In this book, an historical overview of corpus linguistics, related literature of corpus linguistics and its place in language teaching are explained in general terms.

As an introduction, general and specific literature related to corpus linguistics are focused in order to be comprehensive in respect of corpus-based research. Firstly, the background of development of corpus linguistics diversified with a timeline. Three main periodsare explained in detail as; 13th century to 1950s, early modern corpora period and contemporary corpora periods. Then main points computerized corpora and types of corpora are emphasized to provide a comprehensive understanding on the background of fundamentals of corpora. The last sections scrutinize the place and the use of corpora in language teaching by focusing on learner corpo-ra research.

Turkuaz Kongre Organizasyonu
Fuarcılık Yayıncılık Eğitim
Turizm ve Ticaret Ltd. Şti.

A COMPREHENSIVE OVERVIEW OF CORPUS LINGUISTICS IN LANGUAGE TEACHING

M. Pınar BABANOĞLU
Cem CAN

A COMPREHENSIVE OVERVIEW OF CORPUS LINGUISTICS IN LANGUAGE TEACHING



M. Pınar BABANOĞLU & Cem CAN

A COMPREHENSIVE OVERVIEW OF CORPUS LINGUISTICS IN LANGUAGE TEACHING

TURKUAZ
Kongre Kitabı Yayıncılık

A COMPREHENSIVE OVERVIEW OF CORPUS LINGUISTICS IN LANGUAGE TEACHING

M. Pınar BABANOĞLU (Orcid 0000-0001-8166-974X)
Cem CAN (Orcid 000-0002-0324-5527)

TURKUAZKEY AKADEMİK YAYINLAR
Araştırma Kitapları Dizisi: Eğitim Bilimleri 2022/35
Research Book Series: Educational Sciences 2022/35

ISBN: 978-625-6379-34-3

Birinci Basım Aralık 2022 © *Copyright 2022*

First Edition December 2022 © Copyright 2022

All responsibility for the content of the book belongs to the authors of the chapters.
All rights of this work are reserved. Without the written permission of Turkuaz Congress Organization Fuarcılık Publishing Education, Tourism and Trade Limited Company; It cannot be adapted to radio or television in any way, made into a game, film, electronic book, CD or magnetic tape, reproduced partially or completely by photocopying or any other means, reproduced copies cannot be kept or distributed.

Cover Design: Turkuazkey (Image taken from Freepik)



Turkuaz Kongre Organizasyonu Fuarcılık Yayıncılık Eğitim Turizm ve Ticaret Limited Şirketi
İletişim: turkuazkey@hotmail.com
I.Anafartalar Mah. Konuk Cad. No:6-10 Şehzadeler/MANİSA

A COMPREHENSIVE OVERVIEW OF CORPUS LINGUISTICS IN LANGUAGE TEACHING

M. Pınar BABANOĞLU
Cem CAN

TURKUAZ
Kongre Eğitim Yayıncılık

PREFACE

In this book, an historical overview of corpus linguistics, related literature of corpus linguistics and its place in language teaching are explained in general terms.

As an introduction, general and specific literature related to corpus linguistics are focused in order to be comprehensive in respect of corpus-based research. First-ly, the background of development of corpus linguistics diversified with a timeline. Three main periodsare explained in detail as; 13th century to 1950s, early modern corpora period and contemporary corpora periods. Then main points computerized corpora and types of corpora are emphasized to provide a comprehensive understanding on the background of fundamentals of corpora. The last sections scrutinize the place and the use of corpora in language teaching by focusing on learner corpo-ra research.

This study was produced from M.Pınar BABANOĞLU's doctoral thesis.

Writers
M. Pınar BABANOĞLU & Cem CAN

CONTENTS

Preface i
List of Tables v
List of Figures vi
Introduction 1
1 Historical Overview of Corpus Linguistics 3
1.1 Early Corpora and Pre-electronic Period: Thirteenth Century –1950s..... 5
1.2 Generative Grammar vs. Corpus Linguistics 9
1.3 Early Modern Corpora: 1950s - 1980s 18
1.4 First Computer-based Corpus: Brown Corpus and Beyond..... 20
1.5 Contemporary Corpora: 1990s-Present Day 28
1.6 Future of Corpora 32
2 The Scope of Corpus Linguistics.....34
2.1 Corpus Linguistics: Theory or a Methodology?.... 35
2.2 Corpus-based vs. Corpus Driven Approaches..... 36
3 Aspects of Corpora 39
3.1 Processing and Annotation of Corpus 42
3.1.1 Part-of- Speech Tagging..... 43
3.1.2 Lemmatisation 45
3.1.3 Parsing 47
3.1.4 Semantic Annotation 49
3.1.5 Prosodic Annotation 50
4 Corpus Analysis 53
4.1 Concordance 54
4.2 Frequency 55
4.3 Wordlists..... 58

4.4 Keyword and Keyness 61
5 Types of Corpora 65
5.1 General Corpora 65
5.2 Specialized Corpora 66
5.3 Written and Spoken Corpora 67
5.4 Synchronic and Diachronic Corpora 68
5.5 Historical Corpora 69
5.6 Multilingual Corpora 70
5.7 Monitor Corpora 72
5.8 Learner Corpora 73
5.8.1 International Corpus of Learner English (ICLE) 75
5.8.2 Learner Corpora Analysis 78
5.8.3 Contrastive Interlanguage Analysis (CIA) 79
5.8.4 Computer-aided Error Analysis 81
5.8.5 Studies on Learner Corpora 82
6 Corpus-based Research in Language Learning and Teaching 87
6.1 The Applications of Corpora in Language Learning and Teaching 91
6.2 Data-driven Learning 96
Summary 100
References 101

