batches including variable numbers of essays. The batches and essays are applied a certain coding system to identify them. Each batch of essays has been given 5-character code. First two is the code for nationality: e.g. BG for Bulgarian or TR for Turkish. These two letters are followed by institution code: e.g. NI for Nijmegen or CU for Çukurova University. If there are more than one batch within an institution, it was given a single letter code. For example:

FRUC3 refers to French sub-corpus, Universite catholique de Louvain, 3rd batch

TRCU1 refers to Turkish Sub-corpus, Çukurova University, 1st batch

In the following tables, general breakdown of three national sub-corpus as Turkish, Japanese and Spanish are presented. The batch codes are given in the first column. Essay codes which are made up of batch number are presented in the second column. Third column indicates the breakdown of sun-corpus in terms of text types whether they are argumentative (L), literary (L) or other (O). The forth and fifth columns show the number of essays and the words included in each batch.

Table 30

Turkish Sub-corpus (adapted from Granger et.al., 2009, p.37)

Batch	**Essay Codes**	**Text Types**			**Number of**	**Number of**
		A	**L**	**O**	**Essays**	**Words**
TRCU1	TRCU1001-1177	177	0	0	177	128,297
TRKE2	TRKE2001-2072	72	0	0	72	50,228
TRME3	TRME3001-3031	31	0	0	31	21,007
	Total	**280**	**0**	**0**	**280**	**199,532**

A= Argumentative, L= Literary, O= Other

As presented in Table30, Turkish sub-corpus contains 280 essays and 199,532 words. Major of essays are argumentative essays gathered from three institutions: University of Çukurova, the University of Mustafa Kemal and Mersin University.

Another sub-corpus which is one of the main data groups of the present study is Japanese sub-corpus. In Table 31, general breakdown of Japanese sub-corpus is illustrated:

Table 31

Japanese Sub-corpus (adapted from Granger et.al., 2009, p.33)

Batch	**Essay Codes**	**Text Types**			**Number of**	**Number of**
		A	**L**	**O**	**Essays**	**Words**
JPAI1	JPAI1001-1002	2	0	0	2	1,451
JPDO1	JPDO1001	1	0	0	1	679
JPFJ1	JPFJ1001	1	0	0	1	622
JPHI1	JPHI1001-1005	5	0	0	5	2,887
JPKO1	JPKO1001-1020	20	0	0	20	10,762
JPKO2	JPKO2001-2031	31	0	0	31	18,871
JPKW1	JPKW1001-1002	2	0	0	2	1,160
JPKY1	JPKY1001-1002	2	0	0	2	1,206
JPMI1	JPMI1001-1002	2	0	0	2	1,123

(Table 31 Contunied)

JPMJ1	JPMJ1001-1002	2	0	0	2	1,057
JPMU1	JPMU1001-1002	2	0	0	2	972
JPNH1	JPNH1001	1	0	0	1	538
JPOK1	JPOK1001	1	0	0	1	837
JPRI1	JPRI1001-1002	2	0	0	2	1,109
JPSE1	JPSE1001	1	0	0	1	834
JPSH1	JPSH1001-1004	4	0	0	4	2,422
JPST1	JPST1001-1002	2	0	0	2	1,248
JPSW1	JPSW1001-1039	39	0	0	39	18,338
JPSW2	JPSW2001-2021	21	0	0	21	11,691
JPSW3	JPSW3001-3031	31	0	0	31	16,532
JPSW4	JPSW4001-4032	31	0	0	31	16,735
JPTF1	JPTF1001-1043	43	0	0	43	23,011
JPTK1	JPTK1001-1002	2	0	0	2	987
JPTM1	JPTM1001-1028	28	0	0	28	16,793
JPWA1	JPWA1001-1019	19	0	0	19	9,433
JPWA2	JPWA2001-2009	9	0	0	9	4,291
JPWA3	JPWA3001-3020	20	0	0	20	10,097
JPWA4	JPWA4001-4012	12	0	0	12	7,257
JPWA5	JPWA5001-5029	29	0	0	29	14,649
JPWA6	JPWA6001	1	0	0	1	649
	Total	**366**	**0**	**0**	**366**	**198,241**

A= Argumentative, L= Literary, O= Other

The Japanese sub-corpus contains 366 argumentative essays for a total number of 198,241 words. Twenty-one institutions contributed to ICLE project, some of which provided only one or two essays.

Third sub-corpus is the Spanish sub-corpus which comprises 251 essays with 198,131 words in total. Table 32 displays the Spanish sub-corpus:

Table 32

Spanish Sub-corpus (adapted from Granger et.al., 2009, p.36)

Batch	Essay Codes	Text Types			Number of Essays	Number of Words
		A	L	O		
SPAL1	SPAL1001-01010	0	10	0	10	17,764
SPM01	SPM01005-01021	14	3	0	17	12,762
SPM02	SPM02001-02015	15	0	0	15	9,120
SPM03	SPM03001-03054	53	0	0	53	30,569
SPM04	SPM04001-04057	55	0	0	55	39,626
SPM05	SPM05001-05022	22	0	0	22	16,258
SPM06	SPM06001-06015	0	15	0	15	23,225
SPM07	SPM07001-07025	24	0	0	24	11,854
SPM08	SPM08001-08016	0	16	0	16	20,391
SPM09	SPM09001-09008	0	8	0	8	6,444
SPM10	SPM10001-10006	16	0	0	16	10,118
	Total	**199**	**52**	**0**	**251**	**198,131**

A= Argumentative, L= Literary, O= Other

Although the majority of essays in Spanish sub-corpus are argumentative essays (199), there are a number of literary essays (52) as well.

In the present study, three sub-corpora of ICLEv2 have been used as learner data: Turkish sub-corpora (TICLE), Japanese sub-corpora (JPICLE) and Spanish sub-corpora (SPICLE). The design and components of each sub-corpora described above. ICLEv2 was described as the corpora of learner language. In order to compare the data gathered from these sub-corpora of EFL learners, we need a NS corpus as well. NS corpus of the study is Louvain Corpus of Native English Essays (LOCNESS) which is described below.

3.2.2. Reference Corpus 1: LOCNESS

Corpus-based L2 studies usually depends on contrastive approach which requires to compare the learner corpus (as L2 data) with a native reference corpus (as L1 data) in order to gain insight on quantitative differences between L1 and L2. As explained in Contrastive Interlanguaage Analysis (CIA) L1-L2 comparisons bring out features of L2 properties though a linguistic item (for CIA, see section 2.3.3.8.1.1.). Louvain team who carried out ICLE project has collected a corpus of essays written by English students named Louvain Corpus of Native English Essays (LOCNESS) as the mirror image of the ICLE to ensure the comparability with the ICLE data (Granger, 2009). LOCNESS was designed as a control corpus to enable comparison between learners and native speakers.

In many CIA studies, LOCNESS corpus was used as L1 reference database to compare with learner corpora. Granger and Tyson (1996) used LOCNESS to compare L2 learners' connector usage. Grangers' (1997b) study of participle clauses also compares ICLE and LOCNESS results. Several studies based on comparisons also relied on LOCNESS corpus for example Virtanen (1998) studied direct questions; Rinbom (1998) vocabulary; Altenberg and Tapper (1998) adverbial connectors; Lorenz (1999) intensifiers; Petch-Tyson (2000) demonstratives; Aijmer (2003) modality and Ädel (2008) meta-discourse with utilizing LOCNESS as the control corpus.

LOCNESS corpus consists of native English argumentative essays written by British and (mainly) American students. It is currently contains approximately 300.000 words. The content of LOCNESS is as follows:

- British pupils' A level essays: 60,209 words
- British university students essays: 95,695 words
- American university students' essays: 168,400 words

Total number of words: 324,304 words

(http://www.uclouvain.be/en-cecl-locness.html)

LOCNESS corpus compiled by the Centre for English Corpus Linguistics at the Catholic University of Louvain, Belgium and made available for public use in 1998. The texts of corpus consist of essays gathered from British and American native speakers during the period of 1991-1995. The corpus contains four components as essays of British A-level students, essays of British university students, argumentative essays of American students and literary-mixed essays of American students. Texts types of the corpus contain examination papers, timed essays and free essays. Reference tools were used in some timed and free essays whereas not used in examination papers. The length of essays is around 500 words similar to that of ICLE corpus. The age of participants ranges from 17 to 23 and a small number of older ages. Wide variety of essay topics were in the selection in social problems such as water pollution, nuclear power, sex, violence, gender roles and in campus-related issues such as values and consequences of school interaction, controversy in the classroom or prayer in schools.

Table 33

General Distribution in Selected Component of LOCNESS Corpus.

Institution	**Codes**	**Number of Essays**	**Number of Words**
Marquette University	ICLE-US-MRQ	46	54,285
Indiana University at Indianapolis	ICLE-US-IND	27	13,382
Presbyterian College, South Carolina	ICLE-US-PRB	6	12,447
University of South Carolina	ICLE-US-SCU	53	52,885
University of Michigan	ICLE-US-MICH	43	16,502
	TOTAL	**175**	**149,501**

In the present study, the particular component of LOCNESS, namely the argumentative essays of American students were selected as the control group against ICLE corpus to make comparison as L2-L1. In Table 33. above, the general distribution of selected component (argumentative essays of American students) of LOCNESS which was utilized in the current study is illustrated. The data from LOCNESS has been extracted via WordSmith Tools software.

3.2.3. Reference Corpus 2: TUC

Turkish University Corpus (TUC) is the other native reference corpus which has been used in the present study. TUC contains argumentative essays of native Turkish university students in Turkish language. TUC was begun to be compiled at 2011 led by Cem Can who was the coordinator and the compiler of Turkish national sub-corpus of ICLE (TICLE). Texts in TUC were gathered from two institutions; University of Çukurova and Kahramanmaraş Sütçü İmam University In Table 34., the general distribution of TUC corpus is presented:

Table 34

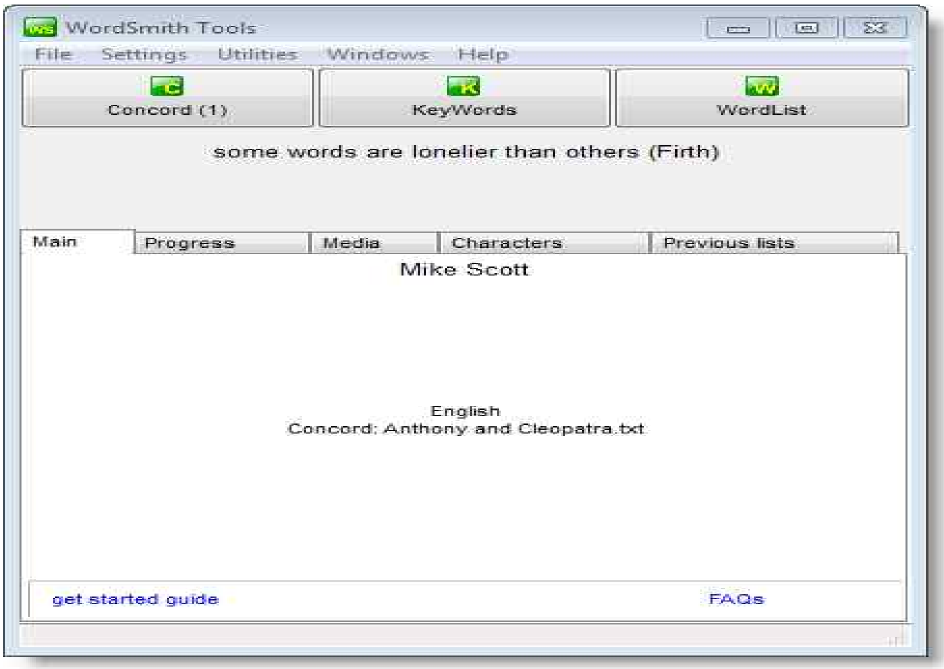
General Distribution of TUC Corpus

Institution	**Codes**	**Number of Essays**	**Number of Words**
University of Çukurova	TUC-CU	75	45,119
Kahramanmaraş Sütçü İmam University	TUC-KSU	108	62,757
	TOTAL	**183**	**107,876**

In Table 34, the total number of TUC is 107,876 words of texts gathered from 183 participants. University of Çukurova includes 75 participants' text of 45,119 words and Karamanmaraş Sütçü İmam University 108 text of 62,757 words. Design criteria of TUC are based on ICLE corpus which provides comparability of argumentative texts of Turkish university students in their L1. Essay topics of ICLE were translated and presented to participant to submit argumentative essays in their mother tongue. Since the participant of TUC are university students (as in ICLE and LOCNESS), average of age ranges between 19-24 and a few of older age of participants. TUC has not been made available for public research yet.

3.2.4. WordSmith Tools 5

WordSmith Tools is a software package for analysing the lexis of texts and corpora, developed by Mike Scott (1996, 1998, 1999, 2004). WordSmith Tools basically can be used to produce *frequency* lists, to run *concordance* searches and calculate collocations for particular words, and to *find keywords* in a text and examine their distribution (Baker, Hardie and McEnery, 2006). Figure 25 shows the main screen of WordSmith Tools:



some words are lonelier than others (Firth)

Mike Scott

English
Concord: Anthony and Cleopatra.txt

Figure 25. The main screen view of wordsmith tools controller (adapted from http://www.lexically.net/wordsmith/step_by_step/index.html).

WordSmith Tools provides *concordance* program which lists the occurrence of given word or phrase in a corpus (Scott, 2001). The point of a concordance is to be able to see lots of examples of a word or phrase, in their contexts.

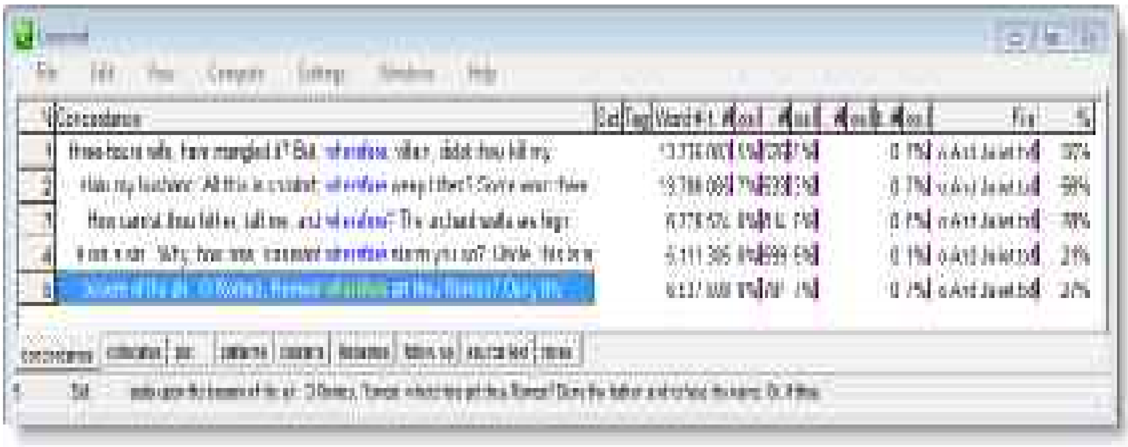
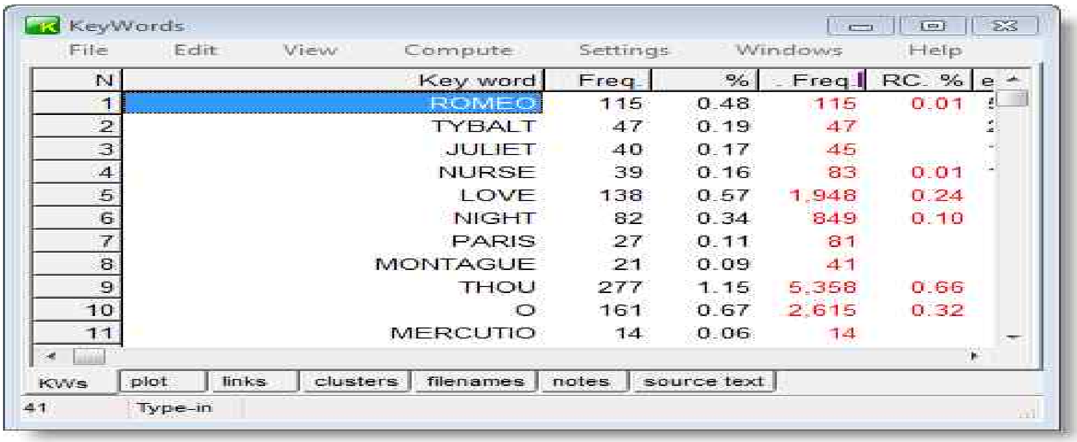


Figure 26. A screenshot of concordance menu in wordsmith tools (adapted from http://www.lexically.net/wordsmith/step_by_step/index.html)

After the selection of Concordance menu from main screen of WordSmith Tools in Figure 25. (should be loaded with a particular corpus), the word or phrase is entered as *wherefore* in Figure 26, then all lines including *wherefore* are listed in the screen. As can be seen, there are five examples of *wherefore* in concordance list.

WordSmith tools also offers a program, known as *KeyWords* . The key words are words which occur unusually frequently in comparison with some kind of reference corpus. Figure 27 presents the *KeyWords* page in software:



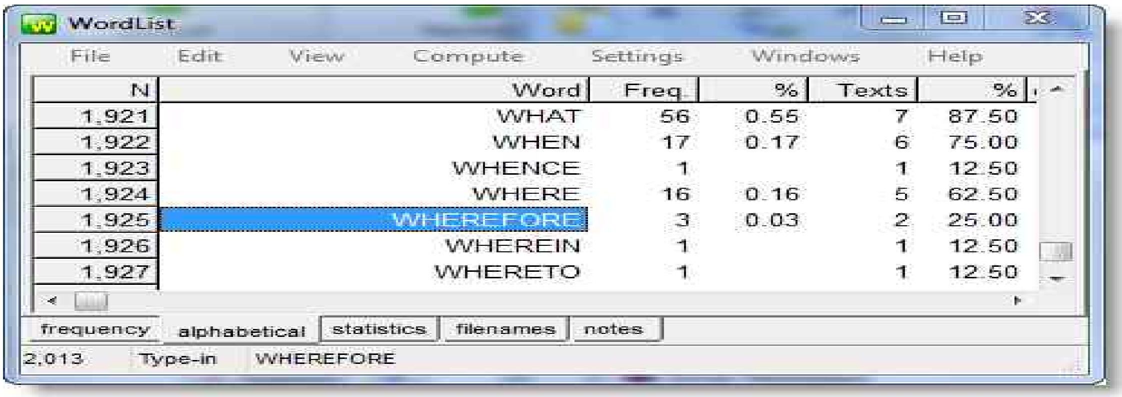
N	Key word	Freq.	%	Freq.	RC. %	e
1	ROMEO	115	0.48	115	0.01	!
2	TYBALT	47	0.19	47		:
3	JULIET	40	0.17	45		-
4	NURSE	39	0.16	83	0.01	-
5	LOVE	138	0.57	1,948	0.24	
6	NIGHT	82	0.34	849	0.10	
7	PARIS	27	0.11	81		
8	MONTAGUE	21	0.09	41		
9	THOU	277	1.15	5,358	0.66	
10	O	161	0.67	2,615	0.32	
11	MERCUTIO	14	0.06	14		

41 Type-in

Figure 27. A screenshot of key word page in wordsmith Tools (adapted from http://www.lexically.net/wordsmith/step_by_step/index.html).

