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

RECURRENT PHRASES IN LEARNER ENGLISH: A CORPUS DRIVEN APPROACH

Aysel ŞAHİN KIZIL

A PhD DISSERTATION

ADANA, 2013

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

RECURRENT PHRASES IN LEARNER ENGLISH: A CORPUS DRIVEN APPROACH

Aysel ŞAHİN KIZIL

Supervisor: Asst. Prof. Dr. Abdurrahman KİLİMCİ

A PhD DISSERTATION

ADANA, 2013

To Directorate of the Institute of Social Sciences of Çukurova University

We certify that this dissertation is satisfactory for the award of degree of Doctor of Philosophy in the subject of English Language Teaching.

Chairperson: Asst. Prof. Dr. Abdurrahman KİLİMCİ
Supervisor

Member of Examining Committee: Assoc. Prof. Dr. Yasemin KIRKGÖZ

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

Member of Examining Committee: Asst. Prof. Dr. Adnan BİÇER

Member of Examining Committee: Asst. Prof. Dr. Fahritdin ŞANAL

I certify that this dissertation conforms to the formal standards of the Institute of Social Sciences. ……/……/2013

Prof. Dr. Yıldırım Beyazıt ÖNAL
Director of the Institute

PS. The uncited usage of the reports, charts, figures, and photographs in this dissertation, whether original or quotes for mother sources, is subject to the Law of Works of Art and Thought No: 5846.
Not: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge, ekil ve fotoğrafların kaynak gösterilmeden kullanımı 5846 sayılı Fikir ve Sanat Eserleri Kanunu'ndaki hükümlere tabidir.

ÖZET

ÖĞRENİCİ İNGİLİZCESİNDE TEKRARLANABİLİR ÖBEKLER: DERLEM TEMELLİ YAKLAŞIM

Aysel ŞAHİN KIZIL

Doktora Tezi, İngiliz Dili Eğitim Anabilim Dalı
Danışman: Asst. Prof. Dr. Abdurrahman KİLİMCİ
Kasım, 2013, 215 sayfa

Derlem dilbilim alanında kaydedilen gelişmeler dilin tekrarlanan doğasının, diğer bir deyişle, rutin olarak kullanılagelen yapıların anlaşılmasına büyük ölçüde katkıda bulunmuştur. Ana dil üzerine yapılan çalışmalar ikinci dil edinimi alanından araştırmacılara esin kaynağı olmuş ve son zamanlarda, tekrarlanabilir öbekler ikinci dil edinimi alanında giderek daha da fazla ilgi çekmeye başlamıştır. Karşılaştırmalı Aradil Analizi (CIA) tekniğinin gelişimi de aradilde daha önce keşfedilmeyen bir takım özellikleri ortaya çıkarmıştır.

Karşılaştırmalı Aradil Analizi çerçevesinde (CIA), derlem temelli tekrarlanabilir kelime öbekleri yöntemi kullanılan bu çalışmada, Türk öğrenicilerin yazılı ve sözlü İngilizcede kullandıkları iki, üç, dört, beş ve altı kelimeden oluşan öbekler araştırılmıştır. Araştırmanın temel hedefleri, yabancı dil olarak İngilizce öğrenen Türk öğrenicilerin aradilde kullandıkları tekrarlanabilir öbeklerin ortaya çıkarılması ve bu öbeklerin anadili İngilizce olan kullanıcıların yazılı ve sözlü İngilizcede kullandığı öbeklerle nitel ve nicel bakımdan karşılaştırmasını yapmaktır.

Çalışmanın veri kaynağını Uluslararası Öğrenici İngilizcesi Derlemi (ICLE) ve Louvain Uluslararası Aradil Konuşma İngilizcesi Veritabanı (LINDSEI) oluşturmaktadır. Türk öğrenicilerin aradil özelliklerini araştırmak için, bu derlemlerin, Türk öğrenicilerden toplanan verilerle oluşturulan alt derlemleri (TICLE ve LINDSEI-TR) kullanılmıştır.

Elde edilen bulgular, Türk öğrenicilerin aradilde kullandıkları tekrarlanabilir kelime öbekleri ile ilgili bazı ortak özellikler ortaya koymuştur. Öte yandan, Türk öğrenici İngilizcesinin anadili İngilizce olan kişilerin İngilizcesiyle karşılaştırılması, yazılı ve sözlü İngilizcede sıklıkla kullanılan kelime öbekleri ile ilgili olarak hem benzerliklerin hem de farklılıkların olduğu karmaşık bir tablo ortaya koymuştur. Ulaşılan sonuçlar, birinci dilin aradil üzerindeki etkisi, eğitime dayalı unsurlar ve kesit girişimi (register interference) kavramları kapsamında tartışılmıştır.

Anahtar Kelimeler: Bilgisayarlı Öğrenici Derlemi (CLC), Tekrarlanabilir Öbekler, Derlem Dilbilim, Derlem Temelli Analiz, Yazılı Öğrenici Derlemi, Sözlü Öğrenici Derlemi, Öğrenici Derlemi Araştırması, Karşılaştırmalı Aradil Analizi.

ABSTRACT

RECURRENT PHRASES IN LEARNER ENGLISH: A CORPUS DRIVEN APPROACH

Aysel ŞAHİN KIZIL

PhD Dissertation, English Language Teaching Department
Supervisor: Asst. Prof. Dr. Abdurrahman KİLİMCİ
November, 2013, 215 pages

Insights from corpus linguistics have contributed considerably to the understanding of recurrent nature of language, that is, language use, to a great or lesser extent is marked by routine or recurrence. Studies designed on native language have inspired the researchers from the field of second language acquisition, and investigations of recurrent word combinations, the way words co-occur with other words, have recently gained more and more attention in the field of second language acquisition. Advent of Contrastive Interlanguage Analysis (CIA) has highlighted unprecedented features that characterize interlanguage.

Within the framework of CIA, adopting a corpus-driven recurrent word combination method, the present study has focused on the 2-to 6- word combinations in both spoken and written interlanguage to investigate Turkish learners' tendencies in designing their discourse in English. The main objectives are to explore the use of recurrent phrases in Turkish EFL learners' interlanguage and to compare and contrast them with native speakers' use of recurrent phrases across written and spoken language in terms of both quantitative and qualitative variation.

The primary source of material for this study is two major corpora: International Corpus of Learner English (ICLE) and Louvain International Database of Spoken English Interlanguage (LINDSEI). For non-native data, Turkish components of these corpora have been used.

The overall findings have revealed some common aspects of Turkish learners' interlanguage in terms of recurrent word combinations. Comparisons with native speaker language have painted a complex picture of similarities and differences regarding frequently used word combinations both in spoken and written language. The results are discussed referring to the possible effects of first language, instructional factors and register interference.

Keywords: Computer Learner Corpus (CLC), Recurrent Phrases, Corpus Linguistics, Corpus-Driven Analysis, Spoken Language Corpora, Written Language Corpora, Learner Corpus Research, Contrastive Interlanguage Analysis (CIA)

ACKNOWLEDGEMENTS

The completion of this dissertation marks an important milestone in my professional career, and a number of people have a special part to play in bringing this about. It is a pleasure to thank those people who made this dissertation possible.

First, I would like to thank my supervisor, Asst. Prof. Dr. Abdurrahman KİLİMCİ, who has all along been an incredibly supportive mentor. He is genuine and kind, and he made valuable insights to my research as an accomplished linguist. It is his congenial guidance that inspired me and sharpened my focus. I am truly indebted to him.

I also wish to express my gratitude to the members of dissertation committee: Assoc. Prof. Dr. Yasemin KIRKGÖZ, Assoc. Prof. Dr. Ahmet DOĞANAY, Asst. Prof. Dr. Adnan BİÇER and Asst. Prof. Dr. Fahritdin ŞANAL for the time they have dedicated to the reading of my thesis and all the valuable comments they have made.

I would like to thank Elif BİNBOĞA, a great friend, for her being available at times when I need relieving and stimulating comments for ten years. She is more than a friend for me. I thank her being on the other end of the phone for the last four years. Even when I have not contacted with her for weeks, I have always felt her support.

My heartfelt thanks go to my family for their invaluable support and encouragement. I am grateful to my mother who not only raised me, taught me, and loved me but also believed in me and supported every step I took in my life. My deepest appreciation is due to my father, the first man I loved, Mustafa ŞAHİN, who has always been the source of my motivation and inspiration. Although he is not physically with us anymore, I believe he is watching me up there and still feeling proud of me. This dissertation is written in his memory.

Last but certainly not the least, I would like to thank the most precious person in my life- my little daughter- Neva, for her being happiest side of my life. I apologise her for spending too much time in front of a computer during this work instead of being with her.

Aysel ŞAHİN KIZIL

Adana / 2013

TABLE OF CONTENTS

Pages

CHAPTER I
INTRODUCTION

CHAPTER II
LITERATURE REVIEW

CHAPTER III
METHODOLOGY

CHAPTER IV
FINDINGS AND DISCUSSION

CHAPTER V
CONCLUSION

LIST OF ABBREVIATIONS

BNC	**:** British National Corpus
CA	**:** Contrastive Analysis
CEA	**:** Computer-aided Error Analysis
CIA	**:** Contrastive Interlanguage Analysis
CL	**:** Corpus Linguistics
EFL	**:** English as a Foreign Language
ELT	**:** English Language Teaching
ESL	**:** English as a Second Language
ICE	**:** International Corpus of English
ICLE	**:** International Corpus of Learner English
L1	**:** First (native) Language
L2	**:** Second Language
LINDSEI	**:** Louvain International Database of Spoken English Interlanguage
LINDSEI-TR	**:** Louvain International Database of Spoken English Interlanguage, Turkish Component
LL	**:** Log-likelihood
LOCNEC	**:** The Louvain Corpus of Native English Conversation
LOCNESS	**:** Louvain Corpus of Native English Essays
NNS	**:** Non-Native Speakers
NS	**:**Native Speakers
SLA	**:**Second Language Acquisition
TICLE	**:**Turkish International Corpus of Learner English
TL	**:** Turkish Learner

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF APPENDICES

Pages

CHAPTER I

INTRODUCTION

1.0. Introduction

Defined, in its general sense, as the study of language based on the examples of actual language use, corpus linguistics has its origins in Firthian tradition of 50s which emphasizes that language research should consider the context in which language is used, as well as its frequent joining patterns. In Firth's words "we must take our facts from speech sequences, verbally complete in themselves and operating in contexts of situations which are typical, recurrent and repeatedly observable" (cited in Diniz, 2007p.35). Firth's view of language has been influential in the theoretical underpinning in which corpus linguistics is framed today. Working within the framework of an approach suggested by Firth, researchers such as Sinclair, Hoey, Halliday- often dubbed neo-Firthians- have contributed to a great extent both to the scope and focus of research in corpus linguistics and to the compilation of corpora and the use of corpus based methodologies (McEnery& Gabrielatos, 2006).

Together with the growing number of corpus studies in many fields, corpus linguistics has currently manifested itself as a fruitful field of study. Although there is an on-going debate on whether it is a branch of linguistics or a methodology, it has been argued that corpus linguistics is "not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject" (Leech, 1992).

Corpus linguistics is an empirical approach in that it examines and draws conclusions from samples of actual language use. It has also a theoretical status in that examinations of language facts "lead to the formulation of hypotheses and generalizations, which are then unified in a theoretical statement" (Gabrielatos & McEnery, 2006 p.2). Another main characteristic of modern corpus linguistics is the use of computers which have enabled researchers to store, access and analyse large amounts of language data. Through specially designed software, computers facilitate quantitative (e.g. word frequency) as well as qualitative (e.g. patterns of use) analyses of language (O’Keeffe, McCarthy, & Carter, 2007).

Foregoing features of corpus linguistics have been acknowledged in many areas. As stated by Gilquin, Granger and Paquot, (2007), corpus linguistics and corpus based research has played "a key role in most language-related fields from lexicography to language teaching through natural language processing and literary criticism" (p. 320). In the same vein, Meyer (2004) regards corpora as valuable resources for descriptive, theoretical and applied discussions of language.

The practical and theoretical potential of computer assisted corpus analysis has recently been recognized in the field of Second Language Acquisition (SLA) as well. Given the fact that studies of language acquisition have always had an empirical basis, corpus based research has been found to be particularly useful to objectively investigate interlanguage, a term coined by Selinker (1972) to refer to a linguistic system based on the observable output that results from the learner's production of a target language form. With the purpose of researching interlanguage through usage-based descriptive and quantitative as well as qualitative analyses, a number of researchers have begun developing what are called *learner corpora* which refer to electronic collections of speech or writing of foreign or second language learners in a variety of language settings.

Although learner corpus compilation is a relatively recent activity, a number of projects have already been started (even some were completed) since 1990s. Most of the learner corpora compiled so far contain data about learner English, usually based on written language and cover the learners of one L1 group (Nesselhauf, 2004). Some of the biggest learner corpora to date are Hong Kong University of Science and Technology (HKUST) Learner Corpus containing 25 million words, TeleNex Student Corpus containing 3 million words of written language, the Chinese Learner English Corpus (CLEC), containing 1.2 million words and the Uppsala Student English Project (USE) with about 1 million words of different types of essays. Among a number of learner corpora, the International Corpus of Learner English (ICLE) is especially noteworthy since, unlike the most of existing learner corpora focusing on one L1 group, ICLE contains data from learners with different L1s. It currently consists of 4, 5 million words of argumentative essays written by university students of English with 16 different L1 backgrounds. One strength of ICLE stems from the fact that it has a reference corpus, the Louvain Corpus of Native English Essays (LOCNESS), compiled from the native speakers (NSs) under the same task conditions, which make it an efficient comparable base for learner English. For spoken learner corpora, the number of projects is fairly few due to time-consuming and labour-intensive transcribing process required in the creation of spoken corpus (Aijmer, 2004; Behrens, 2008; Huang, 2011). Currently, the biggest one is the Louvain International Database of Spoken English Interlanguage (LINDSEI), which contains interviews with advanced learners of various L1s. To date, it covers 11 different mother tongue backgrounds. Like ICLE, LINDSEI has also a comparable corpus of NSs, The Louvain Corpus of Native English Conversation (LOCNEC).

The creation of learner corpora has led to a number of studies in the area of *Contrastive Interlanguage Analysis* (CIA) (Granger, 1998a), which involves comparing learner data with native speaker (NS) data or making comparisons among various interlanguages that individuals from different first language backgrounds develop. A wide range of research carried out within this framework has added considerably to "our hitherto somewhat patchy knowledge of the different stages of interlanguage development" (Gilquin et al., 2007 p. 322).

Research topics focusing on learner corpora based on CIA are highly variable ranging from high frequency vocabulary to recurrent word combinations. For example, Altenberg and Granger (2002) investigate the lexical and grammatical patterning of high frequency verbs using the French and Swedish components of ICLE. The results of the study point to remarkable differences between native and non-native students. The researchers propose overgeneralization of the main English pattern and influence of L1 as the main reasons for the difference. Aijmer (2002) focused on the range and frequency of some modal words in native English writing and English L2 writing of advanced students. Her study covered Swedish, German and French L1 groups. Based on the relevant components of ICLE, the study reveals a global overuse of modal auxiliaries by all the L2 writers. Aijmer (2002) explains the reasons of overuse referring to developmental factors, L1 effect and register interference. Likewise, H. Chen (2010) studies modality in L2 writing by examining how epistemic modality is used by non-native speaker (NNS) writers and native speaker (NS) writers using CLEC for the learner data. The overall results show a notable difference in the total frequency of the epistemic modality. Learner corpora have also been exploited in the studies of collocations. Lombard (1997) analyses written English of Mandarin L1 speakers using a corpus of 78,000 words. She finds that collocations produced by NNSs are significantly different from those of NSs. Based on the argumentative essays written by German and

Austrian learners of English, Nesselhauf (2005) investigated collocations in learner English. She finds evidence for extensive erroneous use of collocations by learners.

Regarding the studies on Turkish learners, research interests cover the topics of lexical richness, stance adverbials and use of connectors, linking adverbials among others. Kilimci (2001), for example, investigated the lexical profile of EFL learners through corpus query techniques. Written corpus gathered from both NS and NNS students made the data source. The results revealed that written language of Turkish learners is remarkably different from NS writing in terms of lexical variety. In a similar vein, Şanal (2007) analysed the Turkish learners’ lexical complexity and richness in their written English. Comparisons of NNS essays with NS writing indicated that learner writing is less complex in lexical diversity and density than NS writing. Kilimci (2008) explored how appositive linking adverbials function in the construction of argument both syntactically and semantically in the argumentative writings by nonnative speakers (NNS: Turkish, German, French) and native speakers (NS: LOCNESS - Louvain Corpus of Native English Essays). Distinctive interlanguage features of the Turkish learners of English and features shared by all or several learner groups were highlighted. In a recent study, Can (2012) examined the usage of stance adverbials by Turkish learners through the comparison of NNS and NS written language. The findings showed that Turkish learners rely on limited number of stance adverbials. Can (2012) points to instructional factors as a possible reason of limited use of stance adverbials.

All these studies together with many others focusing on various aspects of learner English (Ädel & Römer, 2012; Aertselaer, 2008; Axelsson & Hahn, 2001; De Cock, Granger, Leech, & McEnery, 1998; Dutra, 2004; Ebeling, 2011; Granger & Tyson, 1996) occupy a crucial place in defining interlanguage. However, a closer look at the relevant literature reveals that the studies are mostly based on written corpora, which pose restrictions in painting a complete picture of learner language. Therefore, the literature on corpus based interlanguage analysis implies the need for studying spoken and written performance of learners conjointly to have a better understanding of interlanguage (Hunston, 2002; Mukherjee, 2009).

1.1. Background of the Study

Corpus studies of learner language so far, especially those designed within the framework of Contrastive Analysis (Corder, 1981) have outlined new directions in SLA research. The findings of the studies on NSs’ performance have revealed that naturally produced language is dominated with recurrent word combinations. This fact gave rise to the question of phraseology in language and contributed to the establishment of phraseology as a field in its own right. Forming the theoretical background of the present study as well, phraseology, in its most general sense, refers to the "study of the structure, meaning and the use of word combinations" (Cowie, 1998 p.12). Together with the increasing number of researchers in this field, this definition has evolved to a great extent as word combinations could be in many different shapes and forms. Considering the scope of phraseology, Gries (2008) has recently defined it "as the study of co-occurrence of a lexical item and one or more additional linguistic elements whose frequency of co-occurrence is larger than expected on the basis of chance" (p.6).

Depending on the fact that the phrase is the basic level of language representation where "form and meaning meet with greatest reliability" (Ellis, 2008 p. 6), it is now widely acknowledged that focusing on recurrent phrases could shed light on the nature of language itself. In other words, study of recurrent phrases provides us with a clear understanding of what is typical or as Béjoint (2000) states "of the tendencies in the encoding of text by native speakers" (p. 216). This is especially significant since these tendencies are the part of the mastery of the language as pointed out by Wray and Fitzpatrick (2008) who regard the knowledge of phraseology as a facilitating element for language fluency.

One of the earliest studies on phraseology is reported by Altenberg (1998) who analysed the phraseology of spoken English by focusing on recurrent word combinations from the London-Lund corpus. The most striking result of his study is that NSs have a large stock of word combinations that can be described as preferred ways of saying things. He states that "the use of routinized expressions is evident at all levels of linguistic organization and affects all kinds of structures, from entire utterances operating at discourse level to smaller units acting as single words and phrases" (p.56). In other words, prefabricated word combinations pervade lexical, grammatical and pragmatic levels of linguistic organization and NSs retrieve and productively assemble these expressions instead of generating every word sequence through the application of rules of the syntax.

These findings literally figure in the formulation of Sinclair's *idiom principle* that is one of the substantial principles of modern corpus linguistics (Gries, 2008). According to this principle, "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments (Sinclair, 1991, p.110), and it contrasts with *open choice principle* that states "at each point where a unit is completed (a word or a phrase or a clause), a large number of choices opens up and only restraint is the grammaticalness (ibid. p.109). In this relatively new description of language, syntax is assigned a secondary role and "is only brought into service occasionally as a kind of glue to cement the preconstructed phrases together (McCarthy, 2006).

Drawing from the literature on NSs’ language use, researchers in the field of SLA have started to look beyond the word in describing learner language and focus on recurrent word combinations or phrases (Granger, 1998b; Howarth, 1998; Wray & Perkins, 2000). Corpus based research contrasting learner language with NS performance has shown that recurrent multiword expressions that come so naturally to NS pose difficulty for non-native users (De Cock, 2004; Nesselhauf, 2005). Recurrent phrases are usually easy to understand but they hinder language production for the learners. NNSs construct their spontaneous speech by combining individual words. This results in producing unnaturally sounding language although it is grammatically correct. Kjellmer (1991, p. 124) summarizes this as "their building material is individual blocks rather than prefabricated sections".

The notion of recurrent phrases through corpus based methodology has been investigated under various terminologies such as lexical bundles (Biber, Conrad, & Cortes, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 2007; Hyland, 2008a), recurrent sequences (De Cock, 2004), recurrent word combinations (Altenberg, 1998), multi-word constructions (Liu, 2012) and lexical chunks (Ishikawa, 2009). De Cock's study (2003) is among the earliest attempts to investigate the recurrent phrases in learner language. Her primary focus is spoken English produced by French learners. Waibel (2007) conducted an investigation on phrasal verbs in written interlanguage produced by German and Italian learners, which provides insights for the recurrent phrases in learner language. Ping (2009) investigated lexical bundles in written English of Chinese students and compared the findings with NS writing. Likewise, Ishikawa

(2009) compared the high-frequent word combinations in English essays written by Japanese learners of English with those used by NSs. Focusing on Lithuanian learners of English, Juknevičienė (2009) investigated the word combinations in the essays written by students at three different proficiency levels. Chen and Baker (2010)report a study of recurrent phrases through a written corpus of learner language. More recently, Ädel and Erman (2012) carried out a study on recurrent word combinations in academic written English of Swedish learners in comparison with NSs' written performance. These and some other studies provided evidence to the claims that recurrent word sequences or more specifically, recurrent phrases are the source of difference between the NS and NNS language, and thus, should be considered a significant aspect of EFL, which needs further research focusing on learners from a variety of first language backgrounds. Additionally, a closer look at extant studies implies another gap in the literature of recurrent phrases in learner English: In much of the research of this type, the focus is on either NSs’ spoken performance or written corpora of language learners have been chosen as the medium. However, the systematic studies of phraseology of English as a Foreign Language (EFL) learners' English based on both spoken and written corpora are few in number (Adolphs & Durow, 2004; Wei, 2009). Regarding the Turkish learners of English that make up the L1 variation of the present study, no investigation on recurrent phrases has been reported at the time. Spoken English of Turkish learners, in particular, seems to be an untouched area of research, exploration of which potentially has a lot to offer to the literature of learner corpora.

Given the background sketched above, the present study that focuses on spoken and written interlanguage of Turkish learners of English within the framework of Contrastive Interlanguage Analysis (CIA) is planned to make a contribution to the gap related to having a holistic perspective considering both spoken and written corpora in interlanguage analysis.

1.2. Purpose of the Study

Using a written corpus of argumentative essays (TICLE[1]), spoken corpus of informal interviews by Turkish EFL learners (LINDSEI-TR[2]) and a parallel written

[1] TICLE was compiled as research project (Project no:EF2004BAP8) funded by the Comission of Scientific Research Projects, Çukurova University.

(LOCNESS) and spoken NS corpus (LOCNEC), this study is based on the quantitative and qualitative characterization of recurrent phrases. To this end, the study aims to;

- identify the recurrent phrases in both native speakers (NS) and Turkish EFL learners' (TL) written and spoken corpora
- compare and contrast Turkish learners' (TLs) and native speakers' (NSs) use of recurrent phrases across written and spoken corpora in terms of both quantitative and functional variation.

Achieving such objectives will make it possible to gain insights into the spoken and written performance of Turkish EFL learners and will form a base in defining interlanguage characteristics of Turkish learners with respect to their both writing and speaking skills. Comparison with the native speakers' writing and speech is thought to help identify the deviations, if any, from native norms.

1.3. Research Questions

Following questions have guided the present study:

1. What are the major recurrent sequences of two- to six-word combinations Turkish learners tend to use in their spoken discourse?
2. To what extent are these recurrent sequences in Turkish learners' speech similar to and/or different from those in native speaker speech?
3. What are the structural and functional features of recurrent sequences of two or more word combinations prevalent in spoken interlanguage of Turkish EFL learners?
4. What are the major recurrent sequences of two- to six-word combinations Turkish learners tend to use in their written interlanguage?
5. To what extent are these recurrent sequences in Turkish learners' written interlanguage similar to and/or different from those in native speakers' written language?

[2] LINDSEI-TR as a research project (Project no: EF2013BAP22) is funded by the Comission of Scientific Research Projects, Çukurova University. Data compilation and evaluation is still in progress.

6. What are the structural and functional features of recurrent sequences of two or more word combinations prevalent in written interlanguage of Turkish EFL learners?
7. To what extent do the learner and native speaker recurrent word combinations in the written corpora overlap with or differ from those in the spoken corpora in syntactic and functional terms?

1.4. Significance of the Study

Most of the learner corpus projects were launched in the last decade and the research drawing from them is relatively limited (Tono, 2000). However, there are notable efforts in terms of exploiting the potentials of learner corpora in both written and spoken medium (Aijmer, 2002a, 2004; Altenberg, 2002; De Cock, 2004; McEnery & Kifle, 2002; Wei, 2009) –though the spoken interlanguage related studies are fairly new. As for the Turkish context, the learner corpus research mainly has centred on written interlanguage (Kilimci, 2001; Şanal, 2007; Kilimci, 2008; Kilimci & Can, 2009; Can, 2009, 2012).

The significance of this study basically stems from being one of the first attempts to analyse both the spoken and the written performance of Turkish EFL learners holistically through relatively large spoken and written corpora of interlanguage. First, this study sets out to delineate the Turkish learners' written and spoken interlanguage in terms of recurrent phrases. This characterization of interlanguage may offer general insights into the development of lexical competence of Turkish EFL learners, which may lead to enhancement of EFL learning outcomes. The insights could point out directions for further research in the field with regard to development of writing and speaking skills and vocabulary. Second, the findings of this study could be useful in informing language teaching practices in the relevant context. Comparable data obtained from the other corpora of learner and native language may highlight the specific difficulties of Turkish learners and hence, development of a more focused and efficient teaching practice could be possible. Third, the findings could be exploited in developing learner corpus-informed syllabuses and materials particularly for teaching speaking skill to the students. Finally, this study is thought to be significant as the results may lead other researchers to carry out studies on learners of different L1 background, which could add on the interlanguage literature.

1.5. Key Terms

Computer Learner Corpora are electronic collections of spoken and written texts produced by foreign or second language learners assembled according to explicit design criteria (Granger, 2009).

Contrastive Interlanguage Analysis (CIA) is a method of language analysis involving quantitative and qualitative comparisons between native and non-native speakers data or between different non-native groups (Granger, 2009)

Corpus is a large collection of natural texts compiled according to a set of predefined criteria to represent, as far as possible, a language or a language variety as a source of data for linguistic research (Sinclair, 2004a).

Corpus Linguistics is the study of language through a whole system of methods and principles of how to apply corpora in language studies (McEnery, Xiao, & Tono, 2006)

English as a Foreign Language (EFL) is the language learned mostly in classroom setting without a considerable access to the speakers of the language being learned

Filled Pauses are silent pauses filled with vocal activity or noise i.e. non-verbal sounds produced during speech (Eriksson, 2012)

First Language (L1) is the language that is acquired in early childhood, mother tongue

Interlanguage (IL) is "the separate linguistic system based on observable output which results from a learner's attempted production of a target language norm" (Selinker, 1972 p.214)

Phrase: A longer unit of meaning in connected language comprising at least a few words in length (Stubbs, 2002)

Phraseology is "the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance" (Gries, 2008 p.6)

Recurrent Phrases refer to any continuous strings of words occurring more than once in identical form (Altenberg, 1998).

Second Language (L2) or Foreign Language (FL) is the language other than the native language and used interchangeably in this dissertation (R. Ellis, 2008)

1.6. Overview of the Thesis

The present dissertation consists of five chapters organized under the following titles: introduction, literature review, methodology, findings and discussions and conclusion.

The first chapter provides brief information about corpus linguistics, learner corpora and its place in second language acquisition (SLA) research. Main focus of the study is presented in the background of the study and then the purpose and significance of the study is explained. Finally, key terms are given and this section concludes with a summary of the chapter.

The second chapter of the dissertation presents a detailed review of the literature on which the study is based. It consists of five main sections. The first section provides an overview of corpus linguistics touching upon its historical evolution and exploitation of corpora in various fields of language study. The second section explains the connection between corpus linguistics and SLA. The third section introduces major learner corpora projects around world and presents detailed information about the corpora employed in this specific study. In the fourth section, basic approaches to learner corpora in linguistic analysis are presented and an extensive review of previous research on learner corpora is provided. In the fifth section, the field of phraseology as the theoretical framework of the study is introduced with an emphasis on its place in corpus linguistics and SLA. This section ends with the review of existing research on phraseology covering the studies on recurrent phrases in written and spoken corpora, which makes the foundation of the study.

The third chapter provides information about the methodology followed in the present study. The corpora under investigation (LINDSEI, LOCNEC, TICLE and LOCNESS), description of the software used in the study, quantitative and qualitative analyses procedures employed in the study are presented in detail.

The fourth chapter consists of findings obtained through the analysis of the corpora. The tendencies of Turkish EFL learners in using recurrent phrases in spoken and written language are presented and then their language features are compared with the recurrent phrases in native language through the analysis of the reference corpora. The results are accompanied with discussions referring to the relevant literature.

The last chapter of the dissertation is the conclusion section in which the results are summarized and implications for language teaching and future studies are given.

1.7. Chapter Summary

This chapter briefly introduces the corpus linguistics and discusses its role in SLA. The background of the study gives information about the phraseology and emphasizes the need for the research on the recurrent phrases in written and spoken learner language. The purpose and significance of the study is also provided together with the research questions. Finally, it provides the definition of the key terms and overview of the thesis.

CHAPTER II

LITERATURE REVIEW

2.0. Introduction

This chapter presents the literature about corpus linguistics with a focus on its role in researching second language acquisition. First, it gives information briefly about development of corpus linguistics and its benefits for various subfields of linguistics. Then, the importance of using corpora in the field of SLA and existing learner corpora are explained in detail. The next section introduces phraseology which makes the theoretical framework of the present study. The last section aims to explain the connection between phraseology, corpus linguistics and SLA emphasizing the significance of studying recurrent word combinations in learner language. Finally, the existing research on recurrent phrases in interlanguage is summarized.

2.1. Corpus Linguistics: An Overview

The origins of corpus linguistics can be traced back to the acknowledgement of the concept of *observable data* in language research. Under the influence of the positivist and behaviourist approaches, the linguists at the beginning of the last century became concerned to account for the observable data and language in context. In 1950s, with the contribution of the such linguists as Harris, Fries and Hill among others, the notion of collecting real language data was placed at the core of what linguists study as pointed out by Leech (1992), who states that "a corpus of authentically occurring discourse was the thing that the linguist was meant to be studying" (p.105). Fries' attempts to study grammars of written and spoken American English (1940 and 1952 respectively) based on actual language use has been considered as among the earliest ventures in historical evolution of corpus linguistics (Tognini-Bonelli, 2010).

With Chomsky's criticisms of language performance at the beginning of the 60s, the developmental continuity of corpus linguistics was interrupted for a while (Tognini-Bonelli, 2010). Chomsky held a different position regarding the observable data in general and corpus linguistics in particular. His objection to corpus linguistics mainly comes from his well-known distinction between competence and performance, which is later revised as Internal and External language. While competence refers to internalised knowledge of language, performance is the external evidence of language competence, and its usage on particular situations. Performance is mostly affected by a number of factors other than competence. Therefore, to Chomsky (1965), linguists should consider competence as the base rather than performance while describing language. He also notes that "like most facts of interest and importance … information about the speaker-hearer's competence … is neither presented for direct observation nor extractable from data by inductive procedures of any known sort"(Chomsky, 1965 p.18). As obviously seen, this position is in stark contrast with the theoretical assumptions of corpus linguistics since corpus linguistics mainly deals with performance rather than competence. The aim of corpus linguists is to describe actual language use through observable data rather than identify linguistic universals (McEnery & Wilson, 2001; Tognini-Bonelli, 2010). Being very influential on the linguists of the era, Chomsky's views caused a change in paradigm, leading researchers to consider introspection as a more accurate and less time-consuming way of analysing language. Corpus linguistics studies were not given proper attention. Therefore, the development of corpus linguistics slowed down.

However, despite these factors affecting corpus studies negatively, there were still some noteworthy attempts in forming corpora of various types, which Leech (1992) considers as the second stage of the evolution of the corpus linguistics. The Brown Corpus that had been developed in collaboration with Swedish, Norwegian and Dutch universities is among the pioneers of corpus based language studies. In 1975, Lancester-Oslo-Bergen Corpus of British English (LOB) appeared, and it is followed with the publication of the Collins Birmingham University International Language Database (COBUILD) in 1987. In fact, appearance of the text book collection through computerized corpora attests this continuity (Léon, 2005).

Although the creation of foregoing corpora are thought to be milestones in corpus linguistics, the revival of interest in Firthian principles of language study in the last quarter of past century is, no doubt, influential in the theoretical underpinning of corpus linguistics today (McEnery & Gabrielatos, 2006). In Firthian approach to language investigation, the notions of observable data, actual language use and language patterns are highlighted to a great extent. Language is seen as a social phenomenon which is observable in discourse and text. In Firth's words, "we must take our facts from speech sequences, verbally complete in themselves and operating in contexts of situation which are typical, recurrent, and repeatedly observable" (in Diniz, 2007, p.35). The meaning is regarded as function and not only words but also grammatical structures have meaning. Key discussions of Firth's approach has been successfully connected to what corpus linguistics offer at its core by Sinclair and Halliday, who are often called neo-Firthians (McCarthy, 2006). These researchers among others have played a leading role in developing corpus linguistics for both language pattern research and dictionary making (Cowie, 1998).