List of Tables

Table 1: The Revival of Corpus Linguistics (Johansson, 1991:312, adapted from McEnery & Wilson, 1996: 18) 20
Table 2: The Composition of Brown Corpus and LOB Corpus (adapted from Johansson, 2008: 36)..... 22
Table 3: A Sample of Rank Order Frequency List (adapted from Evison 2010:124)..... 57
Table 4: A Sample of Comparison of Rank Frequency (adapted from Evison, 2010: 126)..... 58
Table 5: Positive Keywords in Sociology and History Texts (Adapted from Evsion, 2010: 127)..... 63
Table 6: Studies on Learner Corpora 84

List of Figures

Figure 1: Timeline of corpus linguistics 3
Figure 2: A citation slip from the OED (adapted from Meyer, 2008:9)..... 8
Figure 3: Phrase structure rules of recursion (McEnery a& Wilson, 1996:14)..... 13
Figure 4: An example of POS tagging from LOB corpus (adapted from http://ucrel.lancs.ac.uk/annotation.html). 43
Figure 5: Example of lemmatization from SUSANNE corpus (adapted from McEnery and Wilson, 1996: 54)..... 46
Figure 6: Parsing example from BNC corpus (adapted from McEnery and Wilson, 1996: 43). 48
Figure 7: An example of full parsing from the Lancester-Leeds Treebank (adapted from McEnery and Wilson, 1996:45) 49
Figure 8: An example of semantic annotation (adapted from http://ucrel.lancs.ac.uk/annotation.html)..... 49
Figure 9: An example of prosodic annotation from London-lund corpus (adapted from McEnery and Wilson, 1996: 55)..... 51
Figure 10: A sample concordance of 'witnessed' (adapted from Baker, Hardie and McEnery, 2006: 43). 55
Figure 11: A screenshot of a word list from antconc software (adapted from http://wmtang.org/corpus-linguistics/)..... 60
Figure 12: A screenshot of a wordlist sample by WordSmith Tools Software (adapted from Scott, 2010: 148)..... 61
Figure 13: Types of multilingual corpora (adapted from Aijmer, 2008:276)..... 71

Figure 14: A Screenshot of Learners' Native Language list in ICLEv2 Software (Granger et al., 2009). 78
Figure 15: Contrastive interlanguage analysis (adapted from Granger et al., 2009:p.40) 79
Figure 16: The concern of mainstream language teaching (Mark, 1998, adapted from Granger, 2002:6) 89
Figure 17: The focus on learner output (Mark, 1998, adapted from Granger, 2002:7) 90
Figure 18: The relationship between corpus linguistics and language teaching (adapted from Römer, 2008: 113). 91
Figure 19: Applications of corpora in language teaching (adapted from Römer, 2008: 113) 93
Figure 20: An extract from touchstone student book 3 (McCarthy, McCarten and Sandiford, 2006a, cited in McCarten, 2010:421) 95

CONTENTS

List of Tables iii
List of Figures iv
Introduction 1
1 Historical Overview of Corpus Linguistics 3
1.1 Early Corpora and Pre-electronic Period: Thirteenth Century –1950s 5
1.2 Generative Grammar vs. Corpus Linguistics 9
1.3 Early Modern Corpora: 1950s - 1980s 18
1.4 First Computer-based Corpus: Brown Corpus and Beyond 20
1.5 Contemporary Corpora: 1990s-Present Day 28
1.6 Future of Corpora 32
2 The Scope of Corpus Linguistics 34
2.1 Corpus Linguistics: Theory or a Methodology? 35
2.2 Corpus-based vs. Corpus Driven Approaches 36
3 Aspects of Corpora 39
3.1 Processing and Annotation of Corpus 42
3.1.1 Part-of- Speech Tagging 43
3.1.2 Lemmatisation 45
3.1.3 Parsing 47
3.1.4 Semantic Annotation 49
3.1.5 Prosodic Annotation 50
4 Corpus Analysis 53
4.1 Concordance 54
4.2 Frequency 55
4.3 Wordlists 58

4.4 Keyword and Keyness 61
5 Types of Corpora 65
5.1 General Corpora 65
5.2 Specialized Corpora 66
5.3 Written and Spoken Corpora 67
5.4 Synchronic and Diachronic Corpora 68
5.5 Historical Corpora 69
5.6 Multilingual Corpora 70
5.7 Monitor Corpora 72
5.8 Learner Corpora 73
5.8.1 International Corpus of Learner English (ICLE) 75
5.8.2 Learner Corpora Analysis 78
5.8.3 Contrastive Interlanguage Analysis (CIA) 79
5.8.4 Computer-aided Error Analysis 81
5.8.5 Studies on Learner Corpora 82
6 Corpus-based Research in Language Learning and Teaching 87
6.1 The Applications of Corpora in Language Learning and Teaching 91
6.2 Data-driven Learning 96
Summary 100
References 101

List of Tables

Table 1: The Revival of Corpus Linguistics (Johansson, 1991:312, adapted from McEnery & Wilson, 1996: 18) 20
Table 2: The Composition of Brown Corpus and LOB Corpus (adapted from Johansson, 2008: 36)..... 22
Table 3: A Sample of Rank Order Frequency List (adapted from Evison 2010:124)..... 57
Table 4: A Sample of Comparison of Rank Frequency (adapted from Evison, 2010: 126)..... 58
Table 5: Positive Keywords in Sociology and History Texts (Adapted from Evsion, 2010: 127)..... 63
Table 6: Studies on Learner Corpora 84