In the list above in Figure 23, based on the play Romeo and Juliet in comparison with all the Shakespeare plays, lots of names of the main characters, some pronouns like **thou**, plus theme words like **love** and **night** are seen. The numbers beside key words show the how frequent each one in the texts (http://www.lexically.net/wordsmith/step_by_step/index.html).

A *Word list* helps the language researcher identify the common words in a corpus, information which is useful for example when determining which lexical items to teach and which to ignore, or when the materials writer is attempting to ensure that new vocabulary is met more than once in a textbook (Scott, 2001). Figure 28 shows a word list in WordSmith Tools WordList program:



N	Word	Freq.	%	Texts	%
1,921	WHAT	56	0.55	7	87.50
1,922	WHEN	17	0.17	6	75.00
1,923	WHENCE	1		1	12.50
1,924	WHERE	16	0.16	5	62.50
1,925	WHEREFORE	3	0.03	2	25.00
1,926	WHEREIN	1		1	12.50
1,927	WHERETO	1		1	12.50

2,013 Type-in WHEREFORE

Figure 28. A screenshot of wordlist page in wordsmith tools (adapted from http://www.lexically.net/wordsmith/step_by_step/index.html).

WordList shows how often each word occurs in the text files, what that is as a percent of the running words in the text, and how many text files each word was found in.

WordSmith Tools has been used in several studies by means of describing and analyzing various issues in corpus research (Scott, 1997; Granger and Tribble, 1998; Sardinha, 1999; Barbara and Scott, 1999; Xiao and McEnery, 2005, Astrid and Johnson, 2006; Gilquin and Paquot, 2007, 2008, cited inhttp://lexically.net/wordsmith/corpus_linguistics_links/papers_using_wordsmith.htm) . In the present study, WordSmith Tools 5 was utilized in concordance analysis of adverbial connectors.

3.3. The Log-likelihood Statistics

Log-likelihood (LL) is a test for statistical significance, similar to the Pearsons' Chi-square measure that is often used in corpus analysis, for example for collocation, keyword or frequency analysis. This test is sometimes called ***G-square*** or ***G score.*** In statistical analysis of texts, to test the frequency distributions, LL test is a reliable alternative to Pearsons' Chi-square (Dunning, 1993). LL test considers word frequencies weighted over two different corpora. It measures higher or lower frequencies than expected. G2 score or LL is Log-likelihood value is as p value in Pearsons' Chi-square (McEnery, Xiao& Tono, 2006). Dunning (1993) states that:

> For text analysis and similar problems, the use of likelihood ratios leads to very much improved statistical results. The practical effect of this improvement is that statistical textual analysis can be done effectively with very much smaller volumes of text than is necessary for conventional tests based on assumed normal distributions, and it allows comparisons to be made between the significance of the occurrences of both rare and common phenomenon.
>
> (1993, p.65)

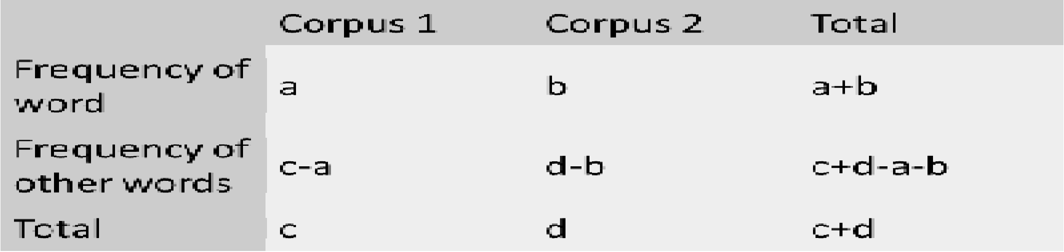
Like Pearsons' Chi-square, LL compares the observed and expected values for two datasets. These two concepts of LL are as follows:

Obverved values: actual frequencies extracted from corpora.

Expected values: the frequencies that one would expect if no factor other than chance were affecting the values. The greater the difference between the observed and the expected values, the less likely it is the difference has arisen by chance.

Dunning (1993) states that the chi-squared value becomes unreliable when the expected frequency is less than 5 and possibly overestimates with high frequency words and when comparing a relatively small corpus to a much larger one. So Dunning (1993) suggest the log-likelihood ratio as an alternative to Pearson's chi-squared test.

Log-likelihood is calculated by constructing a contingency table. Figure 24. presents a contingency table which shows the calculation of LL:



	Corpus 1	Corpus 2	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Tctal	c	d	c+d

Figure 29. Contingency table for LL calculation (adapted from Ryson & Garside, 2000, p. 3).

In Figure 29, value 'c' corresponds to the number of words in corpus one, and 'd' corresponds to the number of words in corpus two (N values). The values 'a' and 'b' are called the observed values (O). We need to calculate the expected values (E) according to the following formula (Ryson & Garside, 2000, p. 3):

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

In this case N1 = c, and N2 = d. So, for this word, E1 = c*(a+b) / (c+d) and E2 = d*(a+b) /(c+d). The calculation for the expected values takes account of the size of the two corpora, so it is not needed to normalise the figures before applying the formula. Then the LL value can be calculated according to this formula (Ryson and Garside, 2000, p. 3):

LL = 2*((a*log (a/E1)) + (b*log (b/E2)))

or following formula also shows the similar calculation for LL value :

$G_2 = {}_2\Sigma x_{ij}(\log_e x_{ij} - \log_e m_{ij})$

Where ***xij*** are the data cell frequencies, ***mij*** are the model cell frequencies, **log*e*** represents the logarithm to the base **e**, and the summation is carried out over all the squares in the table (Oakes 1998, cited in Baker, Hardie & McEnery, 2006, p.110).

Contingency table (or frequency table) as presented in Figure 24. which has two rows and two columns has 1 degree of freedom. In LL, critical values with 1 df are 3.84, 6.64 and 10.83 for the significance levels of 0.05, 0.01 and 0.001. For example, 28.841 (1 df) is greater than 10.83 for the significance level 0.001..So we are more than 99.99 percent confident that difference in the frequencies observed in two corpora is statistically significant. If the LL results is greater than 6.64; the difference between two corpora happening by chance is less than 1%. If LL value is 3.84 or more, the probability of it happening by chance is less than 5%. Or it is $p<0.05$. Or 95% certainty of results (McEnery, Xiao & Tono, 2006).

LL ratio measurement has been taken into account by many researchers in the corpus linguistics field. Ryson & Garside (2000) consider LL measurement for corpus comparison by frequency profiling. Scott (2001) also uses LL in his keywords procedure. Ryson et. al. (2004) discusses the reliability of LL value against chi-squared statistic in word frequency comparisons. They conclude that, in order to extend applicability of the frequency comparisons to expected values of 1 or more, the use of the log-likelihood statistic is preferred rather than chi-squared statistic, at the 0.01% level, indeed the trade-off for corpus linguists is that the new critical value is 15.13 (Ryson et. al., 2004).

In respect of L2 corpora analyses, LL calculation was used in significant studies. For instance, Granger and Ryson (1998) compare automatic profiling in ICLE and LOCNESS as two similar sized corpora. They state that they used LL calculation beside chi-squared statistic since LL does not suffer the same problems as chi-squared does with unbalanced sample sizes (Granger and Ryson, 1998). Narita et. al. (2004) studied adverbial connectors through ICLE and LOCNESS comparison (as an inspiration for this present study) and they used LL calculation for the analysis of two corpora. However, in contrast to Granger and Ryson (1998), Narita et. al. (2004) states that they used LL since two corpora were differed in size.

In the present study, LL calculation was selected as a statistical measurement. There are more than two corpora (LOCNESS, TICLE, JPICLE, SPICLE and TUC) in different sizes (as in Narita et.al., 2004) have been used in the study so that there it was very likely to see differences in same cases. Therefore, as noted above, LL calculation might be the reliable measurement for the comparison of five corpora.

3.4. Data Analysis

The data has been gathered through identifying the connector usage in LOCNESS, TICLE, SPICLE, JPICLE and TUC. The elicited from corpora has been processed by frequency analysis and then log-likelihood (LL) analysis by means of comparing the data groups. In the first phase, overall frequency analysis of all corpora has been done in order to see the total of connector usage. Next, frequency connectors were analysed in semantic categories across corpora, and lastly, individual connectors were investigated within their frequency. In order to confirm the identified frequency differences statistically, LL analysis has been applied to all comparisons.

Overall, categorical and individual frequency results then compared by LL among corpora: LOCNESS frequency results compared with orderly TICLE, SPICLE and LOCNESS in respect of L1 vs. L2 comparison. In order to see the condition among learner groups, overall connector frequency of TICLE, SPICLE and LOCNESS has been compared by LL ratio by means of L2 vs. L2 comparison. In addition, as Turkish L1 production of Turkish students, overall frequency of connectors in TUC corpus has been compared with TICLE as the L2 production of L1 Turkish students to see whether there is a difference (for an L1 transfer trace) within same participant group in different tasks.

3.5. Chapter Summary

The present chapter describes the design, methodology, data types, tools and data analysis procedure of the present investigation. At first, design and the methodology of the study were explained as a start. Next, the backbone of the study, the instruments were described in detail, namely five corpora: ICLE with its three sub-corpora, TICLE, SPICLE and JPICLE; LOCNESS and TUC corpora. WordSmith Tools was explained as the software for concordancing procedure of connectors in corpora. Log-likelihood analysis was mentioned which has been utilized in the present study as the statistical instrument. Lastly, the data analysing procedure was given as the focus of the methodology.

CHAPTER IV

RESULTS AND DISCUSSION

4.0. Introduction

The present chapter reports the findings from data analysis of five corpora, namely LOCNESS, TICLE, SPICLE, JPICLE and TUC. As outlined in the last chapter, data analysis covers the processing of 157 types of English adverbial connectors (under seven categories) across five corpora via WordSmith Tools and ICLE software. After concordance procedure through software, total frequency analyses of connectors in corpora were applied and then statistical analyses for categorical and individual frequencies of connectors were measured across four corpora (LOCNESS, TICLE, SPICLE and JPICLE) by means of L1-L2 and L2 –L2 comparison. Turkish equivalents of connectors have been identified over TUC in order to compare the results with TICLE corpus to compare Turkish learners' usage of connectors in their L1 and L2 writings. This chapter presents the series of frequency and statistical processes which set a unique quantitative analysis for adverbial connectors across five corpora.

4.1. Results

4.1.1. Overall Frequency of Adverbial Connectors Across Corpora

The first step in the analysis, the connectors used in corpora have been gathered and calculated regardless of their category. Each of adverbial connectors from 157 types was identified over LOCNESS, TICLE, SPICLE and JPICLE. By concordancing via software, all instances of adverbial connectors have been found over corpora, then the number of each instance of connectors were calculated and lastly a total frequency of each connector has been obtained. The frequency calculation was made in order to determine the proportion of adverbial connectors in L1 and L2 corpora, thus, the frequency results might be compared each other.

Initially, when compared to non-native speakers in frequency, native speakers fall into the lowest amount as 1277, whereas it is 2590 in TICLE, 1851 in SPICLE and 2844 in JPICLE. As a matter of fact, this means English adverbial connectors have been overused in all three learner corpora when compared to native speakers in frequency analysis. This condition is illustrated in a Figure 30 below:

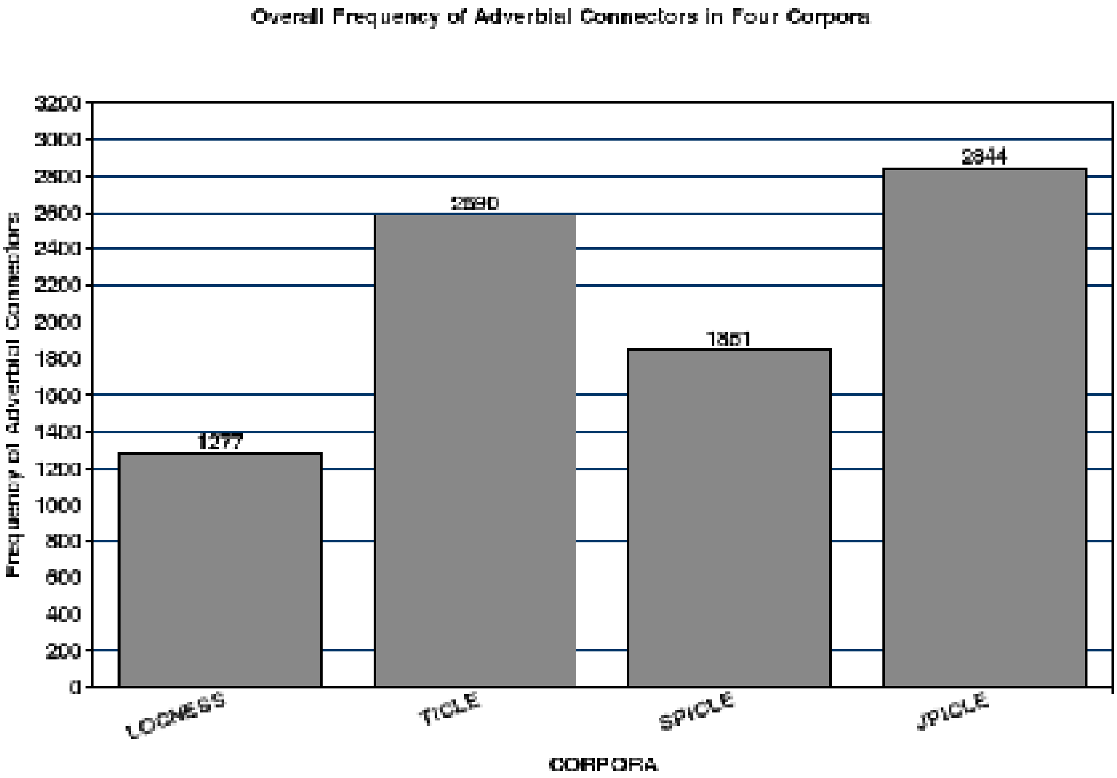


Figure 30. Overall frequency distribution of adverbial connectors in four corpora.

Figure 30 presents a clear picture of frequency differences among four corpora. As can be seen, the lowest frequency belongs to LOCNESS which represents most accurate usage as being the native language data. The highest frequency of connector usage is in JPICE, and then comes TICLE and SPICLE corpora. The significant overuse seems to be in Japanese learners which contains more than two times (2844) from native speakers (1277). There is a difference of 1567 between LOCNESS and JPICLE in number. Similarly, Turkish learners also used approximately more than two times of native speakers' connector usage. As the last L2 data, the lowest number of connector usage is in SPICLE which is the closest to native speakers in terms of usage frequency.
In order to see the frequencies of connectors in four corpora, it is essential to see the amounts within the total size of words and tokens. In table 1., a general distribution of English adverbial connectors across four corpora is presented in Table 36 :

Table 36

Overall Comparative Frequency Distribution of Adverbial Connectors in Four Corpora

	LOCNESS (L1)	TICLE (L2)	SPICLE (L2)	JPICLE (L2)
Corpus Size in words	168,400	171,145	180,367	168,360
Connectors (n)	1277	2590	1851	2844
n per 10,000	76	151	103	169
T/t ratio (%)	0.75	1.51	1.02	1.68
Number of connector types used	79	86	110	79

n= raw frequency of connectors

T/t ratio= Type/token ratio; percentage of number of connectors (types) in total of words (tokens) in each corpus

In Table 36, total frequency of adverbial connector usage in four corpora is given by means of total number of connectors, proportion of connectors per 10.000 and total number of connectors types. Altenberg and Tapper (1998) and Tanko (2004) also examined connectors' rate for per 10,000 words across native and learner corpora. As the corpora sizes are similar in this study, the identification of connectors in every 10,000 words might give a clearer view of possible differences in total connector frequency in each corpus. Accordingly, as shown in Table 1, there is a variable condition in terms of frequency of connectors among corpora. The lowest number of connectors in total belongs to native data in LOCNESS corpus. Native speakers used connectors 1277 times, in other words, they used 76 connectors in every 10,000 words. However, L1 English data is the most accurate usage since the language item under consideration is an English language structure so that the usage of connectors in LOCNESS is accepted as the correct forms. This means all other L2 data overused the connectors in their L2 writings. For example, Turkish learners used 2590 connectors in total and there are 151 connectors per 10,000 words. In SPICLE, which has the closest frequency to native speakers, 1851 connectors were used in total. The most significant overuse seems to be by Japanese learners who used 2844 connectors as the highest frequency among all corpora. In addition, Japanese learners used 169 tokens per 10,000 words two times more than native speakers' 76 tokens. Turkish learners' number of tokens is also doubles native speakers as well (151 vs.76). Although Spanish learners have the lowest adverbial connector frequency among other L2 corpora, they nevertheless overused connectors when compared to L1 corpus. Type/token ratio of connectors also represents the percentage of each connector within all words in a corpus, that is, the number of connectors falls into per 100 words. Accordingly, the T/t ratio of connectors used in LOCNESS is 0.75 whereas it is 1.51 in TICLE, 1.02 in SPICLE and 1.60 in JPICLE. This means that Turkish learners used 1.51 of their every 100 words as an adverbial connector, Spanish learners used 1.02 and Japanese learners used 1.60 of their every 100 words as an adverbial connector. Again, Japanese learners had the highest percentage in T/t ratio in total.