All these contributions affected the establishment of the corpus linguistics traditions of "(a) trawling through texts to find all examples of a particular piece of language, (b) writing dictionaries based on attested usage, (c) analysing language based on actual informant data" (McCarthy & O'Keeffe, 2010 p. 5). However, it is the spread of computers for personal use and the revolution in hardware and software in the 1980s and 1990s which really enabled contemporary corpus linguistics to emerge. Gradual proliferation of relatively small sized computers in 1990s allowed groups and individuals to initiate quite ambitious corpus projects. The parallel growth of the internet and relevant technologies facilitated data transfer and instant access to huge quantities of texts stored in electronic form. Concomitantly, advances in recording tools yielded positive effects on the creation of spoken corpora (McCarthy & O'Keeffe, 2010). Additionally, the advent of such software specially designed for corpus analyses as Scott's *WordSmith Tools* (1996-) and Barlow's *Monoconc* (1996) along with the others has led to the appearance of corpus linguistics as we know it today. Granger (1998) summarizes role of computers by pointing out advantages in two main aspects:

> A first major advantage of computerization is that it liberates language analysts from drudgery and empowers [them] to focus their creative energies on doing what machines cannot do. More fundamental, however, is the heuristic power of automated linguistic analysis, i.e. its power to uncover totally new facts about language (p.3)

In the same vein, comparing pre-electronic corpora with the computer based corpora, Oostdijk (1991) states that

> Unlike earlier corpora, the corpora that are currently used are computer readable and lend themselves to automatic analysis. As a result, larger quantities of data can be processed at a greater speed, while consistency in the analysis is warranted through the use of a formalized description contained in the grammar (p.4).

This historical evolution has been echoed in the definitions of corpus linguistics provided in the literature. A closer look at various definitions offered to date points out an agreement on the inclusion of such notions as *machine-readable, observable data* etc… in the definition. However, what is debated is related to addressing corpus linguistics as a tool, a method, a methodology, a methodological approach, a discipline, a theory, a theoretical approach, a paradigm or a combination of these (Taylor, 2008). Leech (1992) defines corpus linguistics as "not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject" (p.106), and describes characteristics of corpus linguistics as a new paradigm by emphasizing its focus on linguistic performance rather than competence; linguistic description rather than linguistic universals; both qualitative and quantitative models of language and its empiricist view of scientific inquiry. Likewise, Stubbs (2002) regards corpus as not a merely tool but an important concept in linguistic theory stating that "corpus linguistics provides a new point of view for studying language and the point of view allows new things to be seen" (p.220). Teubert (2005) has a similar stance as he defines corpus linguistics as "a theoretical approach to the study of language" (p.2). Gries (2006) favours a methodological conceptualization by defining corpus linguistics as "the analysis of naturally occurring data" and "a methodological paradigm within applied and theoretical linguistics" (p.4). McEnery, Xiao and Tono (2006) note that "corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and it certainly has a theoretical status" (p.8). In spite of the on-going debate on defining scientific nature of corpus linguistics, linguists from various fields share a common belief: "that it is important to base one's analysis of language on real data –actual instances of speech or writing- rather than on data that are retrieved or made-up" (Meyer, 2004 p.8).

Depending on the evidence of actual language use provided, corpora have both theoretical and practical uses, which make them invaluable resources for *descriptive, theoretical* and *applied* discussions of language (Meyer, 2004). As corpus linguists, in one aspect, are interested in counting and categorizing structures occurring in a corpus, and prioritize descriptive adequacy, the results of their studies offer a lot in terms of language descriptions. Starting from these descriptions, corpus linguists could also use corpora effectively to test out linguistic hypotheses; thus, contribute to the evolution of language theories. Leech (1992) emphasizes that what is discovered in a corpus can be utilized as the basis for the relevant theoretical issue. Aarts's (1992)study on small clauses could prove to be good example: Using the London Corpus, Aarts (1992) provided a complete description of small clauses in English and addressed to certain relevant controversies; hence, contributed greatly to the theoretical discussions. Additionally, as corpus linguists deal with texts or parts of texts, they are able to contextualize their analysis of language; therefore, corpus linguistics has a significant role in applied discussions of language. Meyer's (2004) corpus based study on elliptical coordination provides evidence for such a role. Using a-96,000-word corpus including different types of speech and writing, Meyer (2004) investigated why certain types of elliptical coordinations are less frequent in speech than writing. His findings pertaining to elliptical coordinations that place processing burdens on the hearer/reader are likely to foster applied discussions of language.

With all these significant additions, corpora have currently been acknowledged in many disciplines of linguistics. Discourse Analysis is one of the areas where corpus linguistics has been adopted "as a means of looking at language patterns over much larger datasets" (McCarthy & O’Keeffe, 2010 p.9). Through the use of special search techniques (e.g. wordlists, concordances, keyword etc…), corpora could automate many of the process of Discourse Analysis. Historical linguistics is another area to which corpus linguistics is applied. By means of collecting longitudinal corpora, it is likely for the researchers to study the linguistics development of a language. "Such corpora allow corpus linguists not only to study systematically the development of particular grammatical categories but to gain insights into how genres in earlier periods differed linguistically” (Meyer, 2004 p.21). Literary and translation studies also have benefitted from corpus linguistics. As corpus linguistics facilitates the comparison of patterns across languages by comparing source and target texts, it has offered insights in translation and literary areas. Pragmatics, the study of language in use, apparently makes a perfect match with corpus linguistics. Advent of corpus linguistics freed pragmatists from relying on "intuited data" (McCarthy & O’Keeffe, 2010 p.10). In the relevant literature, there are a number of studies focusing on pragmatic markers including deictics, hedges, and discourse markers in both spoken and written contexts using corpora. The use of corpora has also yielded fruitful results in terms of comparing pragmatic features across different languages. Johansson (2006) for instance investigated the use of *well* in English and compared its equivalents in Norwegian and German. Corpora have also had impact on the area of sociolinguistics as corpora especially the spoken ones automatically produce data on language use in relation with "sociolinguistic variables such as age, gender, level of education socio-economic background" (O'Keeffe, McCarthy, & Carter, 2007 p.20). Introduction of the corpus linguistics to the area of language teaching has resulted in the publications of comprehensive practical materials. Through various types of corpora, it is possible to obtain information on the structure and usage of many different grammatical constructions, which makes a sound basis for writing a reference grammar of target language. Greenbaum's (1996) Oxford English Grammar based on British component of the International Corpus of English, and Biber, Johansson, Leech, Conrad, and Finegan's, (2007) Longman Grammar of Spoken and Written English based on Longman Corpus follow this tradition. Additionally, lexicographers have utilized corpora to create corpus informed dictionaries. For example, British National Corpus (BNC) is the basis of Longman Dictionary of Contemporary English (Meyer, 2004).

In sum, over the past four decades, corpus linguistics has greatly evolved, and corpora today have increasingly been accepted as essential resources in linguistic investigation. Despite the on-going debate on its scientific categorisation, it is a fact that corpus linguistics has revolutionized nearly all the branches of linguistics from lexicography through sociolinguistics to language teaching. And, the field of second language acquisition is no exception.

2.2. Corpus Linguistics and Second Language Acquisition Research

The main goals of the Second Language Acquisition (SLA) research is to determine the learners' second language (L2) knowledge and to describe L2 acquisition process through uncovering the principles that shape and constrain this process(R. Ellis, 2003; Lakshmanan & Selinker, 2001). Research conducted within the field of SLA has shown that language used by learners contains a considerable amount of deviant forms from both learners' first language and the target language. This finding has led to the formulation of *interlanguage theory* (Selinker, 1972). Selinker (1972) defines interlanguage as "the separate linguistic system based on observable output which results from a learner's attempted production of a target language norm" (p. 214). The nature of interlanguage is systematic, dynamic and influenced by the learners' previously acquired languages (Lightbown & Spada, 2006). According to Selinker (1972, p. 215), five main cognitive processes are influential on the development of interlanguage:

1) First Language Transfer: The knowledge of first language (L1) interferes with that of L2.
2) Transfer of Training: Rules or subsystems occurring in training procedures might affect interlanguage.
3) Strategies of Second Language Learning: Strategies being learned with regard to any language items can be applied to the others to reduce the target language to a simpler system.
4) Strategies of Second Language Communication: Communication management strategies might result in avoiding the problematic items, which might influence interlanguage development.
5) Overgeneralization: Rules or semantic features of the target language may be overextended to any language items.

In order to understand the mechanisms of second language acquisition, researchers have to rely on interlanguage performance data that are the unique way of accessing the invisible underlying principles (Stubbs, 2001). Lakshmanan and Selinker (2001) state that "interlanguage competence cannot be examined directly. Instead, information about the nature of interlanguage competence can only be derived indirectly through an examination of interlanguage performance data" (p. 393).

While researching the interlanguage, researchers have traditionally drawn on a variety of data types ranging from spontaneous speech through elicited data to judgements. R. Ellis (2008) identifies three major categories of interlanguage data as illustrated in Figure 1.

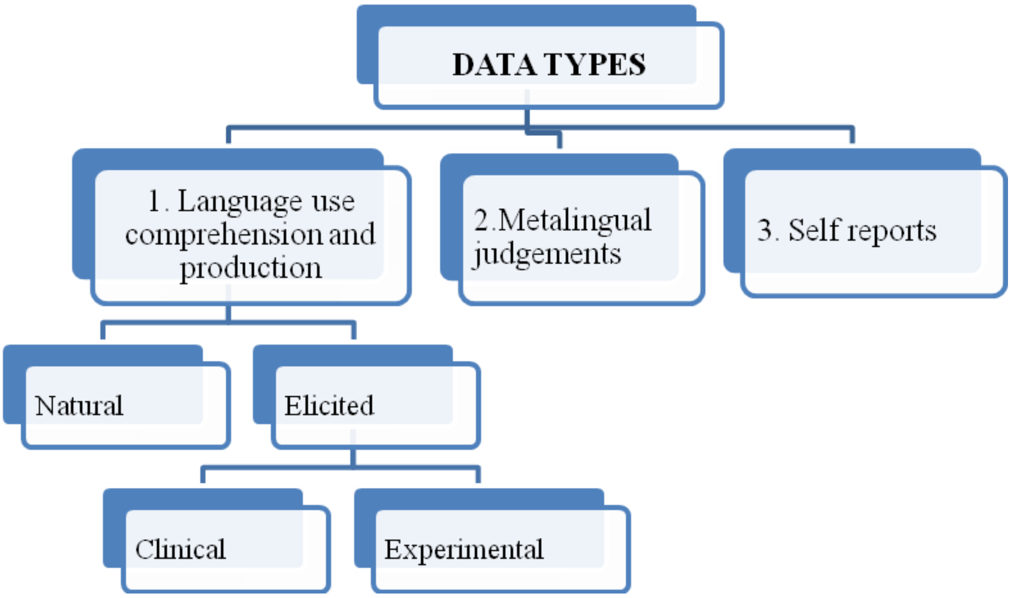


Figure 1. Main data types employed in SLA research. Adopted from "The Study of Second Language Acquisition” by R. Ellis 2008 p.670, Copyright 2008 by Oxford University Press

1) Language use data that reflects learner's attempts to use the second language in comprehension or in production. 2) Metalingual data which refers to judgements or learner's intuition about L2. 3) Self report data which is based on exploring learner's strategies through questionnaires or think-aloud tasks.

Different kinds of research in the field of SLA have made use of different data types. For example, case studies on the order and sequence showed a clear preference for natural language use. Cross-sectional studies (e.g. morpheme studies) favoured clinically or experimentally elicited data. Research that set out to test the SLA theories heavily relied on metalingual judgements. Self-report data are given priority in the studies on individual differences (R. Ellis, 2008). Regarding the language produced by learners as the central source of evidence for mental processes, Myles (2005) states that “the success of SLA research relies on having access to good quality data (p.374). While aforementioned data types ensure quality to some extent, each has its own drawbacks, which requires reconsidering their uses in SLA research. Table 1 summarizes the major drawbacks of traditional data types used in SLA along with the potential solutions offered by corpus data.

Table 1

Strength and Drawbacks of Traditional Data Types in SLA and Potential Contributions of Corpus Data

Data Type	Strength	Drawbacks	Potentials of Corpus Data
Natural Language Use	1. gives information about what learners actually do with the L2 2. provides authentic data (R. Ellis, 2008)	3. is time-consuming and difficult to collect 4. the amount of data may not be sufficient to make quantitative analysis possible 5. poses problems in generalizability of the results due to small and unrepresentative sample (MacWhinney, 2000) 6. does not allow systematic inquiry (Beebe & Cummings, 1996)	7. provides essentially authentic data due to its focus on actual language use (Sinclair, 1996) 8. enables to work on large amounts of data quantitatively in a relatively less time due to technological tools (Granger, 1998a) 9. makes systematic inquiry of learner language possible through strict design criteria (Granger, 2002)
Elicited Language Use	1. provides systematic data (R. Ellis, 2008)	2. may result in inadequate information relating to specific language features (R. Ellis, 2008) 3. remains small and limited in scope (e.g. containing very focused data aiming to answer a specific research question) 4. is usually not accessible to the research community as a whole (Myles, 2005)	5. larger amounts of data obtained through corpus compilation maximizes the chance of any language feature being present (Myles, 2005) 6. As corpora contain data in electronic form, they are easily and effectively made available to research community (Granger, 2002)
Metalingual Judgement Self-Report data	1. provide information about the language learners do not use (R. Ellis, 2008) 2. enable to uncover some of the affective and cognitive factors in L2 learning.	3. are difficult to obtain reliable judgements 4. cause response biases 5. are difficult to determine exactly what it is learners judge when they evaluate sentences 6. affected by the learners' skill in performing tasks. 7. may not report what learners actually are doing. (R. Ellis, 2008) 8. are less objective and less generalizable (Gilquin & Gries, 2009)	9. as it is difficult to validate the metalingual judgement data, learners' actual use of language is prioritized in SLA (R. Ellis, 2008). Actual language use is among the basic strengths of corpus data. (Granger, 2002)

When the drawbacks summarized in Table 1 are taken into account, it can be inferred that the field of SLA should make use of a different type of data. Or in Granger's words (1998a) “there is clearly need for more and better quality data and this is particularly acute in the case of natural language data” (p.5).

Much current SLA research favours the experimental and introspective data seemingly dismissive of natural language use data (Granger, 2002). One of the reasons for this is the difficulty of controlling variables having impact on interlanguage performance data in non-experimental settings. Since it is difficult to include a large number of informants in the experimental design, SLA research is traditionally based on a relatively narrow empirical base.

The drawbacks of using limited empirical base are pointed out by Gass and Selinker (2008) who note that "it is difficult to know with any degree of certainty whether the results obtained are applicable only to learners studied or whether they are indeed characteristics of a wide range of subjects" (p.55). Likewise, MacWhinney (2000) notes that "conducting an analysis on a small and unrepresentative sample may lead to incorrect conclusions" (p.3). These limitations consequently pose problems in terms of generalizability of the results. Therefore, it has been admitted that in order to have sound and generalizable claims about SLA process, SLA researchers need data sources which illustrates natural language use by a wide number of learners with the variables tightly controlled (Granger, 2009). As Myles (2005) rightly notes “time has now come, though, to test some of the current hypotheses on larger and better constructed datasets, as has happened in L1 acquisition” (p.376).

When the foregoing literature is considered, the use of corpus data apparently is a timely arrival for SLA research. The exact innovation that corpus linguistics has brought about the field of SLA is to provide researchers with a type of data source which enable them to rely on larger and carefully gathered datasets in their investigations. Computer learner corpora which refer to electronic collections of spoken and written texts produced by foreign or second language learners assembled according to explicit design criteria (Granger, 2009)contain data from hundreds (sometimes thousands) of learners, and can therefore lay claim to greater representativeness than previous SLA studies. Granger (2009) points out that "one of the main assets of learner corpus research is that it brings to the SLA field a much wider empirical basis than has ever previously been available” (p. 16).

Development of learner corpora is crucial in the field of SLA for a number of reasons. To start with, computer learner corpora typically are in the category of natural language use data, which has been particularly useful in SLA research as shown by the fertile learner corpus studies over the past decade (see Myles, 2005; Pravec, 2002). Emphasizing the authenticity of the learner corpora data, Sinclair (1996) explains that “all the material is gathered from the genuine communications of people going about their normal business” unlike data gathered “in experimental conditions or in artificial conditions of various kinds”. Additionally, learner corpora have proven to be beneficial when their size is considered (Granger, 2004). As the data are stored electronically, it is likely to collect large amount of data in a practical and quick way. This facilitates having claims in terms of "representativeness of the data and generalizability of the results" (Granger, 2004, p. 125). Moreover, the potentials of learner corpora in controlling the variables make learner corpora superior to the previous data sources in SLA research. It is a fact that learner language is highly variable and is influenced by a number of linguistic, situational and psycholinguistic factors. If these variables are not controlled properly, the reliability of the findings in language learner research becomes limited. Regarding the matter of variability in the previous SLA research, Gass and Selinker (2008) comment that "there is often no detailed information about learners' themselves and the linguistic environment in which production is elicited" (p. 57). Learner corpora, however, have the potential to provide researchers with information on such variables. As given in the definition, learner corpora are assembled according to strict design criteria and it is quite likely to identify and include variables beforehand. This makes corpora potentially promising data sources for SLA.

2.3. Learner Corpora: Basic Features

Proven to be useful in researching learner language, learner corpora as a field of scientific inquiry started as recently as the late 1980s. One reason of this late emergence is that, until recently, data collection and analysis necessitated too much time and effort on the part of the researcher (Granger, 2002, 2004). With the technological advances that have gained impetus since 80s, the work of gathering data in very large quantities, storing them on the computer and analysing them automatically or semi-automatically using available linguistic software has been made possible. Therefore, corpora of non-native varieties have begun to appear.

Although learner corpora are roughly defined as the electronic collections of learners' written and spoken performances, Sinclair (1996) proposes a comprehensive definition integrating the distinguishing properties of learner corpus: "Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance".

As clear in this definition, one of the default values of learner corpora is its emphasis on authenticity since they tend to be gathered from genuine communications of people (Sinclair, 1996). As stated above, with this feature, learner corpora remain superior to language use data elicited in experimental settings. However, the notion of authenticity should be taken in a different sense in the context of learner English. Granger (2002) states that "even the most authentic data from non-native speakers is rarely as authentic as native speaker data, especially in the case of EFL learners, who learn English in the classroom"(p.8). Learner English in classroom context involves some kind of artificiality since they do not use the target language while going about their normal business. Therefore, authenticity of learner corpus data results from the authentic classroom activity (Nesselhauf, 2004). That's to say; "in as far as essay writing is an authentic classroom activity, learner corpora of essay writing can be considered to be authentic written data, and similarly a text read aloud can be considered to be authentic spoken data (Granger, 2002 p.8).

Another basic property of learner corpora which makes corpus data invaluable is that texts are computerized. Nesselhauf (2004) highlights the significance of computers in learner language research:

> computerized data can be distributed more widely, so that results are more easily comparable and also more easily verifiable than if each researcher (or each small group of researchers) uses a different set of data for their analyses.(p.130)

As the data are stored on a computer, researchers can automatically perform the functions of *count*, *sort*, *compare* and *annotate,* which are of high relevance with SLA research. This automation help researchers save time and efforts especially while working with large data sets. Through the function of *count*, precise figures in terms of frequency of linguistic items in various texts can be obtained. Granger (2009) notes that "frequency is an aspect of language that plays a major part in many linguistic applications which require knowledge not only of what is possible in language but what is likely to occur (p.4). The automatic *sort* function using concordancing programs give SLA researchers a view of learners' lexico-grammatical patterning of words. The *annotate* function which refers to "the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written data"(Garside et.al., 1997 cited in Granger, 2004) help researchers carry out detailed investigations on a range of aspects of learner language including lexical analysis of interlanguage, use of grammatical categories like prepositions, modals, passives or the use of discourse markers in interlanguage.

Furthermore, learner corpora contain continuous stretches of words rather than separate words or sentences. Granger (2002) considers this feature as among the distinguishing properties of learner corpora. Therefore, learner corpora include both erroneous and correct use of learner language, which, in essence, provides a very rich data base for researching various aspects of learner language. With a comparable native corpus, the over and underuse of some language features can be studied beside the deviant and consistent use.

Strict design criteria followed in compilation of any learner corpus is the next significant property of learner corpus data. Because of the great amount of variation in EFL/ESL, design criteria are crucial in the case of learner data, and “a random collection of heterogeneous learner data does not qualify as a learner corpus” (Granger, 2002 p.9). Some of the learner corpus-related criteria to be considered in compilation are represented in Figure 2.

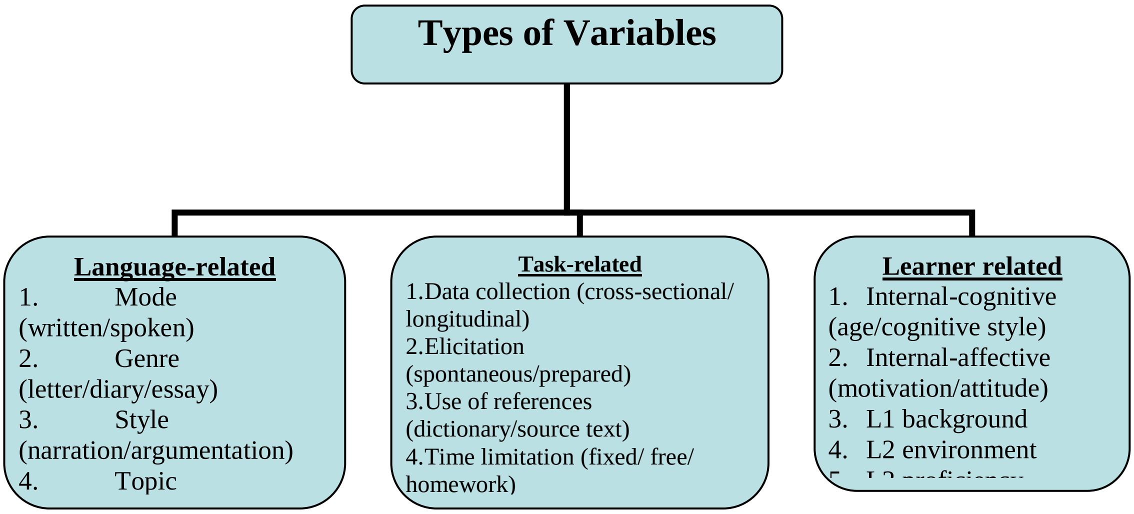

Types of Variables
Language-related
1. Mode (written/spoken)
2. Genre (letter/diary/essay)
3. Style (narration/argumentation)
4. Topic
Task-related
1.Data collection (cross-sectional/ longitudinal)
2.Elicitation (spontaneous/prepared)
3.Use of references (dictionary/source text)
4.Time limitation (fixed/ free/ homework)
Learner related
1. Internal-cognitive (age/cognitive style)
2. Internal-affective (motivation/attitude)
3. L1 background
4. L2 environment

Figure 2. Major design considerations in learner corpus compilation**.** Adopted from "Computer Learner Corpus Research: Current Status and Future Prospects" by S. Granger 2004 p. 126, Applied Corpus Linguistics. A Multidimensional Perspective Copyright 2004 by Rodopi

Using a number of criteria enables researchers to investigate learner language with respect to the learners' proficiency level, their L1,the medium, text type, the learning environment in which the language was acquired, the age and sex of the learners, the years of acquisition, the influence of other foreign languages and any other information that the corpus provides.

2.3.1. Learner Corpora around the World

Learner corpora are compiled for various purposes; in some cases, testing or improving some aspects of SLA theory (e.g. theories about L1 transfer or order of acquisition) could be a driving force and in others, designing better instructional tools and materials can make the purpose behind the corpus data gathering. Depending on the purpose of the research, the processing of the learner corpus data changes as well. Types of processing available are shown in Table 2.

Table 2

Types of Processing of Learner Data

Extra-textual information	Header information (learner/ language/ task variables)
Level of transcription	Orthographic (+ phonemic/ phonetic for spoken corpora)
Level of annotation	Sentence-boundary disambiguation Tokenisation POS tagging Lemmatisation Parsing (Treebanking) Semantic tagging (word senses/ semantic relationships and categories) Discourse tagging (apologies/greetings/politeness/?? moves/acts??/etc.) Error tagging Prosody annotation Anaphoric annotation

Adopted from "Learner corpora: design, development and applications" by Y. Tono, 2003 Proceedings of Corpus Linguistics p.801, Copyright 2003 by Ucrel

While collecting learner data, the type of processing of corpus should be made clear and shared with the research community so that the relevant data enable researchers to carry out various comparisons (e.g. comparison of native data with non-native data). Granger (2002) notes that full details about the type of processing should be documented in a way that it will ease the process of compiling sub-corpora as well. Similarly, Tono (2003) states that if the data are gathered without documentation of extra-textual information or level of annotation along with variables, then the resulting corpus will be of little value.

Despite the relative youth of learner corpora, its significance and potentials have been acknowledged, and quite a number of learner corpora have already been compiled or in the process of compilation. Table 3 presents some of the major learner corpora developed so far.

Table 3

Major Learner Corpora Around the World

Learner Corpus	Subjects/Task/Size	Annotation	Comparison
International Corpus of Learner English (ICLE)	-University EFL 3/4 year students -16 nationalities -Written essays -4,5 million	Error tagged Pos tagged	NNS vs NNS (different L1s) NS vs NNS
Louvain International Database of Spoken English Interlanguage (LINDSEI)	-50 interviews + -11 nationalities -3/4 year students -100,000	orthographic	NNS vs NNS (different L1s) NS vs NNS
Longman Learners Corpus (LLC)	-All levels -Written essays -10 million -Commercial	Pos Tagged	NNS vs NNS
The Hong Kong University of Science and Technology Learner Corpus (HKUST)	-Chinese undergraduate students -Written academic texts -25 million words	Error tagged	NS vs NNS
The ISLE corpus of non-native spoken English	-20 minute speech -German and Italian intermediate learners of English	Orthographic Phone-stress	NS vs NNS
Cambridge Learners Corpus (CLC)	All levels 10 million Commercial	Pos tagged Error tagged	NNS vs NNS
Indianapolis Business Learner Corpus (IBLC)	- US univ. business students - business writing - plain text	- Plain text	NNS vs NNS (different L1s)
Chinese Learner English Corpus (CLEC)	-Chinese students from five L2 proficiency levels -written texts -1 million words	-Error tagged	NS vs NNS
PELCRA University of Lodz, Poland	-Polish learners of English at different levels of L2 proficiency -written texts	- Plain text	NS vs NNS
USE Uppsala University, Sweden (USE)	-Swedish university students of advanced level -written academic texts	- Plain text	NS vs NNS
TeleNex Student Corpus	secondary school &university students written texts 3 million words	- Plain text	NS vs NNS

As seen in Table 3, there are great efforts to create corpora of learner language. Among the available ones, HKUST is regarded as probably the biggest learner corpus containing around 25 million words, and it is still growing. It comprises different academic text types written by Chinese undergraduate students. TeleNex and CLEC are other examples of big written corpora. TeleNex contains about 3 million words of composition produced by secondary school students from Honk Kong. CLEC also is made up of compositions (1.2 million words) by secondary school and university students. USE includes about 1 million words of written texts by Swedish undergraduate students. IBLC is among the few specialized, non-academic learner corpora containing 200,000 words of business letter written by L1 Japanese business people. ICLE, LINDSEI, LLC and CLC make up the much smaller group of corpora containing language of learners with different first language backgrounds. The latter two are commercial corpora that contain around 10 million words. ICLE and LINDSEI are, on the other hand, non-commercial corpora created in academic setting, and are notable as they are among the few non-commercial corpora representing a large number of learners with different mother tongues (Nesselhauf, 2004).

The learner corpora presented in table 3 is only a small reflection of a myriad of learner corpora that have been or are being compiled and exploited by researchers. However, a closer look at even these several corpora implies that “there is still great scope for further corpora and for improvement of the existing ones” (Nesselhauf, 2004 p.132). Granger (2004) suggests to evalute current learner corpora according to two major dimensions: learner and task. The learners represented in the corpora are generally the learners of English as a Foreign Language (EFL) who learn the language in an environment with a restricted access to the speakers of the language being learned (Gass & Selinker, 2008). Regarding the proficiency level of the learners, it is observed that intermediate and advanced levels dominate the current learner corpora scene. However, it should be noted that the labels of intermediate or advanced are a bit vague as “one researcher’s advanced category may correspond to another’s intermediate category”(Gass & Selinker, 2008 p.37). The general tendency in defining learners’ proficiency levels is to use external criteria (e.g. third year university students) as in the case of ICLE and LINDSEI.

With regard to tasks employed in the compilation of learner corpora, it is seen that tasks eliciting written language data are more frequently used than tasks requiring spoken language. In other words, most of the current learner corpora focus on written interlanguage than spoken medium. It is because of the fact that collecting and transcribing spoken language is a very hard and time-consuming work that requires a clearly planned collaborative projects as in the case of LINDSEI (Granger, 2004). As regards the discourse, English for General Purposes (EGP) is dominantly taken as the base rather than English for Specific Purposes (ESP); the Indianapolis Business Learner Corpus (IBLC) is an exception in this respect. When the possibility of comparison they allow is considered in table 3, it is observed that most of the current learner corpora are cross-sectional (i.e. gathered from different categories of learners at a single point in time). Longitudinal corpora (i.e. gathered from the same learners over time) are very few. The reason is that compiling such corpora is very difficult as it requires following a learner population for a very long time. Housen's (2002)Corpus of Young Learner Interlanguage that consists of EFL data from European School pupils at different stages of development and from different L1 backgrounds is an exception. What is preferred instead by the researchers dealing with developmental interlanguage features is the use of ‘quasi-longitudinal’ data (i.e. they collect data from a homogeneous group of learners at different levels of proficiency) (Granger, 2002). Finally, a look at the annotation column of table 3 shows that that building pos tagged and error tagged corpora is on the increase (Granger, 2004). Considering the available corpora, it seems that besides corpora for more L1s, there is a need for corpora representing different proficiency levels, different registers and different varieties of English (Nesselhauf, 2004).

As seen from the literature presented above, learner corpora with its inherent features potentially provide a very rich data source to have an understating of learner language. Due to the size and authenticity of the corpus data, it helps researchers to gain insights into what learners are actually doing when they use L2. As it contains continuous stretches of words, it gives detailed information about both erroneous and correct use of learner language. Design criteria identified in the literature and the types of processing of learner data available to learner corpora enable researchers to evaluate learner language from different point of view. Granger (2008) summarizes well that learner corpora should be seen “as one highly versatile resource which SLA/FLT researchers can usefully add to their battery of data types” (p.20). An examination of existing learner corpora shows that although the field of learner corpora is growing quickly, the current learner corpora are “only a beginning” (Nesselhauf, 2004 p. 132), which implies a number of areas requiring further efforts to develop well-designed corpora. Among the few relatively well-planned learner corpora are ICLE and

LINDSEI, which make the data source of the present study as well. Following section briefly introduces these two learner corpora.

2.3.1.1. ICLE: International Corpus of Learner English

The ICLE project is one of the best known and prevalently used learner corpora in investigations of interlanguage. Nesselhauf (2004) regards ICLE as “probably the only existing sizeable non-commercial learner corpus containing data from learners with different L1s” (p.129). Being the first learner corpus developed in an academic setting, ICLE as a project was launched in 1990, and since then, has been expanded with the collaboration of a large number of universities internationally. Currently, it consists of around 4,5 million words and comprises argumentative essays written by university students of English. Now, it has 16 sub-corpora representing learner groups from Bulgarian, Czech, Finnish, Dutch, Chinese, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish and Tswana L1 background (Granger, Dagneaux, Meunier, & Paquot, 2009). The research goals of ICLE are two-fold. Firstly, it aims to collect dependable evidence on learner language which allows cross-linguistic comparison in order to determine universal and language-specific features of interlanguage. The comparison also enables to determine to what extent interlanguage-features are affected by factors in the learner’s cultural or educational background. The second research goal of ICLE is “to investigate aspects of ‘foreign-soundingness’ in non-native essays which are usually revealed by the overuse or underuse of words or structures with respect to the target language norm” (Pravec, 2002 p.83). For the investigation of overuse/underuse phenomena, its reference corpus LOCNESS comprising the written essays by native students makes a powerful comparable base.

One strength of ICLE lies in its control on a number of variables as it was built in accordance with a set of very strict design criteria (Granger et al., 2009). Figure 3 presents variables considered in the compilation of ICLE.

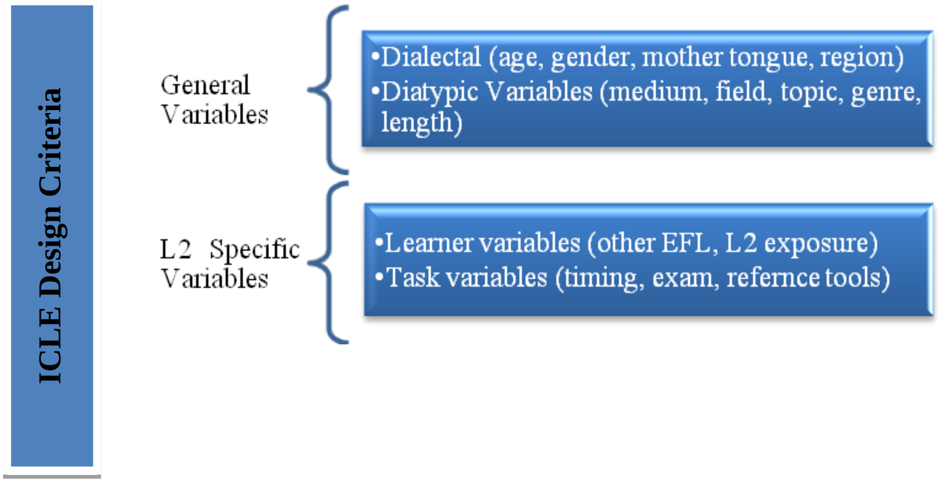

ICLE Design Criteria
General Variables
•Dialectal (age, gender, mother tongue, region)
•Diatypic Variables (medium, field, topic, genre, length)
L2 Specific Variables
•Learner variables (other EFL, L2 exposure)
•Task variables (timing, exam, refernce tools)

Figure 3. ICLE design criteria. adopted from "Computer Learner Corpus Research: Current Status and Future Prospects" by S. Granger 2004 p. 126, Applied Corpus Linguistics. A Multidimensional Perspective Copyright 2004 by Rodopi

The Turkish sub-corpus TICLE was compiled by Kilimci and Can (2009) and included in the second version of ICLE in 2009. Foregoing variables were also considered in the compilation of TICLE as well. Further description of TICLE and its comparable corpus LOCNESS is provided in sections 3.2.1 and 3.2.2 respectively.