List of Figures

Figure 1: Timeline of corpus linguistics 3
Figure 2: A citation slip from the OED (adapted from Meyer, 2008:9)..... 8
Figure 3: Phrase structure rules of recursion (McEnery a& Wilson, 1996:14)..... 13
Figure 4: An example of POS tagging from LOB corpus (adapted from http://ucrel.lancs.ac.uk/annotation.html). 43
Figure 5: Example of lemmatization from SUSANNE corpus (adapted from McEnery and Wilson, 1996: 54)..... 46
Figure 6: Parsing example from BNC corpus (adapted from McEnery and Wilson, 1996: 43). 48
Figure 7: An example of full parsing from the Lancester-Leeds Treebank (adapted from McEnery and Wilson, 1996:45) 49
Figure 8: An example of semantic annotation (adapted from http://ucrel.lancs.ac.uk/annotation.html)..... 49
Figure 9: An example of prosodic annotation from London-lund corpus (adapted from McEnery and Wilson, 1996: 55)..... 51
Figure 10: A sample concordance of 'witnessed' (adapted from Baker, Hardie and McEnery, 2006: 43). 55
Figure 11: A screenshot of a word list from antconc software (adapted from http://wmtang.org/corpus-linguistics/)..... 60
Figure 12: A screenshot of a wordlist sample by WordSmith Tools Software (adapted from Scott, 2010: 148)..... 61
Figure 13: Types of multilingual corpora (adapted from Aijmer, 2008:276)..... 71

Figure 14: A Screenshot of Learners' Native Language list in ICLEv2 Software (Granger et al., 2009). 78
Figure 15: Contrastive interlanguage analysis (adapted from Granger et al., 2009:p.40) 79
Figure 16: The concern of mainstream language teaching (Mark, 1998, adapted from Granger, 2002:6) 89
Figure 17: The focus on learner output (Mark, 1998, adapted from Granger, 2002:7) 90
Figure 18: The relationship between corpus linguistics and language teaching (adapted from Römer, 2008: 113). 91
Figure 19: Applications of corpora in language teaching (adapted from Römer, 2008: 113) 93
Figure 20: An extract from touchstone student book 3 (McCarthy, McCarten and Sandiford, 2006a, cited in McCarten, 2010:421) 95

Introduction

This work includes the descriptive, 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 corpora in terms of foreign language teaching.

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 2 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

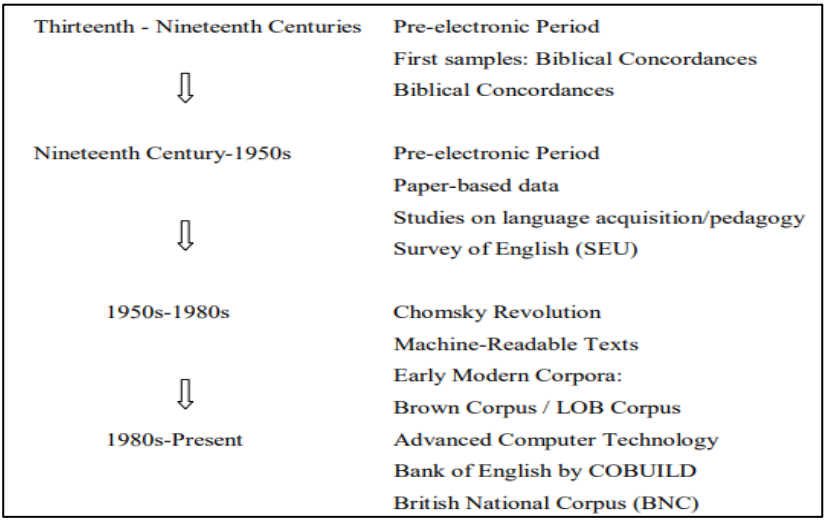
empirical, based on what is observed in corpus. McEnery &Wilson (2001) claims that an empirical research can be carried out by using any written or spoken text and such individual texts from the basis of many kinds of linguistics analysis. Therefore, it can be accepted that a corpus-based study may serve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

In respect of language teaching, 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.

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.:

Figure 1: 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 14 were the signs of the revival of corpus linguistics. Starvik (2007:12) 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”.

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: 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 15 of books, essays and poems (Meyer, 2002). For instance, to show the widespread usage, Jespersen illustrates indefinite pronoun from extensive lists as:

Example 1.

God send **euery one their harts** desire. (Shakespeare, Much Do About Nothing, III, 4.60, 1623)

(adapted from Meyer, 2008:5)

Example 2.

Each had **their** favorite. (Jane Austen, Mansfield Park, 1814)

(adapted from Meyer, 2008,: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:

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, 1992:7)*

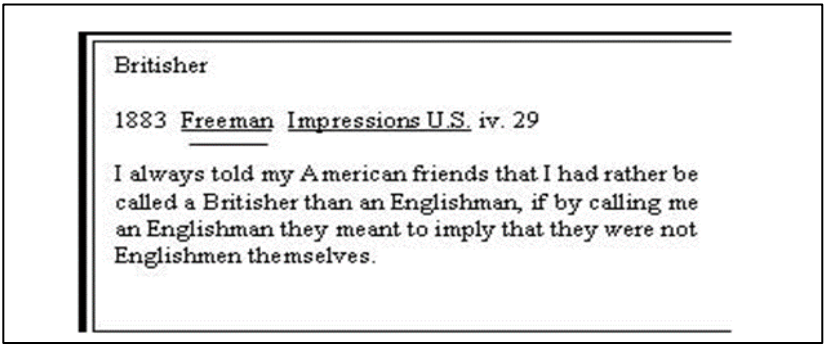
This old manual data formation of corpus-based study has a long honorable tradition in linguistics. Jespersen's methodology improved by Charles Carpenter Fries (1957, in Meyer, 2008) who provides grammatical descriptions in 'The Structure of English':

> *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 1951:3, cited in Starvik, 2007:13)*

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:

Figure 2: A citation slip from the OED (adapted from Meyer, 2008:9)



Dictionary of English and OED are the first dictionaries based upon preelectronic corpora in eighteenth and twentieth centuries. However, the first most significant and influential pre-electronic corpus was the Survey of English Usage (SEU) 17 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, 1996:2). 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.