Bar graphic below shows the number connector types used in four corpora:

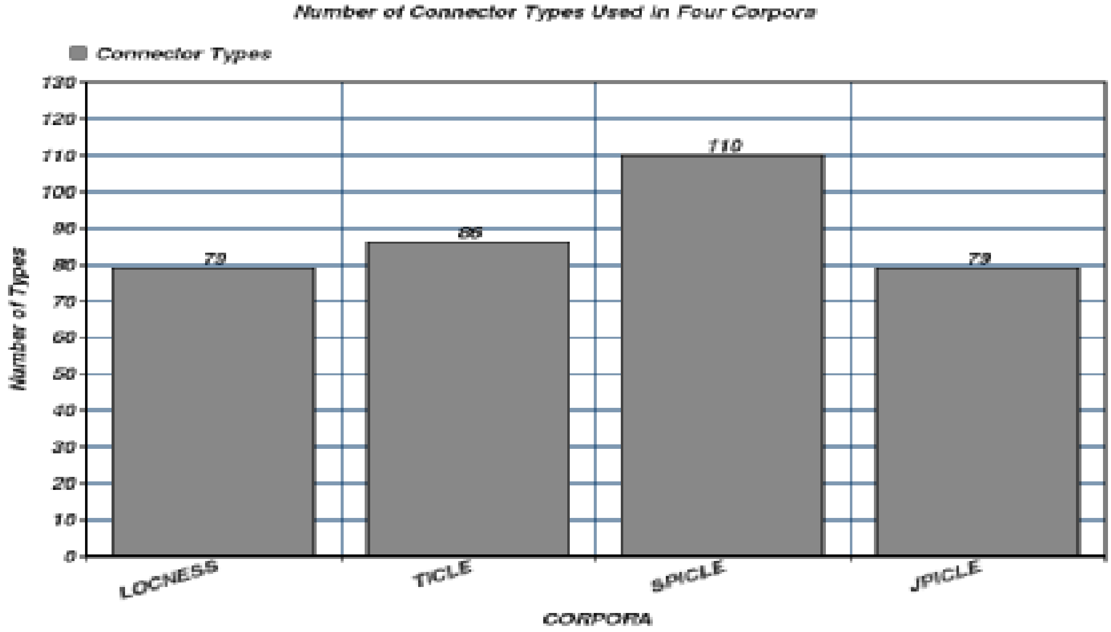


Figure 31. Number of connectors types in four corpora.

In respect of connector types, native speakers and Japanese learners have the same number as 79 of 157 different types. In TICLE, it is a bit much more connector types were have been used (86 types) whereas Spanish learners used significantly more connector types.

SPICLE includes 110 different types of connectors although it has the lowest token frequency (1851) among other learners. Thus, the number of connector types varies regardless of L1 or L2 data.

In addition to frequency analysis, there is a need to see the significant values of overuse or underuse in corpora by means of statistics. LL ratio statistics is the measurement which supports the difference between frequencies of certain items observed in corpora. LL calculation not only indicates the overuse or underuse between corpus but also shows that whether the difference aroused from overuse or underuse is statistically significant or not. In the study, LL ratio have been used to compare the native corpus and learner corpora (L1 vs. L2) and then to learner corpora each other (L2 vs. L2).

Firstly, LL calculation has been made in overall frequency of adverbial connectors identified in all corpora. Initial LL analysis was made between LOCNESS and TICLE corpora to test the overall frequency difference occurred between them. Table 1 shows the LL ratio of connectors in comparison with TICLE and LOCNESS.

Table 37

LL Ratio of Overall Adverbial Connectors in TICLE and LOCNESS

	TICLE (O1)	%1	**LOCNESS** (O2)	%2	**LL Ratio** (*p < 0.05)
Connectors	2590	1.51	1277	0.76	+ 433.83

O1 is observed frequency in Corpus 1
O2 is observed frequency in Corpus 2
%1 and %2 values show relative frequencies in the texts.
+ indicates overuse in O1 relative to O2,
- indicates underuse in O1 relative to O2

The LL value is handled by a contingency table (see chapter 3) in which corpora size and observed item frequency are calculated. In Table 37, O1 and O2 refer to overall frequency of adverbial connectors observed in TICLE and LOCNESS. On the other hand, %1 value includes the relative frequency of connectors in the texts, i.e., 1.51 relative frequencies in TICLE means there are approximately 1.51 connectors fall into every 100 words in TICLE. In the same way, relative frequency of LOCNESS revealed 0.76 connectors per 100 words. According to the result, LL ratio measurement indicates an overuse in TICLE with an + 433.83 LL value ($p < 0.05$). There is a significant difference between two corpora in terms of connector frequency ($p < 0.05$), so the overuse in TICLE relative to LOCNESS has been approved by LL calculation.

Next comparison is between SPICLE and LOCNESS which might be another step in comparing overall frequency of connector usage by means of L1 vs. L2. Table 38 presents LL ratio of the overall adverbial connector frequency in SPICLE and LOCNESS below:

Table 38

LL Ratio of Overall Adverbial Connectors in SPICLE and LOCNESS

	SPICLE (O1)	%1	**LOCNESS** (O2)	%2	**LL Ratio** (*p < 0.05)
Connectors	1851	1.03	1277	0.76	+ 70.21*

O1 is observed frequency in Corpus 1
O2 is observed frequency in Corpus 2
%1 and %2 values show relative frequencies in the texts.
+ indicates overuse in O1 relative to O2,
- indicates underuse in O1 relative to O2

As expected from frequency differences, the overuse in SPICLE in contrast to LOCNESS revealed +70.21 LL value which is statistically significant ($p < 0.05$). Relative frequency per 100 words in each corpus also shows a difference between two corpora (1.03 connectors in SPICLE and 0.76 connectors in LOCNESS per 100 words).

Third L1 vs. L2 comparison in order to see the overuse or underuse statistically is between JPICLE and LOCNESS. Initial frequency analyses showed the most significant frequency differences in JPICLE when it compared to LOCNESS. Table 39. shows LL value between JPICLE and LOCNESS:

Table 39

LL Ratio of Overall Adverbial Connectors in JPICLE and LOCNESS

	JPICLE (O1)	%1	**LOCNESS** (O2)	%2	**LL Ratio** (*p < 0.05)
Connectors	2844	1.69	1277	0.76	+ 611.48*

O1 is observed frequency in Corpus 1
O2 is observed frequency in Corpus 2
%1 and %2 values show relative frequencies in the texts.
+ indicates overuse in O1 relative to O2,
- indicates underuse in O1 relative to O2

LL value between JPICLE and LOCNESS revealed a high amount of overuse as +611.48 ($p < 0.05$) as expected from the highest frequency difference between them. The higher the LL value means the more significant overuse as in Table 1. The relative frequency between two corpora also explains the significant difference between them.

In sum, LL values by means of L1 vs. L2 supported the overuse in learner corpora which was observed in frequency analyses. LL calculation also showed the the highest and the lowest significant differences between L1 and L2s. For instance, the highest LL value between an L1-L2 is in JPICLE (+611.48) which indicated a very significant overuse by Japanese learners. Second significant difference is between TICLE and LOCNESS as 433.83 LL value also marks a considerable overuse in Turkish learners when compared to native speakers. Although the LL value between SPICLE and LOCNESS is the lowest among other, the LL ratio of 70.21 between them is another statistical support of overuse in learners.

The next step in comparing corpora is analysing the statistics of frequency difference between learners' usage of connectors. LL ratio is applied to see the statistical significance of frequency differences or over/underuse among learners. As the major concern of the study, TICLE corpus which represents the Turkish EFL learners L2 productions were compared with other EFL learners by means of LL ratio for frequency differences.

First LL calculation has been made between TICLE and SPICLE which normally indicated overall frequency difference (2590 in TICLE and 1851 in SPICLE). As can be seen in Table 40, LL value is +165.39 which indicates a significant difference and a high overuse in terms of Turkish learners. Relative frequency is 1.51 per 100 words in TICLE whereas it is 1.03 per 100 words in SPICLE.

Table 40

LL Ratio of Overall Adverbial Connectors in TICLE and SPICLE

	TICLE (O1)	%1	**SPICLE** (O2)	%2	**LL Ratio** (*p < 0.05)
Connectors	2590	1.51	1851	1.03	+ 165.39*

O1 is observed frequency in Corpus 1
O2 is observed frequency in Corpus 2
%1 and %2 values show relative frequencies in the texts.
+ indicates overuse in O1 relative to O2,
- indicates underuse in O1 relative to O2

LL ratio confirmed the overuse of Turkish learners against Spanish learners. This result is then going to be re-examined by calculation of LL values of individual connectors which might give an explain for the significance of overuse by Turkish learners.

Next L2 vs. L2 comparison has been made between Turkish EFL learners and Japanese learners in terms of overall adverbial connector usage. The observed overall frequency in TICLE is 2590 whereas it is 2844 in JPICLE so that there is an expected overuse in JPICLE or underuse in TICLE due to the frequency difference in number. Table 41 displays the LL value of overall frequency of connectors in TICLE against JPICLE:

Table 41

LL Ratio of Overall Adverbial Connectors in TICLE and JPICLE

	TICLE		**JPICLE**		**LL Ratio**
	(O1)	%1	(O2)	%2	(*p < 0.05)
Connectors	2590	1.51	2844	1.69	-16.41*

O1 is observed frequency in Corpus 1
O2 is observed frequency in Corpus 2
%1 and %2 values show relative frequencies in the texts.
+ indicates overuse in O1 relative to O2,
- indicates underuse in O1 relative to O2

When TICLE compared with JPICLE in LL ratio, it indicated an underuse in LL value as -16.41 which is statistically significant. Although revealing a relatively less value, there is a certain underuse in Turkish learners when compared to Japanese learners.

Last L2 vs. L2 comparison was made between SPICLE and JPICLE or vice-versa. As noted above, LL value of TICLE between SPICLE indicated overuse in TICLE since there was also frequency difference between them. The similar condition is relevant for SPICLE and JPICLE comparison because there is also a considerable amount of overall frequency difference between them (1851 connectors in SPICLE and 2844 connectors in SPICLE). Table 42 shows the LL ratio of underuse in SPICLE against JPICLE:

Table 42

LL Ratio of Overall Adverbial Connectors in SPICLE and JPICLE

	SPICLE (O1)	%1	**JPICLE** (O2)	%2	**LL Ratio** (*p < 0.05)
Connectors	1851	1.03	1277	1.69	-285.59*

O1 is observed frequency in Corpus 1
O2 is observed frequency in Corpus 2
%1 and %2 values show relative frequencies in the texts.
+ indicates overuse in O1 relative to O2,
- indicates underuse in O1 relative to O2

The revealed underuse in SPICLE is -285.59 LL value which is considerably more than -16.41 LL ratio of underuse in TICLE-JPICLE comparison. The LL values of overall frequency of adverbial connectors used in learner corpora are shown in Table 43:

Table 43

LL Ratio of Overall Frequency of Adverbial Connectors among TICLE, SPICLE and JPICLE

L2 vs. L2	LL Ratio (*p < 0.05)	Overused/Underused
TICLE-SPICLE	+165.39*	Overused in TICLE
TICLE-JPICLE	-16.41*	Underused in TICLE
SPICLE-JPICLE	-285.59*	Underused in SPICLE

In respect of L2-L2 comparison, the LL ratio showed overall overuse or underuse among corpora. The same process has been applied to comparison of LL values individual connectors to see the details of differences at item level.

In sum, it can be inferred that there is an overall overuse in learners when compared to native speakers. Indeed the English adverbial connectors mostly overused by Japanese learners and Turkish learners. Among learner group, Spanish learner data is moderately closer one to native speakers in the use of adverbial connectors. That is, initial frequency analysis indicates overuse of adverbial connectors by L2 learners by means of L1 vs. L2 comparison. The connector types used in four corpora indicates a diversity as there are similarities and differences between L1-L2 data in type number.

Another overall frequency comparison of L1 vs. L2 has been made between TUC (L1 Turkish data) and TICLE (L2 English data of Turkish natives) in order to examine any probability of L1 transfer in Turkish EFL learners, namely influencing from connector system in Turkish while writing an essay in English.

Table 44

Overall Comparative Distribution of Adverbial Connectors in TUC, TICLE and LOCNESS

	TUC	**TICLE**	**LOCNESS**
	L1 Turkish	L2 English	L1 English
Corpus Size in words	107,876	171,145	168,400
Connectors (n)	2462	2590	1277
n / 10,000	228	151	76
T/t ratio (%)	2.28	1.51	0.76
Number of connector types used	163	86	79

n= raw frequency of connectors

T/t ratio= Type/token ratio; percentage of number of connectors (types) in total of words (tokens) in each corpus

Table 44 shows that Turkish students use much more connectors in their essays in Turkish than in English. Although TUC and TICLE corpora differ in size, frequency of connectors in two corpora revealed that Turkish students have been used 2462 connectors in the essays in Turkish while they have been used 2590 connectors in the essays written in English. By means of frequency per 10,000 words, 228 connectors have been used in Turkish essays by Turkish students whereas 151 connectors have been used in every 10,000 words in their essays in English. In other words, Turkish students used 2.2 per cent connectors in total in their L1 essays and 1.51 per cent in their L2 English essays. That is, Turkish students have no problem with the connector usage while writing an essay in Turkish when compared their essays in English. On the other hand, Turkish student have used 163 different types of connectors in their L1 essays whereas 86 different types in their L2 English essays. That is, Turkish students use a wide range of connector types in their L1 essays than in their L2 English essays.

When the frequency of native Turkish students' connector usage compared to native English students'', Turkish natives use much more connectors in their essays than native English students (2.2% in total vs. 0.76% in total). Overall connector frequency of TUC and LOCNESS indicate nearly twice as much difference in Turkish natives (2462 vs. 1277) as well as in connector types (163 vs. 79) although there are more adverbial connector types in English (175 types) (Quirk et. al., 1985) than in Turkish (163 types). That is, native Turkish students tend to use more connectors than native English students which might be a factor that influence Turkish students' L2 writing by means of L1 transfer.

In order to test the significance of frequency differences between TUC and TICLE, LL ratio has been calculated which is presented in Table 45:

Table 45

LL Ratio of Overall Adverbial Connectors in TUC and TICLE

	TUC		**TICLE**		**LL Ratio**
	(O1)	%1	(O2)	%2	(*p < 0.05)
Connectors	2462	2.28	2590	1.51	+ 210.84

O1 is observed frequency in TUC
O2 is observed frequency in TICLE
%1 and %2 values show relative frequencies in the texts.
+ indicates overuse in TUC relative to TICLE,
- indicates underuse in TUC relative to TICLE

LL ratio of overall connectors between TUC and TICLE supports the overuse in TUC with 210.84 plus LL value against TICLE. That is, although there seems a fewer connector frequency in TUC (2462) than in TICLE (2590), it revealed overuse in TUC when total word number in TUC and TICLE (107,876 vs. 171,145) has been regarded.

When overall frequency of connectors in TUC is measured by LL against frequency of connectors LOCNESS f as another L1 data, the overuse in TUC has been confirmed in terms of statistically significance. Table 46 show the TUC and LOCNESS LL ratio results.

Table 46

LL Ratio of Overall Adverbial Connectors in TUC and LOCNESS

	TUC (O1)	%1	**LOCNESS** (O2)	%2	**LL Ratio** (*p < 0.05)
Connectors	2462	2.28	1277	0.76	+1093.76

O1 is observed frequency in TUC
O2 is observed frequency in LOCNESS
%1 and %2 values show relative frequencies in the texts.
+ indicates overuse in TUC relative to LOCNESS,
- indicates underuse in TUC relative to LOCNESS

As shown in Table 46, overall LL results indicated a very significant difference between TUC and LOCNESS on behalf of overuse in TUC with +1093.76 LL value. This supports the high degree of connector usage by Turkish students in the essays written in their native language in respect of both in frequency and in type diversity. As noted above, this high proportion of usage in the essays in Turkish language may be an explanation of overuse of connectors TICLE when compared to LOCNESS.

4.1.2. Frequency of Semantic Categories of Connectors across Corpora

Another important issue in connector is that to what extend learners use adverbial connectors to mark the semantic relations (Altenberg and Tapper, 1998). In order to investigate this matter, adverbial connectors have been analyzed within their semantic categorizations over corpora. The main categories of adverbial connectors (Quirk, 1985) were calculated in terms obtaining the total frequency of use. In graphic below presents the frequency of semantic categories of connectors across four corpora:

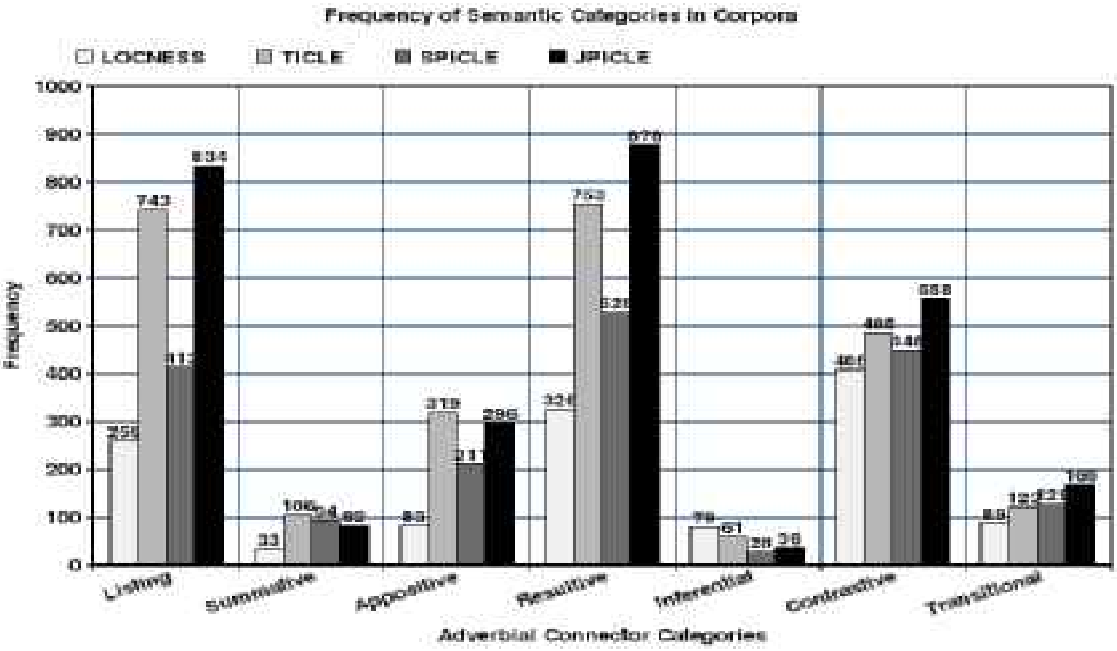


Figure 32. Overall frequency of semantic categories in four corpora.