2.3.1.2. LINDSEI: Louvain International Database of Spoken English Interlanguage

LINDSEI is the first large scale corpus of spoken learner English. It is designed so as to be easily comparable with an already existing written corpus (ICLE). It was launched in 1995 at the Université Catholique de Louvain. The aim was to create a data source for the investigation of oral production of advanced learners of English from different mother tongue backgrounds, which allows cross-linguistic comparison. The first component included spoken data gathered through interviews by French learners of English. With the inclusion of other L1s, the project was expanded. Currently, LINDSEI has a total of 20 sub-corpora with different L1 backgrounds.11 of them have been completed (Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish) and the others (Arabic, Basque, Brazilian-Portuguese, Czech, Finnish, Lithuanian, Norwegian, Swedish, Taiwanese and Turkish) are in progress (Gilquin, 2012). It has a comparable native corpus LOCNEC making the NS and NNS comparison possible as in the case of this study.

Like ICLE, LINDSEI was also formed according to strict and explicit design criteria. Figure 4 displays the variables considered in the compilation of LINDSEI.

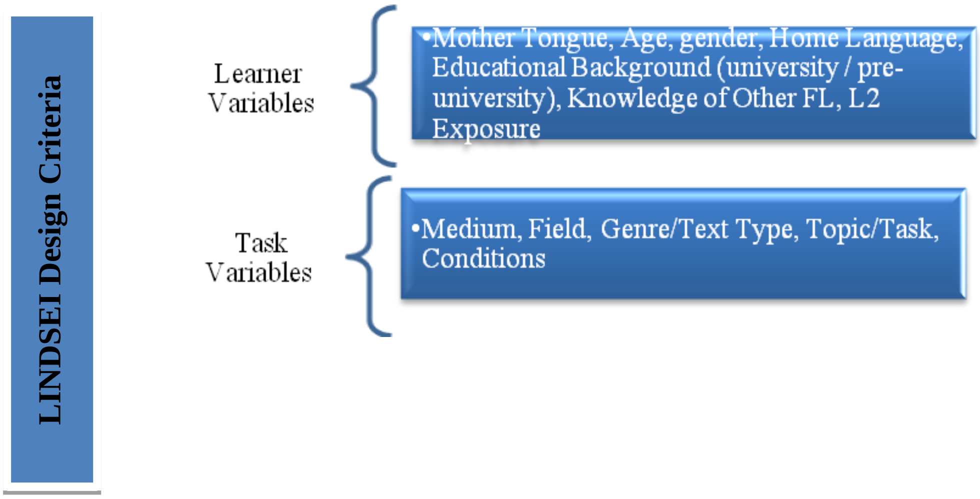

LINDSEI Design Criteria
Learner Variables
•Mother Tongue, Age, gender, Home Language, Educational Background (university / pre-university), Knowledge of Other FL, L2 Exposure
Task Variables
•Medium, Field, Genre/Text Type, Topic/Task, Conditions

Figure 4. LINDSEI design criteria

The Turkish sub-component of LINDSEI, LINDSEI-TR[3], was compiled at Çukurova University with the attendance of the third and fourth year students studying at the department of English Language Teaching. LINDSEI-TR included a total of 58 interviews of 81,711 words with an average length of 1,408 words per interview. Further description of LINDSEI-TR and LOCNEC is given in sections 3.2.3 and 3.2.4 respectively.

2.4. Linguistic Analysis of Learner Corpora

Linguistic exploitations of learner corpora are usually centred on two methodological approaches. One is the Computer-aided Error Analysis (CEA) which deals with errors in interlanguage and employs computer tools to tag, retrieve and analyse them (Granger, 2002). The other commonest approach is Contrastive Interlanguage Analysis (CIA) (Granger, 2002, 2008; Paquot, 2010) that forms the methodological base for the present study as well.

[3] LINDSEI-TR was compiled in collaboration with Université Catholique de Louvain. Data gathered are still under evaluation, and this study presents the preliminary results.

2.4.1. Computer-aided Error Analysis (CEA)

Computer-aided error analysis originally dates back to 1970s when Corder (1976) emphasized the significance of analysing the learner language noting that errors could provide deeper understanding about how languages are learned and learners' grasp of language at any given point during the learning process (Kırkgöz, 2010). Error analysis (EA) of 1970s was criticized because of its approach to errors in analytical process. Granger (2002) states that "former EA was characterized by decontextualization of errors, disregard for learners' correct use of the language and non-standardised error typologies" (p.14). Advent of corpus linguistics and analysis of learner language through computers have greatly changed the way learner errors are analysed and current computer-aided error analysis emerged.

Today's EA analyses the errors in context including both the context of use and the linguistic context. Erroneous instances of a linguistic item can be seen in more than one sentence, in a paragraph or even in a whole text together with the correct usages. The available corpus linguistics procedures enable standardized error tagging; that's error categories are well defined and fully documented (Granger, 2002). All these advantages are made possible through the use of computers.

Computer-aided error analysis basically involves two methods: the first is related to selecting an error-prone linguistic item (e.g. a phrase or a syntactic category) and analysing the whole corpus to retrieve all occurrences of misuse of the target item by using text retrieval software tools. Although this method is limited to items that the researcher considers to be problematic, the fact that it is very fast and practical makes this method advantageous. The second method is to devise a standardized system of error tagging. Then, the researcher tags all the errors in a learner corpus or the errors in a particular category of interest are tagged thoroughly with the help of an error editor. Although it is very time-consuming and labour-intensive, the advantage of this method is that it enables the researcher to discover learner difficulties that he is not aware of.

2.4.2. Contrastive Interlanguage Analysis (CIA)

Granger (2009) describes CIA as a method that "consists in carrying out quantitative and qualitative comparisons between native and non-native speakers data or between different varieties of non-native data" (p.12) as illustrated in figure 5.

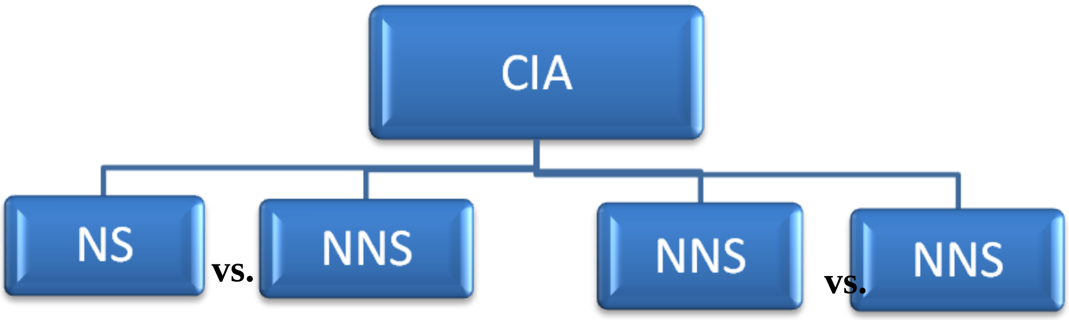


Figure 5. Contrastive interlanguage analysis (CIA). Adopted from “The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation” by S. Granger 2009 Corpora and Language Teachingp.27, Copyright 2009 by John Benjamins Publishing Company

The first type of comparison (NS vs NNS) aims to uncover the non-native features in learners’ writing and speech through detailed comparison of linguistic features in native and non-native corpora. Granger (2009) mentions about two things to be considered while carrying out NS vs NNS comparison. One is the selection of control corpus and the other is the level of proficiency of the native speakers, which should be on a comparable base. Even though Hunston (2002) states that one of the drawbacks of the NS NNS comparison is that the CIA approach assumes that learners view native speakers’ language as the standard norm, it is a fact that CIA helps us to understand what learners do and “what native/expert speakers actually do rather than what reference books say they do” (p. 212). Through NS vs. NNS comparison, distinguishing features of both groups can be identified (H. Chen, 2010). In other words, NS vs NNS comparisons allow seeing the features of non-nativeness in learner language giving a comprehensive view of not only errors but also instances of over and underrepresentation of words, phrases and structures. According to Granger (2009), such a comparison that helps assess the extent of the deviation "is essential since the aim of all foreign language teaching is to improve the learners' proficiency which in essence means bringing it closer some NS norm(s)" (p.13).

The second type of comparison involves comparing different learner populations (NNS vs NNS) with the purpose of improving the knowledge of interlanguage. This comparison enables researchers to identify commonalities and differentiations between various learner populations, and is therefore helpful in determining developmental process of interlanguage. Paquot(2010) clarifies this as:

> Comparisons of different interlanguages (e.g. the English of French speakers compared to that of Dutch speakers), on the other hand, make it possible to assess whether these features are peculiar to one language group (and thus possibly due to the influence of the learner's mother tongue), or shared by several learner populations (and therefore likely to be developmental or due to other causes such as teaching methods) (p.70).

The potential and usefulness of CIA have been shown in a wide range of studies, and has added considerably to "our hitherto somewhat patchy knowledge of different stages of interlanguage development (Gilquin, Granger, & Paquot, 2007, p.322).

2.4.3. Previous studies Based on Learner Corpora

Studies based on learner corpora could be broadly divided into two groups. The first is the studies that aim at investigating quantitative differences between native and non-native language, and the second are those of which purpose is the description of the interlanguage features in its entirety (Tono, 2003).

With regard to quantitative differences, research topics include, among many others, adverbial connectors (Granger & Tyson, 1996; Altenberg & Tapper, 1998; Narita, Sato, & Sugiura, 2004), high frequency vocabulary (Shirato & Stapleton, 2007; Altenberg, 2002; Ringbom, 1998), modals (Aijmer, 2002;McEnery & Kifle, 2002), article system (Diez-Bedmar & Papp, 2008), passive construction (Xiao, 2007), use of progressive tenses (Axelsson & Hahn, 2001) and use of discourse markers (Grant, 2010; Polat, 2011). All these studies have led to insights on foreign-soundingness of interlanguage by bringing out the words, grammatical items or syntactic structures that are either overused or underused by learners.

The findings of the studies are generally discussed by referring to the cognitive processes such as L1 transfer, general learner strategies, paths of interlanguage development, intralingual overgeneralizations, input bias and genre/register influences that are effective in the development of interlanguage. In one of the earliest CIA studies, Altenberg and Tapper (1998) examined the use of adverbial connectors in advanced Swedish learners' of written English and compared it with the use in comparable native English writing. The Swedish component of ICLE has been used for the learner corpus and LOCNESS (Louvain Corpus of Native English Essays) has served as the control corpus for the study. The main conclusion of the study is that the advanced Swedish learners underuse the focused conjunctions, the reason of which the researchers explain in connection with general language development and the instructional factors. Using a corpus generated from adult Japanese EFL learners and British National Corpus (BNC), Shirato and Stapleton (2007) have investigated lexical frequency in learners' spoken English. The findings of the study suggest that NNS underused the lexical items representing interactive functions in the following categories: discourse markers, modal items, adjectives for specific evaluations, some interactive words, delexical verbs, hedges, face and politeness, and vagueness. On the other hand, NNS overused some high frequency and auxiliary verbs and some common adjectives with broad denotation. According to the researchers, the reason of these findings can be linked to the impact of training in the research setting. The input bias in interlanguage development has also been underlined in a recent study conducted by Grant (2010) who investigated the use of discourse markers in the spoken interlanguage of New Zealander learners of English. The spoken components of Wellington Corpus (WCS) and BNC have been utilized in comparison. The results have revealed that NNS use identified discourse markers differently not only in terms of frequency but also in terms of communicative functions.

As for the Turkish context, one of the earliest studies within the framework of CIA is carried out by Kilimci (2000) employing a corpus of written English of Turkish students as the experimental corpus and LOCNESS as the control corpus. The aim is to determine the lexical richness in the written production of Turkish learners. The frequency analyses on ten most frequent words show that Turkish learners underuse the article *the*, the preposition *of* and the demonstrative *that,* which the researcher links to the effect of first language. The overall results indicate that Turkish learners' written interlanguage carries the characteristics of informal speech. The use of prepositions in written discourse is another area of research in Turkish context. Kilimci (2002) investigated the constructional and functional properties of prepositions in the essays by advanced Turkish EFL learners and compared them with prepositions in the writings of the native speakers of American English (NS). Results revealed that NNS use prepositions more frequently than NS, and they mostly prefer to use prepositions as complementation of an adjective. NNS are also found to have a tendency towards the use of the prepositions, such as *in*, *from*, *to*, mostly collocating with such high frequency verbs as *come, go, give, live* in verb + preposition patterns. In another study, Kilimci (2003) examined the stance and attitude in advanced Turkish learners' written discourse. From a broader perspective, interlanguage of Turkish EFL learners is compared with Turkish with that of German and French learners in terms of functions and constructions of linking adverbials to explore common interlanguage features. Adverbials in written language have also been the topic of some other studies (Bayrakci, 2004; Özhan, 2012). More recently, Can (2012) conducted a study on the use of stance adverbials in NS and NNS written language. TICLE and LOCNESS served as the data source. The results point to less variety of stance adverbials with higher frequency in Turkish learners' interlanguage compared with NS written discourse.

As mentioned above, the second strand in CIA research is the description of overall interlanguage characteristics at a fixed stage or at different developmental stages. Granger and Rayson (1998) for instance, focused on the automatic profiling with the purpose of revealing the stylistic characteristics of written production of EFL learners compared with native texts. The French speaking learners' corpus from ICLE serves as the experimental corpus, and LOCNESS is the control corpus. The researchers produced word frequency profiles that can demonstrate significant patterns of the over and underuse of major word categories. The findings show that NNS writers overused determiners, pronouns and adverbs significantly, while conjunctions, prepositions and nouns were underused to a great extent. They also conclude that written interlanguage of advanced learners display many of the stylistic features of spoken, rather than written English. Granger and Rayson (1998) propose the input bias and developmental paths as two main reasons to account for this interlanguage characteristics. The characterization of interlanguage development through CIA has also been the focus of the study conducted by Housen (2002). Through a complex analysis of the formal and functional development of verbal system in learners, Housen (2002) aimed at investigating how second language learners of English acquire the English verbal system. A corpus named Corpus of Young Learners Interlanguage gathered from Dutch-speaking and French-speaking learners has been collected for the study. After coding the data for morphosyntactic form, agreement values, tense, aspect, and inherent aspect etc., the analyses were carried out in terms of overuse/underuse of verb categories. Based on the data, Housen (2002) described three formal stages. Stage 1 is invariant default forms. Verbs appear as invariant forms, typically the unmarked base form, but high frequency irregular forms also occur. Stage 2 is non-functional variation. The order of emergence of forms is bare form of the verbs, progressive form of the verbs which is followed by past and participle form of the verbs. Stage 3 is more target-like use of verb morphology to encode tense, aspect and agreement. The patterns of underuse and overuse decrease with increasing proficiency, although there is still variation among different verb forms. The overall results reveal general patterns in the development of the English verbal system and also the variability in development. The patterns of language emerging from the study of actual production data reflect a variety of influences, including language processing, L1 influence, and frequency of forms in the input. With the purpose of discovering whether there is a common pattern of interlanguage development across distinct populations of advanced learners, Cobb(2003) conducted a replication study through a-250000-words written corpus of learner English. Native speaker data were taken as needed from Brown corpus and BNC. One of the findings of the study is that learner language in writing is vague depending on a lack of employment of more nuanced precise vocabulary. There is an overuse of basic vocabulary accompanied by an underuse of more varied richer vocabulary. One explanation offered by Cobb (2003) for this written interlanguage is that learners are using high frequency zone vocabulary which is generally common to spoken language. In Cobb's words, "the learners are employing this restricted lexicon of speech" (p. 403), and this causes learners to produce texts which are vague in meaning and poorly judged in terms of overall writing quality. His findings also suggest that second language learners work through identifiable acquisition sequences. However, "the sequences are not the –ing endings or third person –s, (Cobb, 2003 p.419) as previously thought, but involves more the areas of lexical expansion, word combinations, phraseology, discourse, etc… . Pointing out the advantages of using corpus data in investigating learner English, Cobb (2003) concludes that there is need for further research within CIA focusing on phraseology of learner English beyond morphology and syntax.

In consequence, researching interlanguage through corpus linguistics is quite a promising area which has a lot to offer to the field of SLA. The very nature of the corpus data enables the researcher to approach a wider range of topics and provides a much more diversified view of learner language. As discussed above, in tackling interlanguage topics, the method of CIA has manifested itself as a propitious way, and the studies conducted so far within CIA framework have yielded noteworthy results in terms of interlanguage characteristics. Based on the literature sketched above and the studies along with the others not taken here due to practicality issues (e.g. De Cock, Granger, Leech, & Tony McEnerey, 1998; Ringbom, 1998; Petch-Tyson, 1998; Biber

&Reppen, 1998; Howarth, 1998; Granger, 1998b; Altenberg & Granger, 2002),one probable conclusion to be reached is that there is a shift of attention towards phraseology in interlanguage investigations. The traditional emphasis on syntax and morphology has progressively given way to attention to phraseology, a hitherto neglected aspect of learner language (Cobb, 2003; Granger, 2009). The following section delineates phraseology in connection with learner language and presents the relevant studies.

2.5. Phraseology

One main contribution of the studies within corpus linguistics is the discovery of a pervasive syntagmatic phrasal organization in language use. Analyses based on large collections of authentic texts searchable at the comfort and practicality of computers have demonstrated that much language use is routine (Stubbs, 2007). Moreover, comparison of NNS's language performance with that of NS has revealed that although a particular utterance by NNS is grammatically correct, it still sounds foreign/ unnatural. Gilquin (2011) states that "most learners of a foreign language will be familiar with the experience of being told that a sentence is perfectly grammatical but that a native speaker would never use it" (p.1). One explanation to such situations lie in the field of phraseology which has considerably grown in popularity over the last thirty years or so (Gilquin, 2011; Gries, 2008).

Pioneering figures who contributed greatly to the emergence of phraseology as we know it today are Firth, Fries and Harris. Firth's frequently quoted statement "you shall know a word by the company it keeps" (in Ellis, 2008) has made an important starting point for the field of phraseology. The meaning of this quotation is "at the core of structural linguistics which explored language as a self-contained relational structure whose elemental constructions derive their forms and functions from their distributions in texts and discourse" (N. C. Ellis, 2008 p.1). Fries made a distinction between lexical and structural meaning. In his view of language, language acquisition is described as the learning of an inventory of patterns as arrangements of words with their associated structural meanings. Regarding form and meaning inseparable, Harris developed discovery procedures for phonemes and morphemes. The essence of his view of language is that languages are self-organizing systems in which syntactic features and meaning of a word are determined in relation to associated words. Harris also underlines the exposure to usage in learning the patterns of a language(N. C. Ellis, 2008; Sinclair, 2008).

The idea that lexical and structural meaning is interdependent is further developed by Sinclair who criticised the traditional approach to lexis and grammar (Altenberg & Granger, 2002b).Traditional linguistic favours paradigmatic rather than syntagmatic dimension in analysing language. Text is fundamentally thought to consist of "a series of relatively independent choices of one item after another" (Sinclair, 2004 p. 140). However, analysis of large collections of texts has shown that language contains a wide range of word combinations or multi-word units with varying degree of fixedness. Focusing only on the paradigmatic level in language analysis prevent researchers from understanding the language in full terms. Sinclair (2004) explains this as follows:

> A word gives information through its being chosen (paradigmatic) and at the same time it is part of the realization of a larger item (syntagmatic); in order to observe either of these, however, we lose sight of the other. Unless the requirements of the context are precisely stated, the word as a paradigmatic choice will be invested with far too much independent meaning; on the other hand when observed purely as a component of a larger syntagmatic pattern, it can have very little freedom, and therefore can give very little information (p.141).

In order to have a balanced picture of the language under investigation, both paradigmatic and syntagmatic levels should be considered in language studies, or in Sinclair's (2004) words "The meaning of a text can be described by a model which reconciles the paradigmatic and syntagmatic dimensions of choice" (p.141). This inter-relation of syntagmatic and paradigmatic levels of language is one of the key features in the new corpus based studies. Although lexical studies, for a long time, are based on paradigmatic relations, with the recent revival of interest in phraseology and the development of corpora, now the attention is focused on the analysis of co-occurrences (Altenberg & Granger, 2002b).

Until recently, the field of phraseology has remained as a neglected area in language studies. Sinclair (2008) proposes two main reasons why the developmental process for phraseology has progressed so slowly. The first reason is that phraseology underlines the syntagmatic patterns that do not depend on possible alternatives unlike most grammars. It emphasizes the notion of combinations in language description. The second reason, according to Sinclair (2008), is that phraseology does not make a sharp division between grammar and lexis/semantics. Regarded as a subfield of lexicology for a long time, phraseology has "strong but fuzzy borders with syntax, semantics and morphology (Granger, 2005 p.165), which has caused researchers to vary in their opinions about what to include in the field of phraseology.

Due to these unclear borders, various authors have defined phraseology differently. Glaser (1988) gives the definition of phraseology as "the linguistic description of set expressions whose meanings cannot be derived from the meaning of their parts (as cited in (Mckenny, 2006 p.25). Defining phraseology as the study of the structure, meaning and use of word combinations, Cowie (1998) observes that "studies of collocations have pushed the boundary that roughly demarcates the phraseological more and more into the zone of formerly thought of as free" (p.19). Emphasizing the difficulty of delimiting the borders of phraseology and classifying the types involved, Altenberg (1998) states that "phraseology embraces the conventional rather than the productive or rule governed side of language, involving kinds of composite units and pre-patterned expressions such as idioms, fixed phrases and collocations" (p.101). In the same vein, Hunston (2002) defines phraseology as "the tendency of words to occur in a preferred sequence in naturally occurring language data" (p.138). It comprises all aspects of preferred sequencing as well as the occurrence of fixed phrases. More recently Gries (2008) proposes six parameters to be included in the definition of phraseological units with the purpose of having an explicit path in phraseological research. These are nature and number of elements, frequency of occurrence, distance between elements, lexical and syntactic flexibility, semantic unity and non-compositionality. According to Gries (2008), phraseology is "the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance" (p.6).

As seen in the foregoing definitions, phraseology binds syntax, lexis, semantics, and social usage. Research conducted recently within these areas has carried phraseology from the periphery to the core of linguistic concerns, and its place in linguistic theory has begun to be discussed. Goldberg (2003) is among the leaders who brought phraseology back to the centre of language investigations. In her Construction

Grammar, (Goldberg, 2003; 2006) argues that all grammatical phenomena can be explained as learned pairings of form (including morphemes, words, idioms, partially lexically filled and fully abstract phrasal patterns) and their associated semantic and discourse functions. In Functional Linguistics, Langacker (2000) also has emphasized the associations between particular lexico-grammatical patterns and their systemic functions (in N. C. Ellis, 2008).

Investigations in Cognitive Linguistics have clearly demonstrated how language draws on memory and language patterns. They also show that use and function of language interact with language structure. Gries (2008) analyses the connection between *symbolic units* in cognitive grammar and phraseological research by discussing the concept of *symbolic unit* in cognitive linguistics in relation with the parameters proposed to research the phraseology of a language. He comes to the conclusion that "in terms of what they consider the central units of language analysis, Cognitive Grammar and phraseology research are nearly maximally compatible" (p.14) even though the terminology is not alike.

Similarly, usage-based and constructionist theories of language acquisition have contributed to the revival of phraseology in linguistic investigations. Studies designed especially in first language acquisition have highlighted the significance of phraseological analysis as they have shown that language acquisition starts with phrases and is rich in sequential order; that's to say, formulaic phrases before phonemes, holophrases before words, words before simple sentences, simple sentences before lexical categories and so on (Gries, 2008).

Research in such areas as psycholinguistics and frequency-based theories of language has clearly revealed that language users are sensitive to the frequencies of occurrence of different constructions in the language (N. C. Ellis, 2008). Focusing on the effect of frequency and repetition that ultimately bring about form in language, researchers have found that collocations and formulaic sequences are processed more fluently than openly constructed language (N. C. Ellis, Frey, & Jalkanen, 2009; N. C. Ellis, Simpson-Vlach, & Maynard, 2008a, 2008b; N. C. Ellis, 2002). Sufficient frequency of occurrence is assumed to be a necessary condition for the status of linguistic expressions. The notions of *exposure* and *use* have been emphasized in determining the linguistic system of speakers and hearers (Goldberg, 2006).

By way of interim summary, it can be concluded that investigations and publications in various fields of linguistics has given rise to the advent of phraseology in its current version. Although scholars from different fields have used different terminology and definitions, a closer review has demonstrated that there is a high degree of compatibility between phraseological research and various linguistic theories (Gries, 2008). The area where the degree of both theoretical and practical compatibility with phraseology is notably high is corpus linguistics. Following section presents the relation between phraseology and corpus linguistics along with second language acquisition.

2.5.1. Phraseology, Corpus Linguistics and Second Language Acquisition: An Intersection

It is now widely acknowledged that phraseology lies at the core of a wide range of research areas and all the studies contribute to a better understanding of language in terms of description, acquisition or teaching. Still, what has enabled phraseological research to gain impetus in the present day is the development of corpus linguistics. It has grown in importance in most fields of linguistics, and "it is currently the single most frequently used method employed in the study of phraseology" (Gries, 2008 p.15).

There are several reasons explaining this predominance of corpus linguistics in the field of phraseology. To begin with, corpora provide researchers with the frequencies of occurrence and co-occurrence of the elements in question, thus enabling researchers to have sound statistical data. Granger (2004) appreciates this by stating "frequency lists of two or more word combinations are of great value to the researchers interested in phraseological/routine aspects of interlanguage" (p.127).

Another reason for the overlap between phraseological and corpus-based research is the notion of *pattern* which is central in contemporary corpus linguistics. Hunston and Francis (2000) define patterns of a word as "all the words and structures which are regularly associated with the word and contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it" (p.37). What is striking in this definition is that the concept of *pattern* in corpus linguistics is virtually the same as the definition of phraseology as well as the *symbolic unit* in Cognitive Linguistics and that of *constructions* in usage based theories of language. All these intersections "testify strongly to the fact that phraseology is one of the key concepts in both theoretical linguistics and in the method of corpus linguistics" (Gries, 2008 p.18).

A final overlap between corpus linguistics and phraseological research lies in the prevalent view of language proposed by Sinclair (1991). On the basis of his observations in L1 corpora Sinclair (1991), a leading figure in corpus linguistics, claims that there is no distinction between phraseological patterns and meaning, as well as no distinction between lexis and grammar. In his view of language, Sinclair (1991) argues that there are two principles that organize language: the idiom principle and the open-choice principle. In the open-choice principle, it is stated that language should be regarded as the result of a number of complex choices. This principle is mainly based on the model called *slot and filler* which says that language is composed of a number of slots and the language user has a series of choices to complete them. In Sinclair's (1991) words

> Texts have a series of slots which have to be filled from a lexicon which satisfies local constraints. At each slot, virtually any word can occur. Since language is believed to operate simultaneously on several levels, there is a very complex pattern of choices in progress at any moment (p.109).

While filling the slots, lexical and semantic considerations do not represent a major constraint for the choices. The only constraint is the grammaticalness (Barnbrook, 2007).

In addition to open-choice principle, Sinclair (1991) also proposes idiom principle as "we wouldn't produce normal texts simply by operating the open-choice principle" (Sinclair, 1991 p. 110). Idiom principle accounts for syntagmatic relations between words which cannot be explained in terms of grammar. A lot of phraseological research has shown that words tend to appear together, forming a great number of phraseological units that gain new meaning through their combinations. According to idiom principle, "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair, 1991 p. 110). A series of phrases and semi-fixed phrases are expected to be encountered in specific registers and they should be studied as chunks since the language users apply a co-selection process in producing language. For example, the definite article in the phrase *on the other hand* should be considered as a component of the phrase rather than a grammatical item as its use is not a matter of choice. In this phrase, each constituent is “progressively delexicalized” and the meaning of opposition conveyed by the phrase is spread out across all components of the phrase, rather than limiting itself to an individual word (Hunston & Francis, 2000; Sinclair, 1991).

Drawing from the results of corpus investigations, language is said to be primarily interpreted in the light of idiom principle. Kjellmer (1991) observes that in all kinds of texts, utterances are very largely made with semi-preconstructed phrases. Erman and Warren (2000) assert that about half of the fluent native text is constructed according to the idiom principle. Referring to the prevalence of fixed phrases, Nascimento, Mendes, and Antunes (2006) state that when phrases start to be frequently repeated, these multiword phrases tend to correspond to a conventional way of saying things. The same standpoint is echoed in Mason (2008):

> Most of language will consist of chunks that have occurred before, just as we tend to re-use words and occasionally introduce new coinages. But it is not only the words themselves that we re-use, it is also their contexts, as they are inseparable. And their contexts are effectively multiword units.

Furthermore, phraseological analyses comparing spoken and written corpora suggest that these multiword phraseological units are even more common in spoken language (Biber et al., 2007; Leech, 2000; O’Grady, 2010). Much of communication makes use of fixed expressions retained as formulaic chunks, and the phrase is “the basic level of language representation where form and meaning meet with greatest reliability” (N. C. Ellis, 2008 p.6). Corpus based phraseological analyses also showed that fluent language users rely mostly on memorized language sequences (N. C. Ellis, 2002; Granger & Meunier, 2008a; Sinclair, 1991, 2004b; Wray, 2002). As Sinclair (2008) summarizes, the unit of language and basic starting point for the language investigations is “the phrase, the whole phrase, and nothing but the phrase” (p. 407).

This recognition that most of naturally occurring written and spoken language consists of recurrent phrases many of which are of phraseological nature (Altenberg, 1998; Cheng, Greaves, Sinclair, & Warren, 2008; Kjellmer, 1991; Sinclair, 1991, 2004b; Stubbs, 2001, 2002), and that language users mainly operate on *idiom principle* in producing language has unsurprisingly affected the field of SLA, thus, phraseological trend has recently been considered in the analysis of learner language as well (De Cock et al., 1998). In the literature on learner English, several important reasons form the rationale behind identifying phraseology/recurrent phrases of learner language. Firstly, there is no doubt that learners need many words to deal with everyday requirements in both spoken and written contexts. But the phrases in which words occur are more significant than previously thought (Adolphs & Schmitt, 2003; Biber et al., 2004; Coxhead, 2008). Biber et al., (2004) refer to these phrases as “lexical bundles” and regard them as “basic building blocks of discourse” (p. 371). Wray (2002) sees them as the basis for the development of second language acquisition. Identifying recurrent phrases and comparing them with phraseological units in native speakers’ language is likely to enable tracking the learners’ proficiency development, a central issue in the field of SLA (O’Donnell & Römer, 2009). With regard to the learners’ phraseological knowledge of the target language, Cortes (2004) notes that “use of collocations and fixed expressions has been considered a marker of proficient language use”(p. 398). The same point of view is echoed in Nesselhauf (2005) who assume that increased proficiency correlates with increased use of conventional multi word phrases. Secondly, multi word phrases are essential for fluency in both written and spoken language (Nesselhauf, 2005). Studies in psycholinguistics have shown that human brain is better at memorizing than at processing and “the availability of large numbers of prefabricated units reduces the processing effort and thus makes fluent language possible” (Nesselhauf, 2005 p.3). In his work titled *memory for language*, N. C. Ellis (2001) has suggested that two or more words that co-occur are recorded and treated as a single entity. This is a recursive process which enables language users to encode greater amount of information in short-term memory, thus increasing the efficiency and fluency of communication. Thirdly, the knowledge and the use of phraseological units facilitates comprehension as the language user can understand the meaning without attending to any word (Hunston & Francis, 2000; Nesselhauf, 2005). Non-native users’ selection of word combinations which deviate from the native-like use may be irritable for the listener and may hinder the communication. Therefore, identification of the recurrent phrases in both native and non-native language performance and the analysis of their function could lead to a better understanding and language description, which in turn could be helpful to generate solutions for the language learners.

All in all, as summarized in De Cock (2004), natural language is dominated by recurrence of the words occurring in the same clusters again and again, which is named as the routinized way of expressing things. In his study on the phraseology of spoken English, Altenberg (1998) has voiced the same idea as “the use of routinized and more or less prefabricated expressions is evident at all levels of linguistic organization and affects all kinds of structures, from entire utterances operating at discourse level to smaller units acting as single words and phrases” (p. 120). Comparison of these routinized expressions or in De Cock’s (2004) words “preferred ways of saying things” in native and learner English is likely to yield significant results which could redefine the fields of SLA and ELT as they could shed light to the notion foreign-soundingness and native like selection -the ability of the speaker routinely to convey his meaning by an expression that is not only grammatical but also native like- (De Cock, 2004; Granger, 2004; Shirato & Stapleton, 2007) as well as they have a very valuable contribution to make to pedagogical lexicography within ELT (De Cock, 2004). Appreciating the importance of studying the phraseology of interlanguage through corpus linguistics, various researchers have carried out investigations which are presented in detail in the following section.

2.5.2. Existing Research on Phraseology: Foundations of the Present Study

Review of phraseological research has implied that two main issues differentiating the investigations dominate the literature. One pertains to the approach to identification of the phrasal units, and the other is the vast and various terminology used in the description of the phrasal units.

In the investigations of phraseology in general and the phraseology of learner English in particular, Granger and Paquot (2008) describe two major approaches adopted in identifying phraseological units: a linguistically based approach and a data driven approach. In linguistically based approach that is also called traditional approach, pre identified linguistic criteria are employed to distinguish one type of phraseological unit from another. The multiword units have been placed along a continuum “with the most opaque and fixed ones at one end and the most transparent and variable ones at the other” (Granger & Paquot, 2008 p.28). This approach discards the free combination of multiword units which have only syntactic or semantic restrictions from the realm of phraseology, thus limits the scope of phraseological research (Granger & Paquot, 2008; Granger, 2005). The second approach with regard to identification of the phraseological units is generally referred to as the data-driven or frequency-based approach. It originated with Sinclair’s work on lexicography. Instead of following a top-down approach which requires setting linguistic criteria beforehand, it uses a bottom-up corpus driven approach to identify lexical co-occurrences. This inductive approach generates a wide range of word combinations, which do not all fit predefined linguistic categories. According to Sinclair (2004), it has opened up a “huge area of syntagmatic prospection” (p. 19) covering almost all type of recurrent phrases. Corpus driven studies have shown that many of the multiword units previously considered as outside the scope of the phraseology are central and pervasive in language. In this approach, the emphasis is on phraseological items whatever their nature is, and the researchers are less preoccupied with distinguishing between subcategories of multiword units or setting clear boundaries to phraseological units (Granger & Paquot, 2008; Granger, 2005; Sinclair, 2004b).

In order to successfully integrate phraseology into both theoretical SLA studies and the relevant pedagogical applications, Granger (2005) suggests “reconciling the two approaches” (p.2), which forms the methodological base for the present study as well. Accordingly, starting from a wide notion of phraseology, automatically retrieved recurrent phrases through frequency based/corpus driven approach should be complemented with linguistically defined categories.