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 speakerhearers' 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 19 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

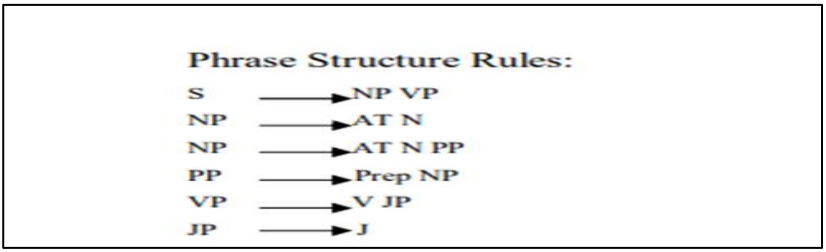
the explanatory adequacy is arguable whether it can be achievable through corpus analysis. In generative grammar, the highest level of adequacy is explanatory adequacy, which is achieved when the description or a theory reaches both descriptive adequacy and abstract principles of Universal Grammar (UG). 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'' (1992:108). 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:

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



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 3

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: 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, 1996:8)*

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, 1962:29, cited in McEnery & Wilson, 2001:11)

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 (2001:12-13) concludes that language is non-finite and corpus is finite, the problem is real and intuition must be considered.

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

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.

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, 1991:312, adapted from McEnery & Wilson, 1996: 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.

1.4 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 Kuc̆era, 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, 1979:117, cited in Johansson, 2008: 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, 1991:7). 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: 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 28 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, 2008:37-38)*

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 Lancaster-Oslo/Bergen 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, 2008:38). 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 International Computer Archive of Modern English (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, 2008:38)*

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, 2008:39)*

The London-Lund Corpus was the most important source for the computerbased 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, 2008:41). 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 31 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." (1982:4). 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.

1.5 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 32 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 34 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).

1.6 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 35 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 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. 36 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., 1998:4)

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.

2.1 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:.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-corpusbased approaches (McEnery, Xiao and Tonio , 2006).

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" (TogniniBonelli, 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 preset 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 (TogniniBonelli 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 corpusdriven 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, 2006:244)

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).

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, 2004:118). 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, 1992:116)*

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:.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).

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.

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:

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

```
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 !_!
```

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).

3.1.2 Lemmatisation

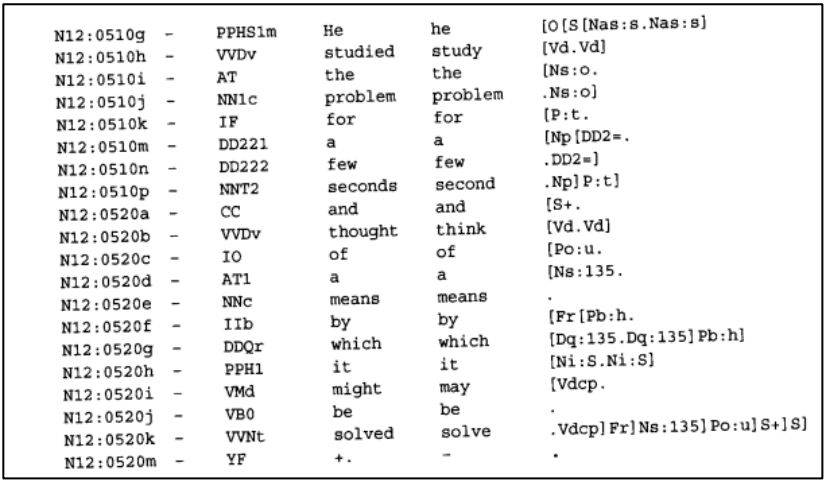
Lemmatisation is another form of automatic annotation that is closely allied to the identification of parts-of- speech and involves the reduction of the words in a corpus to their respective lexemes. Lemmatisation allows the researcher to extract and examine all the variants of a particular lexeme without 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:

Figure 5: Example of lemmatization from SUSANNE corpus (adapted from McEnery and Wilson, 1996: 54)



```
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  -  IIb     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      +.        -       .
```

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

N12:0510h VVDv studied study

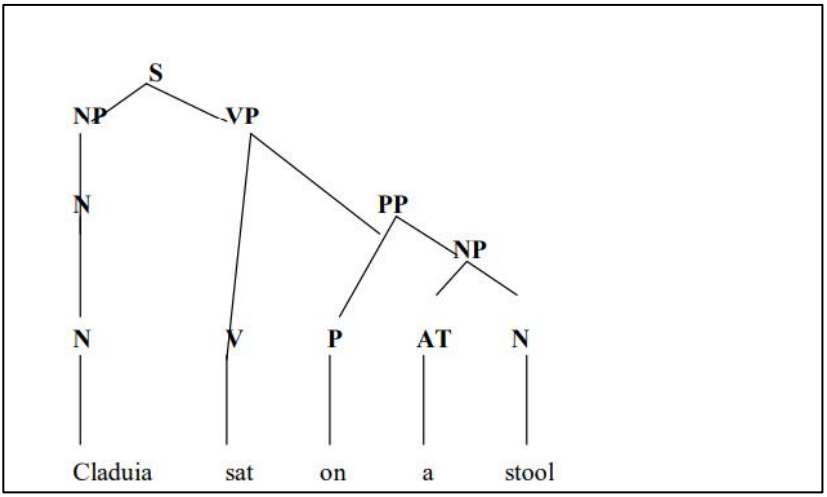
Corpus file POS tag code actual word head word (lemma)

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.