As shown in the graphic, the distribution of semantic categories of adverbial connectors draws similar tendencies in most of the corpora. For instance, adverbial connectors of Listing have been mostly used by Japanese and Turkish learners as in connectors of Resultive category. Japanese learners used 834 connectors from Listing and 878 from Resultive category as the overall highest frequencies in these categories. Turkish learners follow Japanese learners in the use of Listing connectors (753) and Resultive connectors (743). Same condition can be seen in Contrastive category however, the frequency of Contrastive connectors in each corpus is close in total. Appositive connectors were mostly used by Turkish learners as 319 and then by Japanese learners (296). Although similar frequencies in all corpora were seen, Transitional connectors have been used by Japanese learners mostly (169) and then by Spanish learners (129). Native speakers only used Inferential connectors mostly than learners as 79 connectors followed by Turkish learners with 61 Inferential connectors, and this means that Inferential connectors were underused by learners. Again, the bar graphic shows an overuse of certain connector categories by learners. Therefore, except for Inferential connectors, categorical frequency analysis of adverbial connectors confirms the overuse of connectors by learners. Table 47 presents the details of categorical analysis:

Table 47

Frequency Distribution of Semantic Categories of Adverbial Connectors in Four Corpora

	LOCNESS		TICLE		SPICLE		JPICLE	
Category	n	%	n	%	n	%	n	%
Listing	259	20.2	743	28.6	413	22.3	**834**	29.3
Summative	33	2.5	**106**	4.0	94	5.0	82	2.8
Appositive	83	6.4	**319**	12.3	211	11.3	296	10.4
Resultive	326	25.5	753	29.0	528	28.5	**878**	30.8
Inferential	**79**	6.1	61	2.3	28	1.5	36	1.2
Contrastive	408	31.9	485	18.7	448	24.2	**558**	19.6
Transitional	89	6.9	123	4.7	129	6.9	**160**	5.6

n=raw frequency of each category

%=percentage of each category to overall frequency of connectors (in-group T/t ratio)

Table 47 shows the T/t ratio of each connector categories, i.e., the percentage of each connector category in the overall connector frequency in a corpus. When this condition is examined from T/t ratio (within connector groups), corpora show a ranging picture. For instance, in native data, the most frequently used connector category is Contrastive with 408 frequency and constitutes the 31.9 of all connectors in LOCNESS. On the other hand, in TICLE corpus, Resultive connectors have the highest percentage (29.0) in overall connectors. Similar condition for Resultive connectors can be seen in JPICLE and SPICLE as well. That is to say, 30.8% of all connectors in JPICLE and 28.5% of all connetcors in SPICLE consist of Resultive connectors. Thus, Resultive connectors are more common in learner corpora than native corpus.

In terms of categorical choice among learner groups, Listing connectors mostly preferred by Japanese learners. Turkish learners have the maximum usage frequency in both Appositive and Summative connector when compared to other learners and native speakers. Resultive connectors were highly used by Japanese learners, as well as in Contrastive and Transitional connectors. Spanish learners have used all categories in a moderate way in which again they show the closest usage tendency to native speakers'. Inferential connectors have been used mostly by native speakers which mean learners underuse Inferential type of connectors. Accordingly, as the all semantic categories of adverbial connectors out of Inferential connectors were overused by learners.

Another perspective to see the overused or underused semantic categories by learners is based on frequency ratio per 10,000 of each category which then compare with native corpus. In Table 48 below, raw frequency and frequency per 10,000 words of semantic categories in TICLE and LOCNESS are compared to identify the overuse or underuse categorically. For instance, there are 743 Listing adverbial connectors were identified in TICLE and this equals to 43.4 Listing connectors fall into per 10,000 words. In other words, the frequency of ratio of Listing connectors per 10,000 in TICLE is 43.4. In order to see whether there are overuse or underuse in Listing connector frequency per 10,000 words in TICLE, the frequency of ratio (43.4) is needed to be compared with frequency of ratio of Listing connectors per 10,000 in LOCNESS as the native data.

Table 48

Overused/underused Semantic Categories in TICLE in Comparison with LOCNESS

Semantic	**LOCNESS**		**TICLE**		**overuse/underuse**
Categories	n	n/10,000	n	n/10,000	+/-
Listing	259	15.3	743	43.4	+28.1
Summative	33	1.9	106	6.1	+4.2
Appositive	83	4.9	319	18.6	+13.7
Resultive	326	19.3	753	43.9	+24.6
Inferential	79	4.6	61	3.5	-1.1
Contrastive	408	24.2	485	28.3	+4.1
Transitional	89	5.2	123	7.1	+1.9

n= raw frequency of semantic categories

n/ 10,000= frequency of semantic categories per 10,000 words

(-/+) = difference between relevant value and value in Native Corpus (LOCNESS) per 10,000 words; (+) denotes overuse; (-) denotes underuse

The frequency of ratio of Listing connectors in LOCNESS is 15.3 per 10,000 words, and the difference between two corpora (43.4-15.3) is 28.1 which means that Listing connectors have been approximately used 28 times more than LOCNESS in TICLE so that Listing connectors were overused in TICLE.

In order to confirm the overuse/underuse reveal from differences of frequency per 10,000 words in corpus, LL ratio of categorical frequencies has been calculated among corpora. As the native reference, frequency of semantic categories in LOCNESS has been compared with frequency of semantic categories found in TICLE, SPICLE and

JPICLE. The aim is to calculate the categorical frequencies between L1 vs. L2 to support the frequency differences with a statistical test.

LL ratio in semantic categories has been measured between LOCNESS and TICLE to test the occurred overused or underused categories. Table 49 shows LL values of TICLE in comparison with LOCNESS:

Table 49

LL Ratio of Semantic Categories in TICLE and LOCNESS

Semantic Categories	**TICLE n**	**LOCNESS n**	**LL Ratio (*p<0.05)**
Listing	743	259	+236.09*
Summative	106	33	+39.16*
Appositive	319	83	+144.07*
Resultive	753	326	+166.86*
Inferential	61	79	-2.62
Contrastive	485	408	+5.46*
Transitional	123	89	+4.94*

n=raw frequency of connectors in corpus

+ indicates overuse in TICLE relative to LOCNESS,

- indicates underuse in TICLE relative to LOCNESS

As given results in Table 46, LL analysis of frequency of semantic categories of TICLE in comparison to LOCNESS as native corpus confirmed the overuse and underuse in certain categories. As discussed before, Frequency analysis per 10,000 words between TICLE and LOCNESS have indicated overuse in Listing, Summative, Appositive, Resultive, Transitional and underuse in Inferential. LL ratio of categorical comparison has supported the categorical frequency differences and confirmed the overuse and underuse in TICLE. The highest overuse in TICLE seems to be in Listing category with +236.09 and then in Resultive category with +166.86 LL value. The least overuse difference in LL value is in Transitional with +4.94 LL value after Summative (+39.16) and Contrastive (+5.46) connectors. In Inferential connectors, the LL ratio of Inferential category between TICLE and LOCNESS confirmed the underuse revealed in the frequency in per 10,000 words (-1.1) underuse in TICLE against LOCNESS), but with -2.62 LL value which is under p value and not a significant difference. In sum, the overuse and underuse in TICLE which have been identified in frequency analysis per 10,000 words of semantic categories had been supported with LL ratio statistics.

Table 50

Overused/underused Semantic Categories in SPICLE in Comparison with LOCNESS

Semantic	**LOCNESS**		**SPICLE**		**overuse/underuse**
Categories	n	n/10,000	n	n/10,000	+/-
Listing	259	15.3	413	22.8	+7.5
Summative	33	1.9	94	5.2	+3.3
Appositive	83	4.9	211	11.6	+6.7
Resultive	326	19.3	528	29.2	+9.9
Inferential	79	4.6	28	1.5	-3.1
Contrastive	408	24.2	448	24.8	+0.6
Transitional	89	5.2	129	7.1	+1.9

n= raw frequency of semantic categories

n/ 10,000= frequency of semantic categories per 10,000 words

(-/+) = difference between relevant value and value in Native Corpus (LOCNESS) per 10,000 words; (+) denotes overuse; (-) denotes underuse

According to the frequency differences between LOCNESS and SPICLE in Table 50, overused connectors in categories in SPICLE are as follows: Listing connectors have been used 7.5 more than NNS, Summative 3.3; Appositive 6.7, Resultive 9.9, Contrastive 0.6 and Transitional connectors have been used 1.9 more in SPICLE than LOCNESS. Similar to in TICLE, the only underused semantic category is Inferential connectors which were -3.1 less than in LOCNESS.

The frequency of semantic categories in SPICLE also compared with LOCNESS by means of LL ratio to test the frequency comparison. Table 48 below represents LL values between SPICLE and LOCNESS. The frequency of semantic categories of adverbial connectors in SPICLE have revealed similar results as in TICLE; overuse in Listing, Summative, Appositive, Resultive, Contrastive and Transtitional categories and underuse in Inferential category.

Table 51

LL Ratio of Semantic Categories in SPICE and LOCNESS

Semantic Categories	**SPICLE n**	**LOCNESS n**	**LL Ratio (*p<0.05)**
Listing	413	259	+25.83*
Summative	94	33	+26.51*
Appositive	211	83	+49.20*
Resultive	528	326	+35.37*
Inferential	28	79	-28.95*
Contrastive	448	408	+0.13
Transitional	129	89	+4.89*

n=raw frequency of connectors in corpus

+ indicates overuse in SPICLE relative to LOCNESS,

- indicates underuse in SPICLE relative to LOCNESS

LL results also confirmed the frequency differences between SPICLE and LOCNESS in these semantic categories. In contrast to in TICLE, the highest difference in LL ratio between SPICLE and LOCNESS is seen as overuse in Appositive category with +49.20 LL value. The least overuse is in Contrastive category with +0.13 which is not a significant difference as overuse between a native and a non-native group. Again, similar to in TICLE, the underuse in Inferential connectors by Spanish learners have also confirmed with LL ratio with -28.95 value. To summarize, the examined overuse and underuse in frequency analysis per 10,000 words in SPICLE in comparison with LOCNESS have been confirmed with LL statistics.

The frequency results of semantic categories of LOCNESS and JPICLE have also been compared as the third L1 vs. L2 comparison. Table 51 below shows the overused and underused semantic categories in JPICLE when compared to LOCNESS by means of overall frequency and frequency per 10,000 words. Comparison of raw frequency and frequency per 10,000 of semantic categories in JPICLE revealed overuse in six categories (as in TICLE and SPICLE), namely, Listing, Summative, Appositive, Resultive, Contrastive and Transitional. The only underused semantic category according to frequency per 10,000 words is Inferential connectors which also revealed underuse in TICLE and SPICLE as well.

Table 52

Overused/underused Semantic Categories in JPICLE in Comparison with LOCNESS

Semantic	**LOCNESS**		**JPICLE**		**overuse/underuse**
Categories	n	n/10,000	n	n/10,000	+/-
Listing	259	15.3	834	49.5	+34.2
Summative	33	1.9	82	4.8	+2.9
Appositive	83	4.9	296	17.5	+12.6
Resultive	326	19.3	878	52.1	+32.8
Inferential	79	4.6	36	2.1	-2.5
Contrastive	408	24.2	558	33.1	+8.9
Transitional	89	5.2	160	9.5	+4.3

n= raw frequency of semantic categories

n/ 10,000= frequency of semantic categories per 10,000 words

(-/+) = difference between relevant value and value in Native Corpus (LOCNESS) per 10,000 words; (+) denotes overuse; (-) denotes underuse

In JPICLE, the most overused category per 10,000 words is Listing connectors which have been used 49.5 times in JPICLE whereas 15.3 times in LOCNESS revealing a 34.2 overuse in JPICLE than NS group. The least overused connector group is Summative which have been 2.9 more in JPICLE. Inferential connectors have been identified 4.6 per 10,000 words in NNS whereas 2.1 in JPICLE, that is, this connectors have been underused by Japanese learners with -2.5 difference from NS group.

L1 vs. L2 comparison of LL ratio in adverbial connectors categorically has been made between JPICLE and LOCNESS in order to test the frequency differences. Previously, among all corpora, the highest rate of overuse of in sematic categories have been investigated in JPICLE in respect of frequency difference with LOCNESS per 10,000 words. Table 53 shows the LL results of semantic categories between JPICLE and LOCNESS:

Table 53

LL Ratio of Semantic Categories in JPICLE and LOCNESS

Semantic Categories	**JPICLE n**	**LOCNESS n**	**LL Ratio (*p<0.05)**
Listing	834	259	+318.40*
Summative	82	33	+21.57*
Appositive	296	83	+127.02*
Resultive	878	326	+262.92*
Inferential	36	79	-16.47*
Contrastive	558	408	+23.42*
Transitional	160	89	+20.55*

n=raw frequency of connectors in corpus

+ indicates overuse in JPICLE relative to LOCNESS,

- indicates underuse in JPICLE relative to LOCNESS

Frequency differences of semantic categories in JPICLE in contrast to LOCNESS have also been confirmed by LL calculation. The highest LL values between native and learner corpora seem in Listing (+318.40) and Resultive (+262.92) connectors (as seen in TICLE). Similar underuse in Inferential connectors seen in TICLE and SPICLE have also occurred in LL value (-16.47) of SPICLE against native corpora. To conclude, categorical frequency differences between JPICLE and LOCNESS have been confirmed with LL analysis.

In summary, the frequency distribution of semantic categories of adverbial connectors in four corpora draws a similar perspective as to overall frequency distribution. Overuse and underuse revealed in frequency differences in learner corpora have been supported with LL ratio measurement. Traces of categorical overuse and underuse in learner corpora can be examined when compared to native speakers' preference of adverbial connector categories.

4.1.3. Individual Connectors

Besides the overall and categorical evaluation, the adverbial connectors should be handled individually in the material they have been used. The mostly used single connectors can explain the learner attitudes in selection of certain connectors while using an argumentative essay. Firstly, top ten most frequently used adverbial connectors are shown in Table 54:

Table 54

Most Frequently Used 10 Connectors in LOCNESS

Connectors	n	T/t %	n per 10,000
however	175	13.7	10.3
then	126	9.8	7.4
so	125	9.7	7.4
therefore	81	6.3	4.8
also	77	6.0	4.5
for example	54	4.2	3.2
yet	49	3.8	2.9
thus	39	3.0	2.3
first	31	2.4	1.8
though	24	1.8	1.4
Total	**781**	**60.7**	**46**

n= raw frequency of connector in corpus

% T/t= Type/token ratio, percentage of the connector in overall connector types in corpus

In native corpus, top ten adverbial connectors cover the most of connectors used in general. That is, the 60.7 of total connectors consist of these most frequently used connectors. Indeed, these frequent ones such as **however**, **then** or **so** can be seen 46 times in every 10,000 words in LOCNESS.

On the other hand, Top ten adverbial connectors in TICLE is shown in Table 55:

Table 55

Most Frequently Used 10 Connectors in TICLE

Connectors	n	T/t %	n per 10,000
so	438	16.9	25.5
also	208	8.0	12.1
for example	182	7.0	10.6
however	151	5.8	8.8
then	143	5.5	8.3
of course	91	3.5	5.3
therefore	84	3.2	4.9
on the other hand	76	2.9	4.4
moreover	75	2.8	4.3
for instance	63	2.4	3.6
Total	**1511**	**58.0**	**87.8**

n= raw frequency of connector in corpus

% T/t= Type/token ratio, percentage of the connector in overall connector types in corpus

In TICLE, 58% of total connectors include the top ten adverbial connectors. The amount of top ten connectors per 10,000 words is nearly 88 whereas it is only 46 in LOCNESS. When compared TICLE with LOCNESS, there are six connectors are identical in TICLE and LOCNESS such as **so**, **also**, **for example**, **then**, **however** and **therefore** which vary in number. However, connectors in LOCNESS like **first**, **though**, **yet** and **thus** do not exist in TICLE in top ten range. Accordingly, in TICLE, the connectors in the most frequent ten range such as **of course**, **on the other hand**, **moreover** and **for instance** have been overused by Turkish learners. In TICLE, so has been used 438 times which means it can be seen 25 times in every 10,000 words, whereas it can be seen 7.4 times per 10,000 in LOCNESS. On the other hand, **however** which the most frequent connector in LOCNESS placed in the fifth range in TICLE.

In order to test the significance of highly frequent connectors in TICLE, LL ratio was applied by comparing with NS. Thus, the frequency difference of a particular connector between TICLE and LOCNESS has been approved by a statistical measurement. In previous sections, LL ratio utilized to test the significant differences for overall frequency of connectors, for frequency of semantic categories of connectors between NS and learner groups. Now, LL ratio of overused and underused connectors in learner data in comparison NS data was calculated to testify the identified overuse/underuse in the uses of them. Table 56 shows the LL ratio of overused connectors in TICLE in comparison with LOCNESS:

Table 56

LL ratio of Overused Connectors in TICLE

Overused connectors	**TICLE n**	**LOCNESS n**	**LL Ratio (*p<0.05)**
so	438	125	+179.29*
moreover	75	5	+72.37*
for example	182	54	+71.25*
also	208	77	+60.44*
to sum up	33	0	+45.22
of course	91	23	+42.30*
for instance	63	12	+37.20*
on the other hand	76	22	+30.61*
Then	143	126	+0.82
Therefore	84	81	+0.02

n=raw frequency of connectors in corpus

+ indicates overuse in TICLE relative to LOCNESS,

- indicates underuse in TICLE relative to LOCNESS

Overuse of connectors in TICLE has been testified according to the LL rates as can be seen in Table1. The highest LL value is belong to **so** which indicated 179.29 plus value which signs a very significant difference between TICLE and LOCNESS in terms of both frequency and statistics. Other overused connectors like **moreover** (72.37), **for example** (71.25), **also** (60.44), **of course** (42.30), **for instance** (37.20) and **on the other hand** (30.61) revealed high LL values and high significant differences at high rates. Connectors like **then** and **therefore** also revealed significant overuse between TICLE and LOCNESS but with very low rates (not significant). The example below show the rate of overuse of **so** in a single text of a Turkish learner in TICLE:

Example 43

[it was struggle of democracy. **So,** the democracy struggle became initial and women rights]

[..to semi-democratic regimes, **so** the women started to join production and they went out…]

[In my opinion, to get something, need serious struggle. **So,** the men don't give the rights..]