Another issue that ensues the review of phraseological research is the abundance of terms used to describe multiword units that is the essence of phraseology as multiword combinations appear in many shapes. Some of them are *chunks* (Beckner et al., 2009; Lindstromberg & Boers, 2008), *prefabs* or *lexical phrases* (Erman, 2007), *phrasal lexemes* (Moon, 1998), *n-grams* (Forchini& Murphy, 2008), *prefabricated patterns* (Granger, 1998c), *formulas* and *formulaic sequences* (Adolphs& Durow, 2004; Schmitt, 2004; Granger & Meunier, 2008; Wray & Perkins, 2000; Wray, 2002), *clusters* (Hyland, 2008b; Schmitt, Grandage, & Adolphs, 2004), *lexical bundles* (Biber et al., 2004; Y.-H. Chen & Baker, 2010; Cortes, 2004; Juknevičienė, 2009; Stubbs, 2007), *preferred sequences* (De Cock, 2004), *multiword constructions* (Liu, 2012) and *recurrent word combinations* (Altenberg, 1998; De Cock et al., 1998; Ebeling, 2011; Ishikawa, 2009; Rodriguez, 2005; Ädel & Erman, 2012) among others. As the labels vary, focus of the studies, the research methods followed and the type of corpora exploited have varied as well.

2.5.2.1. Recurrent Phrases and Written Corpora

One of the earliest studies on the phraseology of learner English is designed by Howarth (1998) with the purpose of analysing the extent to which NNS deviate from NS phraseological forms. He focused on the written production of non-native postgraduate students, from a variety of language backgrounds, studying at British universities. The data for native speaker analysis came from two corpora: social sciences texts from LOB corpus and a corpus of university texts. In analysing the data, Howarth (1998) drew up a collocational framework of three categories: free combinations, restricted collocations and idioms. Comparison of the corpora showed that the most salient category in NS writing is restricted collocation in which a) substitution is allowed for both noun and verb (e.g. introduce/an amendment), b) the choice of noun is restricted, but some substitution of the verb is allowed (e.g. pay/take) c) there is complete restriction on the choice of verb, but some substitution of the noun (e.g. give the appearance/impression). And, these restricted collocations are the most problematic for learners, accounting for their foreign-soundingness. Howarth (1998) explains that as learners' proficiency develops, they memorize a great number of idioms which are learned as fully lexicalized fixed expression and they mostly have no difficulty in using them. However, they make lexical and grammatical errors in selecting co-occurring words in restricted collocations.

In a similar study, Granger (1998b) investigated the frequent word combinations in the written production of advanced learners of English from French L1 background. Through CIA, she used the French component of ICLE as the source of learner language, and three corpora made the base for native language: International Corpus of English (ICE), LOCNESS and LOB corpus. The analyses revealed that EFL learners tend to overuse active structures. The non-nativelike use of English by EFL learners resulted from overuse of some combinations and underuse of others. For example, in her study, the word combination *and so on* was reported to be used almost ten times as often by non-native speakers as by native speakers On the other hand, another word combination *or whatever* was used 11 times as often by native speakers as by non-native speakers. Likewise, sequences with *say* and *think* (*e.g. I think that, we can say that*)were frequently recurring in learner language while they were not frequent in native data. Granger (1998b) links this overuse of the word combinations to the effect of learners' first language.

Cortes (2004) compared the written productions of university students with published journal articles. Her corpus of over 2 million words consisted of two main disciplines; history and biology. She focused on sequences of three or more words that co-occur frequently in a particular register and used the term *lexical bundles* in their classification. After identification of the most frequent four word combinations in both corpora, she classified them structurally and functionally by using the Longman Spoken and Written English (LSWE) taxonomy developed by Biber, Conrad and Cortes (2003). The four categories in this taxonomy, which is also used in the present study, are referential sequences, text organizers, stance bundles, and interactional sequences. The study revealed that students rarely used the lexical bundles identified in the corpus of published writing. In Cortes's (2004) words, “not only was the frequency of the target bundles [i.e. those used by professional writers] used by students extremely low, but also students’ use of bundles did not always convey those functions typically associated with published academic writing in history and biology” (p. 419). She proposes instructional factors as the reason for students’ rare use of bundles.

The discrepancy between NNS and NS phraseological skill was also studied by Ishikawa (2009) within Japanese context. Adopting CIA, Ishikawa (2009) examined the learners’ overuse and underuse of word combinations consisting of two, three and four words under the label *n-gram*s. The learner corpus was Corpus of English Essays Written by Asian University Students (CEEAUS), and Corpus of English Essays Written by Native Speakers (CEENAS) served as the control corpora for the NSs’ written performance. By statistical comparison, the researcher concluded that Japanese learners tend to overuse the phraseologies including the first person pronouns, implying that they tend to discuss the topics from a subjective viewpoint. Similar to French students in Granger (1998), Japanese learners overuse the phrase *I think* as a conventional and convenient sentence starter. The findings also indicate that Japanese learners mainly rely on high frequency idiomatic expressions such as *not only/but also* and *a lot of*, yet they have a limited variety of expressions with which they are familiar. Ishikawa (2009) evaluates this finding in line with the effect of input bias/instructional factors. Comparisons also show that learners prefer negating sentences simply by using *not,* and underuse prepositional phrases while native students alter the wording in negations, and frequently modify nouns and verbs, which suggests learners’ highly limited lexical variety.

Similar conclusions have been reached in the study designed by Juknevičienė (2009) who compared written English of Lithuanian learners at two different proficiency levels with that of native students. She used three corpora (ICLE Lithuanian component, AFK1 corpus consisting of essays by first year university students and LOCNESS for native data) to retrieve four-word lexical sequences. The structural and functional analysis of the lexical sequences were carried out through LSWE taxonomy by Biber et al., (2004) and Biber (2006). Structural classification of lexical sequences showed that language learners of lower proficiency tend to contain more verb bundles while the corpus of native speaker students has yielded a bigger proportion of noun phrases. When the sequences were functionally classified, it was found that lower level learners overuse phrases serving stance and discourse-organizing functions. However, the written language of native speakers is dominated by referential bundles which is reported as the common feature of academic prose. In comparison with lower level students, the use of referential bundles is proportionally higher in the essays of higher level NNS. Juknevičienė (2009) interprets this finding as an indication that "development of written language skills progresses from spoken to written language" (p.61). The overall results imply that NNS learners rely on a limited set of lexical phrases, and emphasize the need for further research on recurrent phrases of other lengths and research on both written and spoken corpora to get a more accurate picture of differences between NNS and NS language.

In the same vein and adopting a frequency based corpus driven approach, Chen and Baker (2010) identify the most frequent word combinations in three written corpora: a sub-corpus from FLOB (academic prose section); BAWE-CH (Chinese students of English); and BAWE- EN (English students). The analyses based on LSWE taxonomy showed that from a structural point of view, recurrent phrases show resemblances in non-native and native student essays. They both have more verb phrase bundles and discourse organizers than native expert writing, whereas, native professional writers exhibit a wider range of noun phrase bundles and referential markers. The findings regarding the functions of four-word combinations demonstrated that while referential expressions dominate the expert native writing, discourse organizers make the largest part of both native and non-native student writing. Stance bundles, on the other hand, are the smallest part in each of the three corpora.

More recently Ädel and Erman (2012) investigated recurrent word combinations in advanced learner writing by L1 speakers of Swedish and in comparable native speaker writing. Focusing on four-word recurrent phrases, the researchers obtained the data form the Stockholm University Student English Corpus (SUSEC), which includes learner and native English. Through quantitative and qualitative analyses of the functions of the phrases, Ädel and Erman (2012) found out that the native speakers have a larger number of types of lexical bundles, which are also more varied, such as unattended 'this' bundles, existential 'there' bundles, hedging and passive bundles. Other lexical bundles which were found to be more common and more varied in the native-speaker data involved negations. With regard to word combinations in learner English, the non-native student writers produced not only fewer types of bundles but also less varied ones. They have restricted repertoire of word combinations both structurally and functionally. Clearly pointing out the importance of investigating recurrent phrases in learner English through analysing more than 60 four-word sequences in detail, Ädel and Erman (2012) come to the conclusion that new directions for interlanguage analysis should cover two, three and more word combinations, and spoken language should also be analysed to get a fuller picture of learner language.

2.5.2.2. Recurrent Phrases and Spoken Corpora

While researching the phraseology of learner language, researchers mainly have worked on the written corpora as collecting and transcribing spoken data is much more difficult and time consuming (Granger, 2004). Still, there are studies examining spoken performance of learners though relatively limited in number.

De Cock et al., (1998) is a first step towards a corpus driven study of the spoken phrasicon of adult advanced EEL learners. The data consisted of 25 informal interviews with learners of French mother tongue, and 25 informal interviews with native speakers of British English. All the interviews followed the same pattern and transcription rules, and were of similar length. The researchers analysed the recurrent phrases performing pragmatic or discourse structuring functions under the term *formulae.* The rationale behind this study was to test if learners foreign sounding arouse from the lesser use of prefabricated phrases as stated by Kjellmer (1991) who claims that "learners' building material is individual bricks rather than prefabricated sections" (p.124). Using special software, De Cock et.al., (1998) extracted all the recurrent word combinations of 2 to 6 word in length in both native and learner corpora. The comparison through CIA showed that that advanced learners do use prefabricated phrases, in fact use them more than native speakers do. However, what distinguishes learners from NS's, these researchers find, is the small number of formulae advanced learners have at their disposal, and the extent to which these are used and overused. In other words, multiword combinations found in the interlanguage data "(1) are not used with the same frequency, (2) have different syntactic uses, and (3) fulfil different pragmatic functions" (De Cock et al. 1998 p.78), thus displaying a non-nativelike usage pattern rather than necessarily using non-nativelike constructions of form.

Similar findings have been echoed in a later study by De Cock (2004)who investigated the use of recurrent sequences of words in NS and NNS speech both from a quantitative and a qualitative point of view. Using French component of LINDSEI corpus and LOCNEC for NS data, De Cock (2004) concentrated on the major functional differences between native speakers' and advanced learners' preferred ways of saying things. The results of the study showed that "advanced learners' use of frequently recurring sequences of words displays a complex picture of overuse, underuse, misuse of target language NS sequences and use of learner idiosyncratic sequences" (De Cock 2004 p. 243). De Cock (2004) highlights the importance of further contrastive studies of recurrent phrases in learner language for pedagogical theory and application, since "not only do they provide us with real NS usage, but they also bring to light the sequences learners appear to find problematic" (ibid.).

Employing Swedish sub-corpus of LINDSEI, Aijmer (2004) also carried out a study on the spoken interlanguage with the hypothesis that learners may overuse or underuse certain phrases in comparison with native speakers and therefore sound non-native. The focus of the study was on the discourse functions of the multiword combinations. He identified a list of word combinations used as pragmatic markers both in learner corpus and native corpus along with their patterning. The overall results suggested that learners employ vague and uncertain markers to express uncertainty or hesitation and not for face-saving or to signal politeness. Markers are also used as strategies when the learners have communication problems.

A more recent study to be cited here is the one conducted by Grigaliūnienė and Juknevičienė (2011) in Lithuanian context. The researchers set out to get the overall picture of recurrent phrases in the Lithuanian learner speech. Using Lithuanian sub-corpus of LINDSEI and LOCNEC as its comparable version, they analysed the 2 to 5 word combinations. Any recurrent sequence of words was included in the automatically retrieved lists without identifying a cut-off point. The analyses yielded a list of 83 different types of recurrent phrases and Grigaliūnienė and Juknevičienė (2011) interpreted the functions of multiword units mostly referring to the literature on pragmatics. They concluded that the speech of the Lithuanian learners of English is quite formulaic; however, the majority word sequences in the data are semantically transparent and their formulaicity is largely determined by pragmatic functions rather than idiomaticity. They also stress the need for designing further studies based on different L1 background to have strong claims about spoken interlanguage.

In conclusion, the literature on phraseology as sketched above and some of the corpus driven studies briefly presented here disclose a lot of valuable information on the importance of recurrent phrases and how they differ both structurally and functionally in different contexts. A closer look reveal that several key issues have been pinpointed in the literature regarding the studies on recurrent phrases in learner language, which also have inspired and shed light to the design of the present study. Accordingly,

1. Majority of studies on recurrent phrases in learner language have been predominantly based on written language data, which implies the need for considering spoken interlanguage in SLA research (Biber et al., 2004)
2. So as to get a better description of interlanguage, it is important to map out differences between different mother tongue populations (Colson, 2008)
3. Four-word sequences are found to be the most researched length for writing studies, probably because the number of 4-word bundles is often within a manageable size for manual categorization and concordance checks (Ädel& Erman, 2012; Arnon & Snider, 2010; Bal, 2010; Y.-H. Chen & Baker, 2010). However, the literature suggests including recurrent phrases of other lengths (2, 3, 4, 5, and 6-word combinations) into the analyses to have a better description of learner phraseology (Ädel & Erman, 2012; Granger, 2005; Juknevičienė, 2009).
4. Among the existing approaches to the identification of multiword units (linguistically based and corpus driven), automatically retrieved recurrent phrases through frequency based/corpus driven approach is favoured(Granger & Meunier, 2008b)
5. Diverse terminology used to refer to different types of multiword phrases is “a direct reflection of the wide range of theoretical frameworks and fields in which phraseological studies are conducted” (Granger & Paquot, 2008

> p.45). What is emphasized in the literature is the clear definition of the multiword units under investigation.

Therefore, addressing to the gaps and issues identified in the relevant literature, this study sets out to analyse recurrent phrases (2 to 6-word units) which are defined as "any continuous strings of words occurring more than once in identical form"(Altenberg, 1998 p.101) through corpus driven automatic extraction of sequences of words. Within CIA framework, this study makes use of four parallel corpora Turkish component of ICLE for learners' written performance, LOCNESS for native speaker written English; LINDSEI for learners spoken performance and LOCNEC for native speaker comparison. The next chapter will also briefly describe methodological concerns prior to the analysis.

2.6. Chapter Summary

This chapter includes four parts: first, an overview of corpus linguistics is provided; second, the relation of corpus linguistics and SLA has been explained with a focus on learner corpora and relevant studies on learner language; then, the review of literature on phraseology is presented; and finally, the connection between phraseology, corpus linguistics and SLA is elaborated. The chapter ends with the existing studies which combine these three fields of study.

CHAPTER III

METHODOLOGY

3.0. Introduction

This chapter presents a description of the methodology followed to find out answers to the research questions of the present study. This chapter includes four sections. The first section provides information on the nature of the present study through explaining key concepts in the methodology. The second section delineates the corpora employed in the investigation. The next section describes the method followed for the identification and classification of the recurrent phrases and frequency distributions. Finally, section four focuses on the analysis procedures including the description of the software used in data extraction and detailed explanation of the taxonomies used in the analyses.

3.1. Nature of the Study

The present study is aimed at unravelling, describing and comparing and contrasting recurrent phrases in spoken and written English of learners from Turkish L1 background. To have sound claims about the learner language, comparable native speaker data are also included in the study. The investigation covers two-, three-, four-, five- and six-word sequences that occur at certain frequency thresholds.

As in any other empirically based study, the choice of material in a corpus study is crucial for the validity and scope of the results it provides. Setting out to describe and explain the occurrence of recurrent phrases in learner language, several points need to be handled in order to make sure that the chosen material and the method employed are appropriate for getting answers to the research questions. According to Granger (1998), second language acquisition research in general has as its main goal to "uncover the principles that govern the process of learning a foreign/second language", and as this process "is mental and therefore not directly observable, it has to be accessed via the product, i.e. learner performance data" (p. 4). Similarly, Ellis and Barkhuizen (2005) assert that "all researchers who accept the primacy of learner language as data for investigating L2 acquisition accept that learners' use of the L2 in some way reflects their L2 competence/proficiency" (p. 364). As stressed in chapter 2, evidence from authentic material is the foundation of many linguistic studies today, which emphasise that "it is important to base one's analysis of language on real data – actual instances of speech or writing – rather than data that are contrived or 'made-up'" (Meyer, 2004 p.xiii). The present study makes use of such data relying on learner corpora collected for the purposes of investigating learner language. Another point to be considered is the representativeness of the data and the variables at work, which determines the validity of the corpus results. Every language situation includes a variety of variables connected to their subjects and settings, and these variables need to be accounted for if results are to be generalized to a broader language population, as well as to allow for replication and comparison of studies. Bearing these issues in mind, this study employs corpora complied according to very strict and explicit design criteria, which makes it possible to have claims on representativeness and generalizability of the results (Gilquin & De Cock, 2011).

Although the nature of corpora may restrict the number of methods available to investigate it, an electronic corpus generally offers many and diverse possibilities in terms of analysis. As Gries (2010) observes, "branches of linguistics that have been using corpora or text databases have always been among the most quantitatively oriented sub-disciplines of the field" (p. 5), and frequency counts of words or word-combinations are at the centre of most corpus studies. Therefore, the present study adopts a quantitative approach to inquire recurrent phrases in learner writing and speech. Based on Granger's (1998b, 2009) *Contrastive Interlanguage Analysis* which is explained in chapter 2, this study is contrastive in nature, and adopts the quantitative methodology suggested in Altenberg (1998), which is discussed in detail in section 3.3. However, in order to avoid presenting results in decontextualized numbers, this study also integrates qualitative analysis. Qualitative corpus analysis provides rich insights about the language phenomena and allows for classification of the linguistic forms (Hasko, 2011). As mentioned earlier, the qualitative part of the analysis includes an identification of functionality of the recurrent phrases. In brief, this study is a mixed research as it integrates both quantitative and qualitative analysis in the interpretation of the data.

3.2. Material: General Considerations

What distinguishes corpus data from other data types used in SLA research is the authenticity of the texts and the representativeness of the language population under investigation. These properties can be attained through strong control of the variables which have profound influence on the validity of the results. Granger (2008) argues for setting strict design criteria as “learner language is influenced by a wide variety of linguistic, situational and psycholinguistic factors, and failure to control for these factors greatly limits the reliability of findings in learner language research” (p.263). Granger (1998a) furthermore underlines that “it is especially important to have clear design criteria in the case of learner language, which is a very heterogeneous variety: there are many different types of learners and learning situations” (p.7). In order for a corpus analysis to produce valid and reliable results, these factors should be controlled in gathering data. With regard to collecting an interlanguage corpus in particular, Granger (2008) identifies two sets of variables to be considered; one pertaining to *learner* (including age, gender, region, mother tongue, learning context, proficiency level, time spent in an L2 country) and the other to the *task* or *situation* (medium, field, genre/text type, topic, task setting, length). These variables are especially significant to account for in a contrastive analysis, as a contrast may only provide interesting results if the differing variables between the corpora compared are known. Being a contrastive study, this research made use of four corpora as shown in Table 4.

Table 4

Corpora under Investigation Summarized

	Written	**Spoken**
Native	LOCNESS	LOCNEC
Non-native	ICLE (Turkish sub-corpus TICLE)	LINDSEI (Turkish sub-corpus LINDSEI-TR)

International Corpus of Learner English (ICLE) and The Louvain Corpus of Native English Essays (LOCNESS) were employed for the written English; The Louvain International Database of Spoken English Interlanguage (LINDSEI) and The Louvain Corpus of Native English Conversation (LOCNEC) were compared for the spoken English. The following section describes these corpora in detail in connection with the foregoing variables.

3.2.1. ICLE v2 and TICLE

ICLE is a large-scale database of written learner English, and it is the counterpart of LOCNESS (Granger et al., 2009). It is a pioneering project in the collection of learners' writing and probably the one which has engendered most international collaboration based at the Catholic University of Louvain. Initially, it was planned to comprise written learner language of advanced native speakers of French learning English, but soon it was expanded to a joint corpus that can serve as a research base for analysing the very nature of interlanguage across countries and investigating whether any features of learner language are universal or affected by the speakers' L1, previously learned foreign languages, educational background etc… (Pravec, 2002). For this reason, learner and task variables were included in the creation of the corpus.

3.2.1.1. Learner Variables

With regard to learners, six of the eight variables above were clear enough; that is age, gender, mother tongue, region, knowledge of other foreign languages and time spent in an English-speaking country, yet the variables related to learning context and proficiency level are fuzzy (Granger et al., 2009).

The data were collected from the undergraduate students in their third and fourth year of university education, and the average age of the students is 22.30 (Granger et al., 2009). The gender distribution varies among national sub-corpora. Table 5 displays the age and gender proportions in ICLE v2. The average age for Turkish learners is 22.08, and most of the participants are female as shown in Table 5.

Table 5

Age and Gender Distribution in ICLEv2

Sub-Corpus	**Average Age**	**Gender Distribution**		**Sub-Corpus**	**Average Age**	**Gender Distribution**	
		Female	**Male**			**Female**	**Male**
Bulgarian	20.55	83%	17%	**Japanese**	20.06	73%	27%
Chinese	20.49	64%	36%	**Norwegian**	23.94	74%	26%
Czech	22.07	72%	28%	**Polish**	23.39	80%	20%
Dutch	20.75	73%	27%	**Russian**	21.19	84%	16%
Finnish	22.73	85%	15%	**Spanish**	21.72	86%	14%
French	21.70	88%	12%	**Swedish**	27.73	77%	23%
German	23.39	78%	22%	**Turkish**	22.08	81%	19%
Italian	24.59	92%	8%	**Tswana**	22.47	60%	40%
ICLEv2	22.30	76%	24%				

Table 5 also shows 16 mother tongue backgrounds represented in ICLEv2. Beside mother tongue, the languages used at home were asked, and the responses were recorded as the *language at home* variable. The languages spoken at home were classified as first, second and third language according to the proportion of use. Table 6 presents the proportion of the languages used at home for the Turkish sub-corpus (TICLE).

Table 6
Distribution of Mother Tongue and Languages at Home in TICLE

Mother Tongue	Distribution	Per cent (%)
Turkish	276	98%
Other	4	1.4%
The First Language at Home	Distribution	Per cent (%)
Turkish	271	96.8%
Other	9	3.2%
The Second Language at Home	Distribution	Per cent (%)
Turkish	8	2.9%
German	2	0.7%
English	1	0.4%
Other	14	5.0%
No Second Language	255	91.1%
The Third Language at Home	Distribution	Per cent (%)
No Third Language	280	100%

The variable *region* is relevant for the languages that are spoken in more than one country such as Chinese (Mainland China and Hong Kong), Dutch (Belgium and The Netherlands), German (Germany, Austria and Switzerland), Swedish (Sweden and Finland) (Granger et al., 2009). As for the Turkish corpus (TICLE), all the language data were collected in Turkey from the 3rd and 4th year learners of English studying at ELT departments in three universities (Çukurova University, Mersin University and Mustafa Kemal University); therefore, region variable does not apply to TICLE (Kilimci & Can, 2009).

Apart from native language, knowledge of foreign languages other than English was included in the learner variables as well. This variable is significant to find out the factors (if any) influencing the interlanguage development in the learners as any deviation in the interlanguage may result from the knowledge of other languages. When the entire corpus of ICLEv2 is considered, German is in the first rank with the rate of 32% after English and French follows it with the rate of 23%. The same order has been observed in TICLE as well. Table 7 displays TICLE data with regard to other foreign languages being learnt.

Table 7

Knowledge of Other Languages: Distributions and per cents in TICLE

Second Foreign Language	Distribution	Per cent (%)
German	260	92.9%
French	15	5.4%
Other	5	1.8%
Third Foreign Language	Distribution	Per cent (%)
-	254	90.7%
French	10	3.6%
German	5	1.8%
Spanish	1	0.4%
Dutch	1	0.4%
Other	9	3.2

Time spent in an English-speaking country is another learner variable. Large proportion (45%) of learners in the entire ICLEv2 corpus reported they did not stay in an English-speaking country, while 23% reported a period of three months of stay or more and 19% a stay of less than 3 months. 13%of the participants were recorded as unknown (Granger et al., 2009). Table 8 shows the relevant data for TICLE.

Table 8

Time Spent in an English Speaking Country: TICLE

Time spent in an English-speaking country	Distribution	Per cent (%)
-	276	98.6%
3-6 months	2	0.7%
6-12 months	1	0.4%
10 years and over	1	0.4%

As shown in Table 8, almost none of the Turkish learners stayed in an English-speaking country. Only 2 learners reported that they spent 3 to 6 months in a foreign country and only 1 stayed in an English-speaking country for 1 year.

Learning context is a variable which is described as a ‘fuzzy’ variable by Granger et al., (2009). All the learners in ICLE corpus have learned English in a non-English-speaking country, which leads to the use of label of ESL rather than EFL in the description of their English. Granger et al. (2009) state that the line between EFL and ESL can be extremely fuzzy, because the level of exposure to English changes as it may be limited in some countries while extensive in some others. However, what is certain is that learners in the entire ICLEv2 corpus and in TICLE have learned English primarily in classroom setting.

Proficiency level of the learners is the last learner related variable considered in the design of ICLEv2. Proficiency levels are crucial to have a generalizable picture of EFL learners. The initial purpose in gathering the ICLE corpus was to collect the data from advance level learners studying at the third and fourth grade of universities. In order to see if these assumed proficiency level (i.e. the third and fourth year university students are considered advanced learners) is valid for all the learners in the sub-corpora, 20 essays from 16 sub-corpora was randomly selected. The essays were rated by a proficient rater on the basis of Common European Framework of Reference for Languages (CEF) descriptors of writing. Table 9 presents the results.

Table 9

CEF Results-20 Essays per Sub-Corpus

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

According to CEF results, 60% of sample essays were rated as advanced (C1 or C2). The proportion is much higher in some sub-corpora (e.g.100% Swedish) but it can be low as 10% or less in others. Granger et.al., (2009) point out that "although these results need to be firmed up on the basis of more rigorous assessment methods, they are clear indication that some of ICLEv2 sub-corpora are rather in the higher intermediate range while others clearly qualify as advanced" (p.11). With regard to TICLE sub-corpus, 20% of the randomly selected sample essays were rated as advanced while 80% was identified as intermediate and lower intermediate.

3.2.1.2. Task Variables

Main task variables taken into consideration in gathering the ICLE corpus are medium, field, genre/text type, length, topic and conditions. Such criteria as medium (writing), genre (academic essay), field (general English) and length (between 500 and 1000 words) were identified by the ICLE project directors (Louvain Centre for English Corpus Linguistics) while the choice of topic and task setting which include time arrangements, exam conditions and the use of reference tools were left to the national teams carrying out data gathering.

With regard to topics of the essays, ICLE team provided students with a very range of options to aid the students in their writings. The students were free in writing on the topics of their own interests as well. Following is the information about top ten most popular topics in ICLE and their distribution in accordance with the sub-corpora.

Table 10

Top Ten Essay Topics in ICLE

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

Some topics recur in the corpus since many national coordinators used the list of suggested topics provided by the ICLEv2 coordinator team in Louvain (Granger et al., 2009). As seen in Table 10, the written texts forming the Turkish sub-corpora are argumentative essays on university education.

The conditions of task setting are identified as: if the task was timed or untimed, if it was part of an exam, and if the students were allowed to use any reference source to write their essays. Table 11 presents data regarding the variable of task setting for ICLE in general and for TICLE.

Table 11

Task Setting Variables for ICLE and TICLE

Variable	**ICLE**		**TICLE**	
	Number of Essay	**Distribution**	**Number of Essay**	**Distribution**
Essay Type: Argumentative	5554	100%	280	100%
Time:				
Untimed	3662	65.9%	280	100%
Timed	1683	30.3%	-	-
Unknown	209	3.8%	-	-
Reference Source:				
Allowed	2510	45.2%	146	52.1%
Not Allowed	2846	51.2%	134	47.9%
Unknown	198	3.6%	-	-
Exam Condition:				
Written As a Part of an Exam	1594	28.7%	-	-
Not Written as a Part of an Exam	3600	64.8%	280	100%
Unknown	360	6.5%	-	-

Adopted from “İkinci dil edinimi çalışmalarında bilgisayar destekli bir Türk öğrenici İngilizcesi derlemi: ICLE'nin bir altderlemi olarak TICLE by Can, 2009 Dil Dergisi 144 p. 25.

As shown in Table 11, there is a link between timing and exam conditions. Timed essays were generally written as a part of an exam. Most of the essays creating the ICLEv2 corpus were untimed (65.9%) and they were not written under exam conditions (64.8%). Almost half of the participants (51.2%) were not allowed to use a reference source while writing their essays (Granger et al.,2009). As for TICLE corpus, all the essays in TICLE were of argumentative type written without time limit and they were not produced under exam conditions. More than half of the students (52.1%) used a reference source in writing their essays.

Finally, length of the essays varied in accordance with the sub-corpora in ICLEv2. Table 12 displays the length of essays in terms of the number of words both for ICLE in general and for TICLE in particular.

Table 12

Length of Essays in ICLE and TICLE

Number of Words	ICLE		TICLE	
	Number of Essays	**Distribution**	**Number of Essays**	**Distribution**
<=200	52	0.9%	-	-
200< - <=500	1837	33.1%	3	1.1%
500< - <=1000	3274	58.9%	272	97.1%
1000< - <=1500	340	6.1%	5	1.8%
1500< - <=2000	41	0.7%	-	-
>2000	10	0.2%	-	-

Adopted from “İkinci dil edinimi çalışmalarında bilgisayar destekli bir Türk öğrenici İngilizcesi derlemi: ICLE'nin bir altderlemi olarak TICLE by Can, 2009Dil Dergisi 144 p. 26.

According to Table 12, most of the texts in ICLEv2 corpus contain 500 – 1000 words, and almost all the essays (97.1%) in Turkish sub-corpus are 500 – 1000 words in length. The average length of the TICLE texts are identified as 713 words (Granger et al., 2009).

3.2.2. Reference Corpus: LOCNESS

As stated in earlier sections, comparison of interlanguage performance with native speakers’ performance is crucial to have a better understanding of interlanguage properties. Using an experimental and a reference corpus is the essence of Contrastive Interlanguage Analysis (CIA) which forms the methodological base for the present study. This study has made use of Louvain Corpus of Native English Essays (LOCNESS) as the reference corpus.

LOCNESS project started with the purpose of having a mirror corpus of ICLE to ensure the comparability of the ICLE data with the native English data. It was compiled by the Centre for English Corpus Linguistics at the Catholic University of Louvain, Belgium, and it has been used in many studies since 1998. It contains essays written by British and American native speakers during the period of 1991-1995. The LOCNESS corpus consists of four components: essays of British A-level students (60.209 words), essays of British university students (95.695 words), argumentative essays of American students (149.574 words) and literary-mixed essays of American students (18.826 words). The total number of words in the entire LOCNESS corpus is 324.304 (LOCNESS, 2010).

In this study, the component of "argumentative essays of American students" was chosen as a comparison base since the experimental corpus TICLE contains only argumentative essays. The essays in LOCNESS American component are on different topics, but they are all argumentative rather than narrative, descriptive or expository prose. The essays were untimed and students had access to the reference library of their university. (Any direct quotations in these essays were removed from the corpus and marked thus: <*>). With the exception of eight students, the age of the participants ranges from 17 to 22, which is similar to ICLE corpus. The length of essays (500< - <=1000) is also similar to the texts in ICLE. Table 13 shows the institutions where the texts were gathered and the properties of the students and the texts in the selected component of LOCNESS corpus.

Table 13

General Distribution in the Selected Component of LOCNESS Corpus

Institution	**Codes**	**Age Range of the Students**	**Number of Essays**	**Number of Words**
Marquette University	ICLE-US-MRQ	18-21 *	46	54.285
Indiana University at Indiana Polis	ICLE-US-IND	17-20 **	28	13.454
Presbyterian College, South Carolina	ICLE-US-PRB	20-22	6	12.447
University of South Carolina	ICLE-US-SCU	17-21***	53	52.885
University of Michigan	ICLE-US-MICH	18-21	43	16.502
TOTAL			**176**	**149.573**

* Three students are between the ages 30-40

** Two students are between the ages of 30-60

***Three students are between the ages of 30-60

3.2.3. LINDSEI and LINDSEI-TR

LINDSEI is the spoken counterpart to ICLE, and it is the first large-scale corpus of spoken learner English. As a project, it was started in 1995 in the Centre for English Corpus Linguistics, University of Louvain, Belgium. It aimed to provide a corpus containing oral data produced by advanced learners of English from several mother tongue backgrounds. The first component was gathered from the French mother tongue learners of English, and it contained transcripts of 50 interviews with a total of about 1000 words of learner language (Gilquin, 2012). The project has been expanded with the inclusion of other mother tongue backgrounds. To date, a total of 20 sub-corpora with different L1 backgrounds are in the project, and 11 of them have been completed and made available to public use, yet the others are in progress (Gilquin, 2012)

All the sub-corpora follow the same structure with at least 50 interviews made up of pre-identified tasks. The interviews are transcribed and marked-up using the same guidelines (see Appendix A), and each interview is accompanied with a report including information about learner and task variables. This information enables researchers to study the possible impact of certain factors on learner language (Gilquin, 2012).

3.2.3.1. Learner Variables

In gathering the LINDSEI corpus, very strict and explicit design criteria have been followed, which allows for the cross-linguistic contrastive studies. A total of eight variables have been taken into consideration, among which the variable of “mother tongue” has formed the basic division for the sub-corpora. Therefore, it is possible to assume that any differences found between the sub-corpora mainly result from the effect of mother tongue; or in the case of comparison between LINDSEI and its native counterpart LOCNEC, the differences are possible to be attributed to the proficiency of the two populations.

The Turkish component of LINDSEI (LINDSEI-TR) was compiled at Çukurova University from the third and fourth year students studying at the department of English Language Teaching. The external criteria identified by the LINDSEI team state that all the interviewees should be university undergraduates in English; hence, they were labelled as *advanced* in terms of language proficiency. A total of 58 students were interviewed according to the guidelines provided by the project team. Table 14 presents the age and gender information for the interviewees participating LINDSEI Turkish component.

Table 14

Age and Gender Distribution in LINDSEI-TR

	Average Age	**Gender Distribution%**	
LINDSEI Turkish		Male	Female
	22.3	32%	68%

Home language is another variable that was taken into consideration in gathering the learner information. Almost all the interviewees reported Turkish as the home language; and only eight students out of fifty-eight stated that they use language other than Turkish at home.

Regarding interviewees' educational background of English, two questions were asked: one was related to university education and the other covered the years before starting the university. Table 15 shows the distribution.

Table 15

Educational Background in English: LINDSEI-TR

	Years	Number of Students	Distribution %
Before University Education	< 7	4	6.8 %
	7-10	35	60.3 %
	10-13	18	31.3 %
	>13	1	1.7 %
University Education	4	42	72.4 %
	5	15	25.8 %
	>5	1	1.7 %

As seen in Table 15, majority of the interviewees were taught English at least for seven years before attending to the university. The average years of English at school before university for all interviewees are 9.10. When the university education is considered, almost all of the students participating in the interviews took 12-14 years of English instruction. The interviewees are considered advanced students of English, which is a term used to define not their language proficiency but their status.