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 Cladia 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: 43).



The parsed sentence above can be displayed like this:

[S [NP Cladia_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:

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

```
[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_IN
[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]
```

3.1.4 Semantic Annotation

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

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

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

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).

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.:

Figure 9: An example of prosodic annotation from Londonlund corpus (adapted from McEnery and Wilson, 1996: 55).

```
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#* - - -                                 /
```

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: 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).

4 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, 2010:122) 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 49 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.

4.1 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):

Figure 10: A sample concordance of 'witnessed' (adapted from Baker, Hardie and McEnery, 2006: 43).

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	**the** 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.

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).

4.2 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 nonstandard 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 51 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 2010:124)

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: 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).

4.3 Wordlists

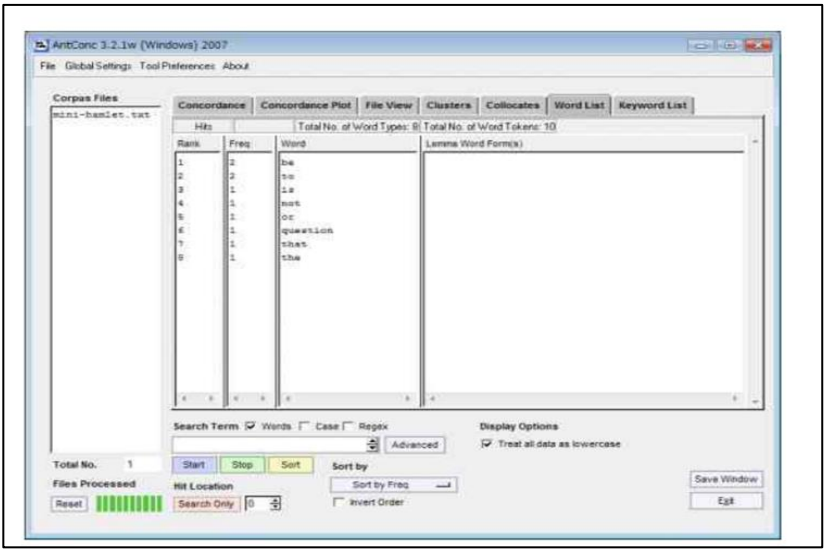
A word list is a list, generally arranged either alphabetically or in frequency order, of all the words in a given corpus with information about the number of times that word occurs in the corpus. The simplest word lists interpret ‘word’ as simply a string of letters; so, for example, the number of occurrences 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:

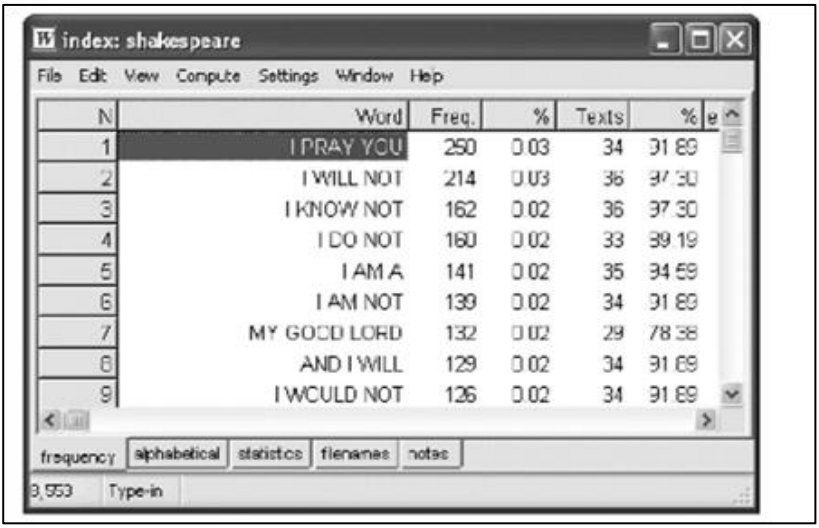
Figure 11: A screenshot of a word list from antconc software (adapted from http://wmtang.org/corpus-linguistics/).



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:

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

4.4 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.ht m)

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: 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 post- 57 modified noun phrases (e.g. the end of the Cold War) and the latter because of its multifunctionality – 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, 2010:128).

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, 2010:3).

5 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.

5.1 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).

5.2 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, 2010:114). 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).

5.3 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.

5.4 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., 2006:153). 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 60 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, *diachronic 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).

5.5 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 61 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).

5.6 Multilingual Corpora

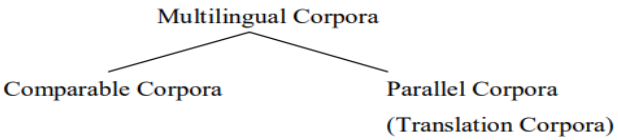
A parallel corpus consists of two or more corpora that have been sampled in the same way from different languages. The prototypical parallel corpus consists of the same documents in a number of languages, which is a set of texts and their translations. Since official documents (technical manuals, government 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, 2010:119). 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 62 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, 2008:276)



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.*
>
> *(2008:276)*

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).