[..healthy problems, social life, cultural activities. . . etc. , **so** the women are the main objects..]

[..,but the man was not punished, **so** it is not enough for woman to get their rights..]

[Also the other problem is being an authority on the other people, **so** the men want to be an authority on the women because of women's physical weakness. **So,** the women should be educated and should join administartion positions.]

[…human rights, children's rights, animals' rights. . . etc, **so** the main problem is not..]

Extracted from <ICLE-TR-CUK-0148.1>

In this one sample in Example 43, eight **so** have been used by learner in his/her text. Indeed, so has been used in one sentence after another which can form a paragraph. That is, the overuse of so in TICLE can be examined in one learners text. Other overused item in TICLE is for example which is shown in a text of a Turkish learner below:

Example 44

[The segregation between women and men in society affects the language in some countries language can change because of gender. **For example**; in Japan women and men use different dialects of the language. "In Muskesgean language Koasat, spoken in Lovisiana, words that end in on /s/ when spoken by men, and in /I/ or /n/ when used by women. **For example** the world meaning "lıft it" is lakawol for women and…

[…the permission is limited until primary or secondary school. **For example;** in Indi the literacy rate is nearly 34 percent among men, 13 percent among women. literacy rate can change region to region like country to country. **For example** in Turkiye the women in the east part of Turkiye are less educated than ……………]

Extracted from <ICLE-TR-KEM-0021.2>

Two separate paragraphs from an essay including the usage of **for example** are seen in Example 44. In the first one, for example has been used two times consecutively to proceed the subject emphasized in the first sentence of the paragraph. Similarly, two for example in the second paragraph have been used as to refer the first sentence. All four connector have been used in the initial positions rather than in the middle. The overuse in **for example** in this sample gives the idea in the overall overuse in TICLE.

In order to test the underused connectors according to frequency analysis, LL calculation has been applied as well. Table 57 presents the underused connectors in TICLE in comparison with LOCNESS according to LL results:

Table 57

LL Ratio of Underused Connectors in TICLE

Underused connectors	**TICLE**	**LOCNESS**	**LL Ratio (*p<0.05)**
yet	16	49	-18.10*
however	141	175	-4.24*
though	12	24	-4.27*

n=raw frequency of connectors in corpus

+ indicates overuse in TICLE relative to LOCNESS,

- indicates underuse in TICLE relative to LOCNESS

The connectors which have been underused in TICLE when compared to LOCNESS according to frequency differences then measured with LL ratio. According to the LL ratio of underused connectors, **yet**, **however** and **though** revealed difference in TICLE against LOCNESS. In other words, the frequency difference by means of underuse of these connectors in TICLE has been confirmed by LL ratio as well. **Yet** has the highly underused connector with 18.10 LL value which means it is not preferred by learners as much as native speakers while writing an argumentative essays. The point is that the all these three underused adverbials in TICLE are Contrastive type of connectors.

Next top ten adverb connectors listing is from SPICLE which is presented in Table 58 below. In SPICLE, top ten most frequent connectors covers the 49% of all connectors and they can be seen 50 times per 10,000 words. Five connector are identical in SPICLE with LOCNESS such as **so**, **however**, **for example**, **then** and **therefore**. The connectors like **on the other hand**, **moreover**, **finally**, **of course** and **for instance** do not exist in top ten range of native speakers' list so that these connectors were overused by Spanish learners

Table 58

Most Frequently Used 10 Connectors in SPICLE

Connectors	**n**	**T/t %**	**n per 10,000**
so	292	15.7	16.1
however	125	6.7	6.9
for example	94	5.0	5.2
then	88	4.7	4.8
therefore	70	3.7	3.8
on the other hand	66	3.5	3.6
moreover	49	2.6	2.7
finally	49	2.6	2.7
of course	47	2.5	2.6
for instance	39	2.1	2.1
Total	**908**	**49.1**	**50.5**

n= raw frequency of connector in corpus

% T/t= Type/token ratio, percentage of the connector in overall connector types in corpus

So is common in both SPICLE and TICLE as the most frequent connector although differing in occurring number, i.e., so can be seen 16 times per 10,000 words in SPICLE whereas 25 times in TICLE. The overused and underused connectors according to frequency results in SPICLE have been measured with LL ratio by comparing LOCNESS in order to see the difference statistically. Interestingly, the view in frequency table given in Table 55 was changed when the LL ratio of connectors were measured. Table 56 shows the overused connectors by means of LL ratio. As can be seen in Table 59 below the picture of mostly used connectors according to frequency results in Table 55 above were replaced with different connectors by means of their LL value. Although the connectors like **then**, **therefore** and **howeve**r seem among mostly used ten connector list in SPICLE, their LL value did not reveal a significant overuse when compared to NS group. The connector **so** is the mostly overused connector similarly to TICLE with the highest LL value (57.82).

Table 59

LL Ratio of Overused Connectors in SPICLE

Overused connectors	SPICLE n	LOCNESS n	LL Ratio (*p<0.05)
so	292	125	+57.82*
moreover	49	5	+38.59*
to sum up	23	0	+30.33*
on the other hand	66	22	+20.11*
finally	49	13	+19.87*
for instance	39	12	+13.26*
besides	41	14	+12.06*
too	42	16	+10.36*
for example	94	54	+8.37*
of course	47	23	+6.83*

n=raw frequency of connectors in corpus

+ indicates overuse in SPICLE relative to LOCNESS,

- indicates underuse in SPICLE relative to LOCNESS

A Summative connector **to sum up** indicated a significant overuse between SPICLE and NS with +30.33 value. Other connectors like **besides** and **too** showed statistically significant overuse whereas **for example** and **of course** confirmed their overuse in frequency difference by significant LL values against LOCNESS.

The mostly overused connector by Spanish learners is so which is presented in a single text from SPICLE below:

> Example 45
>
> [his plays suitable for them to be accepted. **So** , both, Wilde and Shaw represented a way of criticism to Victorian society,…]
>
> [..he thinks is a servant, Miss Hardcastle. **so** we have now three events,…]
>
> […character only for his own interest. **So** , in fact, all the relevant dramatic events come…]
>
> [Wilde tried to criticize thses social bias. **So** we see how love was in a second step,..]
>
> [Joan will reach sainthood, **so** the controverse is presented…]

[he doesn't think sainthood to be possible, (irracionality) **so** maybe this is the source of…..]

Extracted from <ICLE-SP-UCM-0009.8>

The sentences from a text of a Spanish learner in SPICLE include six **so** as a connector. In general, so has been used in separate sentences in one single text in rather than consecutive sentences in a paragraph. Another overused connector in SPICLE is moreover which is shown in a paragraph from a Spanish learner's text:

Example 46

[**Moreover** it is said that many young people mature and shape their character when they do the military service. In contrast they have to be far from their families for a year. **Furthermore** they waste time trying to learn something that they will forget in the future. They have to stop working or studying at an age in which they are changing their way of life. They have to mount guards and to march past in front of people but are they well prepared for defending their country? I don't think so. The defence of a country must be done by a professional army, especially trained for the art of war. **Moreover** there are more and more young people with psychical problems as a result of the military service.]

Extraxted from <ICLE-SP-UCM-0002.7>

In this single paragraph in the text, **moreover** has been used two times in order to reinforce the preceding topic they followed. In addition to **moreover**, another reinforcing connector, **furthermore** has been used as well as moreover which seems to be a sample of repeated use of reinforcing connector in one single text.

In SPICLE, statistically significant underused connectors when compared to LOCNESS are presented in Table 57 following:

Table 60

LL Ratio of Underused Connectors in SPICLE

Underused connectors	**SPICLE**	**LOCNESS**	**LL Ratio (*p<0.05)**
yet	2	49	-57.11*
however	125	175	-12.16*
then	88	126	-9.64*

n=raw frequency of connectors in corpus

+ indicates overuse in SPICLE relative to LOCNESS,

- indicates underuse in SPICLErelative to LOCNESS

The highest LL value was revealed in **yet** as -57.11 which is followed by **however** with -12.16 and **then** with -9.64 LL value. As noted above, **however** and **then** are the connectors which occurred in mostly used ten connectors in SPICLE, but when their LL ratio was compared with NS, they expressed significant underuse by Spanish learners.

The last top ten list contains the most frequent ten adverbial connectors in JPICLE is shown in Table 61:

Table 61

Most Frequently Used 10 connectors in JPICLE

Connectors	**n**	**T/t %**	**n per 10,000**
so	604	21.2	35.8
however	225	7.9	13.3
for example	200	7.0	11.8
then	193	6.7	11.4
therefore	122	4.2	7.2
of course	101	3.5	5.9
first	89	3.1	5.2
too	87	3.1	5.1
also	82	2.8	4.8
on the other hand	76	2.9	4.5
Total	**1779**	**62.4**	**106**

n= raw frequency of connector in corpus

% T/t= Type/token ratio, percentage of the connector in overall connector types in corpus

As can be seen in Table 61 the total number of most frequent connectors in JPICLE is 1779 which covers the 62.4 per cent of all connectors in the corpus. In general, top ten adverbial can be identified as 106 times (as the highest among learners and native speakers) in every 10,000 words in JPICLE sub-corpus. **So** is the most frequent connector in JPICLE similar to SPICLE and TICLE, it can be seen 35 times nearly in every 20 texts (approximately per 10,000 words). When compared to LOCNESS, many of connectors identical except for **of course**, **too** and **on the other hand** which were overused in JPICLE.

Among all learner corpora, JPICLE revealed the highest rates in frequency of connectors in general. In order to test the statistical significance in overuse and underuse, LL calculation has been applied to connectors found in JPICLE by comparing them with LOCNESS. Table 62 shows the LL values of overused connectors in JPICLE in comparison with LOCNESS:

Table 62

LL Ratio of Overused Connectors in JPICLE

Overused connectors	**JPICLE n**	**LOCNESS n**	**LL Ratio (*p<0.05)**
so	604	125	+342.66*
for example	200	54	+89.32*
also	208	77	+60.44*
moreover	64	5	+59.79*
too	87	16	+53.84
of course	101	23	+52.98*
first	89	31	+29.26
then	193	126	+14.19*
therefore	122	81	+8.35*

n=raw frequency of connectors in corpus

+ indicates overuse in JPICLE relative to LOCNESS,

- indicates underuse in JPICLE relative to LOCNESS

LL results of overused connectors in JPICLE against LOCNESS mostly supplied the frequency table of top ten adverbials in JPICLE. As expected, LL values revealed at very high rates since there are high frequency differences between JPCLE and LOCNESS in number. Again, similar to TICLE and SPICLE, **so** is the highest frequent connector in JPICLE and it was confirmed with very high LL ratio as +342.66 which indicates a very significant difference between two group in the use of **so**. The lowest

LL value belongs to therefore with 8.35. According to the LL ratio of various connectors, top ten mostly used ones revealed significant difference against NS group.

In JPICLE, so is the mostly overused connector at as in TICLE and SPICLE, indeed it has the highest frequency rate and LL value among them. Here is an example including intersections from a single essay of a Japanese learner in JPICLE in which so has been significantly overused:

> Example 47
>
> [And it has a function of telephone book, <u>**so**</u> we need not remember telephone number. These days, there are phones that have a function of mobile camera and java, <u>**so**</u> we can play…]
>
> [It has electric waves, <u>**so**</u> it has risks that electronic waves…]
>
> [Above all, we bother people who have apace-maker of their heart. <u>**So**</u> we must keep manners and have common sense. Certainly, we can call whenever we want to do, **so** it is very convenient. But it calls us even when we do not want to answer. It seems as if we are watched and restrained. <u>**So**</u> we can say that it brings….]
>
> [Our mail address of cellular phone is generally our phone number, <u>**so**</u> dealers can easily know our mail address.]
>
> [The charge of cellular phone is much expensive than that of regular phone. <u>**So**</u> we must pay much money.]
>
> [by the evolution of cellular phone and several solutions, <u>**so**</u> it will be more beneficial to society. And it will be able to evolve still more, <u>**so**</u> it will be more convenient for us and will be able to promote the globalization with computer. In addition, many people are depended on it as much as we can not live without it, and it is familiarized as much as many family have at least one phone. <u>**So**</u> it already is one of the necessaries and the world of cellular phone is one of the biggest business in Japan today. <u>**So**</u> I think that a cellular phone is the greatest invention of the twentieth century.]
>
> Extracted from <ICLE-JP-WAS-0013.1>

In this essay, 11 **so** have been used in order to connect separate sentences or to conduct a relation within a paragraph. As can be seen above, one single paragraph contains so three so in sequential sentences and the other (last one) paragraph includes four **so** in sequential sentences or to conclude the paragraph. Thus, the overuse proportion of **so** in JPICLE, which has the highest overuse rate among all learner corpora, can be seen in even one single text of a Japanese learner. Other overused connector by Japanese learners is also which is shown in intersections in a Japanese learner's essay below:

Example 48

[But, if we have our own car, we can go anywhere. **Also** traffic is not congesting. ……….]

[Nature makes us relax. **Also**, it gives me good and fresh air. We can't buy it……………..]

[Some people are probably wearing fashionable wear. **Also** some people eat many kinds of food. But they will cost a lot of many. And then they can't have fresh vegetables. Therefore, they can't keep their health. On the other hand, if we will live in the country, we can grow them in own field. **Also**, there are better relationships in the country …]

[However, the city doesn't have good relationships, I think. **Also**, the city is always in confusion.]

Extracted from <ICLE-JP-SWU-0016.3>

As shown in the example above, this Japanese learner has used also frequently (five times) in his/her argumentative essay. In one paragraph, it has been used closer sentences. In addition to **also**, a wide variety of connectors such as **but**, **therefore**, **and**, **on the other hand** and **however** (which are underlined in the example) have been used even in this very short part of this single text of the Japanese learner which might be an explanation for the high rate of connector usage in Japanese learner when compared to Turkish and Spanish learners.

On the other hand, some of connectors have been underused by Japanese learners when compared to NS group. **Yet** is the common underused connector which has been underused in JPICLE as in TICLE and SPICLE as well. Table 63 presents the underused connectors by means of LL ratio between JPICLE and LOCNESS :

Table 63

LL Ratio of Underused Connectors in JPICLE

Underused connectors	**JPICLE**	**LOCNESS**	**LL Ratio (*p<0.05)**
yet	10	49	-28.08*
instead	2	19	-15.90*
still	6	17	-5.48*

n=raw frequency of connectors in corpus

+ indicates overuse in JPICLE relative to LOCNESS,

- indicates underuse in JPICLE relative to LOCNESS

In Altenberg & Tapper (1998), **yet** is one of the underused by Swedish learners. Other underused connectors by Japanese learners are **instead** and **still** which are Contrastive connectors like **yet**. In their study, Narita et. al. (1996) found that yet and instead were underused by Japanese learners as well.

In summary, adverbial connectors investigated individually across corpora by means of comparing L1 (LOCNESS) vs. L2 (TICLE, SPICLE and JPICLE). Top ten mostly used connectors have been identified in all corpora then frequency comparison has been done among L1 and L2 groups. Mostly used ten connectors in LOCNESS have been compared to mostly used ten connectors in three learner corpora. Next, the overused and underused connectors revealing from frequency differences between NS and NNS groups have been measured by Log-likelihood ratio in order to confirm the differences.

The frequency and LL analysis of individual connectors in three corpora indicates that **so** is the highly frequent connector in TICLE, SPICLE and LOCNESS. On the other hand, **however** is the mostly used connector by native speakers whereas it has been underused in TICLE and SPICLE and overused in JPICLE. Connectors like **for example** and **of course** were commonly overused by all learners whereas also has been overused by Turkish and Japanese learners. Other identically overused connectors are as following: **on the other hand** has been overused in TICLE and SPICLE, **too** seems identically overused in SPICLE and JPICLE. **Then** and **therefore** seem highly overused in JPICLE, slightly overused in TICLE but underused in SPICLE. By means of underused connectors, only **yet** is identical in TICLE, SPICLE and JPICLE.

4.2. Discussion

The total results of frequency analysis of adverbial connector in EFL learner corpora indicated similar conditions to previous studies. Overall overuse of connectors, overuse and underuse of some individual connectors in learner corpora obtained from the frequency analysis indicate identical conditions with many of the previous research.

The present study showed a similar conclusion with connector investigations such as in Bolton et al., 2002; Narita et al, 2004 in terms of overall overuse in connectors. When the results of individual connectors is compared with previous research, the overuse in some connectors also similar with some of them (Granger & Tyson, 1996; Narita et. al., 2004; Tanko, 2004; Chen, 2006; Fei; 2006; Heino, 2010; Can, 2011). Particularly, the connectors like **for instance** and **of course** which were overused by Turkish learners also overused by Swedish learners in Altenberg & Tapper (1998); Heino (2010) and by French learners (Granger & Tyson, 1996). The overused connectors like **for example** and **moreover** by Turkish learners also overused by Japanese learners (Narita et. al., 2004). The overused connectors like **so** and **on the other hand** also reported overuse by Chinese learners (Bolton et. al., 2002). When the underused connectors are regarded, the underuse of **yet** by Turkish learners revealed in the study also reported underused in Altenberg & Tapper (1998) and Narita et. al., 2004. The underused connectors like **however** and **though** identically underused by Turkish learners in this study and by Swedish learners in Altenberg & Tapper (1998). In the study, some L1 influence found in connector usage by Turkish learners which may due to the overall overuse of connectors in Turkish L1 essays. The L1 traces reported in Altenberg & Tapper (1998), however they raletd such transfer to some cross linguistics differences.

The results obtained from the analysis of connector usage by Turkish learners suggest that the overuse and underuse of certain connectors seem common in many EFL learners. This means that the overuse and underuse of adverbial connectors by EFL learners are universal features of connector usage across EFL learners from different L1 backgrounds.