Knowledge of other foreign languages was also included in the learner variables as it is an important factor which could be used to account for any deviation in the interlanguage development. Table 16 displays the knowledge of foreign languages for the interviewees in LINDSEI Turkish corpus.

Table 16

Knowledge of Other Foreign Languages in LINDSEI-TR

Second Foreign Language	Number	Per cent (%)
German	44	75.8 %
French	12	20.7%
Other	2	3.5%
Third Foreign Language	Number	Per cent (%)
-	47	81.3%
French	3	5.1 %
German	3	5.1 %
Dutch	3	5.1 %
Arabic	1	1.7 %
Hungarian	1	1.7 %

Accordingly, German is in the first rank among the languages being learnt by the interviewees, which is followed by French with the per cent of 20. Regarding the third foreign language, almost all the students responded negatively.

The last variable considered is time spent in an English speaking country. Only eight students of all the participants reported that they stayed in an English speaking country beforehand. Five of them spent 3 months while the others stayed abroad only for 1.5 months.

3.2.3.2. Task Variables

Deciding on the type and content of the tasks and the conditions in task setting in eliciting data for learner language is of vital importance in terms of the validity of the results, especially to provide the construct validity that refers to the extent to which a study is measuring what it set out to measure. Ellis and Barkhuizen, (2005) claim that in the collection of learner data, the construct validity "is best established by demonstrating that the performance it taps reflects, as far as possible, the kind of use for which language is designed and acquired" (p. 21). With regard to designing tasks to elicit learner language, Granger (1998a) states that "the artificiality of an experimental language situation may lead learners to produce language which differs widely from the type of language they would use naturally" (p.5). So as to reach valid conclusions about natural language use, then, it seems that researchers should design elicitation tasks which enable learners to use naturally occurring language as much as possible.

However, eliciting natural language use in speech of learners may pose problems for compilers of learner corpora as learners “rarely use the target language to go about their normal business” (Granger, 2008 p. 261). In fact, in any situation where learners speak their second language, they will likely to perform higher level of language-consciousness. Hence, Granger (2008) suggests a “naturalness continuum” for designing elicitation tasks, with informal interviews ranking highest. Through informal interviews which made up the important proportion of LINDSEI corpus, learners are allowed “to choose their own wording rather than being requested to produce a particular word or structure” (Granger 2008, p. 261).

The informal interviews in the LINDSEI corpus lasted about fifteen minutes each, and they were recorded non-surreptitiously. The length of each interview was approximately 2,000 words. For the Turkish component of LINDSEI, the average duration of the interviews was recorded as 13 minutes with the shortest interview being 10.41 minutes and the longest being 18 minutes. The whole Turkish component contains 81,711 words with an average length of 1,408words per interview.

Each interview followed the same set pattern: at the beginning, the interviewee was requested to choose a topic among three topics which were provided in written form just like below:

Topic 1: An experience you have had which has taught you an important lesson. You should describe the experience and say what you have learnt from it.

Topic 2: A country you have visited which has impressed you. Describe your visit and say why you found the country particularly impressive.

Topic 3: A film/play you've seen which you thought was particularly good/bad. Describe the film/play and say why you thought it was good/bad.

The students were allowed a few minutes before the conversation to plan what they were going to say, but they were requested not to make any written notes or to use a dictionary to ensure the spontaneity as it was intended for the spoken productions to be as spontaneous as possible (De Cock, 2004). They were told that they would need to be able to talk about it for a few minutes. After the students had spoken for a while, the interviewer became involved by asking questions related to what the student had said, and by raising general topics such as life at university, hobbies, future plans etc… to start the second part of the interview. The last part of the interview was picture description for which the interviewee was provided with a set of pictures (see Appendix B) which made up a story and asked to retell the story. The interviews were transcribed using a broad orthographic transcription scheme (see Appendix A). Table 17 summarizes the task variables in LINDSEI corpus.

Table 17

LINDSEI Task Variables Summarized

Task Variables	LINDSEI	
Medium	Spoken	Typically lower level of self-monitoring than written registers; spoken corpora have not been extensively collected and analysed previously
Field	Education /Academia	Typical learner environment that is familiar to the interviewees
Genre/Text Type	Informal Interview	Similar to spoken conversation due to the informality. The restrictions makes for valid comparisons with sub-corpora
Topic/Task	A personal topic set beforehand, informal chat prompted by the interviewer, picture description (see Appendix B)	Encourages implicit performance (attention to topic rather than language), few constraints on language use, close to natural linguistic behaviour, enough restrictions for sub-corpora to be comparable
Conditions	No reference tools available; non-surreptitious recording; each interview should last for at least 15 minutes	Absence of reference tools creates a more authentic situation, and promotes continuous language use and topic awareness rather than explicit attention to form

Even though some kind of thematic control was exercised on learner performance through the previously set topics, the overall task may be considered linguistically ‘open’, since the learners are allowed to choose linguistic form, and since extracting certain grammatical features is not the primary aim of the data collection. The continuity of the interview as well as the few constraints on the subjects in terms of content and form, makes it justifiable to consider data from LINDSEI data for investigating characteristics of natural conversation, and to compare findings and explanations in terms of the spoken interlanguage.

3.2.4. Reference Corpus: LOCNEC

LOCNEC, the Louvain Corpus of Native English Conversation, was compiled as part of the interlanguage research project. It is the mirror image of LINDSEI as the comparable corpus of native speaker English. Following the same design principles as LINDSEI, the interviews were carried out at Lancaster University, UK. The interview sessions were recorded non-surreptitiously, and they were not used for any sort of external assessment of the participants who are all university students majoring in English. As to the content of the conversations, the same procedure as LINDSEI was followed. Namely, the interviewees were first introduced general topics identified beforehand, which is proceeded with the follow-up questions depending on what the interviewees had said. The last part of the questionnaire included making-up a story based on the given pictures. The LOCNEC interviews makeup a total number of 161,725 words and learner turns only consist of 118.553 words.

3.3. Method

3.3.1. Quantitative and Qualitative Corpus-Driven Method

Working with the corpus data allows researchers to look at larger bodies of texts at the same time, and investigate the quantitative aspects of language with relative ease. Thus, quantitative view of data both in terms of data size and quantitative searches is the distinguishing feature of corpus studies. However, in addition to quantitative analysis, qualitative evaluations should also be carried out to avoid presenting data as decontextualized numbers. Leech (2000) notes that:

> In representing grammatical differences as used in different subsections of a corpus [or different corpora], we have to make use of quantitative methods. In relating these quantitative differences to factors external to language, on the other hand, we depend on qualitative analysis (Leech 2000: 693).

Granger (1998a) identifies two major approaches to learner corpora: hypothesis-based and hypothesis-finding. Hypothesis-based studies rely on pre-existing ideas, “generated through introspection, SLA theories, or as a result of the analysis of experimental or other non-corpus-based sources of data" (Barlow, 2005 p.344) the hypothesis-finding corpus researcher, on the other hand, "may simply decide to gather data and quantify everything he or she can think of just to see what emerges" (Granger 1998a p.15). Granger (1998b) notes that "this approach [hypothesis-finding approach] is potentially very powerful since it can help us gain totally new insights into learner language" (p.16). Likewise, Gries (2010b) distinguishes between corpus-based and corpus-driven linguistics, stating that similar to hypothesis-finding approach, corpus-driven studies "aim to build theory from scratch, completely free from pre-corpus theoretical premises" and "base theories exclusively on corpus data" (p.328).

Biber (2010) regards "corpus-driven approach" the best known approach used to describe the overall patterns of variation in spoken and/or written language. As opposed to *corpus-based* approach which is used to refer to "a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (Tognini-Bonelli, 2001 p.65), in a *corpus-driven* approach "the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence" (Tognini-Bonelli, 2001 p.84). Hence, the corpus is regarded more than a repository of examples to support predefined theories. Recurrent patterns and frequency distributions are expected to form the basic evidence for linguistic categories. In other words, linguistic constructs emerge from analysis of linguistic co-occurrence patterns in the corpus (Biber, 2010; Tognini-Bonelli, 2001). Even, the absence of a pattern is considered potentially meaningful. This inductive approach could provide a wide range of word-combinations or in Sinclair's words (2004) " it has opened a huge area of syntagmatic prospection" (p.19) covering almost all types of word combinations.

Focusing on the recurrent phrases in learners' written and spoken language, this study has adopted a "corpus-driven recurrent word combination" method in the identification of co-occurrences. Recurrent word-combination method was first proposed by Altenberg (1998) in his work on the phraseology of spoken NS English in London-Lund corpus, and used in a number of studies on learner language (De Cock, 2000, 2004; Juknevičienė, 2009). This method involves automatic extraction of sequences of word forms of length *n* which recur in identical form with frequency greater than *m* from a corpus using specialised software. The length of word combinations and the frequency thresholds are identified by the user. It doesn't presuppose any linguistic category or pre-established sequences, thus, generates a wide range of combinations, which enables researchers to see the learner language from a multiple point of view (Granger & Paquot, 2008) as discussed in chapter 2. De Cock (2004) underlines that recurrent word-combination method is particularly suitable for interlanguage investigation as there are no widely agreed upon list of sequences to start with. Moreover, results obtained through automatic extraction could make powerful starting point "as they arguably lead the researcher to take into consideration a series of frequently used clusters he or she may otherwise have overlooked because of their lack of psychological salience" (p.228).

All in all, within a corpus-driven approach, this study is based on quantitative and qualitative methodology as it is concerned with how many times recurrent phrases occur in native and non-native English corpora, and how they function linguistically.

3.4. Data Analysis

3.4.1. Quantitative Analysis

The initial analysis of this study is quantitative inspired by the recurrent word combination method. This first step identifies the recurrent phrases in the material along with the quantitative deviations or similarities between the native- and non-native corpora. Recurrent phrases and frequencies are extracted from the corpora using WordSmith Tools v5.0 (Scott, 2010).

Analytical software WordSmith v5.0 is an integrated suite of programs for evaluating how words behave in text. It is capable of automatically retrieving recurring words, letters and spaces and of instantly tallying their frequencies. It includes a number of text-handling tools to support quantitative and qualitative textual data analysis. **Wordlist** tool gives information on the frequency and distribution of the vocabulary - single words but also word sequences used in one or more corpora. Wordlists for two corpora can be compared automatically so as to highlight the sequences that are particularly salient in a given corpus, i.e., its keywords or key word sequences. The Concordancer, **Concord** is used to analyse the co-text of a linguistic feature, i.e., its linguistic environment in terms of preferred co-occurrences and grammatical structures. With Keyword tool, the researcher is able to identify key words in a text. More sophisticated tools are currently being developed to help researchers explore large corpora.

Within WordSmith, users can specify a window within which frequencies of occurrences can be calculated. For the present study, two-, three-, four-, five- and six-word sequences that occur at least 12, 6, 4, 3 and 3 times respectively in the NS or NNS corpus were investigated. The investigation covered recurrent phrases of two- to six-word sequences because this study aimed to investigate learner language from a broader perspective. Most of the investigations of recurrent phrases in NS and learner speech and writing are restricted to one specific sequence length (e.g. Ädel and Erman, 2012; Biber, Conrad, and Cortes, 2004; Chen and Baker, 2010; V. S. Cortes, 2002; V. Cortes, 2004; Hyland, 2008; Ping, 2009; Rafiee, 2011 focus on four-word sequences; Adolphs and Durow, (2004) concentrate on three-word sequences find some more references) as discussed in chapter 2. Different frequency thresholds were set for each sequence because the length of recurrent word combinations is inversely related to their frequency (Altenberg, 1998; De Cock, 2004). Following De Cock (2004), the frequency thresholds were also scaled so that approximately 10% - 12% of recurrent sequence types are taken into consideration for each length. This provided to have at least some guarantee that the sequences have some currency in NS and NNS writing and speech (Altenberg, 1998).

In the presentation and discussion of the frequency of word combinations that recur in the corpora under investigation, the following system illustrated in table 18 was adopted from De Cock (2004).

Table 18

Frequency of Recurrence of Investigated Sequences

Symbol	Frequency
--	not recurrent at or above frequency threshold
◇	recurrent sequences occurring less than 10 times per 100,000 words **N.B:**(3-word sequences recur at least 6 times, 4-word sequences recur at least 4 times, 5- and 6-wordsequences recur at least 3 times)
◆	recurrent sequences occurring 10 to 19 times per 100,000 words
◆◆	recurrent sequences occurring 20 to 49 times per 100,000 words
◆◆◆	recurrent sequences occurring 50 to 74 times per 100,000 words
◆◆◆◆	recurrent sequences occurring 75 to 99 times per 100,000 words
◆◆◆◆◆	recurrent sequences occurring over 100 times per 100,000 words

While extracting and analysing recurrent phrases, the notions of type and token are referred as well. In this study, each different sequence of words is considered a different type and each occurrence of a sequence of words a different token. As a part of quantitative analyses, this study has also made use of Log Likelihood statistics to investigate the data considering the overused/underused occurrences. Log Likelihood is a test for statistical significance that is often used in corpus analysis to identify the words or word-combinations that are particularly characteristic of a corpus (Paquot & Bestgen, 2009). Also called G-square or G score, log-likelihood compares the observed and expected values for two datasets like chi square test. However, it uses a different formula to compute the statistic that is used to measure the difference (Baker, Hardie, & McEnery, 2006). If the target sequence occurs more than expected by chance, then this association is given a positive score; if the target sequence occurs in the corpus at chance level, then the value is close to zero; and if the occurrence is less than chance then the association is negatively scored. That's to say, the higher the G-square value, the higher the significance level of the difference is between two frequency scores.

The identification of the recurring phrases makes the quantitative part of the analysis with the purpose of drawing a general picture of recurrence in native and learner writing and speech. This step is followed by a detailed qualitative analysis dealing with structural and functional aspects of recurrent sequences of words.

3.4.2. Qualitative Analysis: Functional and Structural Taxonomies

The structural classification of the recurrent expressions in the Longman Grammar of Spoken and Written English (Biber et al, 2007) has been widely used in the studies on recurring word combinations in the field (Cortes, 2002b, 2004; Hyland, 2008a, 2008b). The present study has also made use of the structural taxonomy offered by Biber et al., (2007). The structural taxonomy for the written language analyses the recurrent phrases into 12 major categories, eight of which are related to phrasal structures and four to clausal structures. The taxonomy for the structural categorization of the word sequences in spoken language consists of 17 categories seven of which are related to "verb phrase fragments", five of which are related to "dependent clause fragments" and five are related to "noun phrase and prepositional phrase fragments". Table 19 and 20 illustrate the structural classification of word combinations in written and spoken language respectively (Biber et al., 2007 pp. 1014, 1050). Example expressions are included as well.

Table 19

Structural Taxonomy for Recurrent Word Combinations in Written Language

Category	Example
1. **PHRASAL**	
a) **Noun Phrase Based**	
➢noun phrase with of-phrase fragment	*the end of the, one of the most important*
➢other noun phrase or noun phrase fragment	*Such a way that, the difference between the, the extent to which the*
b) **Prepositional Phrase Based**	
➢prepositional phrase with embedded of-phrase fragment	*As a result of, as in the case of, from the point of view of*
➢other prepositional phrase fragment	*In an attempt to, in the same way as, in such a way as to*
c) **Verb Phrase Based**	
➢anticipatory it + verb phrase/adjective phrase	*it can be seen that, it should be possible to, it was found that, it is important that*
➢passive verb + prepositional phrase fragment	*is to be found in, are shown in table, can be seen as*
➢copula be + noun phrase/adjective phrase	*is the same as, is similar to that of, may or may not be,*
➢Pronoun/noun phrase + be (+…)	*this is not the, there are a number of*
2. **CLAUSAL**	
➢(verb phrase +) that-clause fragment	*that it is not, has been shown that, should be noted that the*
➢(verb/adjective +) to-clause fragment	*is interesting to note that, is likely to be, does not seem to be*
➢adverbial clause fragment	*as we shall see, if there is a, as shown in figure*
3. **OTHER EXPRESSIONS**	*as well as in, than that of the*

As seen in Table 19, phrasal expressions are analysed in three subcategories: Noun-Phrase (NP) based, Preposition Phrase (PP) based, and Verb Phrase (VP) based. NP-based word combinations include any noun phrases with post-modifier fragments, such as "*the role of the*" or "*the way in which*"; PP-based word combinations refer to expressions starting with a preposition and a noun-phrase fragment or another prepositional phrase fragment, such as "*at the end of*" or "*in relation to*". The last category was VP-based combinations that refer to those with any word combination with a verb component, such as "*in order to make*" or "*was one of the*". Clausal lexical expressions, on the other hand, can be a verb or adjective followed by a to-clause fragment as in the example of "*is likely to be*", or a verb phrase followed by a that-clause fragment such as "*should be regarded that*". Lexical clauses that incorporate that-clause (can be seen that), to-clause (are more likely to), or adverbial clause (if there is a) are categorized in one broad group as clausal.

Table 20

Structural Taxonomy for Recurrent Word Combinations in Spoken Language

STRUCTURE	EXAMPLES
Type1: Verb Phrase Fragments	
➢1st/2nd person pronoun+VP fragment	*you don't have to, well I don't know, you see*
➢3rd person pronoun+ VP fragment	*this is a, it is going to be, that's one of the*
➢Discourse markers + VP fragment	*I mean you know, you know it was*
➢Verb Phrase (with non-passive verb)	*is one of, have a lot of, take a look*
➢verb phrase with passive verb	*is based on, can be said*
➢yes/no question fragments	*are you going to, do you want, does it make sense*
➢Wh-Question fragments	*what do you think, who is that,*
Type 2: Dependent Clause Fragments	
➢1st/2nd person pronoun+Dependent Clause	*I want you to, I don't know if*
➢Wh-clause fragments	*what I want, when we get to*
➢if-clause fragments	*if I have, if you want to*
➢verb/adjective+to-clause	*want to do, to come up with*
➢that-clause fragment	*that this is, that there is a*
Type 3: Noun Phrase and Prepositional Phrase Fragments	
➢Noun phrase with –of phrase fragment	*the end of, one of the, a bit of*
➢Noun phrase with other post-modifier	*the way in, a little bit about*
➢Other noun phrase expressions	*or something like that, a little bit more*
➢Prepositional phrase expressions	*of the things, of them, at the same time*
➢Comparative expressions	
4. OTHER EXPRESSIONS	*on and on, no nono, two and a half*

As Table 20 shows, three main categories are distinguished regarding the structures in spoken language, and most of the sequences do not represent complete structural units as exemplified (Biber et al., 2004). Accordingly, type 1 word sequences incorporate verb phrase fragments. That's to say, it covers word-combinations beginning with a subject pronoun followed by a verb phrase or beginning with a discourse marker (e.g. I think) followed by a verb phrase fragment (e.g. I think it is). The sequences directly starting with a verb, and those worded in question form are all of this type. The second structural type found in conversation incorporates dependent clause fragments in addition to simple verb phrase fragments. Sequences beginning with a complementizer or a subordinator are also of this type. The last category Biber et al., (2004) identified for the structure of the sequences in spoken language is phrasal incorporating noun phrases and prepositional phrases.

In addition to structural classification, in this study, functional categorization of the recurrent expressions has been carried out as well. This classification tried to focus not on the form but the functions that the recurrent phrases served in spoken and written language. The spoken and written English of the learners and native students represent two extremes in the range of register. Therefore, it is expected that a functional taxonomy based on these two registers would identify range of functions connected with recurrent phrases in both registers (Cortes, 2002a). In the explanation of the taxonomy, examples from both registers are included. The taxonomy used in this study was initially designed by Cortes, (2002a)and mainly based on the recurrent phrases identified in Longman Grammar of Spoken and Written English by Biber et al., (1999-2007). It was later revised and improved by Biber and his colleagues (Biber & Barbieri, 2007; Biber et al., 2004). Table 21 presents the functional taxonomy used for the qualitative analyses of the recurrent phrases in spoken and written performance of learner and native speakers (Biber et al., 2004 pp. 384-388).

Table 21

Functional Classification of Recurrent Phrases

Categories		**Example**
	Sub-Categories	
Stance Expressions	**A) Epistemic Stance**	
	Personal	*I don’t know, I think it was, well I don’t know, You know what I*
	Impersonal	*are more likely, the fact that the*
	B) Attitudinal/Modality Stance	
	B1: Desire	
	Personal	*If you want, what do you want, I don’t want to*
	B2: Obligation/Directive	
	Personal	*You have to, you want me to, you need to know*
	Impersonal	*It is important, it is necessary to*
	B3: Intention/Prediction	

(Table 21 Contuniued)

	Personal	*I'm not going to, I was going to, are you going to, we are going to have*
	Impersonal	*It's going to be, going to have*
	B4: Ability	
	Personal	*to be able to, come up with*
	Impersonal	*It is possible, can be used to*
Discourse/Text Organizers	**A) Topic Introduction/Focus**	*Going to talk about, if we look at, I would like to, to look at,*
	B) Topic Elaboration /Clarification	*I mean you know, on the other hand, as well as, has to do with*
Referential Expressions	**A) Identification/Focus**	*one of the things, those of you who, and this is, of the things*
	B)Imprecision / Markers of Vagueness	*Or something like that, and stuff like that, sort of*
	C) Specification of Attributes	
	C1: Quantity specification / Quantifying Sequences	*How many people, the rest of the, there's a lot of, loads of, one of the*
	C2: Tangible Framing Attributes	*The size of, in the form of*
	C3: Intangible Framing Attributes	*In terms of, in the case of, the way in which, the extent to*
	D) Time/Place/Text Reference	
	D1: Markers of Time	*At the same time, at night, during the day*
	D2: Markers of Place	*In front of, in the United States*
	D3: Text deixis	*As shown in, shown in figure n*
	D4: Multifunctional References	*The beginning of the, at the end, in the middle of, the top of*
Special Conversational Functions	**A) Politeness**	*Thank you, thanks a lot, thank you very much*
	B) Simple Inquiry	*What are you doing*
	C) Speech/Thought Reporting	*I said to, I was like oh, so I thought*
	D) Responses	*Yeah definitely, that's it*

As Table 21 shows, the taxonomy distinguishes among three primary functions served by recurrent phrases: (1) stance expressions, (2) discourse/text organizers (3) referential expressions. Stance expressions refer to group of words expressing express the user's attitudes, judgements and perspective that frame some other proposition. Discourse/text organizers, as the name suggests, reflect the relationships between prior and coming discourse, which help to compose and structure the conversation/text itself.

Referential expressions are word combinations that make direct reference to physical or abstract entities, or to the textual context itself. Each of these functions has several sub-categories associated with more specific functions and meanings (Biber et al., 2004).

Stance bundles have been analysed in two sub-categories: Epistemic stance and attitudinal/modality stance. Epistemic stance bundles refer to those that comment on knowledge status of the information in the following proposition: certain, uncertain and probable/possible (e.g. *I don't know, I don't think so*) (Biber et al., 2004). Attitudinal/modality stance phrases express speakers/writers attitudes towards the events or actions as in the example of *I'm going to, I want you to.* Stance expressions can be personal (overtly attributed to the speaker/writer) or impersonal (expressing similar meanings without being overtly attributed to the speaker/writer) as shown in the examples above.

Discourse/text organizing expressions serve two major functions: topic introduction/focus and clarification/elaboration. The first subcategory covers the markers of speech/thought reporting (e.g. *so I thought, I was like oh*), and the subcategory of clarification/elaboration includes markers of contrast, (e.g. *on the other hand)*; makers of cause: (e.g. *due to the fact*); exemplifiers: (*for example, for instance*; etc.).

Referential expressions are the word combinations that identify an entity or refer to some particular attribute of an entity which is thought to be important. It consists of four subcategories: (a) expressions regarding identification/focus (e.g. a film, this is a, one of the things…) are used to state, explain or summarize the main point in the speech/writing. (b) Imprecision expressions or markers of vagueness(e.g. sort of, something like that, things like that…)are used to indicate that a specified reference is not necessarily exact or to show that there are additional references of the same type. (c) Expressions specifying attributes (e.g. a lot of, a little bit, the size of, the nature of, in terms of…) are used to refer to quantity, amount, size and abstract characteristics of the following noun. (d) Expressions referring to time, place and text are, as the label suggests, the markers of place, time being generally multi-functional depending on the context (Biber et al., 2004).

The final category in the taxonomy is related to the spoken medium covering the subcategories of word combinations expressing politeness, simple inquiry, speech reporting and responses. The analysis of such expressions is thought to provide better insights on the spoken performance of the learners and native students.

In sum, being a contrastive analysis based on quantitative and qualitative corpus driven method in nature, this study has focused on two-, three- four- five- and six-word combinations found in the spoken and written interlanguage of Turkish students learning English. Based on frequency analysis, the first step involves identification of the target combinations in Turkish subcomponent of ICLE, which has then been compared with the LOCNESS, the native corpus. In the second step, the frequency analysis has been carried out in the Turkish subcomponent of LINDSEI, of which reference corpus is LOCNEC. After the frequency analysis, recurrent phrases obtained are analysed and classified both structurally and functionally by using the taxonomies referred in the relevant literature. The final step of analysis covers the comparison of spoken and written interlanguage in terms of the structure and functions of the recurrent phrases. The findings are presented in the following chapter.

3.5. Chapter Summary

This chapter introduces the methodological procedures followed in carrying out the present study. The first part explains the nature of the study, and then it continues with the detailed information about the corpora under investigation including the variables relevant to the material. The next part elaborates on quantitative and qualitative corpus driven method. The last section of the present chapter explains the analytical steps followed with detailed information on the taxonomies used in qualitative analysis.

CHAPTER IV

FINDINGS AND DISCUSSION

4.0. Introduction

This chapter reports the findings from the analysis of four corpora through a corpus-driven recurrent word combination method. The analysis conducted is two-fold: (1) the interlanguage analysis of the spoken performance of learners, using LINDSEI-TR and LOCNEC, and (2) the interlanguage analysis of the written performance of learners, using TICLE and LOCNESS. As explained in chapter 3, data analysis covers the processing of two-, three-, four-, five-, and six-word combinations frequently found in the spoken and written interlanguage of Turkish students learning English to gain insights into the spoken and written performance of Turkish EFL learners and to form a base in defining interlanguage characteristics of Turkish learners with respect to their both writing and speaking skills.

This chapter initially presents the most frequent recurrent phrases quantitatively in spoken language of learners, which is followed by a comparison of recurrent phrases in the speech of native students. Then, it continues with the structural and functional classification of the phrases found in the analysis. The same steps of analysis have been followed in exploring the written language of Turkish learners, which is then compared with written language of native speakers. Functional and structural classification is presented for the written language as well. The quantitative and qualitative analyses are accompanied by example sentences extracted from the corpora under investigation. This helps demonstrate the functions performed by these phrases more clearly. The last section presents further analysis of recurrent phrases comparing register-specific word combinations.

4.1. Quantitative Findings: Spoken Corpus

4.1.1. Recurrent Phrases in Spoken Interlanguage: LINDSEI-TR

The first research question that the present study has set out to answer is: “What are the major recurrent sequences of two- to six-word combinations Turkish learners tend to use in their spoken discourse?" Using WordSmith Tools 5.0, simple n-gram searches of LINDSEI-TR were carried out to find out the number of all 2-6 grams occurring in the corpus. Adopting the similar methodology utilised by Altenberg (1998), De Cock, Granger, Leech, & McEnery (1998) and De Cock (2004), frequency thresholds were set since a certain level of frequency is in itself a reason to consider the combinations as interesting from a phraseological point of view (Altenberg, 1998). Frequency thresholds are particularly significant in the context of spoken interlanguage in order to reduce the possibility that repetitions of certain combinations are not confined to one interview or one subject only (De Cock, 2004). Following Altenberg (1998), and De Cock (2004), this study adopts different frequency thresholds for each word-combination length as "the length of recurrent word combinations is inversely related to their frequency" (De Cock 2004: 228). Table 22 presents the overall statistics for LINDSEI-TR.

Table 22

Overall Statistics for LINDSEI-TR

Statistical Categories	**Numbers**	**Statistical Categories**	**Numbers**
Overall Corpus Size	80817	5-letter words	5556
Learner Turns	63924	6-letter words	3579
Tokens	63922	7-letter words	3715
Types	3162	8-letter words	1424
Type/Token Ratio	4.95	9-letter words	1316
Std. Type/Token Ratio	4.97	10-letter words	823
Av. Word Length	3.75	11-letter words	213
1-letter words	5189	12-letter words	67
2-letter words	17118	13-letter words	42
3-letter words	13185	14(+)-letter words	18
4-letter words	11679		

Depending on the focus of present study, all the analyses were carried out focusing on learner turns which were tagged as <B> and </B> in the transcription of the interviews. Following is the presentation of the findings for each length of the recurrent phrases in Turkish learners' data using LINDSEI-TR.

4.1.1.1. 2-Word Combinations in LINDSEI-TR

In the n-gram searches of the two-word combinations (bigrams) in the spoken interlanguage of Turkish students, frequency threshold was set as 12. Table 23 shows the top twenty two-word sequences in in Turkish learners' (TLs) speech (for the expanded list of bi-grams, see Appendix C).

Table 23

Top Twenty 2-word Combinations in TL Speech (LINDSEI-TR), freq. > 12, and Their Raw Frequencies

Rank	Word Sequence	Frequency	Rank	Word Sequence	Frequency
1	eh I	569	**11**	and I	195
2	and eh	425	**12**	the picture	179
3	eh the	304	**13**	eh in	178
4	I think	238	**14**	but eh	173
5	I like	219	**15**	it is	163
6	it was	217	**16**	to eh	161
7	in the	209	**17**	eh he	161
8	want to	208	**18**	don’t know	158
9	eh she	205	**19**	I want	154
10	eh and	196	**20**	I have	151

In the extraction of 2-word combinations, contracted forms (e.g. *don’t, isn’t)*were treated as one word. Filled pauses, <u>eh</u>, <u>er</u>, <u>em</u>, <u>erm</u> and backchanneling, <u>mm</u>, <u>uhu</u>, <u>mhm</u> were also included in the analysis and treated as words. As seen in Table 23, Turkish learners’ (TLs) most frequent 2-word combinations consist of “fragmentary sequences” (Altenberg, 1998, p.102), such as *and I, it was.* Although 2-word fragmentary sequences are generally left out in the studies of recurrent sequences in spoken language due to “their sheer number” (De Cock, 2004 p, 228), research starting from these combinations may prove to be useful as they may illustrate important organizing features of spoken discourse.

Another point observed in Table 23 is that almost half of the two-word combinations in the most frequent top-twenty list include a filled pause, and filled pauses are quite prevalent in the recurrent combinations which fall outside of the top-twenty list as well. In transcribing the filled pauses, four different ways are identified in the LINDSEI guidelines, *eh*, *er*, *em* or *erm*, considering the length and sound profile. The combination of *and* + *[filled pause]*, for example, can be found with all four filled pauses in the corpus; therefore, the combination *and eh*, which occurs in the top-twenty list could be considered as an even more frequent combination compared with other fillers (e.g. *and em* ♦♦♦, *and er* ♦, *and erm* ◊). An overall look at the whole list of 2-word combinations with filled pauses reveals that *eh* is the most frequently used filled pause among Turkish learners, which implies a difference from the frequently used filled pauses in native speakers (NS) speech (Kjellmer, 2003). In his study on the functions of filled pauses, Kjellmer (2003) employed the CobuildDirect, a corpus consisting of 57 million words of spoken American and British English, and he found out that *er* and *erm* are the commonest filled pauses among the native speakers.

Although there are relatively few studies in number investigating the 2-word combinations in learner English, the available ones imply that a comparison between different L1 populations could provide interesting results. Using the Swedish and Norwegian components of LINDSEI, (LINDSEI-SW and LINDSEI-NO) Aas (2011) reported the most frequent ten 2-word combinations in learner English. Research results reported in Aas's (2011) are chosen as a comparison base since that research is one of the most recent studies designed on spoken interlanguage. In addition, the fact that the corpora used and the word lengths covered in Aas's study are in agreement with the present study creates a fruitful area of comparison for the current study.

Table 24

Comparison of the 2-word combinations in LINDSEI-TR with those in LINDSEI-SW and LINDSEI-NO (cf. Aas, 2011 p. 59,61)

Rank	**LINDSEI-TR**	**LINDSEI-SW**	**LINDSEI-NO**
1	eh I (569)	it was (757)	it was (221)
2	and eh (425)	you know (632)	I think (173)
3	eh the (304)	sort of (583)	and eh (157)
4	I think (238)	I mean (444)	and I (150)
5	I like (219)	I was (437)	eh I (137)
6	it was (217)	I think (433)	in the (135)
7	in the (209)	I don't (423)	I don't (122)
8	want to (208)	in the (416)	so I (122)
9	eh she (205)	and I (367)	I I (120)
10	eh and (196)	and then(345)	a lot (108)

Combinations common to at least two corpora in the top ten list underlined.

As presented in Table 24, one of the recurrent phrases common to all interlanguage corpus is the overuse of *I think*. Explaining the characteristics of spoken

English, Biber, Johansson, Leech, Conrad and Finegan (2007) note that “mental verbs, especially *know*, *think*, *see*, *want*, and *mean*, are particularly common in conversation. These verbs report various states of awareness, certainty, perception, and desire. They typically occur with *I* or *you* as subject, and not infrequently occur together in the same utterance” (p.513). This finding about overuse of *I think* by learners is in line with the literature (Aijmer, 2004; Biber et al., 2007; Huang, 2011; Yong, Jingli, & Zhou, 2010). *I think* is also reported to be the single most frequent I + verb combination in the spoken components of the Corpus of Contemporary American English (COCA) (Davies 2008) and the British National Corpus (BNC), which suggests that its overuse is not the result of the LINDSEI contextual factors. Gilquin and De Cock (2011) explains the reason of overuse of *I think* by the learners by stating that this phrase may be holistically stored in speakers’ mind and automatically retrieved in a spontaneous speech. Another point to be taken into consideration is the use of *you know* and *I mean,* which are reported to be very frequent 2-word combinations in the 5 million-word CANCODE spoken corpus (O’Keeffe et al., 2007). *You know* functions as “an important signal of shared knowledge between speaker and listeners, and as well as a topic launcher (O’Keeffe et al., 2007 p.34). Unlike Swedish learners and unlike the findings related to the use of these two bi-grams in the literature, Turkish learners do not frequently use *you know* (occurring 127 times) and *I mean* (occurring only 30 times) in their speech. Referring to the importance of searching these two phrases (*you know* and *I mean*) in spoken language, Grant (2010) notes that awareness of the frequency, position and the functions of these phrases will contribute to student learning.