5.7 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).

5.8 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 (2008:337) 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.*
>
> *(2008:337)*

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:123) 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.".

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) (2002)** 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) 65 consists of 1.1 million words of transcripts from speech by speakers from eleven different L1 backgrounds.

5.8.1 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 writings 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. 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.**

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 110 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: 3).

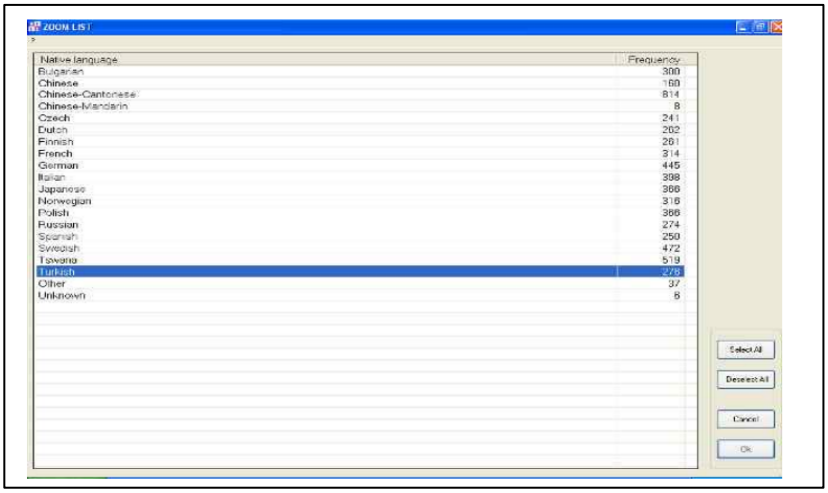
ICLE consists of two main variable groups as task and learner variables. Task variables include six components as medium, genre, field, length, topic, task setting. On the other hand, learner variables contains eight major criteria as age,

gender, mother tongue, region, other foreign languages, stay in Englishspeaking country, leaning context and proficiency level.

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.

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 14, the screenshot of ICLEv2 software which displays the all native language types and their frequency list is presented:

Figure 14: 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: 9).

5.8.2 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 nonnative (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. (2009:40)

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;

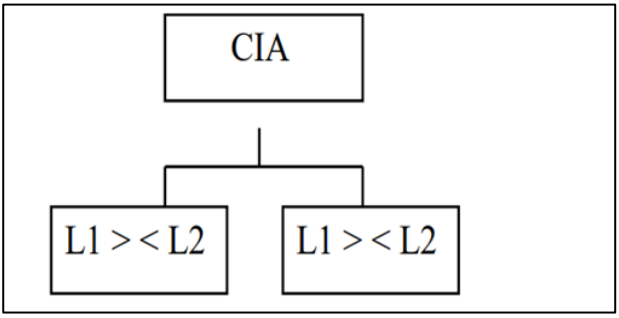
- the two varieties display major differences in frequency patterns, with some linguistics entities being overused and other underused.

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

5.8.3 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 15 illustrates CIA:

Figure 15: 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:

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).

5.8.4 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 69 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).

5.8.5 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- 70 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 some works in learner corpora studied by CEA 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 72 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, 2001:366).

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).

6 Corpus-based Research in Language Learning and Teaching

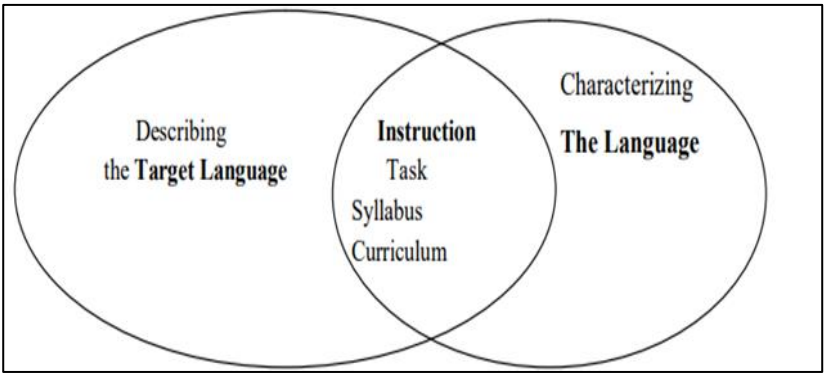
Over the past two decades, corpora and corpus evidence have not only been used in linguistic research but also in the teaching and learning of languages. Applications based on corpus investigation are found in a number of different areas, for example lexicography, translation, stylistics, grammar, gender 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 corpusbased 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, 2008:112). 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: 319)*

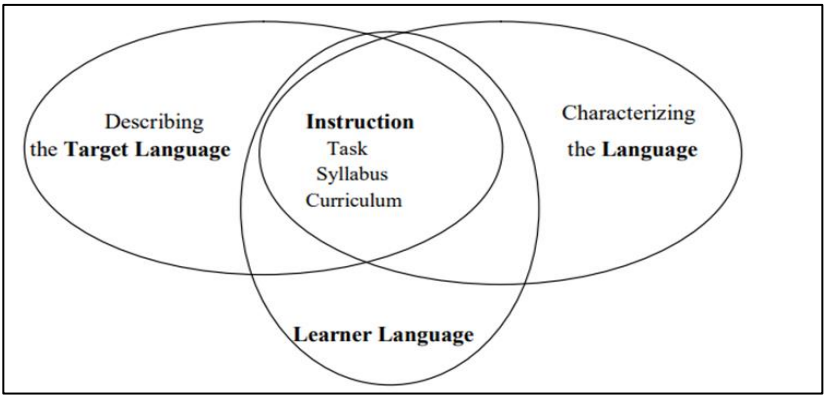
In respect of Second Language Acquisition (SLA) and Foreign Language Teaching (FLT), Granger (2002) states that previous mainstream language teaching approaches great efforts have been made to improve the description of the target language. There has been an increased interest in learner variables, such 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 16:

Figure 16: The concern of mainstream language teaching (Mark, 1998, adapted from Granger, 2002:6)



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, 2002:6). 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 17 presents the focus on learner output (Mark, 1998):

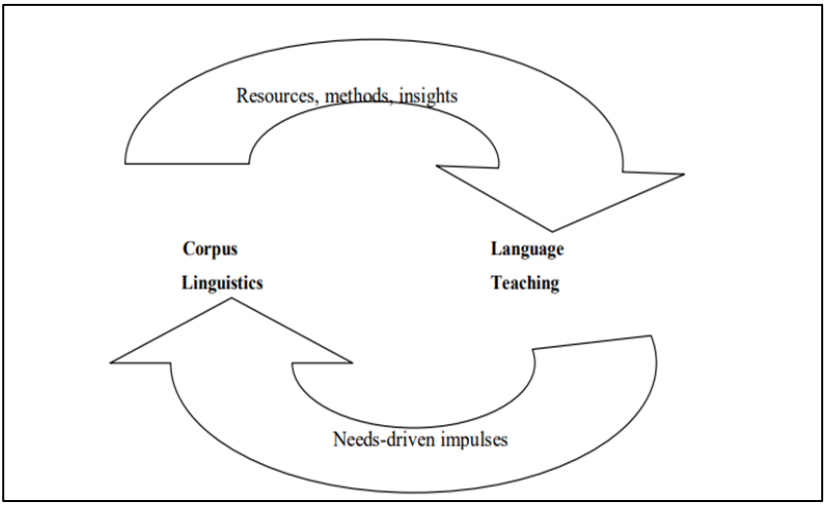
Figure 17: The focus on learner output (Mark, 1998, adapted from Granger, 2002:7)



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:).

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 18 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 18: The relationship between corpus linguistics and language teaching (adapted from Römer, 2008: 113).



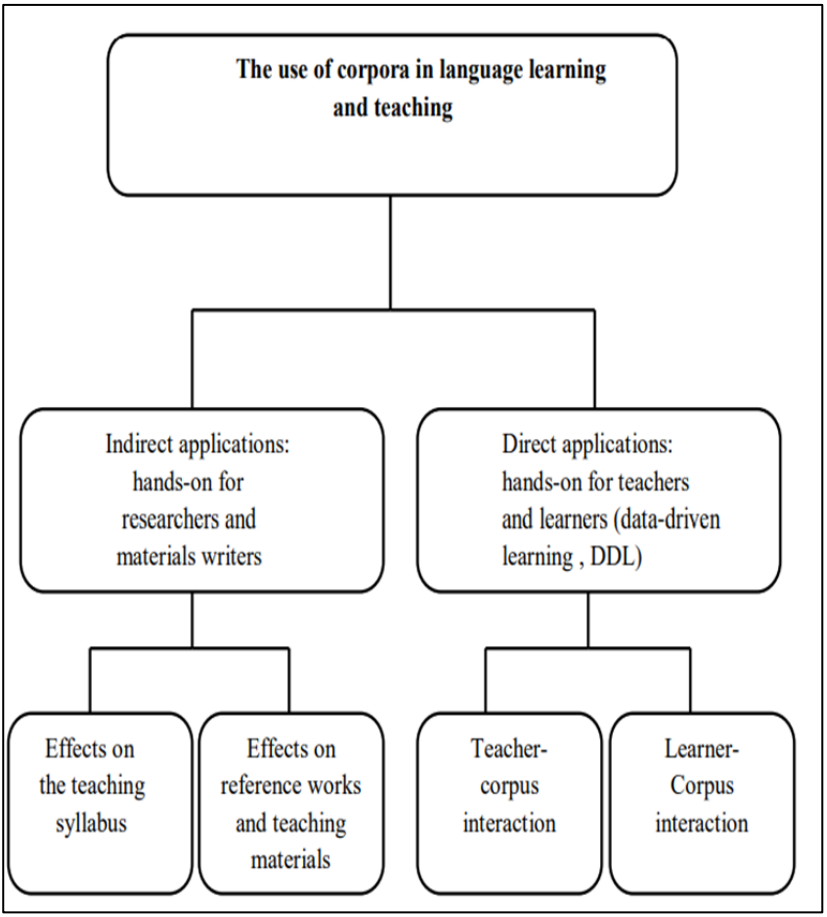
The relationship between corpus linguistic and language teaching is illustrated by Römer (2008) as in Figure 19 in the next section. The requirements of language teaching have a pact on research projects in corpus linguistics and on the development of suitable resources and tools (Römer, 2008).

6.1 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). Resources, methods, insights Needs-driven impulses 76 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 19:

Figure 19: Applications of corpora in language teaching (adapted from Römer, 2008: 113)



As presented in Figure 18, 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.

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 Teachercorpus interaction LearnerCorpus interaction 77 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: 421) 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". 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 4:

Example 4

(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.