4.3. Chapter Summary

The present chapter presents the details in quantitative analysis which has been conducted across four corpora as LOCNESS, TICLE, SPICLE and JPICLE. In the analysis procedure, L1 vs. L2 comparison were regarded in order to focus on CIA as the main analysing system between a native and learner data. Therefore, as the native data, the frequency results of LOCNESS have been compared with learner corpora TICLE, SPICLE and JPICLE. Initially, overall frequency and log-likelihood measurements of all corpora are given to see the total overview. As the L2 vs. L2 phase of CIA, overall frequency of learner corpora have been compared each other as well. Then, the measurements of adverbial connectors as semantic categories were shown in order to examine categorical conditions of connectors among corpora. Next, the frequency and log-likelihood analysis of connectors were taken into consideration individually in addition to mostly used connectors across all corpora. In addition to other comparisons, to investigate L1 transfer specifically, TICLE and TUC corpora have compared by means of overall frequency of connectors.

CHAPTER V

CONCLUSION

5.0. Introduction

In this study, the adverbial connectors in Turkish learners’ argumentative essays have been investigated by means of quantitative analysis. The results compared with native speakers’ and other EFL learners’ usage of connectors. In this section, the evaluation of research questions of the study is presented and then the implications for ELT research and applications are discussed as well as suggestions for further research.

5.1. Evaluation of Research Questions

R.Q. 1: *Which Adverbial Connectors does TICLE corpus contain and how they can be classified?*

The investigation of adverbial connectors in TICLE gave the opportunity to explain the general attitude in adverbial connector usage of Turkish learners. Accordingly, regarding the Turkish learners in TICLE corpus, it can be interpreted that:

- Turkish learners use a wide variety of adverbial connectors in their written texts as 104 different types in 7 categories.
- Turkish learners use approximately 1.5 of every 100 words as adverbial connector.
- Turkish learners mostly use adverbial connectors to indicate a result (by Resultive Connectors) and/or to list (by Listing Connectors) in their sentences.
- Turkish learners mostly use 10 adverbial connectors such as *So*, *Also*, *For exam*ple, *However*, *Then*, *Of course*, *Therefore*, *Moreover*, *First of all*.

R.Q. 2: Do Turkish learners use English adverbial connectors as native speakers in a statistically similar way?

The main focus of the methodology in the present study is to apply the CIA which suggests the comparison between an L1 and L2 in order to investigate the interlanguage properties. Research question 2 seeks interlanguage of Turkish learners by comparing their L2 productions in English with native speakers' L1 written production. In order to answer this second research question, frequency of connectors identified in TICLE has been compared to the frequency of connectors in LOCNESS. The obtained results have been evaluated by means of frequency analysis and then compared through log-likelihood (LL) analysis for the statistical confidence of frequency comparison. According to the frequency and LL analysis, there is certain overuse of connectors in Turkish learners' essays in TICLE when compared to LOCNESS as NS group. The difference in TICLE has been identified in overall, categorical and individual connectors. In addition to overuse in TICLE, the choice of connector types also occurred diversity between TICLE and LOCNESS when the mostly used connectors have been analyzed in both corpora.

Categorically, Turkish learners mostly prefer to use Resultive connectors whereas native speakers mostly use Contrastive connectors in their argumentative essays. Individually, **however** (Contrastive) is the mostly used connector in LOCNESS and **so** (Resultive) is mostly used connector in TICLE. In addition, **moreover**, **for instance** and **on the other** are frequent connectors in TICLE which are missing in top ten connector list of LOCNESS. Connectors like **yet**, **first**, **though** which have been used at high rates in LOCNESS are not have a same performance in top connector list of TICLE. Duo to the significant overuse and different choice of connector in TICLE and LOCNESS, it can be inferred that Turkish learners' use of English adverbial connectors differs as the way of native speakers' use.

R.Q. 3: How is the Turkish EFL learners' use of adverbial connectors different from Spanish and Japanese EFL learners?

The second part of contrastive analysis (in addition to L1 vs. L2) suggests L2 vs. L2 comparison in order to seek the common interlanguage properties. In the study, the frequency analysis of adverbial connectors obtained from three learner corpora, TICLE, SPICLE and JPICLE as well. The frequency results have been compared with each other via LL calculation to specify the significance of differences. Comparison results indicated that certain overuse in TICLE has been found when compared to SPICLE in the overall and categorical measurements. On the other hand, overall and categorical frequency differences revealed underuse in TICLE when compared with JPICLE. That is, Turkish learners use more adverbial connectors than Spanish learners and less than Japanese learners. Accordingly, there is a very significant difference between SPICLE and JPICLE. Thus, connector usage by Turkish learners is somewhere between Spanish and Japanese learners' usage

On the other hand, there are common tendencies have been identified in three learner corpora. For example, Resultive connectors are the mostly preferred connectors by Turkish, Spanish and Japanese leaners. Furthermore, Inferential group of adverbial connectors have been commonly underused in three learner corpora when compared to NS data. In individual connectors, **so** is the mostly used and overused connector in TICLE, JPICLE and LOCNESS. Moreover, as a Contrastive connector, **yet** has been underused by three learner groups commonly.

In summary, it can be inferred that there are frequency differences in the use of adverbial connectors between Turkish EFL learners and other EFL learners. On the other hand, there are similarities in the choice of connectors between Turkish EFL learner and the other EFL learners.

R.Q. 4: What are the sources of divergences in TICLE corpus?

In order to find an answer for fourth research question, another comparison has been made between the different performances of the same group of participant. That is, Turkish learners' L2 English production and L1 Turkish production have been analyzed in terms of adverbial connector usage. TUC corpus which contains L1 Turkish argumentative essays of Turkish students were investigated according to the Turkish connector system. Then the obtained overall frequency of Turkish connector usage has been compared to the overall adverbial connector frequency in TICLE which contains L2 English argumentative essays of Turkish students. An additional comparison also made between TUC and LOCNESS in order to see the general condition of connector usage in different L1 backgrounds. The results indicated an overuse in TUC against TICLE as well as LOCNESS. The slight difference between TUC and LOCNESS might be a result of handling L2 production of Turkish learners. However, there is a fact that L2 English production of Turkish learners in TICLE is already higher rates than LOCNES due to the overuse in TICLE noted in RQ1. Therefore, general performance of connector using of Turkish learners both in their L1 essays and L2 English essays is higher than native English speakers. This divergence is not due to the variety of connectors since there having been 163 connector types in Turkish which can be equivalent to 175 types (Quirk, et. al., 1985). In general, the factor of overuse of connectors is the source of divergence in TICLE. The writing style and writing habits in Turkish might be a reason for such diversity between Turkish and native English students.

a) Are there any signals of L1 transfer?

As noted above, the certain overuse in TUC against TICLE and LOCNESS suggest that Turkish learners have no problem in the use of adverbial connectors in Turkish. Furthermore, the overuse of adverbial connectors identified in TICLE when compared to LOCNESS might be due to the connector usage performance in Turkish language. That is to say, the tendency of connector usage at high rates may affect the writing habits of Turkish learners and they may transfer this tendency and reflect it to their L2 English writing styles.

b) *Are these divergences a property of interlanguage*?

It is accepted that, the general tendency of overuse of adverbial connectors in EFL learners is a common attitude. This fact might be due to the common interlanguage properties which tend the learners to determine certain ways of using adverbial connectors. Although the overuse of connector using is accepted as a way of disguising the poor writing in L2 (Crewe, 1990), the fact of overuse in L1 Turkish data found in the present study needs more different explanation than a basic common interlanguage property. At this point, many explanations may be devoted to the divergences such as L1 transfer, cultural writing styles or cultural writing habits.

5.2. Implications for Language Teaching

Appropriate connector usage by EFL learners is the common point of the related investigations of adverbial connectors. Tankó (2004) states that the process of acquisition of adverbial connectors is more effective when it is controlled by both teacher and the learner. In addition, the students' role is more important since they can discover the characteristics of adverbial connectors by the guidance of reliable ad thorough introduction to adverbial connectors by teachers. Tankó (2004) adds that:

> Information on the variety of adverbial connectors and their frequency in various spoken and written text types can be given on the basis of such sources (e.g. Biber et al. 1999:875–892) that rely on corpus evidence: it is Corpus Linguistics studies that provide the most reliable empirical evidence on the use of adverbial connectors. The teacher can furthermore give valuable feedback concerning the number of adverbial connectors used in student texts as well as make explicit, relevant and therefore effective comments based on particular instances taken from student texts concerning the question of when to use and when not to use adverbial connectors.
>
> (2004, p.179)

Data-driven learning (DDL) (see section 2.4.2.) approach provides empirical base to improve the learning of adverbial connectors by EFL learners. The direct access of learners to corpora via Internet, CD-ROMs or by KWIC (Key Word in Context) concordances including adverbial connectors could help learners to observe paradigmatic presentation of repeated patterns of adverbial connectors as their meaning and the cognitive relation they express, their grammatical function, their genre sensitivity, the linguistic units they span, and the various forms the same adverbial connector can have (Tankó, 2004). By using this approach, Tankó (2004) found that Hungarian students' use of adverbial connectors improved.

Another method in order to create awareness for appropriate adverbial connectors' usage is that developing new EFL materials. Narita et. al. (2006) suggests a computer-based EFL writing tool with a concordance which can help learners to discover the proper use of connectors. In this tool, if the learner enters or selects a connector on the computer screen, a list of sample writings including the that connector could be shown in the KWIC format, then the learner could access the full text to examine the usage of that certain connector (Narita et. al., 2006). In respect of such materials, Narita et. al. (2006) also point out that although further empirical research is needed, repeated exposure to authentic texts of good quality is expected to have a positive effect on EFL writing.

In terms of connector usage by Turkish learners, there are certain conditions can be inferred from the conclusions of the present study. First one is that the usage of adverbial connectors by Turkish learners are generally due to informal register as they generally overuse informal connectors like **so** and **of course**. If the Turkish learners could be aware of register and related connector type, then they can be thought using appropriate connector types for appropriate registers (formal/informal) in English language. In order to create such a learning environment for Turkish EFL learners, EFL teachers can use DDL types of materials and writing tools as noted above.

Another factor that was investigated in the present study is that the possibility of L1 transfer from Turkish language in connector using. Rising awareness of connector systems in both Turkish and English in a comparative way, analyzing cross linguistics differences between these two languages can be effective in the connector usage of Turkish learners. If these conditions can be provided through appropriate teaching methods, e.g. using, searching and analyzing connectors in/over corpora from Turkish and English languages via DDL methods and specifically designed DDL materials, then the balance between two languages can be established in order to use connectors appropriately in both languages.

5.3. Suggestions for Further Research

The present study provides a quantitative approach to the usage of adverbial connectors by Turkish learners by means of comparing native speakers and other learner groups. While doing this, adverbial connectors have been taken to account computing their frequency of use in the sentences within semantic categories they belong to. Future research can evaluate the connectors considering their positions in the sentences they occurred. Next, this study focused on the overuse and underuse of adverbial connectors. Similar studies in the future should be emphasized the misused connectors by learners in order to gain more detailed insight about the usage of connectors. As a last suggestion, academic writing in L2 regarding cultural aspects might be included to a replicated research to see whether there is a relation between interlanguage properties and different cultural attitudes in different L1 backgrounds.

REFERENCES

Aarts, J. (1991). Intuition-based and observation-based grammar In K. Aijmer & B. Altenberg (Eds.) *English Corpus Linguistics* (pp. 44-62). London: Longman.

Ädel, A. (2008). Involvement features in writing: do time and interaction trump register awareness? In S. Gilquin, S. Papp & M.B. Díez-Bedmar (Eds.), *Linking up Contrastive and Learner Corpus Research* (pp. 35-53). Amsterdam, Atlanta: Rodopi.

Aijmer K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 55-76). Language Learning and Language Teaching 6. Amsterdam & Philadelphia: Benjamins.

Aijmer, K. (2008). Parallel and comparable corpora. In A. Lüdeling & M. Kytö (Eds.), *Handbook on Corpus Linguistics* (pp. 275-292). Mouton de Gruyter.

Aksan, Y. & Aksan, M. (2009). Building a national corpus of Turkish: Design and implementation. *Working Papers in Corpus-based Linguistics and Language Education*, 3, 299-310, Tokyo: Tokyo University of Foreign Studies.

Altenberg, B. (1990). Automatic text segmentation into tone units. In J. Starvik (Ed.), *The London-Lund corpus of spoken English: Description and research* (pp. 287-324). Lund: Lund University Press.

Altenberg, B. (2002). Using bilingual corpus evidence in learner corpus research, *in* S. Granger, J. Hung, and S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 37-54). Amsterdam: John Benjamins.

Altenberg, B. & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.) *Learner English on Computer* (pp. . 80-93). London & New York: Addison Wesley Longman.

Anthony, L. (2007). *AntConc 3.2.1w.* Waseda: Waseda University.

Aston, G. (Ed.) (2001). *Learning with Corpora.* Bologna: CLUEB and Houston, TX: Athelstan.

Aston, G., Bernardini, S. & Stewart, D. (Eds.) (2004). *Corpora and Language Learners.* Amsterdam & and Philadelphia, PA: John Benjamins.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics.* Edinburg: Edinburg.

Barber Sardinha, T. (1999). Using KeyWords in text analysis: Practical aspects. *DIRECT Papers,* 42, 1-8. São Paulo and Liverpool

Barlow, M. (2000). *MonoConc Pro.* Houston, TX: Athelstan

Bartsch, S. (2004). *Structural and functional properties of collocations in English. a corpus study of lexical and pragmatic constraints on lexical cooccurrence.* Tübingen:Narr. Retrieved from http://books.google.com.tr/books?id=CMyPT-nDm4sC&prints (last accessed 12/04/2012).

Bauer, L. (1993). Progress with a corpus of New Zealand English and early results. In C. Souter & Eric Atwell (Eds), *Corpus-Based Computational Linguistics* (pp. 1-10). Amsterdam & Atlanta : Rodopi.

Biber D (1987). A textual comparison of British and American writing. *American Speech*, 62, 99–119.

Biber D (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993a). Co-occurrence patterns among collocations: a tool for corpus-based lexical knowledge acquisition. *Computational Linguistics*, 19(3), 531-538.

Biber D (1995*). Dimensions of register variation: a crosslinguistic comparison.* Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Rippen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English.* London: Longman.

Biber, D. and Reppen, R. 2002: What does frequency have to do with grammar teaching? *Studies in Second Language A cquisition,* 24, 199–208.

Bikeliene, L. (2008). Resultive connectors in advanced Lithuanian learner's English writing. *Kalbotyra*, 59(3), 30-37.

Breivik, L. (1999). On the pragmatic function of relative clauses and locative expressions in existential sentences in the LOB corpus. In. H. Hasselgárd & S. Oksefjell (Eds.), *Out of corpora,* (pp. 121-135). Amsterdam: Rodopi.

Bradley, P.T. & Mackenzie, I. (2004). *Spanish: An essential grammar.* London: Routledge.

Bolton, K., Nelson, G. & Hung, J. (2002). A corpus-based study of connectors in student writing. Research from the International Corpus of English in Hong Kong. *International Journal of Corpus Linguistics,* 7 (2), 165–182.

Bondi, M. (2010). Perspectives on keywords and keyness: An introduction. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 1-18). Amsterdam: Benjamins

Bondi, M. & Scott, M. (Eds.) (2010). *Keyness in texts.* Amsterdam: Benjamins

Bonelli, E., & Sinclair, J. (2006). Corpora. In K. Brown, *Encyclopedia of Language and Linguistics* (2nd Edition ed.) (pp. 206-219). Oxford, United Kingdom: Elsevier.

Can, C. (2010). A Turkish learner corpus of English in second language studies: TICLE as a sub-corpus of ICLE. *Language Journal*,144, 16-34.

Can, C. (2011). Conjunctive adverbs in learner English: A usage-based approach. In G. Wlazlak (ed.) *The Dialogue of Language, The Dialogue of Culture*, 9, 92-105, Teacher Training College, Zabrze.

Can, C. (2012). Uluslararası Türk öğrenici İngilizcesi derleminde tutum belirteçleri. *Dilbilim Araştırmaları*, 1, 39-53.

Chambers, A. (2010). What is data-driven learning?. In A. O'Keeffe & M. McCarthy (Eds.), *Routledge Handbook of Corpus Linguistics* (pp. 345-358). London & NY: Routledge.

Cheng, W. (2010). What can a corpus tell us about language teaching?. In A. O'Keeffe & M. McCarthy (Eds.), *Routledge Handbook of Corpus Linguistics* (pp. 319-332). London & NY: Routledge.

Chomsky, N. (1957). Syntactic structures. The Hague: Mouton.

Claridge, C. (2008). Historical corpora. In A. Lüdeling & M. Kytö (Eds.), *Handbook on Corpus Linguistics* (pp. 242-259). Mouton de Gruyter.

Conrad, S. (1999). The importance of corpus-based research for language teachers. *System*, 25, 301-315.

Conrad, S. (2002). Corpus linguistics approaches for discourse analysis. *Annual Review of Applied Linguistics*, 22, 75-95.

Conrad, S. (2005). Corpus linguistics and L2 teaching. In E. Hinkel (Ed.) Handbook of research in second language teaching and learning (pp. 393-409). Mahwah NJ: Erlbaum.

Cook, G. (1989). *Discourse*. Oxford: Oxford University Press.

Crewe, W. J. (1990). The illogic of logical connectors. *ELT Journal*, 44 (4), 316–325.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics* 3(1), 59-80.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), 61-74

Evison, J. (2010). What are the basics of analysing a corpus?. In A. O'Keeffe & M. McCarthy (Eds.), *Routledge Handbook of Corpus Linguistics* (pp. 122-135). London & NY: Routledge.

Fei, D. (2006). The Effect of the use of adverbial connectors on Chinese EFL learners writing quality. *CELEA Journal*, 29 (1), 105–111.

Field, Y. & L. M. O. Yip. (1992). A Comparison of internal conjunctive cohesion in the English essay writing of Cantonese Speakers and native speakers of English. *Relc Journal,* 23(15), 15–28.

Fillmore, C.J. (1992). Corpus linguistics or computer-aided armchair linguistics. In J. Starvik (Ed.), Proceedings of Nobel Symposium: 82. *Directions in corpus linguistics* (pp. 35-60). Berlin: Mouton de Gruyter.

Flowerdew, L. (1998b) Application of learner corpus based findings and methods to pedagogy. In *Proceedings of First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 38-44.), *14-16 December, 1998.*

Flowerdew, L. (1998d) Integrating expert and interlanguage computer corpora findings on causality: discoveries for teachers and students. *English for Specific Purposes* 17(4), 329-345.

Flowerdew L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, A. Henry & R. Roseberry (Eds.), *Small Corpus Studies and ELT* (pp. 123-132). Amsterdam: Benjamins.