All in all, bi-gram searches of LINDSEI-TR show that Turkish learners’ spoken interlanguage is dominated by fragmentary sequences, and not surprisingly by filled pauses like *eh*, *er*, *em*, and *erm.* Compared with the findings in the relevant literature, Turkish learners use some of the bi-grams (e.g. *you know* and *I mean*) less frequently, which gives an idea about fluency and discourse organization of Turkish learners. A comparison of two-word combinations in LINDSEI-TR with those in LOCNEC, the reference corpus, is likely to provide better insights about Turkish learners’ spoken interlanguage, and such a comparison is presented in detail in section 4. 1.2.

4.1.1.2. 3-Word Combinations in LINDSEI-TR

In the extraction of 3-word combinations (three-grams) in spoken interlanguage of Turkish learners, frequency threshold was set as 6 following the relevant literature (Altenberg, 1998; De Cock, 2004). Table 25 lists the top twenty most frequent 3-word sequences in Turkish learners' English. (For the expanded list of 3-grams, see Appendix D).

Table 25

Top Twenty 3-word Combinations in NNS (LINDSEI-TR), freq. >6, and their raw frequencies

Rank	Word Sequence	Frequency	Rank	Word Sequence	Frequency
1	I want to	141	**11**	I think eh	41
2	I don’t know	60	**12**	and eh she	40
3	eh it was	60	**13**	er I want	39
4	and eh I	58	**14**	don’t know	37
5	a lot of	47	**15**	and then eh	36
6	want to be	46	**16**	eh I think	36
7	eh I like	44	**17**	it was very	35
8	to be a	44	**18**	but eh I	34
9	eh and eh	42	**19**	with my friends	34
10	there is a	42	**20**	I didn’t eh	33

Regarding the 3-grams in LINDSEI-TR, what is conspicuous at first glance is the dominant use of filled pauses (e.g. *eh*), as 11 out of 20 sequences include such items. De Cock (2004) has termed filled pauses and repetition phrases as hesitation items. Although hesitation items/ filled pauses were previously thought as “of little phraseological interest” (Altenberg, 1998), it is now widely recognized that they are very important features of spoken language. In Kjellmer's words (2003) “one characteristic of speech is its frequent indication of hesitation or uncertainty” (p.170). Therefore, as mentioned earlier, keeping filled pauses in the frequency search maybe helpful to illuminate some of the functional properties these items may provide in combination with other words and word-combinations.

Use of these filled pauses, potentially, may indicate encoding problems at clause beginnings (De Cock, 2004), and highlights problems of planning pressure in learner language (Kjellmer, 2003). Based on the findings of his study, Kjellmer (2003) claims that “one main function of [a filled pause] thus seems to be to introduce what I will loosely call a new ‘thought unit’, a word, a phrase and sometimes a whole clause” (p.174). In line with Kjellmer’s findings, Tottie (2010) suggests the term “planners” to refer to these filled pauses. Although these planners serve important functions in facilitating conversation, Kjellmer (2003) also asserts that “since we are most of the time unaware of the [filled pauses], their (moderate) use will not normally affect adversely our impression of a speaker’s fluency or eloquence” (ibid. p.191). Thus, it is likely that filled pauses, when overused or ‘misused’ compared with the native speaker norm, will make the listeners become aware of the disfluencies, and that this will have an impact on our impression of a learner’s fluency or eloquence. The finding that Turkish learners use filled pauses too frequently as shown in Tables 23 and 25 suggests that they have significant difficulties in keeping their speech fluent, which may be linked to the lower proficiency level in English. Before making further claims about spoken interlanguage, it will be useful to make comparison between interlanguage of different L1 populations. Table 26 shows 3-word combinations in LINDSEI-SW and LINDSEI-NO as reported in Aas, (2011) .

Table 26

Comparison of the top ten 3-word combinations in the LINDSEI-SW, LINDSEI-NO and LINDSEI-TR (cf. Aas, 2011 p. 59,61)

Rank	**LINDSEI-SW**	**LINDSEI-NO**	**LINDSEI-TR**
1	I don’t know (139)	a lot of (70)	I want to (141)
2	a lot of (94)	I don’t know (56)	I don’t know (60)
3	I think it’s (60)	I went to (33)	eh it was (60)
4	I don’t think (57)	I think it’s (31)	and eh I(58)
5	and it was (51)	it was eh (29)	a lot of (47)
6	I think it (51)	it was a (28)	want to be (46)
7	I think I (46)	and it was (23)	eh I like (44)
8	you have to (40)	you have to (23)	to be a (44)
9	I think so (38)	eh it was (22)	eh and eh (42)
10	it was very (37)	yeah yeahyeah (22)	there is a (42)

Combinations common to at least two corpora in the top ten list underlined.

Compared to the Swedish and Norwegian learners, Turkish learners perform a greater use of filled pauses in their speech. Similarly, in her study of spoken interlanguage of French learners, De Cock (2004) reports overuse of filled pauses as well. This noticeable difference among learner groups could make an interesting base for further research.

As for the conclusion to be reached from this finding, there are a lot of studies treating overuse/underuse of the filled pause as an indicator of proficiency. In a study of interlanguage, Lauttamus, Nerbonne and Wiersma (2008) suggest that features such as filled pauses, repetitions, false starts and repairs are statistically significant determinants that distinguish less proficient learners having acquired an L2 later in life from more proficient learners having acquired their L2 at early ages. Therefore, it is possible to claim that Turkish learners are at lower level of English proficiency due to their frequent use of filled pauses compared with Swedish learners who are reported as having had the highest number of years in terms of learning English and having spent the highest number of months in an English-speaking country in the overall corpora, which are likely to contribute positively to their level of proficiency.

Before ending the discussion on the overuse of filled pauses in Turkish learners' speech, one point to be touched upon is the place of these items in the conversation. Paananen-Porkka (2007) argues that "pauses not only occur at sentence or clause boundaries, but also at word boundaries" (p. 259). De Cock (2004) found out that most of the recurrent phrases containing hesitation items (e.g. filled pauses and repetitions) are clause beginnings in the speech of French learners. This is also the case with Turkish learners as exemplified in the following utterances taken from the LINDSEI-TR:

(1) *<B> (eh) I chose .topic one . (eh) I want to talk about (eh) m= one of my experience </B>*

(2) *<B> (em) .. I was so affected .after the lesson first of all (eh) I have learned (eh) we should we should meet (eh) .their needs <overlap /> (eh) </B>*

(3) <B>*(eh) .. I like spending time with my friends <overlap /> reading (eh) ..and . I like listening to music very much <overlap /> (erm) ... and I like writing (eh) .poems</B>*

(4) *<B>(mm) for example I (erm) I learned that I shouldn't use (eh) grammar teaching method <overlap /> grammar (eh) method (eh) I should . I should use (eh) direct method (eh) in future in my (er) <overlap /> teacher life (eh) because they are (eh) .so (eh) .smallch= (eh) children </B>*

This finding suggests that learners in LINDSEI-TR have encoding problems at clause boundaries or in Altenberg's words (1998) "thematic springboards". This supports the assumption that clause beginnings are major planning points (Biber et al., 2007). The overuse of hesitation items by the learners shows that having to plan a clause in a language other than one's mother tongue increases the planning pressure speakers face at the beginning of a clause. Setting off on a clause is something of a challenge for learners (De Cock, 2004). The extracts above also show that Turkish learners use hesitation items at word and phrase boundaries as well, which suggests that learners presumably have problems finding the words they need to encode their messages. One reason for this might be the limited vocabulary knowledge learners possess. Or another reason could be that the difficulty of expressing oneself in a foreign language causes pressure on the overall encoding process. Further support for learners encoding problems comes from the n-gram searches of the corpus for 4-word combinations. The frequency results of 4-word sequences show that the expression *how can I say* (♦♦) is frequently used by learners and it mainly functions as a communicative strategy for asking for assistance. A detailed explanation for *how can I say* is provided in 4.1.1.3.

Apart from combinations including filled pauses, the most frequent 3-word sequences used by Turkish learners are *I want to* (141), *I don't know* (60), *a lot of* (47) etc… (see Table 25). The finding that these expressions are very common in spoken language is in line with the literature. According to O'Keeffe et al. (2007) *I don't know* is the most frequent tri-gram in the five million word CANCODE spoken corpus. Biber et al., (2007)note that *a lot of* as a very frequent sequence in BNC is the characteristic of casual speech. Studying through MICASE (1.7 million words) and BNC (spoken part 431000 words), Simpson-Vlach and Ellis (2010) report *I want to* as one of the frequently used combination which primarily belongs to spoken register.

4.1.1.3. 4-Word Combinations in LINDSEI-TR

Most of the studies on recurrent phrases in learner language have focused on the 4-word combinations due to the fact that four-word combinations hold three-word combinations in their structures, as in *as a result of,* which contains *as a result* (Cortes, 2004; Hyland, 2008a). Bearing this in mind, this study has analysed the 4-grams found in LINDSEI-TR with a frequency threshold of 4. Table 27 presents the top-twenty most frequent 4-word sequences in spoken interlanguage of Turkish learners. (For the expanded list of 4-grams, see Appendix E).

Table 27

Top Twenty 4-word Combinations in NNS Speech (LINDSEI-TR), freq. >4, and their raw frequencies

Rank	**Word Sequence**	**Frequency**	**Rank**	**Word Sequence**	**Frequency**
1	eh I want to	36	11	I want to talk	15
2	I want to be	32	12	the end of the	15
3	I don’t know but	29	13	want to talk about	15
4	how can I say	27	14	to draw her picture	14
5	I want to eh	22	15	eh I eh I	13
6	I don’t like eh	17	16	eh I think eh	13
7	to be a teacher	17	17	and then eh she	12
8	want to be a	17	18	eh and eh I	12
9	at the end of	16	19	eh for example eh	12
10	and I want to	15	20	eh I like eh	12

As seen in Table 27, the most frequent four-word sequences among Turkish learners include mainly the verb *want* which is one of the basic vocabulary generally taught at very early stages of language learning. Cambridge dictionary (2013) categorizes the *want* as A1 level; thus, it is not interesting that learners know and actively use this word. This very common use of *want* and its combinations confirms Biber et al., (2007) findings that *want* is among the commonest verbs of conversation usually combined with first and second person singular pronoun. Biber et al., (2007) also note that 90 % of the recurrent phrases in spoken language are declarative and interrogative clause segment, and 50 % of them begin with a personal pronoun which is observed in Turkish learners’ four-word sequences as well (e.g. *I want to be, I want to talk, and I want to, eh I like eh*). The expression *I don’t know but* is ranked as the fourth most frequent 4-grams in Turkish learners’ speech, which is also in line with the literature. Diani (2004) states that *but* is one of the words frequently combined with the sequence *I don’t know* along with the phrases *well, I mean, you know.* The similar combinations are echoed in Aijmer (2009) who found that *but* is frequently co-occur with *I don’t know,* yet in both studies these words occur at initial position unlike Turkish learners’ data in which *but* is combined with *I don’t know* at final position. This implies a deviation from the frequent patterns of *I don’t know,* and functional analysis of this combination in LINDSEI-TR, which is given in section 4.2., may provide better insights regarding the spoken interlanguage of Turkish learners.

One of the frequent 4-grams which is worth mentioning here is the sequence *how can I say* that ranked fourth in the top-twenty list. As mentioned earlier, data in LINDSEI-TR suggest that Turkish learners have difficulty in terms of fluency and furthering effortless conversation, and the recurrence of *how can I say* provides further support for learners' encoding problems. *How can I say* as a communication strategy is employed generally in the challenging conversations to smooth the message. However, its function is different in L2 context: Brouwer (2003) called this expression as an "explicit word search marker", and shows two different functions of it: 1. a technique used to produce a mutually recognized reference in talk; 2. to request or invite help. Based on his research on learner language, Jung (2004) found out that the sequence "*how can I say"* in L2 context is used to appeal for assistance and initiate repair. Following excerpts (5) and (6) taken from LINDSEI-TR illustrates both functions:

(5) *<B> K P S S exam (eh) if I .succeed .to pass the exam . (eh) .. if .again if . my (eh)* ***how can I say atama****<laughs></B>*
<A> appoint </A>
<B> appoint </B>
<A> I can be appointed </A>
<B> if I can be <coughs>appointed . (eh) I will I will be a teach= I will work as a teacher </B>

(6) *<B> and I was (eh). I would (eh) go to the first grade of high school (eh)..that (eh) friend of me (eh) was (em) . (eh) .. had (eh) an a ... (eh) he (eh) (eh)* ***how can I say (eh) he got (eh) strangled (eh) in the (eh) river****</B>*

Overall, the frequency analysis of 4-grams in Turkish learner speech seem to suggest that sequences such as *I want to, I don't like, I don't think* and their combinations are their preferred ways of expressing themselves, or as discussed by Biber et al., (2007) "recurrent discourse building blocks" (p.1002) as they show a statistical tendency to co-occur. The preliminary analysis demonstrates that recurrent phrases show somehow deviations when their functions are considered, which makes a qualitative analysis necessary. Functional analysis of recurrent phrases is presented in section 4.2.

4.1.1.4. 5- and 6-Word Combinations in LINDSEI-TR

Studying longer phrases like 5- and 6- combinations are not so prevalent in corpus literature as they are comparatively rare in data sets and mostly contain 2-, 3- and 4-word sequences (Cortes, 2004; Simpson-Vlach & Ellis, 2010). However, this study keeps these longer combinations too in the analysis for the sake of thoroughness. In the extraction of these longer sequences, frequency threshold was set as 3 since only lower frequency thresholds would provide relevant data (De Cock, 2004). Table 28 and 29 present the top twenty 5- and 6-word sequences in Turkish learners' spoken interlanguage. (For the expanded list of 5 and 6-grams, see Appendix F).

Table 28

Top Twenty 5-word Combinations in NNS Speech (LINDSEI-TR), freq. >3, and their raw frequencies

Rank	Word Sequence	Frequency	Rank	Word Sequence	Frequency
1	I want to be a	16	**11**	and I want to be	5
2	I want to talk about	15	**12**	doesn't like the picture	5
3	at the end of the	13	**13**	eh at the same time	5
4	the end of the film	11	**14**	eh first of all I	5
5	want to be a teacher	9	**15**	I like it very much	5
6	eh how can I say	8	**16**	I want to be eh	5
7	the woman in the picture	7	**17**	I will be a teacher	5
8	eh at the end of	6	**18**	she wants him to draw	5
9	eh I want to eh	6	**19**	the picture and eh she	5
10	eh I want to talk	6	**20**	to draw her picture and	5

Table 29

Top Twenty 6-word Combinations in NNS Speech (LINDSEI-TR), freq. >3, and their raw frequencies

Rank	**Word Sequence**	**Frequency**	**Rank**	**Word Sequence**	**Frequency**
1	at the end of the film	9	**11**	don't have any chance I think	3
2	I want to be a teacher	9	**12**	eh she wants him to draw	3
3	eh I want to talk about	6	**13**	I want to be a good	3
4	I want to talk about a	5	**14**	I will be an English teacher	3
5	the woman in the picture is	5	**15**	one day a woman comes to	3
6	eh at the end of the	4	**16**	that there is no one in	3
7	I have a lot of friends	4	**17**	the picture to his to her	3
8	okay I want to talk about	4	**18**	the woman on the portrait was	3
9	the end of the film eh	4	**19**	want to talk about a film	3
10	and at the end of the	3	**20**	when I was in high school	3

As seen in Tables 28 and 29, the frequency of longer combinations is very low probably due to the corpus size which is about 64000 words. Additionally, comparison of the table 28 and 29 clearly shows that many 5-word sequences are incorporated into 6-word sequences. For example, *I want to be a* is a part of *I want to be a teacher* and *eh I want to talk* is a part of *eh I want to talk about*. The recurrent 5-and 6-word combinations also show that most of them are produced depending on the contextual factors of LINDSEI project. That's to say, these expressions are closely related to the tasks set to collect data from the learners. As explained in chapter 3, one of the tasks designed to carry out the interviews with the learners was 'picture description' which made use of the same set of pictures illustrating a woman and a painter. (see Appendix B for the task); therefore, the students uttered many sentences related to this context as in the expressions *the woman in the picture, she wants him to draw, one day a woman comes to* etc… . This makes it difficult to have generalizable and sound claims regarding the longer sequences in Turkish learners' spoken interlanguage.

In conclusion, the overall quantitative analysis of LINDSEI-TR based on frequency of recurrent phrases of various lengths demonstrates that Turkish learners have a repertoire of word-combinations used together more frequently than expected by chance. Not interestingly, the proportion of this repertoire changes in accordance with the length of the combination. Figure 6 is a graphic representation of the frequencies of recurrent phrases of different lengths in the corpus.

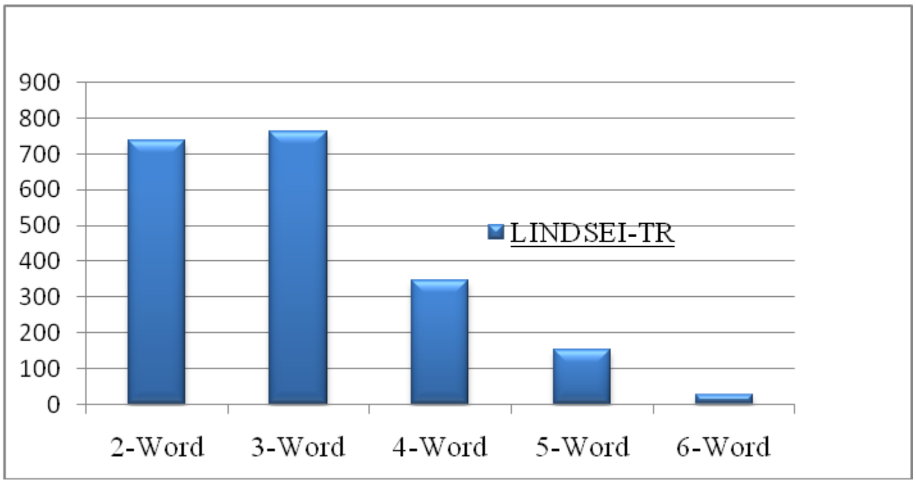


Figure 6. Proportion of most frequent 2-6-word combinations in TLs speech (LINDSEI-TR)

As figure 6 shows, 2-and 3-word sequences make up the largest proportion of all the combinations, and the proportional weight decreases in parallel with the decrease in combination length although data in LINDSEI-TR show that there is a slight difference between 2-word and 3-word combinations. This similar distributional pattern according to length has been echoed in the literature (Aas, 2011; Wei, 2009). This confirms Altenberg's (1998) observation that “continuous recurrent word-combinations in speech tend to be fairly short” (p.103). Biber et al., (2007) make a similar observation, in their investigation on spoken conversation, finding that there are almost ten times as many 3-word combinations as 4-word combinations in their data. The high frequency of shorter combinations is of course partly due to the fact that these may be embedded into longer combinations, such as *I think* in e.g. *I think it’s*, *but I think*, and *I think so*.

The quantitative analysis provides some understanding with the general properties of the Turkish learners’ spoken interlanguage (consider the frequent use of filled pauses, and the communicative strategies); however, a deeper analysis comparing Turkish learners’ speech properties with those of NS could yield better insights. The following section presents the results of such an analysis.

4.1.2. Recurrent Phrases in Native Speech: LOCNEC in Comparison with LINDSEI-TR

As explained in chapter 3, LOCNEC is compiled in the same task conditions by considering the same variables as LINDSEI with the purpose of creating a reference corpus to LINDSEI. In the analysis for the word sequences in NS speech, the same points in terms of frequency thresholds considered in interlanguage speech analysis were taken into account. Figure 7 displays the overall comparison between NS and TLs speech types in accordance with recurrent phrases of different lengths.

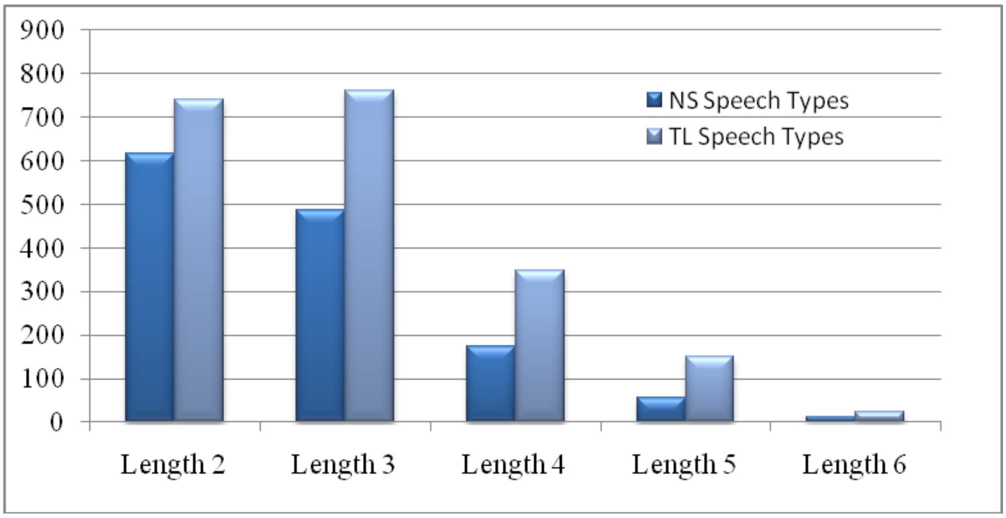


Figure 7. NS speech vs. TL speech types

Figure 7 shows that there are more 2-, 3- and 4-word sequence types in Turkish learners' speech but the difference is not statistically significant. On the other hand, there are significantly more 5-word sequence types (at $p \leq 0.05$) in the TL corpus. There are however slightly fewer 6-word sequence types in the learner corpus but the difference is not statistically significant. The overall results suggest that native and learner language do not solely consist of 'individual building blocks' assembled according to predefined rules and semantic information, but rather appear to be produced partly on the basis of larger, previously encountered and memorized sequences. Within the phraseology context, appearance of many word combinations more frequently than expected by chance supports the notion that "words belong with other words not as an afterthought but at the most fundamental level" (Wray, 2002 p.13) - both in NS and Turkish learners' speech.

The second research question of the present study is "to what extent are the recurrent sequences in NNS speech similar to and/or different from those in NS speech?" which requires a deeper analysis considering the instances of overuse/underuse phenomena. For this reason, using the "compare two wordlist" function of WordSmith Tools, first of all, identical and different 2-5 word combinations have been identified, and then Log Likelihood Statistics were carried out to see if the findings are statistically significant or not. Table 30 displays the most frequent twenty 2-to 5-word combinations in NS speech, which could give an idea about the recurrent phrases common to both NNS speech and NS spoken language (cf. Table 23, 25, 27 and 28). Since search for 6-word combinations in NS speech at the frequency threshold produced too few instances to have sound claims, they were not included in the table.

Table 30

Top Twenty 2- to 5-word Combinations in NS Speech (LOCNEC), freq. >12/6/4/3, and their raw frequencies

Rank	2-Grams	3-Grams	4-Grams	5-Grams
1	it was (393)	I don't know (98)	I don't know I (26)	the I end of the (15)
2	the I (379)	a lot of (65)	the end of the (21)	at the I end of (11)
3	you know (340)	I mean I (56)	at the I end (19)	I don't know I mean (6)
4	yeah yeah(276)	yeah yeahyeah(48)	and things like that (17)	I don't know I don't (5)
5	sort of (275)	the I end (45)	it was it was (17)	yeah I don't know I (5)
6	I mean (258)	it was a (42)	I end of the (15)	yeah that's right yeah yeah (5)
7	I was (252)	and then I (36)	yeah that's right yeah (12)	you know what I mean (5)
8	in the (230)	it was just (35)	yeah yeah yeah yeah (11)	English as a foreign Language (4)
9	and then (204)	of the I (35)	or something like that (11)	I did when I was (4)
10	I don't (198)	in the I (34)	I think it was (10)	I end of the film (4)
11	and I (194)	I went to (33)	a bit of a (9)	I end of the year (4)
12	I think (192)	yeah yeah I (33)	a lot of people (9)	like that I don't like (4)
13	of the (182)	you know I (32)	I thought it was (9)	painting a picture of a (4)
14	yeah I (149)	yeah it was (30)	I'd like to go (9)	what I wanted to do (4)
15	a lot (140)	you have to (29)	in the I end (9)	a look at it and (3)
16	so I (140)	I think it's (28)	the I er the (9)	a lot of people had (3)
17	but I (121)	when I was (28)	erm I don't know (8)	all the rest of it (3)
18	and the (113)	you know you (28)	I don't know if (8)	and all of a sudden (3)
19	don't know (113)	and I was (27)	it was really good (8)	and all that kind of (3)
20	a bit (112)	don't know I (27)	yeah yeah it was (8)	and then I went back (3)

Combinations common to LINDSEI-TR in the top twenty list underlined.

When the top twenty most frequent recurrent phrases in LOCNEC are compared with those in LINDSEI-TR (cf. Tables23, 25, 27 and 28), it is observed that Turkish learners differ to a great extent in terms of their preferred sequences in speech. Only seven similar sequences were found in top-twenty lists of both corpora (e.g. *it was, in the, and I, I think, don't know, I don't know* and *a lot of*). Native speakers frequently make use of word sequences like *you know, sort of, I mean* which are considered as a characteristics of informal spoken English (Aijmer, 2004; Shirato & Stapleton, 2007). These discourse items are the signs of assumption of common ground and social closeness and contributes to the informality of the interaction; therefore, "creates a congenial atmosphere" (Aijmer, 2002 p.209). Their being common in spoken English has also been reported by Romero-Trillo, (2002) and Stenström, (2006). However, as the comparison shows, the use of such discourse items is very rare in Turkish learners' speech. While there are 340 instances of *you know*, 275 instances of *sort of* and 258 instances of *I mean* in NS speech, in NNS spoken language, they are used in 30 and 126 instances respectively, and *sort of* is not used at all. Likewise, the word *yeah* which is reported as the most distinctive keyword of the spoken English (Paquot & Bestgen, 2009) is extremely frequent in NS speech while it is used only about 15 times in the whole corpus of LINDSEI-TR. These preliminary findings suggest that Turkish learners are unfamiliar with the characteristics of spoken English, and they have problems in adopting, memorizing and making use of some of the very common word-combinations in English conversation. This finding related to the big gap between NS and TL speech –even in the comparison of few conversation-specific items- also suggests that a thorough analysis of both corpora considering the overuse/underuse phenomena could provide better insights about spoken interlanguage of Turkish learners.

Table 31 displays the overused 2-5-word sequence types obtained through the comparison of NNS speech and NS language.

Table 31

Overused word-combinations (freq. >12/6/4/3) occurring in both LINDSEI-TR in comparison with LOCNEC, raw frequencies and relevant Log Likelihood statistics)

Word Combinations	**LOCNEC**	**LINDSEI-TR**	
Overused Phrases	**Freq.**	**Freq.**	**Log Likelihood**
and eh	13	425	+489.41
I like	34	220	+151.81
the picture	27	179	+125.28
I want	18	154	+122.90
want to	46	208	+111.52
she is	10	113	+100.98
I have	28	151	+92.67
I can	34	134	+63.48
for me	14	94	+66.27
because of	13	81	+54.64
very much	14	81	+52.14
for example	-	115	+45.17
with my	12	63	+37.93
of course	17	64	+28.97
there are	13	57	+29.77
yes I	17	57	+22.75
of them	17	57	+22.75
to her	10	53	+32.13
I say	11	45	+22.09
in this	10	43	+22.08
I want to	13	141	+124.12
to be a	8	44	+27.37
the picture and	8	30	+13.53
I want to be	-	32	+44.30
how can I say	-	27	+37.38
I want to talk about	-	15	+20.77
the end of the film	-	11	+15.23
the woman in the picture	-	7	+9.69

+ (p<0.05).

In order to decide if the observed overuse combinations are significant findings in statistical terms, Log Likelihood test was applied to the combinations. An overall look at the word combinations given in table 31 reveals that frequency differences are all proven to be statistically significant by the Log Likelihood test (p<0.05). The overused sequences seem to reflect several important tendencies concerning Turkish learners' use of phraseology: Turkish learners overuse various phraseologies including the first person pronouns (*I like, I want, I have, I say* etc…), which implies that they hardly attempt to discuss topics from a more objective viewpoint. This might be partly explained through NNS’s general inclination for reader/listener involvement (Aijmer, 2002a; Ishikawa, 2009). From a structural point of view, learners’ preference for active (*I say, I believe*) rather than passive voice (*it is said*) could be accounted for by the NNS’s general tendency towards the use of active discourse frames (Granger, 1998c). The overuse list additionally shows that learners tend to use idiomatic expressions more intensively than NS. To illustrate, the frequency of *for example, very much, of course* and *because of* rather than their variants (e.g. *too much, sure, due to*) is considerably higher in TL corpus. A possible reason for the overuse of these sequences is that the learners consider them to be safe options. Hasselgren (2002) labels such usages as “phrasal teddy bears” and explains that that NNS tend to rely exclusively on expressions they know well, just as small kids rely on their favourite teddy bears. These phrasal teddy bears are not necessarily related to the learner’s L1, and there seem to be at least a number of them that are treated as such across different L1 groups (Nesselhauf, 2005). The overuse of *for example* (Gilquin et al., 2007) and *of course* (Aijmer, 2004; De Cock, 2004) reported in different studies conducted with learners from different L1 groups confirms this assumption.

The overuse of *of course* deserves special attention since it has been discussed in the similar investigations as one of the overused but at same time misused phrases. In her study, De Cock (2004) finds out that French learners of English generally use *of course* to answer a request for information or to respond to an opinion expressed by another speaker, which may cause the learners to sound over-emphatic or even impolite. A similar tendency for the use of *of course* is observed among Turkish learners as well. Consider the following examples:

(7) *<A> so did you make good friendship <overlap /> there </A>*
<B><overlap /> yeah ***of course*** *</B>*
<A> so did you lose any connection with them or you <overlap /> still continue </A>

(8) *<A>good and could you tell me about your future plans after university </A>*
<B>(em) ***of course*** *I will (er) enter (eh) . I will (eh) .pass my (eh) examination <overlap /> for (eh) being a teacher and then (eh) I have (em) ... (em) five (eh) four sister<?> and one brother (eh) my (eh) sister .had<X> to study our*

<overlap /> school and I will help my <overlap />two .sister and brother .to (eh) study their school </B>

(9) *(mhm) so what have you done (eh)to help other people . is there anything that you are doing now </A>*
*<B> yes **of course** (eh) for example (eh) I I know some children and I (eh). learn them something . about English I give them (eh)lesson because I want to help them or (eh)old people every time I help them </B>*

As seen in the examples, Turkish learners use *of course* inappropriately in the context where the interviewer request for information. According to Oxford Advanced Learner's Dictionary (2000) if this phrase is used as an answer to a request for information, the user would sound impolite because the message it gives is that "you think the answer to the question is very clear and you think the person is stupid to need to ask you" and that "it may sound as if you think the answer to the question is obvious and that the person should not ask". The dictionary recommends the use of "yes that's right/yes I+auxiliary" instead in such cases in order to avoid sounding awkward.

Finally, the overuse of *and eh* which is the top overused sequence in Turkish learners' spoken interlanguage needs to be discussed here. It realizes some functions seemingly idiosyncratic for Turkish learners as shown in the following examples:

(10) *<B> (eh) a film (em) I've seen last (eh) seven pounds I love that film (em) it's (eh) the actor is Will Smith **and (eh)** it was about people's life and it's about life and death (em) so (eh) beautiful I (eh) mostly affect <laughs> from (eh) it ... <laughs></B>*
<A> (eh) who who who are the actors and actresses </A>
*<B> (eh) actor (eh) is Will Smith and (eh) the actress is (eh) Kate Winslet I think **and (eh)** there are (eh) lots of (eh) actors and actress and their life are different from the other people (eh) man (eh) Will Smith (eh) lost his wife and children and he blames himself (eh) so: he saves the other people's life he tries to change their life (em) </B>*

(11) *<B> yes and he moved from another country **and (eh)** there was a woman (eh) and children whose husband was (eh) beating her and (eh) Will Smith gave (eh) gave her (eh) his house (eh) his villa to her **and (eh)** he didn't recognize her before </B>*

(12) *<B> (eh) in my life and there I I saw very different dimensions of life <overlap /> because I it was very different form the life I had I was .used to <overlap /> live different ..geography people etc* ***and (eh)*** *so I had many difficulties (eh) first of all the work was very difficult <overlap /> I worked hard I . I was very tired <overlap /> also I travelled alone .a lot .and (eh) I had many many dangers experiences <overlap /> also there (eh) for example a fight </B>*

(13) *<B> with my friends* ***and (eh)*** *by accident we had to (eh) ..there was a fight and we were passing by so my friends (eh) get into the (eh) fight and the police come the only girl I . was I am and the police asked me what are you doing here or they were very rude my my friends were injured about this* ***and (eh)*** *the fight and I was very afraid I couldn't think or an= do anything <overlap /> just scream cry this was very .. </B>*

(14) *<B> woman asks (eh) I think woman ask <?> what is this (eh) did did I (eh) give my pose like that (eh) and man (eh) a= agrees (eh) with that .and tries again (eh) tries again the portrait of the woman but (eh) this time (eh) he paints the woman's (eh) portrait (eh) .. very beautifully .. woman woman (eh) gives the same . pose again but the painter (eh) makes a beautiful portrait . this time and woman (eh) likes that portrait and tells about a portrait her picture (eh) to his friends her friends <overlap /> and (eh) she lo= loves her portrait* ***and (eh)*** *with pride (eh) she says . she says to her her friends that's my portrait </B>*

Examples (10), (11) and (12) show that the combination *and eh* acts as a stop where learners rest for planning the proceeding utterance, or in Altenberg's words (1998) as a "stepping stone" (p.119). In his investigation on the phraseology of spoken English based on London-Lund Corpus, Altenberg (1998) identifies the function of "stepping stone" for the combination of *but er* in NS speech; however, it seems that in Turkish learner corpus, such a function is realized through the use of *and eh* as it is produced mostly before the rest of utterance is planned in detail. Examples (13) and (14) illustrate another function of this sequence: word search device. While searching for the appropriate word to complete the utterance, Turkish learners apparently coupled the filled pause *eh* with the conjunction *and*. *And eh* with this function is different from *how can I say,* another overused sequence for word searching (which is discussed in section 4.1.1.3.) as it doesn't include direct appeal for assistance.