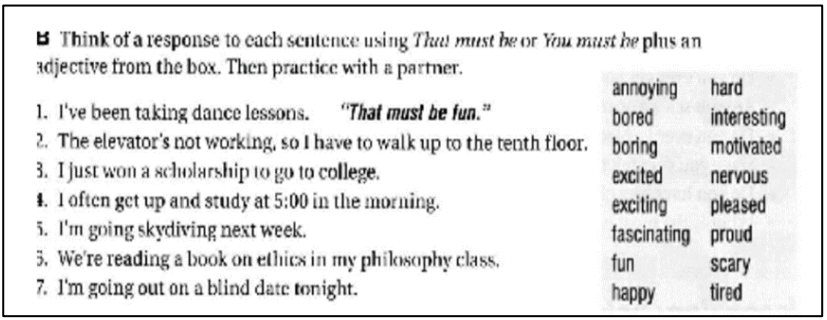
Speaker 1: It's not too bad. Speaker

2: Yeah. It must be tough though. Moving back and forth a lot.

(Cambridge International Corpus, North American English conversation) (adapted from McCarten, 2010:421)

This real life example that contains modal structure can be practiced in examples such as in Figure 19.:

Figure 20: An extract from touchstone student book 3 (McCarthy, McCarten and Sandiford, 2006a, cited in McCarten, 2010:421)



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

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.

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 79 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).

6.2 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 5.:

Example 5

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: 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 81 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: 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.

Summary

In this book, an historical overview of corpus linguistics, related literature of corpus linguistics and its place in language teaching are explained in general terms.

As an introduction, general and specific literature related to corpus linguistics are focused in order to be comprehensive in respect of corpus-based research. Firstly, the background of development of corpus linguistics diversified with a timeline. Three main periodsare explained in detail as; 13th century to 1950s, early modern corpora period and contemporary corpora periods. Then main points computerized corpora and types of corpora are emphasized to provide a comprehensive understanding on the background of fundamentals of corpora. The last sections scrutinize the place and the use of corpora in language teaching by focusing on learner corpora research.

The book covers the essentials of corpus linguistics as a field and its perspectives and utlizations with manners in language teaching with a comprehensive look.

Declarations of interest

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Aarts, J. (1991). Intuition-based and observation-based grammar In K. Aijmer & B.

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 (Eds.) English Corpus Linguistics (pp. 44-62). London: Longman.

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.

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.

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=CMyPTnDm4sC&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 (1995). Dimensions of register variation: a crosslinguistic comparison. 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. and Reppen, R. 2002: What does frequency have to do with grammar teaching? Studies in Second Language A cquisition, 24, 199–208.

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.

Bikeliene, L. (2008). Resultive connectors in advanced Lithuanian learner's English writing. Kalbotyra, 59(3), 30-37.

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. & Scott, M. (Eds.) (2010). Keyness in texts. Amsterdam: Benjamins

Bondi, M. (2010). Perspectives on keywords and keyness: An introduction. In M. Bondi & M. Scott (Eds.), Keyness in texts (pp. 1-18). 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.

Bradley, P.T. & Mackenzie, I. (2004). Spanish: An essential grammar. London: Routledge.

Breivik, L. (1999). On the pragmatic function of relative clauses and locative expressions in existential sentences in the LOB corpus. In. H. Hasselgård & S. Oksefjell (Eds.), Out of corpora, (pp. 121-135). Amsterdam: Rodopi.

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. (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.

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.

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.

Göksel, A. & Kerslake, C. (2005). Turkish, A comprehensive grammar. London: Routledge.

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. & 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. (1993b) The International Corpus of Learner English. The European English Messenger, 2(1), 34. 178

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. (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., 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.

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.

Kornfilt, J. (1997). Turkish. London and New York: Routledge

Lee D. Y. W. (2005). Modelling variation in spoken and written language: the multidimensional 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.

M. Sarıca, N. Sarıca ve A. Karaca (Ed.), XXII. Ulusal Dilbilim Kurultayı Bildirileri,1- 11, Ankara: Yuzuncu Yıl Universitesi Yayınları.

MacWhinney, B. (1996). The CHILDES System. American Journal of SpeechLanguage Pathology, 5, 5–14.

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.

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.

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., & A. Wilson (1996). Corpus linguistics (1st ed.). Edinburg: Edinburg University Press.

McEnery, T., & A. Wilson (2001). Corpus linguistics (2nd ed.). Edinburg: Edinburg University Press.

McEnery, T., Xiao, R., & Tonio, Y. (2006). Corpus-based language studies: An advance resource book. New York: Routledge.

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.

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.

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.

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.

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.

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.

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. (2003). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Lancaster University

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.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A crosslinguistic 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.

Römer, U. (2008). Corpora and language teaching. In A. Lüdeling & M. Kytö (Eds.), Handbook on Corpus Linguistics, (pp. 112-131), Mouton de Gruyter.

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. & Tribble, C. (Eds.) (2006). Textual patterns: Key words and corpus analysis in language education. Amsterdam & Phladelphia: Benjamins.

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.

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.

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. EAGTCWG-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.

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.

Stenström, A.-B. (1987). Carry-on signals in English conversation. In W. Mejis (Ed.) Corpus linguistic and beyond (pp. 87-119). Amsterdam: Rodopi.

Stubbs, M. (1996). Text and corpus analysis. Oxford: Blackwell.

Tang, W.M. (2012). A Very Short Introduction to Corpus Linguistics. Retreived from http://wmtang.org/corpus-linguistics/.

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.

Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (Eds.) (1997). Teaching and Language Corpora. London: Longman.

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.

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