Gass, S. & Selinker, L. (2001) *Second Language Acquisition: An Introductory Course.* London: Lawrence Earlbaum.

Gavioli, L. (1997). Exploring texts through the concordancer: guiding the learner. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.) *Teaching and language corpora* (pp. 83–99). London: Longman.

Gilquin, G. & Granger, S. (2010). How can data-driven learning be used in language teaching?. In A. O'Keeffe & M. McCarthy (Eds.), *Routledge Handbook of Corpus Linguistics* (pp. 359-370). London & NY: Routledge.

Granger, S. (1993b) The International Corpus of Learner English. *The European English Messenger*, 2(1), 34.

Granger, S. (1994). The Learner Corpus: A revolution in applied linguistics. *English Today* 39, (10/3), 25-29.

Granger, S. (1996a) From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.) *Languages in Contrast. Text-based cross-linguistic studies* (pp. 37-51). Lund Studies in English 88. Lund: Lund University Press.

Granger, S. (1996b). Learner English around the World. In S. Greenbaum S. (Ed.) *Comparing English World-wide* (pp. 13-24). Oxford: Clarendon Press.

Granger, S. (1997b). On identifying the syntactic and discourse features of participle clauses in academic English: native and non-native writers compared. In J. Aarts, I. de Mönnink & H. Wekker (Eds.) *Studies in English Language and Teaching* (pp. 185-198).Amsterdam & Atlanta: Rodopi.

Granger, S. (1998). *Learner English on Computer*. London & New York: Addison Wesley Longman.

Granger S. (1998a). Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. Cowie (Ed.) *Phraseology: theory, analysis and applications*, (pp. 145-160). Oxford: Oxford University Press.

Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hong & S. Petch-Tyson (Eds.) *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33), Amsterdam & Philadelphia: John Benjamins.

Granger, S. (2003b). The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly,* 37(3) (special issue on Corpus Linguistics), 538-546.

Granger, S. (2004). Computer learner corpus research: current status and future prospects. In U. Connor & T.A. Upton (Eds.) *Applied Corpus Linguistics: A Multidimensional Perspective* (pp. 123-145).Amsterdam & Atlanta: Rodopi.

Granger, S. (2008). Learner Corpora. In A. Lüdeling & M. Kytö (Eds.), *Handbook on Corpus Linguistics* (pp.259-275). Mouton de Gruyter.

Granger, S. and Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes,* 15, 19-29.

Granger, S. & Rayson, P. (1998). Automatic Lexical Profiling of Learner Texts. In Granger S. (Ed.), *Learner English on Computer* (pp.119-131), London & New York: Addison Wesley Longman.

Granger, S. & Tribble, C. (1998). Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on Computer* (pp. 199-209). London & New York: Addison Wesley Longman.

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

Granger, S., Dagneaux, E. & Meunier, F. (2009). *The international corpus of learner English.Version 2. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Greenbaum, S., Nelson, G. & Weitzman, M. (1996). Complement clauses in English. In J. Thomas & M. Short (Eds.), *Using corpora for language research* (pp. 76-91). London: Longman.

Göksel, A. & Kerslake, C. (2005). *Turkish, A comprehensive grammar*. London: Routledge.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Heino, P. (2010). *Adverbial connectors in advanced EFL learners' and native speakers' student writing*. Student Thesis, University of Stockholm, Stockholm.

Hines, T. C., Harris, J. L. and Levy, C. L. (1970). An experimental concordance program. *Computers and the Humanities,* 4 (3), 161–71.

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.

Hundt, M., Nesselhauf, N. & Biewer, C. (2007). *Corpus linguistics and the web*. Amsterdam & New York: Rodopi.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2006). Corpus linguistics. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd Edition ed.), (pp. 234-248). Oxford, UK: Elsevier.

Johansson, S. (2008). Some aspects of the development of corpus linguistics in the 1970s and 1980s. In A. Lüdeling & M. Kytö (Eds.), *Handbook on Corpus Linguistics* (pp. 33-53). Mouton de Gruyter.

Johns, T. (1993). Data-driven learning: an update. *TELL & CALL*, 3, 4-10.

Johns, T. (1994). From to handout: Grammar and vocabulary teaching in context of data-driven learnin. In T. Odlin (Ed.), Perspectives on pedagogical grammar (pp. 293-313). Cambridge: Cambridge University Press.

Kaiser, S., Ichikawa, Y., Kobayashi, N. & Yamamoto, H. (2001). *Japanese: A comprehensive grammar.* London: Routledge.

Karlsson, F. (2008). Early generative linguistics and empirical methodology. In A. Lüdeling & M. Kytö (Eds.), *Handbook on Corpus Linguistics* (pp. 14-33). Mouton de Gruyter.

Kehoe, A. (2006). Diachronic linguistic analysis on the web with WebCorp. In A. Renouf & A. Kehoe (Eds.) *The Changing Face of Corpus Linguistics* (pp. 297-307). Amsterdam: Rodopi.

Kennedy, G. (1987). Quantification and the use of English: A case study of one aspect of the learner's task. *Applied Linguistics*, 8, 264-286.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Kettemann, B. & Marko, G. (Eds.) (2002). *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19*, *24 July, 2000.* Amsterdam: Rodopi.

Kilgarriff, A., Rychl, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress*. Lorient, France: Universite de Bretagne-Sud, 105–16. Retreived from http://nlp.fi.muni.cz/publications/euralex2004_kilgarriff_pary_smrz_tugwell/ (last accessed 25/04/2012).

Kilimci, A. & Can, C. (2009). TICLE: Uluslararası Turk Oğrenici İngilizcesi Derlemi. M. Sarıca, N. Sarıca ve A. Karaca (Ed.), XXII. Ulusal Dilbilim Kurultayı Bildirileri,1- 11, Ankara: Yuzuncu Yıl Universitesi Yayınları.

Kornfilt, J. (1997). *Turkish*. London and New York: Routledge

Lee D. Y. W. (2005). *Modelling variation in spoken and written language: the multi-dimensional approach revisited*. London: Routledge.

Lee, D. Y. W. (2010). What corpora are available?. In A. O'Keeffe & M. McCarthy (Eds.), *Routledge Handbook of Corpus Linguistics* (pp. 107-121). London/ NY: Routledge.

Leech, G. (1992). Corpora and theories of linguistic performance: in J. Svartvik (ed.), *Directions in corpus linguistics: proceedings of Nobel symposium 82*, (pp. 125-148). Berlin and New York, Mouton de Gruyter.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1-23). London: Longman.

Leech, G. (1998) Learner corpora: what they are and what can be done with them. In S. Granger (ed.), *Learner English on Computer* (pp.xiv-xx). London & New York: Addison Wesley Longman.

Leech, G. (1999). The distribution and function of vocatives in American and British English. In. H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora* (pp. 107-118). Amsterdam: Rodopi.

Leech, G. (2004). Adding linguistic annotation. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 17-29), Oxford: Oxbow Books. Retrieved from http://ahds.ac.uk/linguistic-corpora (accessed at 30/03/2012).

Lorenz, G. (1998). Overstatement in Advanced Learners' Writing: Stylistic Aspects of Adjective Intensification. In S. Granger (ed.) *Learner English on Computer* (pp. 53-66). London & New York: Addison Wesley Longman.

Lu, X. (2010). What can corpus software reveal about language development?. In A. O'Keeffe & M. McCarthy (Eds.), *Routledge Handbook of Corpus Linguistics* (pp. 184-193). London/ NY: Routledge.

MacWhinney, B. (1996). The CHILDES System. *American Journal of Speech-Language Pathology,* 5, 5–14.

McCarthy, M. & Carter, R. (1994). *Language as discourse: Perspectives for language teachers.* New York: Longman.

McCarthy, M. & Carter, R. (1995). Spoken grammar: What is it and how can we teach it?. *ELT Journal*, 49, 207-218.

Milton J. & Tsang E. (1993). A corpus-based study of logical connectors in EFL students' writing. In R. Pemberton & E. Tsang (Eds.), *Studies in Lexis* (pp. 215-246). Hong Kong University of Science and Technology, Hong Kong.

McCarten, J. (2010). Corpus-informed course book design. In A. O'Keeffe & M. McCarthy (Eds.), *Routledge Handbook of Corpus Linguistics* (pp. 413-427). London/ NY: Routledge.

McEnery, T. & Kifle N.A. (2002) Epistemic modality in argumentative essays of second-language writers. In J. Flowerdew (Ed.) *Academic Discourse* (pp. 182-215). London: Longman.

McEnery, T., Xiao, R., & Tonio, Y. (2006). *Corpus-based language studies: An advance resource book*. New York: Routledge.

McEnery, T., & A. Wilson (1996). *Corpus linguistics* (1st ed.). Edinburg: Edinburg University Press.

McEnery, T., & A. Wilson (2001). *Corpus linguistics* (2nd ed.). Edinburg: Edinburg University Press.

Meunier, F. (2002a). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Hung & S. Tyson (Eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 119-143). Amsterdam & Philadelphia: Benjamins.

Mindt, D. (1991). Syntactic evidence for semantic distinctions in English. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik* (pp. 183-196), London/New York: Longman.

Mindt, D. (1997). Mindt, Dieter (1997). Corpora and the Teaching of English in Germany. In Wichmann et al. (Eds.), *Teaching and Language Corpora* (pp. 40-50). London: Longman.

Meyer, C. F. (2002). *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Meyer, C. F. (2008). Pre-electronic corpora. In A. Lüdeling & M. Kytö (Eds.), *Handbook on Corpus Linguistics* (pp.1-14), Mouton de Gruyter.

Myers, G. (1989). The pragmatics of politeness in scientific articles. *Applied Linguistics*, 10, 1-35.

Narita, M., Sato, C. & Sugiura, M. (2004). Connector usage in the English essay writing of Japanese EFL learners. *Proceedings of 4th International Conference on Language Resources and Evaluation. LREC,* 1171–1174.

Neff, J., Ballesteros, F., Dafouz, E., Martínez, F. & Rica, J.-P. (2003). *Formulating Writer Stance: A Contrastive Study of EFL Learner Corpora.* In Archer et al. (eds.) (2003), 562-571.

Nesselhauf N. (2005). *Collocations in a learner corpus*. Amsterdam: Benjamins.

O'Keeffe, A. & McCarthy, M. (Eds.) (2010). Historical perspective: What are corpora and how have they evolved?. In O'Keeffe & M. McCarthy (Eds.), *Routledge Handbook of Corpus Linguistics* (pp. 3-14). London/ NY: Routledge.

Parrish, S. M. (1962). Problems in the making of computer concordances. *Studies in Bibliography*, 15, 1–14.

Partington, A. (1998). *Patterns and meanigs: Using corpora for English language research and teaching*. Amsterdam: Benjamins.

Quirk, R., Greenbaum, S., Leech, G. & Starvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.) *Learner English on Computer* (pp. 41-52). London & New York: Addison Wesley Longman.

Ritchie, W.C., & Bhatia, T. K. (Eds.) (1996). *Handbook of second language acquisition.* San Diego: Academic Press.

Reppen, R. (2001). Elementary student writing development: Corpus-based perspectives. In R. Simpson & J. Swales (Eds.) *Corpus linguistics in North America: selections from the 1999 Symposium* (pp. 211–225). Ann Arbor: University of Michigan Press.

Reppen, R. (2006). Corpus linguistics: second language. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics,* (2nd Edition ed.), (pp. 248-250), Oxford, UK: Elsevier.

Römer, U. (2008). Corpora and language teaching. In A. Lüdeling & M. Kytö (Eds.), *Handbook on Corpus Linguistics,* (pp. 112-131), Mouton de Gruyter.

Rayson, P. (2003). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. *Ph.D. thesis*, Lancaster University

Rayson, P. & Garside, R. (2000). Comparing corpora using frequency profiling. In proceedings of the *workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*. 1-8 October 2000 (pp. 1-6), Hong Kong.

Rayson P., Berridge D. & Francis B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In *Volume II of* G. Purnelle., C. Fairon, & A. Dister (Eds.), *Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004),* *Louvain-la-Neuve, Belgium, March 10-12, 2004*, (pp. 926 - 936), Presses universitaires de Louvain.

Shea, M. (2009). A Corpus-based study of adverbial connectors in learner text. *MSU Working Papers in SLS,* 1, 1, 1-13.

Sigley, R. (2006). Corpora in studies of variation. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd Edition ed.), (pp. 220-226). Oxford, UK: Elsevier.

Say, B., Zeyrek, D., Oflazer, K.,& Özge, U. (2004). Development of a corpus and a treebank for present-day written Turkish. In K. İmer & G. Doğan (Eds.), Proceedings of the Eleventh International Conference of Turkish Linguistics, August, 2002, *Current Research in Turkish Lingustics,* 183-192, Eastern Mediterranean University Press.

Schmied, J. (2006). New ways of analysing ESL on the WWW with WebCorp and WebPhraseCount. In A. Renouf & A. Kehoe (Eds.) *The Changing Face of Corpus Linguistics* (pp. 309-324). Amsterdam: Rodopi.

Scott, M. (1996). *WordSmith Tools*. Oxford: Oxford University Press. ISBN 0-19-458984-6.

Scott, M. (1997). *WordSmith Tools.* version 2. Oxford: Oxford University Press. ISBN 0-19-459283-9.

Scott, M. (1999). *WordSmith Tools*. version 3, Oxford: Oxford University Press. ISBN 0-19-459289-8.

Scott, M. (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. In M. Ghadessy, A. Henry & R. Roseberry (Eds.) *Small Corpus Studies and ELT* (pp. 47-67). Amsterdam: Benjamins.

Scott, M. (2004). *WordSmith Tools*. version 4, Oxford: Oxford University Press. ISBN: 0-19-459400-9.

Scott, M. (2008). *WordSmith Tools.* version 5, Liverpool: Lexical Analysis Software.

Scott, M. (2010). What can corpus software do?. In O'Keeffe & M. McCarthy (Eds.), *Routledge Handbook of Corpus Linguistics* (pp. 136-151). London & NY: Routledge.

Scott, M. (2012). WordSmith Tools Help. Liverpool: Lexical Analysis Software. ttp://www.lexically.net/downloads/version5/HTML/proc_tag_handling.htm (last accessed 30/04/2012).

Scott, M. & Tribble, C. (Eds.) (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam & Phladelphia: Benjamins.

Sinclair, J. (1987). *Looking up: An account of the COBUILD project in lexical computing*. London: Collins.

Sinclair, J. (1996). EAGLES preliminary recommendations on corpus typology. *EAG-TCWG-CTYP/P*. Pisa: ILC-CNR.

Sinclair, J. (2004). *How to use corpora for language teaching.* Amsterdam & Philadelphia: John Benjamins.

Sinclair, J. et al. (Eds.) (2001). *Collins COBUILD English Dictionary for Advanced Learners.* London: HarperCollins.

Stenström, A.-B. (1987). Carry-on signals in English conversation. In W. Mejis (Ed.) *Corpus linguistic and beyond* (pp. 87-119). Amsterdam: Rodopi.

Starvik, J. (1992). Introduction. In J. Starvik (Ed.), Proceedings of Nobel Symposium: 82. *Directions in corpus linguistics* (pp. 7-14). Berlin: Mouton de Gruyter.

Starvik, J. (2007). *Corpus linguistics: 25+ years on*. Amsterdam: Rodopi.

Stubbs, M. (1996). *Text and corpus analysis*. Oxford: Blackwell.

Tankó, G. (2004). The use of adverbial Connectors in Hungarian university students' argumentative assays. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 157-181). Amsterdam: John Benjamins.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam & Philadelphia: John Benjamins.

Tono, Y. (2000). A computer learner corpus based analysis of the acquisition order of English grammatical morphemes. In L. Burnard and T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 123-132). Frankfurt am Main: Peter Lang.

Virtanen, T. (1998b) Direct questions in argumentative student writing. In S. Granger (Ed.), *Learner English on Computer* (pp. 94-118). London & New York: Addison Wesley Longman.

Wei-yu Chen, C. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics*, 11(1), 113–130.

Wikberg, K. (1992). Discourse category and text type classification: procedural discourse and production in the Brown and the LOB corpora. In G. Leitner (Ed.) In Leitner, Gerhard (Ed.), *New directions in English language corpora:*

Methodology, results, software developments (pp. 247-261). Berlin: Mouton de Gruyter.

Wilson, A. (1989). *Prepositional phrase modifiers of nominal and their prosodic boundaries: some data from the Lancaster Spoken English Corpus*. MA Thesis, Lancaster University.

Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (Eds.) (1997). *Teaching and Language Corpora.* London: Longman.

Tang, W.M. (2012). A Very Short Introduction to Corpus Linguistics. Retreived from http://wmtang.org/corpus-linguistics/ (last accessed 24/04/2012).