As an interim summary, the sequences overused by Turkish learners point to both similarities and differences compared with the results of the studies on learner language in various L1 contexts and compared with the NS speech. Turkish learners preferred ways of saying things are, to a great extent, different from those in NS as shown by the really higher Log Likelihood scores, which implies that Turkish learners are not so familiar with the spoken properties of native-like conversation. Their overuse of such sequences as *for example, of course, because of,* and their preference for using active discourse frame is in line with the literature on interlanguage (Aijmer, 2004; De Cock, 2004; Ishikawa, 2009); however, Turkish learners' overuse of *and eh* and *how can I say* as communicative strategies differ from the other learners of L1 groups reported in the literature. While learners typically use the sequence *but er* as a strategy for having time for planning the utterance and appeal for assistance (De Cock, 2000), Turkish learners prefer the combination *and eh* for such a function, which is seemingly an idiosyncratic feature of Turkish learners' spoken interlanguage . Their overuse of *how can I say* as an explicit word search marker could also be claimed as something peculiar to Turkish learners spoken interlanguage.

In order to paint a complete picture of characteristics of spoken interlanguage of Turkish learners, the recurrent phrases underused by the learners should also be considered. Table 32 displays the underused 2-5-word sequence types obtained through the comparison of TL speech and NS language.

Table 32

Underused Word-Combinations (freq. >12/6/4/3 Occurring in both LOCNEC and LINDSEI-TR, Raw Frequencies and Relevant Log Likelihood statistics)

Word Combinations	**LOCNEC**	**LINDSEI-TR**	
Underused Phrases	**Freq.**	**Freq.**	**Log Likelihood**
sort of	275	-	-381.72
yeah yeah	276	5	-383.11
I mean	258	30	207.19
you know	340	126	-102.44
don’t know	113	17	-79.55
well I	78	5	-77.41
cos I	58	-	-80.51
a bit	112	29	-52.31
that’s right	46	-	-63.85
but er	47	6	-65.24
and then	204	114	-25.98
all the	95	29	-37.13
I don’t know	98	13	-73.86
you have to	29	6	-16.49
things like that	26	-	-36.09
sort of things	23	-	-31.93
but I mean	22	-	-30.54
that’s right yeah	19	-	-26.37
well I mean	12	-	-16.66
and things like that	17	-	-23.60
yeah that’s right yeah	12	-	-16.66
I don’t know I mean	6	-	-8.33

- (p<0.05).

Frequency differences presented in Table 32 are all proven to be statistically significant (p<0.05). From an overall view, table 32 shows that there are striking differences between NS and Turkish learners of English in terms of overall discourse organization. The sequences such as *sort of, that’s right, well I, well I mean, things like that* which are regarded as distinctive features of spoken language are underused by the Turkish learners; in fact, they don’t appear at all in the corpus. The expressions like *you know, a bit* and *I don’t know* that are also closely connected to spoken language are used by Turkish learners but they are very infrequent.

According to Shirato and Stapleton (2007) who compared vocabulary in a spoken learner corpus with a native speaker corpus, the word-combinations such as *you know, I mean, sort of, a bit, things like that, I don’t know* play a major role in spontaneous spoken interaction where “speakers have to plan, encode, and actually produce their utterances in real-time (p. 396). Additionally, they enable speakers to talk without too much hesitation or without too many pauses (Aijmer, 2004). McCarthy and Carter (2002)listed seven items including *something like that, sort of, a bit* that are inherently used by NS. Aijmer (2002a) listed more than 40 variants of the foregoing combinations by examining the London-Lund Corpus and demonstrated that these expressions are ubiquitous in the native speaker's spoken data. Despite the frequency and importance of these multi-word devices in spoken English, almost none of these variants occurred in the corpora used in this study.

The fact that Turkish students have not used most of these discourse items at all sounds interesting when this finding is compared with the findings of the studies on interlanguage. Focusing on French learners, De Cock (2004) finds out that discourse items like *I don't know, well* and its combinations, *sort of* are quite common in the learner language although their pragmatic functions are different from those in NS. Similar findings have been reported for Swedish and Norwegian learners as well depending on the comparison of LOCNEC and LINDSEI-SW and LINDSEI-NO. (Aas, 2011; Aijmer, 2004). Aas (2011) claims that use of such discourse items at a similar frequency with NS contributes a lot to perceived aspects of non-nativeness in advanced learner English speech. De Cock (2004) and Aijmer (2004) also propose the level of proficiency as the main reason for the overuse/underuse of these items implying that use of these conversation-specific word combinations frequently is the indication of higher level of English proficiency. In this respect, Turkish learners could be regarded less proficient in English. Another reason for the underuse of these sequences may be the learners' insufficient and imprecise use of the resources available to them or having difficulty in retrieving them automatically. If these combinations are easy to retrieve from the mental lexicon in real-time processing situations due to their entrenchment as formulaic patterns, it is likely that they appear often. This causes the learner language to be foreign-sounding. This finding together with Turkish learners' extremely frequently use of filled pauses as discussed in section 4.1.1.2 confirms the claim that they are unfamiliar with the discourse organization in English, which, in turn, bears crucial implications from a pedagogical viewpoint.

In conclusion, the quantitative results discussed so far have demonstrated that Turkish learners' spoken interlanguage differ greatly from the language of NS. First of all, their discourse is dominated by the frequent use of filled pauses and fragmentary sequences referring to the encoding problem of the learners, and this will affect the perceived fluency and non-nativeness of their speech. Second, the analysis of overused word sequences points to “phrasal teddy bears” effect at work in discourse organization of the Turkish learners, which means that they intensively rely on the expressions they know well. Finally, although they show similarities in the use of *I think* with the NSs, their use of other discourse items is strikingly different in quantitative terms. They either underuse or do not use at all most of the inherent sequences of spoken English, which results in a wide discrepancy between TL and NS in both frequency of recurrent phrases and in discourse organization. All these conclusions reached suggest that some pedagogical issues discussed in detail in chapter 5 should be considered to facilitate real-world interactional competence of Turkish learners.

4.2. Qualitative Findings: Spoken Corpus

As stated in chapter 1, one of the objectives of the present study is to delineate the spoken interlanguage of Turkish learners. Although the findings presented so far offer important insights into the learner speech, relying on only quantitative variations is likely to limit the deeper understanding of spoken interlanguage. This necessitates evaluating the data from a qualitative point of view as well. In order to answer the research question “What are the structural and functional features of recurrent sequences of two or more word combinations prevalent in spoken interlanguage of Turkish EFL learners?”, the data described in quantitative terms has been reconsidered and re-analysed. Only the 2-,3-, 4- and 5-word combinations are included in this new stretch of analysis since frequency search for longer combinations produced restricted results as shown in the foregoing section.

In the analysis for structural and functional features of recurrent phrases, previous frequency thresholds adopted in quantitative analysis has been changed. Following the studies in similar vein, different frequency thresholds were set in order to turn the data into a manageable size considering that this study incorporates 2-word sequences as well (Aas, 2011; Biber et al., 2004). Accordingly, frequency thresholds were set at 50, 15, 5 and 5 for the identification of 2-, 3-, 4- and 5-word combinations respectively. The search retrieved 129 2-gram types, 128 3-gram types, 195 4-gram types and 27 5-gram types from LINDSEI-TR; 85 2-gram types, 73 3-gram types, 84 4-gram types and 7 5-types from LOCNEC.

4.2.1. Structures of Recurrent Phrases in LINDSEI-TR in Comparison with LOCNEC

The qualitative analysis is inspired by Biber et al.,'s (2007)classification of word combinations according to their structural form, and pragmatic and textual function. Based on the Longman Spoken and Written English Corpus (LSWE) consisting of 40 million words from Britain and the United States in four registers (spoken conversation, fiction, news and academic prose), Biber et al., (2007) offer a taxonomy for the structural classification of word sequences in spoken language. Accordingly, there are three broad types of word sequences employed in spoken language: verb phrase fragments, dependent clause fragments and noun/prepositional phrase fragments, which are further divided into sub-categories (see section 3.4.2. for a detailed description of the taxonomy). Figure 8 is the graphic representation of the overall distribution of main structural types of word combinations in Turkish learners' spoken interlanguage and NS speech.

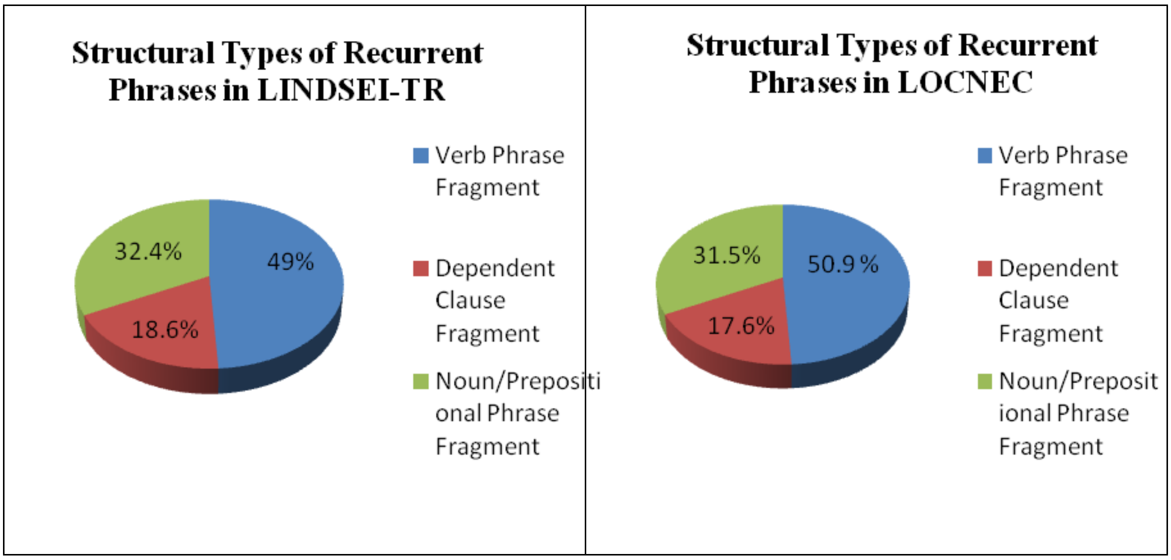


Figure 8. Distribution of major structural types in LINDSEI-TR and LOCNEC

As shown in Figure 8, structural types of word combinations show a very similar distributional pattern in the learner and native language. Verb phrase fragments in both corpora make up the biggest proportion in the overall structural types. This category includes such combinations as pronouns+VP fragments, active and passive verb phrase fragments, yes/no questions forms and wh- question fragments. This structural profile of data in Turkish learners' speech and NS speech is, to a great extent, similar to the data distribution across the structural types reported in Hernández, (2013). Based on three learner corpora compiled from both native and non-native speakers of English, Hernández, (2013) found out that the highest percentage of word combinations incorporate verb phrase fragments. This finding lends support to Biber et al.' (2004) who note that in conversation, almost 90% of all the word combinations include verb phrases, and even "50% of them begin with a personal pronoun + verb phrase" (e.g. I think it is, I don't know) (p.380). The second largest group of word combinations incorporates noun and prepositional phrases. Dependent clause fragments are the least frequently used structures in spoken language. These appear to suggest that verb, noun and prepositional phrase fragments are the single most important building blocks for on-going discourse, whether in native speaker or in learner spoken communication. Therefore, it is possible to claim that conversation is fundamentally phrasal rather than clausal, which is not surprising when the cognitive load of making full sentences in real-time production is considered. Although Turkish learners show similarities in general distribution of the structural types of word sequences, a detailed analysis dividing these combinations into more specific sub-categories reveals several differences between these two speaker groups. Table 33 displays the major recurrent phrases further categorized according to their structures with examples from both TL and NS corpus. (see Appendix G for the expanded list)

Table 33

Structural Categories of Recurrent Phrases in LINDSEI-TR and LOCNEC with Example Combinations

STRUCTURE	**EXAMPLES FROM LINDSEI-TR**	**Percentage (LINDSEI)**	**EXAMPLES FROM LOCNEC**	**Percentage (LOCNEC)**
1. Verb Phrase Fragments				
(connector +)1st/2nd person pronoun+VP fragment	***I think*** *(and I think)* ***I like*** *(and I like)* ***I want*** *(so I want, I want to)* ***I have*** *(eh I have, I have to)* *I can say,* ***you know*** ***I don't*** *(I don't know, I don't have any, I don't like, I don't know but,)* ***we don't know,*** ***I'm trying to***	***61.2 %***	***Ithink*** *(yeah I think)* ***you know*** *(you know you)* ***I mean*** *(yeah I mean)* ***I don't*** *(I don't like, I don't know)* ***I used to,***	***60.3 %***
(connector +) 3rd person pronoun+ VP fragment	***it was*** *(and it was, it was very)* ***it is*** *(it is a, it is not)* ***she doesn't like,*** ***she wants,****(she wants to)*	***22.5.%***	***it was*** *(and it was, it was just)* ***so that was,***	***28.3.%***
Discourse markers + VP fragment	*I think it is, of course I want*	**2.1. %**	*sort of you know, you know it was,*	**3.7%**
Verb Phrase (with non-passive verb)	*wants to, like it, talk about, have a lot of,*	**11.8%**	*paint a picture, know it,*	**3.7%**
verb phrase with passive verb	------		*was very impressed*	**1.8%**
yes/no question fragments	*can I say*	**1.0%**	----------	
Wh-Question fragments	*how can I say*	**1.4%**	*what else did I*	**2.2%**
2. Dependent Clause Fragments				
1st/2nd person pronoun+Dependent	*she wants him to, I don't know*	**16.6%**	*I know but, you know but*	**15%**

(Table 33 Contuniued)

Clause	*but, I don't know why*		*I, you know it was*	
Wh-clause fragments	*when I was a child, when I came to*	**11.1%**	*I don't know what, you know what, when I was there,*	**70%**
if-clause fragments	-------		-------	
verb/adjective+to-clause	*I want to do, want to have, she wants to be,*	**55.5%**	*like to go, like to see, to go to,*	**15%**
that-clause fragment	*I can say that, that there is*	**16.6%**	----	
3. Noun Phrase and Prepositional Phrase Fragments				
Noun phrase with –of phrase fragment	*of them, the end of the, a lot of, end of the film, the name of*	**22.2%**	*a bit of, a couple of, a lot of, awful lot of end of, sort of things, kind of*	**67.8%**
Noun phrase with other post-modifier	*plans for the future, good experience for me*	**5.6%**	----	
Other noun phrase expressions	*the picture, my friends, the woman, her picture*	**36.1%**	*a look at, two and a half, and things like that, or something like that*	**17.9%**
Prepositional phrase expressions	*in high school, for me, in a different way, from my family, in the picture, at the picture*	**36.1%**	*in the, at all, at the, at the moment*	**14.3**
Comparative expressions	------			
4.OTHER EXPRESSIONS	*once upon a time,*		*yeah yeah yeah,*	

The first thing seen through the detailed categorization of the recurrent phrases is the very frequent use of *first person pronoun+verb* combination, which is largely to be expected partly because of text- and task types, where the speaker is urged to talk about himself/herself. However, when the verbs combined with the pronouns are taken into account, a general tendency is observed in both corpora: that's the use of *know/think/want* after the *first person pronoun*. This corresponds to Biber et al.'s, (2007)observation, based on their conversational data, that “most of the sequences made up of following elements occur as recurrent [phrases] in conversation: I/you+know/think/want” (p.1001) which is often followed by a complement clause, as in the examples of *I don’t know but, I don’t know why,*(◊) form LINDSEI-TR and *you know what I* (◊) from LOCNEC. Secondly, detailed structural classification highlights the fact most of the 3-and 4-word combinations in the data consist of a highly frequent 2-word combinations. And some of these 2-word sequences are placed too high on the frequency rank mainly because of the high frequencies of longer combinations. Take the 2-word combination *I don’ t* (♦♦♦♦) in LINDSEI-TR which is embedded in *I don’t know* (♦♦♦) *and I don’t like* (♦) which together make up almost all of the instances of *I don’t*. The last similarity between two corpora lies in the use of *if-clause fragments* that do not appear in both corpora. This is most probably the result of contextual properties of data gathering tasks which don’t require the use of such clausal fragments.

A closer look at the sub-categories of structural types ascertains several differences between NS and TL speech as well. The different rates in terms of the use of *third person pronoun+VP fragment*, especially in the use of pronoun *it*, in NS (28.3%) imply that unlike non-native students, native students do not rely as much on their personal experiences. A similar findings is also reported by Hernández (2013) comparing spoken language of NS and NNS students. The structural sub-categorization of *dependent clause* indicates another difference in the use of *verb+to-clause* fragments and *wh-clause* fragments by NS and NNS students. While non-native students’ speech is dominated by verb+to-clause fragments (55.5%), native speakers use such structures at the rate of 15%. On the other hand, there is a big gap in the use of wh-clause fragments by NS and TL. NS employ *wh-clause fragments* intensively (70%) in their clausal constructions, which is in line with (Biber et al., 2007) who list wh-clause fragments in the frequent sequences of conversational English and demonstrate that native speakers often use such fragments as “ utterance launchers, presenting a personal stance dative to the information in the following complement clause” (p.1003) as illustrated in the following examples (16) and (17) taken from LOCNEC.

(16) *B> modern Eng= yeah like my major* ***what I want to do****is .modern English language <\B>*

(17) <B>erm great big <XX> which was brilliant ..erm ..the a l= a lot of ***what I liked about it*** *A= America in general was that some of the things were exactly how I'd seen . on T V and in the films <\B>*

However, Turkish learners use the wh-clause fragments only ate the rate of 11.1%. Comparing the use of clausal fragments in spoken and written texts, Kaltenbock (2004) points to "extra processing effort required by clausal constructions"(p.223) if they are not stored as automatically retrievable sequences. Accordingly, it is possible to claim that Turkish learners have small repertoire of automatically retrievable wh-clause fragments, they are likely to process such fragments on the basis of grammatical rules, which is relatively difficult in real time production. Educational background which is based on grammar rules rather than colloquial sequences make a possible explanation for this finding, which, in turn, bears pedagogical implications.

4.2.2. Functions of Recurrent Phrases in LINDSEI-TR in Comparison with LOCNEC

In addition to the structural analysis of recurrent phrases found in NS and TL speech, a functional analysis of these word combinations was carried out as well. Based on the taxonomy designed and used in a number of studies on both written and spoken language by Biber et al., (2007), the recurrent phrases in Turkish learners' speech and NS spoken language are classified into four broad categories: stance expressions, discourse organizers, referential expressions and special conversational expressions. Figure 8 illustrates the distributional pattern of recurrent phrases in terms of their functions in both TL and NS corpora. Biber et al., (2007) further divide these broad categories into sub-classes in accordance with the precise functions the word sequences perform. In chapter 3, these sub-categories were explained in detail (see section 3.4.2). However, the analysis of LINDSEI-TR and LOCNEC in terms of the functions of recurrent phrases demonstrated that not all sub-categories identified in the original taxonomy are found in the recurrent phrases. For example, phrases expressing impersonal stance or personal/impersonal ability do not appear in both corpora. Therefore, the original taxonomy offered by Biber et al., (2007) has been modified by deleting some categories and by adding a new category of function for the recurrent phrases. It should also be noted that some combinations appear in more than one category as they may perform multiple functions in different context, such as e.g. *I don't know,* which "does not have a single function but is characterised by its broad spectrum of uses"(Aijmer, 2009 p.156). Figure 9 represents the distribution of major functional categories across TL and NS speech.

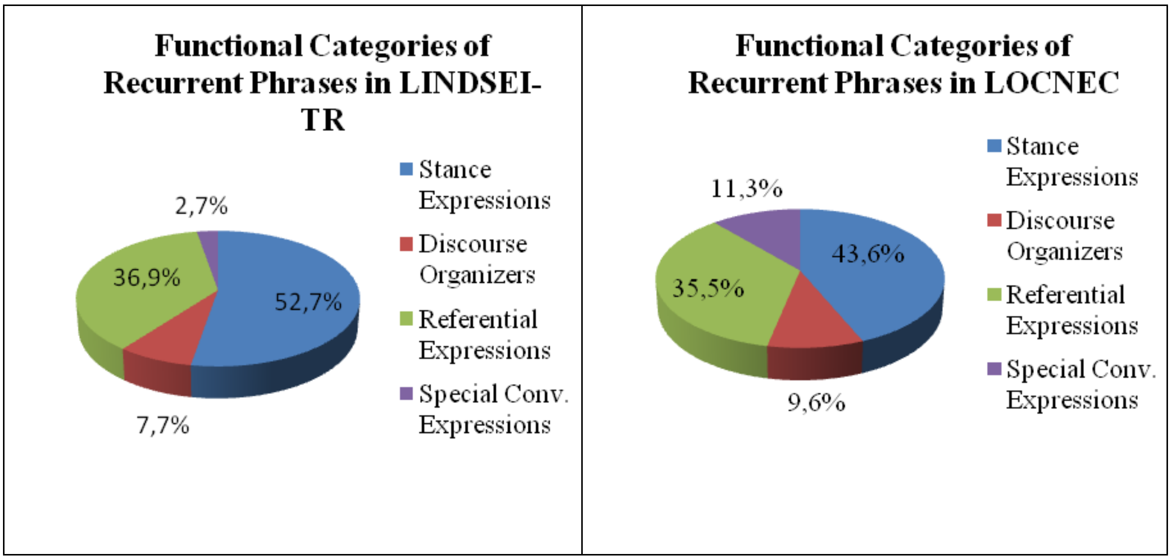


Figure 9. Distribution of major functional categories in LINDSEI-TR and LOCNEC

As seen in Figure 9, stance expressions that cover the word combinations expressing the user's attitudes, judgements and perspective which frame some other propositions have the largest proportion in both corpora. Stance expressions are evaluated in two groups: Epistemic stance and attitudinal stance which have personal and impersonal variations. Data analysis show that both TLs and NS preferred to use personal stance expressions in conveying their messages. This high proportion of personal stance expressions in spoken language is also observed by Biber and Conrad, (2004) who state that "the most striking aspect of conversation's use of word combinations is the high proportion of personal stance expressions" (p.67). Similar findings were also reported by Biber and Barbieri, (2007); Biber et al., (2007) and Hernández, (2013) concluding that personal stance expressions make up more than 60% of the typical conversation in English. Referential expressions have the second largest proportion, which is followed by the discourse organizers and special conversational expressions respectively. Special conversational expressions are more widely used by native speakers than Turkish learners, which points to Turkish learners unfamiliarity with the conversational English.

Table 34 further displays the functions of major recurrent phrases in LINDSEI-TR and LOCNEC with examples extracted from transcribed texts.

Table 34

Functional categories of recurrent phrases in LINDSEI-TR and LOCNEC

Categories		**Example from LINDSEI**	**Example from LOCNEC**
	Sub-Categories		
Stance Expressions	**A) Epistemic Stance**		
	Personal	***I think*** *and I think* *I think it is,* *I think that* *you know* *I don't know,* *I don't know but* *I don't know why*	***Ithink*** *I think it's,* *I think I,* *yeah I think* ***you know*** *you know it's,* *you know you,* *and you know,* *you know I,* *you know what I mean* *I don't know,* *I don't know I mean,* *I don't know I don't,* *I don't really know* *I don't know but* *I don't know if* *I thought it was*
	B) Attitudinal/Modality Stance		
	B1: Desire Personal	***I want*** *so I want,* *I want to,* *eh I want,* *and I want,* *I want to,* *eh I want to,* *I want to be,* *I want to eh,* *and I want to,* *so I want to,* *I want to do,* *I want to have,*	***I would like to,*** ***I wanted to*** *I want to be a,* *I don't like,* *I'd like to go*

(Table 34. Contuniued)

		I want to go, *I don’t like,* ***I would like***	
	B2: Obligation/Directive		
	Personal	*I have to,*	***you have to,*** ***I had to,*** ***you have to go***
	B3: Intention/Prediction		
	Personal	*I will* *I will be*	*I am going to* *I was going to*
Discourse/Text Organizers	**A) Topic Introduction/Focus**	*I want to talk* *firstly I want to*	*I want to talk about,*
	B) Topic Elaboration /Clarification	*for example*	*but I mean,* *I mean I was,* *what I mean,* *I mean it’s a*
	C) Topic Closing/Turn Yielding	*I don’t know*	*I don’t know*
Referential Expressions	**A) Identification/Focus**	*in this picture,* *of the film, the woman on the*	*that kind of thing,* *that’s the only,* *and it was,*
	B)Imprecision / Markers of Vagueness	*I don’t know*	*and things like that* *Or something like that,* *sort of* *sort of you know* *things like that* *you know sort of* *sort of thing* *sort of like*
	C) Specification of Attributes		
	C1: Quantity specification / Quantifying Sequences	*a lot of things* *there are lots of* *there is a woman* *have a lot of*	*there’s a lot of* *one of the* *a lot of* *a bit of* *a couple of* *two and a half* *there’s a lot of*

(Table 34 Contuniued)

	D) Time/Place/Text Reference		
	D1: Markers of Time	*at the same time,* *at the end of* *in my life* *in the first year*	*all the time* *at the moment* *in the morning* *all the time yeah*
	D2: Markers of Place	*in high school* *in the picture*	
	D3: Text deixis	*as I said* *as I said before*	
	D4:Multifunctional References	*at the end,* *in the middle* *first of all*	
Special Conversational Expressions	**A)Speech/Thought Reporting**		*I thought it was* *I thought that was*
	B) Responses	*okay okay okay*	*yeah that's right yeah* *yeah yeah yeah yeah* *yeah yeah it was* *yeah yeah that's right* *yeah yeah that's it*

A closer look at Table 34 shows that epistemic stance occupied a large place in NSs communication as they use more variations than TLs, whereas Turkish learners employed stance expression related to personal desire to a great extent. Regarding the intention/prediction expressions in attitudinal stance, both NS and TLs chose the personal expression rather than impersonal, which sounds meaningful when the fact that one of the tasks in data gathering was directly about the future plans of the interviewees is considered. However, the learner groups differ in terms of the phrasal units they chose to express their intentions and plans. Extracts (18), (19) and (20) from LOCNEC and (21), (22) and (23) from LINDSEI-TR exemplifies the difference.

(18) *<A> what are you doing in the: in the: .theatre group <\A><B> I'm helping out with lighting <\B><A> oh yes <\A><B>erm they don't really need me but **I'm** just basically **going to:** to learn how to <\B>*

(19) *<A> well you can see it .next week I think <\A>*
*<B> yeah **I'm going to I'm gonna** see it then <\B>*

(20) *<B>so .at that time . I think it taught me a .an important lesson because at that time I'd made lots of plans about what* ***I was going*** *to do in the future you know </B>*

(21) *<A> and so: (eh) now what are your plans for future what are you going to do after you graduate from this university </A><B> (eh) I . I answer* ***I will*** *answer the (eh) some exams <X> exams (eh) for example K P D S exam and (eh) A L E S exam (eh) in (eh) exam (eh) in . they are (eh) n= (eh) next week* ***they will be****<XX> next week </B>*

(22) *<A> okay and: last question what are your plans after university .or hopes </A><B> (em) .. I try to* ***I will*** *try to (eh) pass the <foreign> K P S= K P= K P S S</foreign>exam (eh)* ***I will I will*** *be a teach=* ***I will*** *work as a teacher </B>*

(23) *<B> it was very nice .I really don't want to graduate from <starts laughing> university <stops laughing> but this year* ***I will be*** *..yes that is all </B>*

As shown in utterances (18), (19) and (20), NS use combinations with *going to*(♦♦) to express their intentions, and this function of *going to* has also been identified in the studies on functional units of conversation (Biber & Conrad, 2004; Biber et al., 2007). However, Turkish learners use *will* (♦) to talk about their plans and intentions as shown in the extracts (21), (22) and (23). A possible reason for the choice of different linguistic units for the same function could be the effect of first language and their lack of knowledge regarding the functional distinction between *be going to* and *will*. In Turkish, while talking about future events, there is no distinction between the planned/intentional actions and unplanned events in terms of tense construction.

Another difference is observed in the choice of discourse organizers in topic elaboration/clarification. While NS prefer to use *I mean* and its combinations to clarify the previously stated idea, this discourse item is not so frequent in Turkish learners' data. This confirms the earlier findings by De Cock, (2004) and Huang, (2011)who also found that *I mean* is underused by non-native speakers of English. What is interesting is that in Turkish learners' speech, *for example* seems to serve the same function as illustrated in the extract (24)

(24) *<B> (eh) so he he argues but (eh) when his father (eh) learned that (eh) he he is ill (er) ..he accepts . his son (eh) . I I am affected (eh) from this film because (eh) it is very similar to my family . (eh) so (er) this was very sad film (eh) .. </B>*
<A> (uhu) </A>
<B> ***for example*** *(eh) my father (eh) have argued (eh) with his father so (eh) he tells all the time (eh) .what he feels about it </B>*

As is seen, the learner in his second turn explains why the film he is talking about is very similar to his family by using *for example* just before the clarification.

Additionally, the analysis of NS and TL speech revealed that a new sub-category of discourse organizers which is not proposed in the original taxonomy should be added to the classification of discourse items. When the use of *I don't know* is analysed in detail, it has been observed that it has an additional function apart from personal epistemic stance. Consider the following utterances (25) and (26) from LINDSEI-TR and (27) from LOCNEC.

(25) *<B> (eh) well (eh) . I can say that (eh) Turkish people are (eh) more (eh) friendly than (eh) Polish people because (erm) in fact .. for example I stayed there (eh) and no friends (eh) came and (mm) .. we didn't go: (eh) for example to parties so much with friends with classmates they were (em) they weren't so: . smiling</B>*
<A> (uhu) </A>
<B> (eh) maybe because of the climate ***I don't know****</B>*
<A> (uhu) </A>

(26) *<B> offers from the places I worked before </B>*
<A> (uhu) okay </A>
<B> they ask me but ***I don't know*** *but ...</B>*
<A>(uhu) okay how okay how do you spend your time what are your hobbies </A>

(27) *<B> yeah I think that might have something to do with it (erm)****I don't know*** *I've just always felt more comfortable in Ireland and that's maybe where I . I fit in and <\B>*

As seen in the examples, *I don't know* functions as a topic-closing sequence or as Aijmer (2009) demonstrates it has a floor-yielding function in the conversation. Aijmer (2009) further states that *I don't know* in the potential topic closing function may not always be followed by a new turn since the current speaker may choose to continue as shown in (26). What is more, Aijmer (2009) notes that this function of *I don't know* is especially common in interviews, which explains the occurrence of this function of *I don't know* in LINDSEI-TR and LOCNEC.

Regarding the referential expressions, what is conspicuous at first glance in table 34 is the sub-category of imprecision/markers of vagueness. Vagueness tags are the indicator of intersubjectivity and they have a crucial role in informal spoken communications, signalling an assumption of shared experience and social closeness (Aijmer, 2002b; De Cock, 2004). While the instances of vagueness tags for imprecision are very frequent in NS, they haven't appeared in NNS spoken language at all. This is the case with the learner groups in Aas's (2011) and De Cock's (2004) studies. Therefore, it could be claimed that the underuse of vagueness tags or even their not being used in Turkish learners' speech is an idiosyncratic feature of spoken interlanguage. Lack of imprecision in an informal conversation is a reason explaining foreign-soundings of the speakers. Thus, it is likely that Turkish learners' speech sounds non-native as they do not organize their discourse using the characteristics of the informal talk. This finding along with the use of filled pauses too frequently and the significant differences from the NS in terms of frequently used combinations (see sections 4.1.1.2 and 4.1.2) confirms the claim that Turkish learners are unfamiliar with the spoken English. The last category of the functions of recurrent phrases in NS and NNS lend further support to the findings above. As seen in table 31, Turkish learners produced repetitive *okay* for the response function; however, native speakers use a number of variations of *yeah,* which is a distinctive item in spoken English.

In conclusion, this study has set out to explore Turkish EFL learners interlanguage characteristics and compare and contrast the Turkish EFL learners' and the native speakers' use of recurrent phrases across spoken and written corpora in terms of both quantitative and functional variation. The foregoing findings and discussions are one part of realizing this objective. The overall results revealed that although there are some similarities between the Turkish learners' speech and NS speech, Turkish learners' spoken interlanguage has some unique features and differ markedly from the NS in many areas: 1) Filled pauses such as eh, em, er, erm are rather prevalent in the recurrent combinations in their speech. The most frequent one is *eh* often combined with the conjunction *and*, which implies a difference from the frequently used filled pauses in NS speech (Kjellmer, 2003). The overuse of hesitation items in learner data could be regarded as the indication of encoding problems Turkish learners have due to either limited lexical knowledge or general planning pressure that stems from the difficulty of expressing themselves in a foreign language. 2) While the use of *I think* by Turkish learners is in line with the NS and other learner groups reported in the literature (Aijmer, 2004; Biber et al., 2007; Huang, 2011; Yong et al., 2010), their underuse of *you know* and *I mean* is an important area of difference, which gives an idea about discourse organization of Turkish learners of English. 3) Regarding the 4-word combinations, Turkish EFL learners show similarities to NS in their choice of common mental verbs (*want* as in *and I want to, like* as in *eh I like* or *know* as in *I don't know but*), yet they differ strikingly in the use of *how can I say* as an explicit word search marker, which is seemingly peculiar to Turkish learners. 4) Quantitative analysis of the longer combinations suggests that Turkish learners have a repertoire of word combinations occurring more frequently than expected by chance, and their interlanguage do not solely consist of individual building blocks. 5) The overused and underused sequences in comparison with NS show that Turkish learners preferred ways of saying things are strikingly different from that of NS as shown by the high Log Likelihood statistics. This suggests that Turkish learners are unfamiliar with characteristics of the native-like conversation. 6) The overuse of such word combinations as *for example, because of, and eh* and the inappropriate use of *of course* are among the idiosyncratic features of Turkish spoken interlanguage. 7) The comparison of NS and NNS in the structural properties of their spoken language demonstrated that Turkish learners are identical in terms of using verb phrase fragments while they show significant differences in the wh- fragments which are commonly found in NS speech. 8) Finally, the functional comparison of recurrent phrases in NS and NNS speech indicated that even though they are almost the same in the use of stance expression in their speech as the NS, Turkish learners perform different tendencies in realizing the functions of "intention/prediction, topic-closing, imprecision and special conversational responses". One explanation of these differences could be the effect of first language, or instructional background of the Turkish learners may make a reason for these occurrences as touched upon in the foregoing sections.

Whatever the reason is, these findings bear significant pedagogical implications, which is presented in chapter 5.

To realize the objective of the present study in full terms and to have relatively sound claims about the interlanguage characteristics of Turkish EFL learners, written data collected from Turkish learners have been analysed as well. What follows is the presentation of the findings regarding the written interlanguage of Turkish learners in comparison with native written language.