Relevant Websites

Brown Corpus (n.d.). In Wikipedia online. Retrieved from http://en.wikipedia.org/wiki/Brown_Corpus

ICAME (n.d.). ICAME online. Retrieved from http://icame.uib.no/

BNC (2010). BNC online. Retrieved from http://www.natcorp.ox.ac.uk/corpus/index.xml

ICE (n.d.). ICE online. Retrieved from http://ice-corpora.net/ice/index.htm

Metu Corpus (2009). Metu Corpus online (data file). Retrieved from http://www.ii.metu.edu.tr/corpus

WebCorp (2012). WebCorp online. Retrieved from http://www.webcorp.org.uk/live

UCREL (1993-2010). UCREL online. Retrieved from http://ucrel.lancs.ac.uk/annotation.html

Scott, M. (2012). WordSmith Tools program. Retrieved from http://lexically.net/wordsmith/index.html

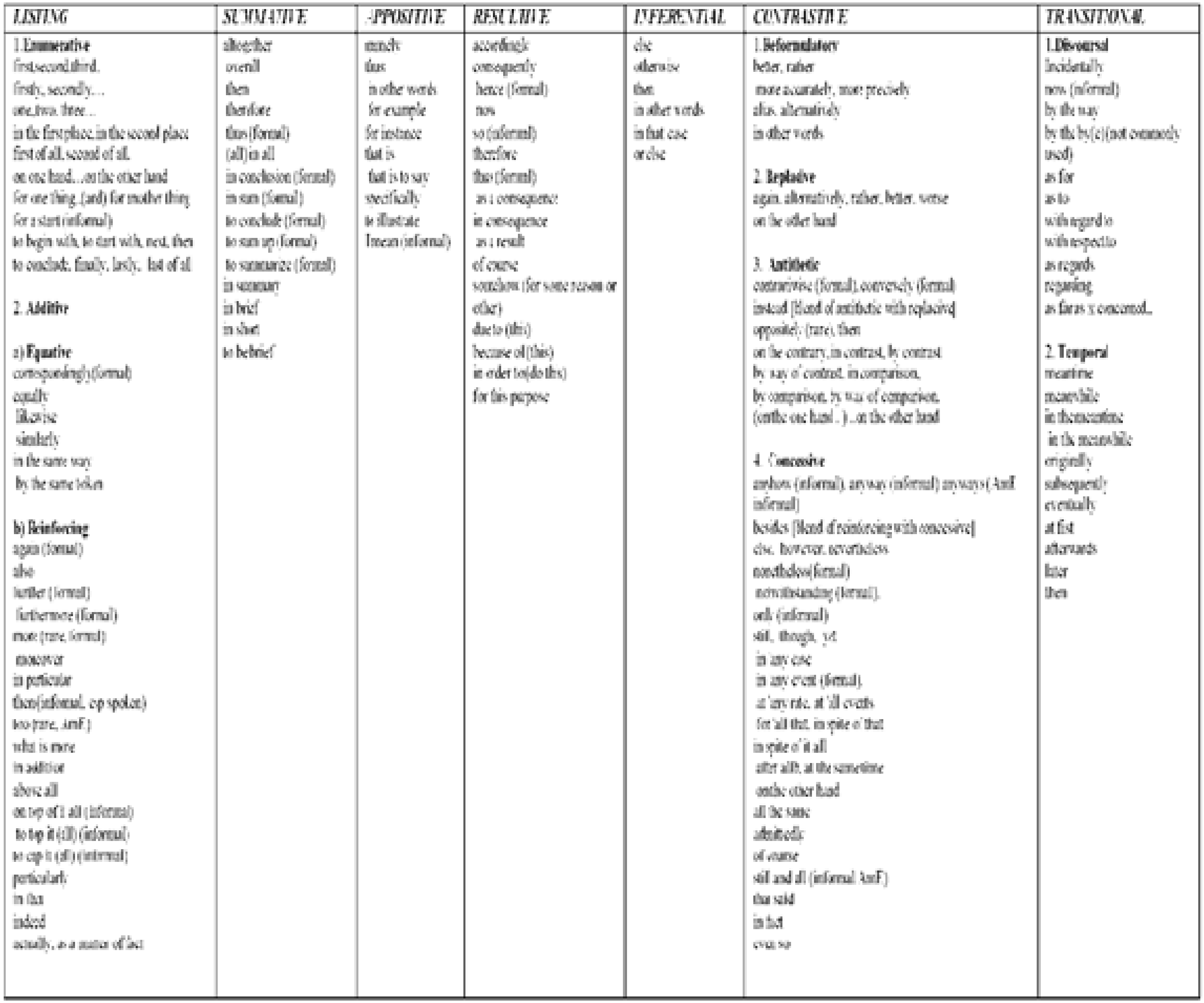
COBUILD (2012). Collind Corpus and Bank of English (data file). Retrieved from http://www.mycobuild.com/about-collins-corpus.aspx

Turkish Connectors (2012). TDK online dictionary. Retrieved from http://www.tdk.gov.tr/

Spanish Grammar (n.d.). Spanish Booster online. Retreived from http://www.spanishbooster.com/SpanishConjunctions.htm

APPENDICES

APPENDIX A
LIST OF ADVERBIAL CONNECTORS IN ENGLISH
(Based on Quirk et. al., 1985)



LISTING	*SUMMATIVE*	*APPOSITIVE*	*RESULTIVE*	*INFERENTIAL*	*CONTRASTIVE*	*TRANSITIONAL*
1.**Enumerative** first,second,third. firstly, secondly… one,two, three… in the first place, in the second place first of all, second of all, on one hand…on the other hand for one thing ,(and) for another thing for a start (informal) to begin with, to start with, next, then to conclude, finally, lastly, last of all 2. **Additive** a) **Equative** correspondingly (formal) equally likewise similarly in the same way by the same token b) **Reinforcing** again (formal) also further (formal) furthermore (formal) more (rare, formal) moreover in particular then(informal, esp spoken) too (rare, AmE) what is more in addition above all on top of it all (informal) to top it (all) (informal) to cap it (all) (informal) particularly in that indeed actually, as a matter of fact	altogether overall then therefore thus (formal) (all) in all in conclusion (formal) in sum (formal) to conclude (formal) to sum up (formal) to summarize (formal) in summary in brief in short to be brief	namely thus in other words for example for instance that is that is to say specifically to illustrate I mean (informal)	accordingly consequently hence (formal) now so (informal) therefore thus (formal) as a consequence in consequence as a result of course somehow (for some reason or other) due to (this) because of (this) in order to(do this) for this purpose	else otherwise then in other words in that case or else	1.**Reformulatory** better, rather more accurately, more precisely alias, alternatively in other words 2. **Replacive** again, alternatively, rather, better, worse on the other hand 3. **Antithetic** contrariwise (formal), conversely (formal) instead [blend of antithetic with replacive] oppositely (rare), 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 4. **Concessive** anyhow (informal), anyway (informal) anyways (AmE informal) besides [blend of reinforcing with concessive] else, however, nevertheless nonetheless(formal) notwithstanding (formal), only (informal) still, though, yet in any case in any event (formal), 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 (informal AmE) that said in fact even so	1.**Discoursal** Incidentally now (informal) by the way by the by(e) (not commonly used) as for as to with regard to with respect to as regards regarding as far as x concerned.. 2. **Temporal** meantime meanwhile in the meantime in the meanwhile originally subsequently eventually at first afterwards later then

APPENDIX B
LIST OF TURKISH EQUIVALENTS OF ADVERBIAL CONNECTORS IN ENGLISH

1) LISTING (LİSTELEME BAĞLAÇLARI)

1.1. Enumerative (Sayıma Ait Bağlaçları)

İlk önce Önce	First, Firstly
Öncelikle İlk başta İlk olarak	In the first place, First of all
Önce…., Sonra….	For one thing…for another..
Başlarken Başlangıçta*	To Begin with, To start with
Daha sonra Ondan sonra	Next, Then
(En) Sonunda Son olarak Velhasıl	To conclude, Finally, Lastly, Last of all

*Connectors which can function in different categories

1.2. ADDITIVE (EKLEMELİ BAĞLAÇLAR)

1.2.1.Equative (Eşitlik Belirten Bağlaçlar)

Buna paralel olarak Aynı şekilde Benzer (bir) şekilde Bunun gibi Aynen Aynı ölçüde Eşit olarak	Correspondingly, Equally Similarly, Likewise, In the same way, By the same token

1.2.2. Reinforcing (Pekiştireç Bağlaçlar)

Ayrıca *Aynı zamanda Yine Tekrar Bir kere daha Hem (de) Bir de	Again, Then, Also, Too,
Hatta Üstelik Dahası Ayrıyeten Buna ek olarak Yanısıra	Further, Furthermore, Moreover, More, What is more, In addition
En önemlisi Herşeyden önemlisi	Above all
Üstüne üstlük Buda yetmezmiş gibi	On top of it all, To top it all, To cap it all
Özellikle	In particular, Particularly
Aslında Asıl Adeta Gerçekten (de) Doğrusu Zaten Nitekim Gerçekte	In fact, Actually, Inded As a matter of fact

*Connectors which can function in different categories

2. ÖZETSEL BAĞLAÇ (SUMMATIVE)

Öyleyse O halde O zaman Bu durumda Böylece Böylelikle	Then, Therefore, Thus,
Sonuç olarak Son olarak En nihayetinde Velhasıl Özetle Özetlemek gerekirse Özetle anlatmak gerekirse Sözün özü	In conlusion, In sum, To conlude, To sum up, To summarize, In summary, Overall, (All) in all)
Kısacası Uzun sözün kısası Uzun lafın kısası	In brief To be brief In short

*Connectors which can function in different categories

3. APPOSITIVE (EŞLEMELİ BAĞLAÇ)

Yani Şöyle ki Bu demek oluyor ki Bir başka deyişle Başka bir deyişle Diğer bir deyişle ..adlandırılan Demek istediğim	Namely, That is, That is to say, In other words I mean

Mesela	For example
Örneğin	For instance
Örnek olarak	(specifically)
Örnek vermek gerekirse	(To illustrate)
Örnek verecek olursak	
Sözgelimi	

*Connectors which can function in different categories

4. RESULTIVE CONNECTORS (SONUÇSAL BAĞLAÇ)

Buna bağlı olarak	
Böylece	
Böylelikle	Accordingly, Therefore,
Bu doğrultuda	Hence, Thus,
Bu yüzden	So, Consequently,
O yüzden	Because of this
Onun için	Due to this
Bu nedenle	In order to do this
Bunun için	
Bu sebeple	
Dolayısıyla (Dolayısı ile)	
Bundan dolayı	
Bu sebepten dolayı	
Bu deneden ötürü (dür ki)	
nedeniyle	
*Sonuç olarak Bunun sonucu olarak	As a consequence ,In consequence , As a result
Elbette (ki)	
Tabi(i) ki	Of course
Pek tabi	
Her nedense	
Öyle yada böyle	Somehow , For this purpose

* Connectors which can function in different categories

5. ÇIKARIMSAL BAĞLAÇ (INFERENTIAL)

Aksi takdirde	Else,
Aksi halde	Or else,
Yoksa	Otherwise
Bunun dışında	In that case,
*O halde	Then, In other words
O zaman	

*Connectors which can function in different categories

6. KARŞITSAL BAĞLAÇ (CONTRASTIVE)

6.1. Yeniden Düzenleyici Bağlaç (REFORMULATORY)

Daha doğrusu … ziyade	Better, Rather, More accurately, More precisely,
	Alternatively, Alias, In other words (????)

*Connectors which can function in different categories

6.2. Yer Tutan Bağlaç (REPLACIVE)

Daha doğrusu Daha kötüsü/beteri Öte yandan /taraftan_Diğer yandan /taraftan	Rather, Alternatively, Better, Worse, On the other hand,Again

*Connectors which can function in different categories

6.3. Zıtlık Bağlaç (ANTITHETIC)

(tam)Aksine (tam)Tersine Bilakis Buna karşılık	Contrariwise, ,Oppositely, Conversely, On the contrary, In contrast, By contrast, By way of contrast, Then, Instead
Nazaran Kıyasla	In comparison, By comparison, By way of comparison
Bir yandan…diğer bir yandan Bir taraftan…diğer bir taraftan/öte yandan/diğer yandan	On one hand…on the other hand

*Connectors which can function in different categories

6.4. Ödünleyici Bağlaç (CONCESSIVE)

Neyse Her neyse Nasıl olsa Bir şekilde (herhangi bir şekilde)	Anyhow, Anyway, Anywise,
Bunun yanısıra Veyahut Bunun dışında Diğer taraftan (yandan)- Öte yandan (taraftan) - Bir taraftan	Besides, Else (on the other hand)
Ama Fakat Lakin Ancak Halbuki Oysa ki	However, Only, Yet, Still
(Buna) (Herşeye) (-e/-a) rağmen	Nonetheless, Nevertheless Notwithstanding,

Yine de (gene de)	Though, In spite of hat, In spite of it all, That said, Even so, In fact
Ne olursa olsun Herşeye rağmen	In any case, In any event At any rate, At all events, For all that, After all
Bununla birlikte (beraber) *Aynı zamanda	At the same time, All the same,
(Hiç) Kuşkusuz *Elbette Şüphesiz *Pek tabi Tabi(i) ki	Of course, Admittedly
?	Still and all (??)

*Connectors which can function in different categories

7. TRANSITIONAL (GEÇİŞKEN BAĞLAÇ)

7.1. Discoursal (Söylevsel Bağlaç)

Bu arada Yeri gelmişken	Incidentally, By the way, By the by, (now)
..Hakkında …Hususunda …Bakımından …ile ilgili …İlişkin …Konusunda -Dair -Gelince	As for, As to, With regard to, With respect to, As regards, Regarding
Kadarıyla	As far as X concerned

*Connectors which can function in different categories

7.2. Temporal (Zamansal Bağlaç)

Bu sırada Bu arada	Meantime, Meanwhile, In the meantime, In the meanwhile
(Daha /Ondan) Sonra Daha sonraları	Then, Later, Afterwards, Subsequently
İlk başta Başlangıçta* İlk olarak İlkin İlk önce	At first Originally
Er ya da geç Eninde sonunda	Eventually

*Connectors which can function in different categories

APPENDIX C

The institution codes processed in the analysis of TICLE

Nation	Essay Codes	CODE
Turkey	University of Çukurova	CU
Turkey	Mersin University	ME
Turkey	Mustafa Kemal University	KE

APPENDIX D

The institution codes processed in the analysis of JPICLE

Nation	Essay Codes	Code
JAPAN	Keio University	KO
JAPAN	Kooriyama Women's University	KW
JAPAN	Kyoto University	KY
JAPAN	Miyagi University of Education	MI
JAPAN	Meiji University	MJ
JAPAN	Musashi University	MU
JAPAN	Nihon University	NH
JAPAN	Okayama University	OK
JAPAN	Rikkyo University	RI
JAPAN	Seijyo University	SE
JAPAN	Shinshu University	SH
JAPAN	Shonann Institution of technology	ST
JAPAN	Showa Women's University	SW
JAPAN	Tokyo University of Foreign Studies	TF
JAPAN	Tokai University	TK
JAPAN	Tamagawa University	TM
JAPAN	Waseda University	WA

APPENDIX E

The institution codes processed in the analysis of SPICLE

Nation	Essay Codes	CODE
Spain	Universidad Complutence de Madrid	M
Spain	Universidad de Alcala	AL

APPENDIX F

Corpus collection guideline for compiling TUC

Yönergeler

1. Aşağıdaki anketi doldurunuz.
2. Verilen kompozisyon konularından bir tanesini seçiniz.
3. Seçtiğiniz konu üzerine Türkçe olarak **en az 500 kelimeden oluşan** kurallara uygun bir kompozisyon yazınız.
4. Yazdığınız kompozisyonu ve anket bilgilerinizi **turkishcorpus@gmail.com** adresine yollayınız.

Anket

Adınız-soyadınız:

Öğrenci no ve şubeniz:

Yaşınız :

Cinsiyetiniz :

Ana diliniz :

Yaşadığınız Ülke:

Bildiğiniz diğer diller:

Kompozisyon Konuları

1. Suç hiç kimseye yarar sağlamaz. Çünkü eninde sonunda yakalanırsınız.
2. Cezaevi sistemi eskidi. Uygar toplumlar suçlularını cezalandırmaktansa onları rehabilite etmelidir.
3. Çoğu üniversite diploması teoriktir ve öğrencilerini gerçek hayata hazırlamaz. Bu yüzden hiçbir değeri yoktur.
4. Bir erkeğin yada kadının mali geliri (kazancı), yaşadığı topluma yaptığı katkılarla uyumlu olmalıdır.
5. Batı toplumunda sansürün yeri.
6. Marx dinin kitlelerin afyonu olduğunu söylemiştir. 20 yy. sonlarında yaşıyor olsaydı dini televizyonla değiştirirdi.
7. Tüm ordular tamamen profesyonel askerlerden oluşmalıdır. Askerlik sisteminde değer kavramı yoktur.

8. Körfez savaşı sonrası bize insanın ülkesi için savaşmasının hala önemli (kutsal) bir şey olduğunu gösterdi.
9. Feministler, kadınların davasına yarardan çok zarar vermişlerdir.
10. Hayvan Çiftliği romanında George Orwell şöyle yazmıştı "Bütün insanlar eşittir; fakat bazıları diğerlerinden daha eşittir." Demiştir. Bu görüş günümüzde ne kadar doğrudur?
11. Eski bir şarkı sözüne göre para bütün kötülüklerin anasıdır.
12. Avrupa; Egemenliğin yitirilmesi mi yoksa yeni bir toplumun doğuşumu?
13. 19 yüzyılda Victor Hugo "Doğanın insanlara seslendiğini ancak insanların onu önemsemediğini düşünmek ne acı" demiştir. Bu tümcenin bugün hala geçerli olduğunu düşünüyor musunuz?
14. Bazı insanlara göre bilim, teknoloji ve sanayileşme ile yönetilen günümüz dünyası da düş ve hayal gücüne artık yer yok. Bu konuda düşünceniz nedir?

CURRICULUM VITAE

A. PERSONAL INFORMATION

Name : M. Pınar BABANOĞLU

Date / Place of Birth: 02/07/1973 Adana

E-mail : mpinarbabanoglu@hotmail.com

B. EDUCATIONAL BACKGROUND

Date	**Degree**	**University**	**Field**
2008-present	Doctor of Philosophy	Çukurova University, Social Sciences	English Language Teaching
2004-2007	Master of Arts	Çukurova University, Social Sciences	English Language Teaching
1993-1997	Bachelor of Arts	Çukurova University	English Language Teaching

C. JOB EXPERIENCE

Date	**Title**	**Institution**
2010-2011	Coordinator	International Office, Osmaniye Korkut Ata Uni., Osmaniye
2008-2010	Lecturer	English Language Teaching Department, The Faculty of Education, Çukurova Uni., Adana
2004 -2004	English Teacher	12 Şubat İlköğretim Okulu, Kahramanmaraş
2001-2004	English Teacher	Remzi Oğuz Arık İlköğretim Okulu, Kozan, Adana

D. ACADEMIC WORK

PAPERS PRESENTED

1. "The acquisition of English dative alternations by Turkish adult learners"
   17th International Conference on Foreign/Second Language Acquisition, University of Slaski, Poland, May 2005 (Joint Paper)
2. "The Use of Lexical Collocations by Turkish ELT and EFL Students"
   22nd International Conference on Foreign/Second Language Acquisition, University of Slaski, Poland, May 2010

RESEARCH

Research visit for Doctoral study, Aston University, Birmingham, United Kingdom, 18th June- 9th September 2011

COURSES/SEMINARS/CONFERENCES

1. Academic Writing, Çukurova University, February 2005
2. 6th International Corpus Linguistics Conference, Birmingham University, Birmingham, United Kingdom, 20th - 22nd July 2011
3. Aston Corpus Summer School, Aston University, Birmingham, United Kingdom, 1st -5th August 2011
4. TEFL Course, Birmingham, United Kingdom, 21st -22nd August 2011