4.3. Quantitative Findings: Written Corpus

4.3.1. Recurrent Phrases in Written Interlanguage

The fourth research question this study aimed to answer was “what are the major recurrent sequences of two- to six-word combinations Turkish learners tend to use in their written interlanguage? To answer this question, simple n-gram searches in the Turkish sub-component of ICLE were carried out. Different frequency cut-off points were adopted for different length of combinations as “the length of recurrent word combinations is inversely related to their frequency” (De Cock 2004: 228). Following the literature(Y.-H. Chen & Baker, 2010; Cortes, 2002b) and considering the size of the corpus, frequency cut-off points were identified as 12, 6, 4, 3 and 3 for 2-, 3-, 4-, 5- and 6-word combinations respectively. Table 35 displays the overall statistics for the TICLE.

Table 35

Overall Statistics for TICLE

Statistical Categories	Numbers	Statistical Categories	Numbers
Overall Corpus Size	182772	5-letter words	21346
Tokens used for word lists	182514	6-letter words	14608
Types	8292	7-letter words	12942
Type/Token Ratio	4.54	8-letter words	9473
Std. Type/Token Ratio	4.55	9-letter words	6682
Av. Word Length	4.53	10-letter words	4755
1-letter words	6108	11-letter words	2126
2-letter words	31458	12-letter words	1052
3-letter words	37207	13-letter words	587
4-letter words	32684	14(+)-letter words	178

As seen in Table 35, TICLE consists of a total of 182,772 words and WordSmith Tools excluded 258 tokens from the analysis. The overall search produced 8292 types (distinct words) with a type/token ratio of 4.55.

4.3.1.1. 2-Word Combinations inTICLE

As stated earlier, most of the researchers working with written corpora have focused on the 3- or 4-word combinations due to the assumption that longer sequences are more likely to hint idiosyncrasies. Still, n-gram searches including bigrams are also appreciated as they can make a fruitful starting point in terms of making assumptions about the nature of learner language, revealing interesting data and suggesting hypotheses that can be followed in future research (Granger, 2008). Therefore, this study has analysed the frequency of 2-word combinations as the first step to delineate the characteristics of Turkish learners' written interlanguage. Table 33 exhibits the top-twenty 2-word sequences in TL written language. (see Appendix H for the expanded list)

Table 36

Top Twenty 2-word Combinations in TL Written Language (TICLE), freq. > 12, and Their Raw Frequencies

Rank	Word Sequence	Frequency	Rank	Word Sequence	Frequency
1	of the	1223	**11**	there are	338
2	in the	884	**12**	is a	333
3	it is	794	**13**	there is	316
4	they are	470	**14**	as a	314
5	to the	420	**15**	should be	307
6	to be	417	**16**	and the	281
7	do not	375	**17**	the other	276
8	is the	374	**18**	they can	271
9	for the	354	**19**	in a	267
10	is not	346	**20**	the world	263

The top-twenty list shows that the most frequent bigrams in TICLE are grammatical words or function words, which is not surprising as they generally occupy the top positions in any corpus, whether it is a learner or a native corpus. Ebeling (2011)states that "in combination with other function words or content words, function words are important building blocks in the phraseology of a language (p. 54). Their being so frequent is reported in Granger's (2008) comparison of French, German and Czech learners of English based on the relevant components of ICLE. Top bigrams used by these three learner groups are almost the same as Turkish learners with the exception of *there is, there are* and *should be* ranking highest in TICLE; *will be* in the French component and *have to* in the German part of ICLE, which could make an interesting base for future research focusing on grammatical units across different interlanguages. Only sequence out of grammatical units occurring in the top-twenty list is the expression *the world*. Its being so frequent could be linked to the essay topic Turkish students preferred to write on. As explained in chapter 3, Turkish learners chose the topic “most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value” among the given topics. While discussing such a matter, it is quite natural to refer to instances in the real world.

When the whole frequent bigrams list excluding grammatical units is examined, it is seen that such 2-word combinations as *for example (♦♦♦♦♦), because of (♦♦♦♦♦), the same (♦♦♦♦♦), the other (♦♦♦♦♦), of course (♦♦♦♦♦), I think (♦♦♦♦♦)* are high up on the list of bigrams in academic writing of Turkish learners, and this may, in fact, point to “characteristic trait of academic writing as regards lexical choice” (Ebeling, 2011 p.56). Still, the frequency list of bigrams seem to provide less evidence to have strong claims about Turkish learners written interlanguage as it contains a few defining features including content words. Thus, it is necessary to examine the longer combinations, which is presented in the following section.

4.3.1.2. 3-Word Combinations in TICLE

N-gram searches for the 3-word combinations in TICLE retrieved more than 1500 different sequences. Table 37 displays the top-twenty trigrams occurring in Turkish learners’ written interlanguage (see Appendix I for the expanded list).

Table 37

Top Twenty 3-Word Combinations in TL Written Language (TICLE), freq. > 6, and Their Raw Frequencies

Rank	Word Sequence	Frequency	Rank	Word Sequence	Frequency
1	a lot of	115	**11**	the most important	93
2	it is not	112	**12**	most of the	87
3	men and women	112	**13**	the real world	75
4	on the other	107	**14**	as a result	74
5	they do not	107	**15**	there is a	73
6	in order to	105	**16**	there are some	70
7	there is no	99	**17**	it is a	69
8	in the world	96	**18**	it is the	67
9	one of the	95	**19**	do not have	62
10	the other hand	95	**20**	that it is	61

The first thing observed from the table is the sharp decrease in the frequency rates of trigrams. While the most frequent bigram *of the* occurs 1223 times in the corpus, the most frequent trigram occurs115 times in the same data. This is because of the fact the length and frequency are inversely related (De Cock et al., 1998; De Cock, 2004). That is to say, the longer the combination is, the relatively lesser the frequency it has. Secondly, top-twenty 3-word combinations of Turkish learners' written language are consistent with the findings reported in the literature. Using a corpus of student writing (BAWE) and academic prose part of BNC, Ebeling (2011) listed most frequent 15 trigrams occurring in both corpora. Almost half of the trigrams used by Turkish students in their writings are in agreement with those 3-word combinations found in BAWE and BNC academic prose (e.g. *it is not, on the other, in order to, there is no, the other hand, there is a, it is a, that it is*). Biber et al. (2007) also identify the combinations *in order to, one of the, part of the, the number of; in the presence; the use of; the fact that, there is a, there is no* as the most common three-word sequences in academic prose. Four of these sequences (*in order to, there is no, one of the, there is a*) have also appeared in the top-twenty list of the Turkish learners most frequent trigrams. A close look at the whole list of trigrams shows that except the sequence *in the presence* which doesn't occur at all, the other combinations identified by Biberet.al., (2007) are also frequent in Turkish learners' written language. This finding suggests that these 3-word combinations are important building blocks of academic prose, and the essays written by Turkish students of EFL bear similarities to the conventions of academic prose in terms of preferred word combinations. Another observation that can be made on the basis of the 3-word combinations is that Turkish learners' essays mostly include present tense forms (e.g. it is not, there is no, they do not, there is a, there are some etc…). This tendency is pointed by Biber et al., (2007), whose corpus findings indicate that along with conversation, academic prose shows "a strong preference for present tense forms" (p.455). Finally, a closer look at the content words in the top- twenty list reveals that they do not have very specific content (e.g. *men and women, in the world, the most important*) though they give some clues about the contents of the essays when interpreted with the knowledge of topics chosen by the students. This suggests that analysing Turkish learners' written English based on structural patterns (e.g. pronoun/noun phrase + be +…) instead of exact lexical occurrences would be more rewarding in identifying the characteristics of Turkish learners' written interlanguage. Such an analysis is provided in section 4.4.1.

4.3.1.3. 4-Word Combinations in TICLE

The four-word scope is ''the most researched length for writing studies, probably because the number of 4-word combinations is often within a manageable size" (Chen & Baker, 2010 p.32), and because it is proposed that they contain the smaller combinations, are more common than 5-word sequences and present a wider range of structures and functions(Cortes, 2004; Hyland, 2008b). Compared with the 2- and 3-word sequences, the search for 4-grams retrieved lesser types (802 different sequences). Table 38 shows the top-twenty most frequent sequences found in TL writing. (See Appendix J for the expanded list)

Table 38

Top Twenty 4-Word Combinations in TL Written Language (TICLE), freq. > 4, and Their Raw Frequencies

Rank	**Word Sequence**	**Frequency**	**Rank**	**Word Sequence**	**Frequency**
1	on the other hand	94	**11**	do not prepare students	27
2	one of the most	41	**12**	degrees are theoretical	27
3	for the real world	40	**13**	prepare students for the	26
4	is one of the	37	**14**	not prepare students for	24
5	between men and women	36	**15**	at the same time	23
6	students for the real	34	**16**	by the help of	23
7	as a result of	30	**17**	most university degrees are	23
8	are a lot of	30	**18**	do not want to	22
9	there are a lot	30	**19**	to be able to	22
10	all over the world	29	**20**	both men and women	21

As Table 38 indicates, the most frequent 4-word sequence in TICLE is *on the other hand,* which is identified as the commonest 4-word combination in the literature. Inthe Longman Grammar of Spoken and Written English (LSWE), Biberet.al., (2007) state that the two most common four-word sequences are *in the case of* and *on the other hand* in academic prose. Although *on the other hand* is very frequently used in the writings by Turkish learners, *in the case of* is underrepresented as it occurs only in four instances. The other top-twenty sequences also overlap with the findings in the literature. Among them, *at the same time, one of the most, is one of the, as a result of* and *to be able to* are highlighted as building blocks of written English in various studies (Bal, 2010; Cortes, 2002b; Ebeling, 2011). According to Biber et al., (2007), in academic prose, more than half of the all 4-word combinations are parts of noun phrases or prepositional phrases, and function words (articles, prepositions and complementizers) and they generally appear as the ending word of these sequences. This combination is observed in Turkish learners' written English as well, suggesting that their essays have the similar properties to academic prose. Although Turkish learners' 4-word sequences are mostly in line with those identified in the literature, the combination of *by the help* of is worth mentioning here. The use of *by the help of* has not been reported in any corpus studies. Several studies that found the word *help* in clusters generally report it to be combined with the preposition *with (e.g. with the help of)* (Ädel& Erman, 2012; Bal, 2010). Searching the word *help* with the preposition *by* in the dictionaries (Longman Dictionary of Contemporary English, Cambridge Dictionary, Oxford Collocation Dictionary) produced no results; rather the preposition *with* is recommended for the combination. Using the simple search tool of BNC, the combinations *by the help of* and *with the help of* have been compared. The result shows that while *with the help of* occur in 834 instances, the sequence *by the help of* appear only 6 times which is extremely low in a corpus of millions of words. This demonstrates that the combination *by the help of* is really infrequent in English; and Turkish learners interestingly chose to use this infrequent sequence instead of combining it with the preposition *with* as shown in the example sentences from TICLE.

(28) *In kinder gardens, children who are five or six, are learning English,* ***by the help of*** *computers.*

(29) *Nowadays, it is very popular to find friends* ***by the help of*** *computers*

(30) ***By the help of*** *the computers we can find the most detailed forms of what we need to learn in a very short time, by only using some keys on the keyboard of the computer.*

All these suggest that *by the help of* is somewhat idiosyncratic for Turkish students' written interlanguage. One possible reason for this could be related to overgeneralization which is underlined as one of the significant factors affecting interlanguage development in the interlanguage theory of Selinker (1972). Overgeneralization refers to cases in which rules or semantic features of the target language may be overextended to any language items. When the instances of preposition *by* in TICLE are examined, it is seen that Turkish learners know its use semantically similar to the preposition *with.* Extract (31) exemplifies the use of *by* in the meaning of *with*; (32) and (33) illustrate the misuse of *by* similar to *with*.

(31) *These movies are full of violence scenes. In these movies people fight with each other* **by** *different guns, swords and kill each other*

(32) *Since husband is always occupies with internet, he doesn't pay attention to his wife very much and for that reason wife wants to get divorce from her husband or vice versa. At the end, this event finishes* **by** *divorce unfortunately*

(33) *Only one computer can be loaded millions of information. Whenever you want to learn something you can do it* ***by*** *a computer*

As shown in example sentences, the learners have learned the meaning of *by* as something similar to *with*, and they have extended it in the expressions where *with* would be better. Thus, it is possible that they have overextended this meaning of *by* to the combinations including *help*.

4.3.1.4. 5- and 6-Word Combinations in TICLE

Since the automatic retrieval of 5-and 6-word sequences produced relatively lesser types, they are decided to be handled together under the same title. As stated in the literature and as shown in the foregoing sections related to spoken corpus, the 5-and 6- word sequences generally incorporate 2-, 3- and 4- word combinations; therefore, it is difficult to reach distinctive conclusions. However, these longer combinations are kept in the present study with the purpose of having analysed the data thoroughly. Tables 39 and 40 display the top-twenty 5- and 6-word combinations respectively found in TL written English.

Table 39

Top Twenty 5-Word Combinations in NNS Written Language (TICLE), freq. > 3, and Their Raw Frequencies

Rank	**Word Sequence**	**Frequency**	**Rank**	**Word Sequence**	**Frequency**
1	there are a lot of	30	**11**	are theoretical and do not	15
2	prepare students for the real	26	**12**	degrees are theoretical and do	14
3	students for the real world	26	**13**	on the other hand the	14
4	do not prepare students for	24	**14**	theoretical and do not prepare	14
5	is one of the most	22	**15**	on the other hand some	12
6	university degrees are theoretical and	21	**16**	there is no need to	12
7	most university degrees are theoretical	20	**17**	when we look at the	12
8	one of the most important	20	**18**	as a result of this	11
9	not prepare students for the	18	**19**	by the help of the	11
10	and do not prepare students	16	**20**	equality between men and women	11

Table 40

Top Twenty 6-word Combinations in NNS Written Language (TICLE), freq. > 3, and Their Raw Frequencies

Rank	**Word Sequence**	**Frequency**	**Rank**	**Word Sequence**	**Frequency**
1	prepare students for the real world	19	**11**	that most university degrees are theoretical	8
2	do not prepare students for the	18	**12**	one of the most important inventions	7
3	not prepare students for the real	18	**13**	do not prepare students for real	6
4	most university degrees are theoretical and	16	**14**	prepare students for the real life	6
5	and do not prepare students for	15	**15**	money is the root of all	5
6	are theoretical and do not prepare	14	**16**	the freedom of the press is	5
7	degrees are theoretical and do not	14	**17**	a lot of women working as	4
8	theoretical and do not prepare students	14	**18**	are a lot of women working	4
9	university degrees are theoretical and do	14	**19**	don t prepare students for the	4
10	is one of the most important	11	**20**	every human being has a right	4

As can be seen in Tables 39 and 40, most of the Turkish learners' 5- and 6-word combinations are contextual (e.g. *university degrees are theoretical and do, don t prepare students for the, prepare students for the real world etc…*). In other words, they recur in the data as they are required by the topics they chose to write on. However, there are some longer combinations which are not necessarily contextual. The combination *one of the most important* and *as a result of this* are among them, and they are also identified by Biber et al., (2007) as the recurrent building blocks of academic prose. The combination of *when we look at the* seems idiosyncratic for Turkish learners as it has not been identified in the literature. In their comprehensive study of written and spoken English, Biber et al., (2007) found that only four combinations begin with an adverbial clause in academic prose, and the adverb which is mostly preferred is *as* as in the examples of *as we have seen, as shown in the following figure* etc… Regarding the use of first person plural pronoun *we*, Biber et al., (2007) state that the use of *we* rather than *I* is the indication of the impersonal writing; however, when *we* is used to include the reader, then the writing becomes more personal. The use of *when we look at the* by Turkish learners is seemingly an effort to include the reader as seen in the following extracts from TICLE.

(34) *When modern computers were first used in 1950s, they weren't very common, but* ***when we look at the*** *end of the 20th century, we can see them everywhere around us, even in nearly every homes.*

(35)*... they are essentially in front of our eyes.* ***When we look at*** *our government we see that the rate of woman deputies' number stands very little when compared with the rate of man deputies' number.*

(36) *And* ***when we look at the*** *business world, we see that only few employer provide crèches for young children in order to encourage women to work for them, thus they discourage women from working.*

In most of the instances in which *when we look at the* appears, it is followed by the combination *we+see* suggesting that learners assume a shared experience, knowledge or beliefs. This assumed shared ground through the combinations of inclusive *we* is one of the features of learner writing distinguishing it from the expert writing (Luzón, 2009).

All in all, n-gram searches of TICLE demonstrate that Turkish learners' written language has a stock of recurrent phrases put together more frequently than expected by chance. It is not surprising that the proportional weight of shorter sequences is higher than that of longer sequences. Figure 10 summarizes the distributional proportion of recurrent phrases according to the length of sequences.

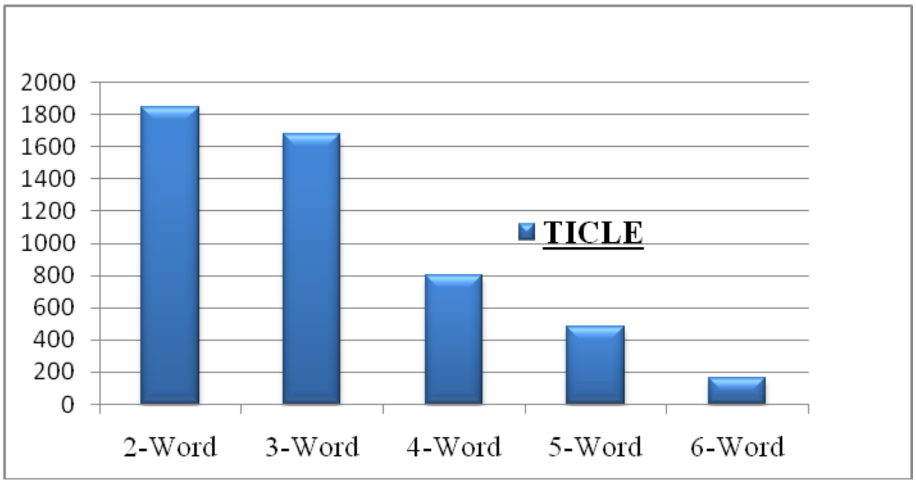


Figure 10. Proportion of most frequent 2-6-word combinations in TLs writing (TICLE)

As summarized in Figure 10, bigrams and tree grams have the largest proportion in Turkish learners written English. The number of recurrent sequence types and tokens decline as the length of combination increases, and the sharp decline between the 3-word and 4-word sequences is notable. This finding is in line with the literature (Guan & Zheng, 2005). Biber and Conrad (2004) found that three word sequences are more frequent than four word combinations. The same observation is echoed in Biber et al., (2007) who note that "there are almost ten times as many three- word lexical bundles as four-word lexical bundles, in both conversation and academic prose. Similarly, there are about ten times as many four-word lexical bundles as five-word lexical bundles" (p.993).

The overall quantitative analysis of 2-to 6-word combinations in Turkish learners' written interlanguage suggest that their building blocks to compose written texts are mostly consistent with the findings in the literature although there are some differences seemingly peculiar to Turkish learners. Grammatical units make up the biggest part in the most frequent 2-word combinations and the word sequences such as *for example, because of, the same* could be regarded as the characteristic lexical choice of Turkish learners written English. With respect to the combinations of 3- and 4-word sequences, Turkish learners are found to use most of the sequences identified in the literature (Biber et al., 2007); and 5-and 6-word sequences are mostly contextual closely related to the topics chosen by the learners. The exceptional combinations in written interlanguage of Turkish learners are *in the case of* which occur less frequently; *in the presence of* which does not appear at all; *by the help of* which points to some kind of overgeneralization affecting the interlanguage development; and *when we look at* which suggests that written interlanguage of Turkish learners is relatively personal compared with expert writing.

4.3.2. Recurrent Phrases in Native Language: LOCNESS in Comparison with TICLE

Apart from identifying the major recurrent word combinations in Turkish learners written interlanguage, this study has set out to find the similarities and differences (if any) between learner and native writing. The following research question was formulated: “To what extent are recurrent sequences in NNS written interlanguage similar to and/or different from those in NS written language?” To answer this question, LOCNESS has been analysed in terms of recurrent word sequences. As explained in earlier chapters, LOCNESS was compiled considering the same learner and task variables as ICLE; therefore, it serves as a comparison base for investigating NS and NNS varieties. In the analysis of word combinations in LOCNESS, the same frequency thresholds which were set for TICLE were adopted. Figure 11 presents the overall comparison between NS and NNS written language types in accordance with recurrent phrases of different lengths.

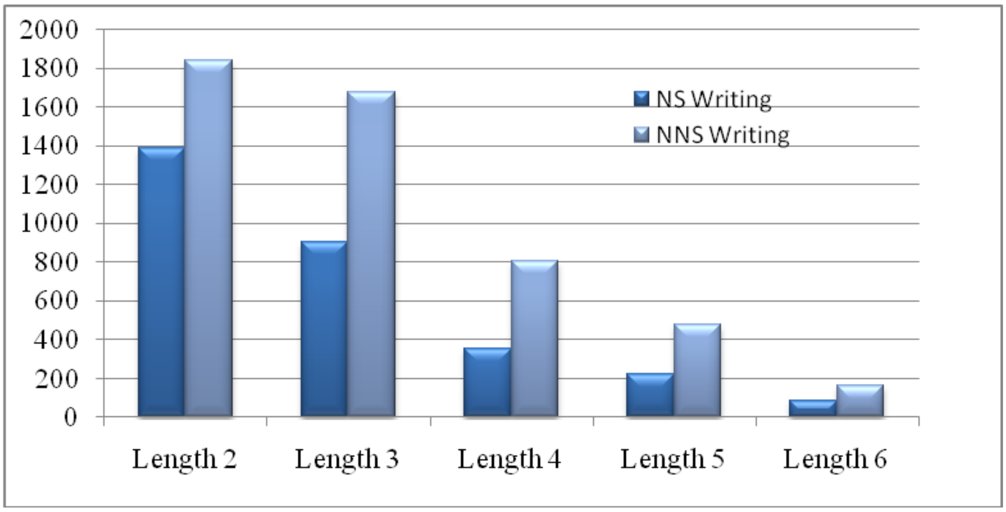


Figure 11. Types in NS writing vs types in TLs writing

As Figure 11 shows, there are more recurrent phrase types at length 2, 3, 4, 5 and 6 in NNS writing than NS writing but the difference is not statistically significant (at $p \leq 0.05$) at all lengths. The use of more recurrent phrases in learner language in comparison with native language has been reported in some other studies as well. In their investigation of Chinese learners and native writing, Guan and Zheng (2005) found that Chinese learners use more recurrent phrases in their writing. Overall conclusion to be reached based on figure 11 is that NS and TL writing do not simply contain individual words, but are made up partly on the basis of larger memorized sequences, as is the case in NS and TL spoken language shown in section 4.1.2. Within the phraseology context, this finding lends further supports to the claim that “words belong with other words not as an afterthought but at the most fundamental level” (Wray, 2002 p.13) both in NS and NNS writing.

Table 41 shows the most frequent twenty 2-to 6-word combinations in NS writing, which could give an idea about the recurrent phrases common to both NNS and NS written language (cf. Table 36, 37, 38 and 39).

Table 41

Top Twenty 2- to 5-Word Combinations in NS writing (LOCNESS), freq. >12/6/4/3, and their raw frequencies

Rank	**2-Grams**	**3-Grams**	**4-Grams**	**5-grams**
1	of the (1174)	the united states (108)	in the United States (56)	lowering the drinking age would (14)
2	in the (873)	the fact that (75)	prayer in public schools (25)	the root of all evil (14)
3	to the (456)	the death penalty (71)	on the other hand (24)	the teaching of new age (13)
4	it is (443)	one of the (67)	the invention of the (23)	due to the fact that (12)
5	to be (430)	in order to (66)	the joy luck club (21)	of prayer in public schools (12)
6	on the (290)	ethnic American literature (57)	one of the most (20)	one of the most important (9)
7	that the (288)	in the united (56)	is one of the (18)	the invention of the airplane (8)
8	for the (278)	because of the (52)	the end of the (18)	against the teaching of new (7)
9	is a (272)	the right to (52)	of the # century (17)	case against the teaching of (7)
10	and the (270)	the use of (52)	the death penalty is (17)	in favor of capital punishment (7)
11	is the (245)	that it is (51)	as a result of (16)	is one of the most (7)
12	as a (235)	be able to (50)	in the case of (15)	teaching of new age ideas (7)
13	is not (215)	as well as (49)	lowering the drinking age (15)	the case against the teaching (7)
14	with the (215)	should not be (46)	of ethnic American literature (15)	at the end of the (6)
15	they are (211)	it is not (42)	of the United States (15)	in the united states and (6)

(Table 41 Contuniued)

16	in a (190)	this is a (40)	the drinking age would (15)	of the joy luck club (6)
17	this is (187)	there is no (38)	of the death penalty (14)	teaching of new age beliefs (6)
18	of a (185)	in public schools (36)	root of all evil (14)	the drinking age would allow (6)
19	should be (171)	it is a (36)	the root of all (14)	an example of this is (5)
20	can be (170)	all of the (34)	to the fact that (14)	

The combinations common to both NNS and NS corpora in the most frequent 20 word sequences are underlined in the table.

When the top twenty most frequent recurrent phrases in LOCNESS are compared with those in TICLE (cf. Tables 36, 37, 38 and 39), it is observed that there is an overlap to a great extent in the bigrams used by learners and native students while the proportion of similarity declines in other lengths. This is because bigrams in both corpora mostly comprise of function words. The comparison of 3 grams shows that both NNS and NS use the combinations of *one of the, in order to, that it is, it is not, there is no, it is a* more frequently expected by chance lending further support to the claim that these are among the basic building blocks of written discourse. The combinations *the fact that, the use of* and *in the case of* used very frequently by the NS are noteworthy. Although these combinations are also identified by Biber et al., (2007) as the common lexical choices of written language, they appear relatively less in Turkish learners data (i.e. the fact that ♦♦, the use of ♦♦ and in the case of ◊) suggesting a difference between NS and NNS writing.

To have a deeper analysis of the differences and similarities between NNS and NS writing, the word sequences in two corpora were compared, and the overused and underused word combinations were identified. Table 42 displays the overused 2-5-word sequence types in Turkish learners writing.

Table 42

Overused Word-Combinations (freq. >12/6/4/3) Occurring in TICLE in Comparison with LOCNESS, Raw Frequencies and Relevant Log Likelihood Statistics)

Word Combinations	**LOCNESS**	**TICLE**	
Overused Phrases	**Freq.**	**Freq.**	**Log Likelihood**
can not	18	242	+210.74
the students	32	249	+172.09
they can	44	271	+162.92
we can	32	231	+153.17
kind of	14	119	+86.06
of course	26	126	+63.28
I think	48	163	+56.69
lot of	31	115	+44.45
the woman	12	163	+142.42
they are	211	470	+79.95
the real world	-	75	+97.44
in the world	21	96	+45.82
the most important	20	93	+45.07
I want to	-	49	+63.79
day by day	-	47	+61.19
in my opinion	9	60	+37.95
to sum up	-	38	+49.47
as a result	23	74	+23.99
we can see	-	28	+36.45
there are some	12	70	+40.55
first of all	9	60	+37.95
on the other hand	24	94	+38.57
by the help of	-	23	+29.94
in addition to this	-	16	+20.83
we can say that	-	15	+19.53
there are a lot of	-	30	+39.06
when we look at the	-		+15.62
there is no need to			+15.62

+ (p<0.05)

For each word sequences in Table 42, Log Likelihood test, which was explained in section 3.4.1 was carried out to make sure if the observed frequencies are statistically significant. The frequency differences are all proven to be significant in statistical terms (p<0.05). The overused word sequences in Table 42 together with the whole list of overused combinations seem to reflect several noteworthy tendencies related to the Turkish learners' use of phraseology in their written English. Initially, Turkish learners overuse the combinations inclusive *can* (as in *they can, we can, we can say that, we can see* etc…), which refers to a degree of uncertainty in arguing in English. In the literature on learner language, a set of modal verbs were identified (i.e. can, would, could, must, have to, should, may, might, ought to, shall), which points to an area of difference between NS and NNS in terms of discourse organization (Aijmer, 2002a). Among them, *can* along with *could*, *may* and *might* is categorized as a possibility modal. One explanation of Turkish learners' overuse of *can* in their preferred sequences may be that it results from their uncertainty in arguing in English. Consider the following examples from TICLE.

> **(37)** *We* ***can*** *make classification of 20th century inventions or discoveries and* ***can*** *give the meaning of 20th century inventions as "the improved models of the things invented in the past. " We <u>may</u> classify the first group as "the inventions in medical science*
>
> **(38)** *The most important invention of the 20th century is the internet. It has a great effect on our life in every field. We* ***can*** *make use of internet almost in everything. It has many advantages on our life as well as disadvantages. As in every technological invention internet also* ***can*** *be misused. Of course, to decrease the misuse depends on us.*

In (37), the student writer explains his/her stance regarding the 20th century but does this with uncertainty. The use of *may* in his following sentence verifies this feeling of uncertainty. In (38) where a replacement of *can* with a passive structure would be better, the use of *can* further exemplifies an uncertain point of view about the internet.

Learners' being uncertain in their arguments is also reported in Aijmer (2002b) who studied the use of modal verbs by learners from different L1 groups. Although the learners in Aijmer's (2002b) study are regarded similar in terms of their overuse of modal verbs referring to possibility, they are different in the choice of special modal auxiliaries of possibility. Accordingly, *can* is overused by German learners; *may* is overused by French learners and *might* is overused by all learner groups in the study. Depending on this finding, Aijmer (2002b) claims that the overuse of *might* is a common feature of learner language. However, the findings of the present study does not support this claim as *might* is not found among the overused combinations, even the sequences inclusive *might* is underrepresented in Turkish learner data. This suggests a difference in the written interlanguage of Turkish learners. One possible reason of Turkish learners' preference of *can* over the other possibility modals could be linked to the instructional factors. In many language teaching materials/books, the modal *can* is generally taught before the other modal verbs and reinforced throughout the instructional process. Yet, another possible reason of the overuse of the *can* in word combinations could be the influence of speech. Although *can* is reported to be common both in academic prose and conversation, its use in conversation by NS is markedly more frequent than in writing (Biber et al., 2007), and NS exclusively prefer *could, may* and *might* to mark possibility in academic prose. Therefore, Turkish learners' overuse of *can* to express uncertainty and possibility may be an indication that they are not using other modal devices expressing possibility preferred by native writers.

Second, Turkish learners seem to prefer using the idiomatic combinations such as *of course, kind of* and *a lot of* more intensively than NS. Foreign learners' overuse of *of course* in writing is highlighted by Narita, Sato, and Sugiura (2004) and Narita and Sugiura (2006) for Japanese learners, by Granger and Tyson (1996) for French learners and by Altenberg and Tapper (1998) for Swedish learners. And the finding that *of course* is overused by in Turkish learners written interlanguage verifies the claim that *of course* is one of the building blocks of learner writing suggesting a difference between NS and NNS writing. Additionally, the overuse of *of course* bears some implications about the speech-like design of writing by Turkish learners. Gilquin and Paquot (2008) states that *of course* as an amplifying adverb is more commonly used in conversation and less preferred in professional writing. Therefore, it could be claimed that Turkish learners' interlanguage writing contains some features of spoken language.

The overuse of the combination *kind of* provides further evidence for the above claim that Turkish learners' written interlanguage includes phrases commonly used in conversation by native speakers. The sequence *kind of* has been identified as a recurrent phrase of spoken English in a number of studies (De Cock, 2004; Qi & Ding, 2011; Simpson-Vlach & Ellis, 2010). Interestingly, Turkish learners overuse *kind of* in their written language while *kind of* is underrepresented in their spoken language (as it appeared only in 20 instances in LINDSEI-TR). This finding may point to an unawareness of register in Turkish learners of English.

A final comment that could be made depending on the overused sequences by Turkish learners is that their written interlanguage is characterized by high writer visibility, which points to a difference from professional writing. Although personal opinions are present in professional writing, they are conveyed through impersonal structures such as *it is reasonable to, it is worth noting* etc… (Gilquin & Paquot, 2008). However, as shown in Table 42, Turkish learners heavily rely on personal structures such as *I think, in my opinion, I want to* in expressing their personal opinions. Learners intensive use of *I think* in writing is also reported by Granger (1998) for French learners, by Aijmer (2002b) for Swedish and by Neff, Ballesteros, Dafouz, Martínez and Rica (2007) for Spanish learners, which refers to a common characteristics of learner writing. 39, 40, 41, 42 and 43 exemplify the use of personal devices in opinion giving in TICLE.

(39) *You can think the best way is abortion. But we don't know the baby wants to be died. Maybe he wants to live, see daylight, be a mother, a father , a teacher. . . . . So* ***in my opinion*** *abortion should not be made except personal reasons. Because* ***in my opinion,*** *it has two reasons. First is, we kill a baby. Second is, we are preventing the babies' justices. So we don't consider important their feelings.*

(40) *Whose mistake is this? or Do we have to look for a responsible for divorce?* ***In my opinion*** *there is no need to look for responsibles, when we go back in history, we will see that before women start working the avarage of divorce was so much lower than today. Today when It is compared with past there are more divorces,* ***In my opinion*** *the reason is that after women start to work and earn their own money they felt themselves secure.*

(41) ***I think*** *one of the most crucial inventions which marked its name on this century is "computer". When* ***I think*** *of its mechanical structure I am getting lost easily, because it has got very complex and confusing structure which is not easy to understand at least for me.*

(42) ***I think*** *sex equality is a very general and common subject and it interests everybody. Maybe you won't be agree with me but First of all* ***I want to say*** *that I don't believe sex equality. Of course, there should be equality between them but in real time there isn't because their duties are different*

(43) *Firstly,* ***I want to talk*** *about men's being stronger physically than women. There is no problem about strength of men and also it may be useful for women as when they can't success doing something*

As clear in the sentences above, Turkish learners tend to make their presence strongly felt in the writing. Their use of explicit expressions in (39), (40) and (41) are in many cases unnecessary, and professional writers simply omit them in their writing

(Gilquin & Paquot, 2008). Their use of combinations with *I want to* as in (42) and (43) also makes them too visible in introducing new topics and ideas, which is also different from the tendency of professional writers who generally prefer more impersonal devices such as *this article examines* (Gilquin& Paquot, 2008).What is more, the expressions *I think, in my opinion, I want to talk about* are frequently used in speech rather than academic writing as discussed in section 4.1.1.1. Therefore, this finding lends support to the claim that Turkish learners' written English bears similarities to the spoken English. A further analysis of learner writing based on implied register interference is provided in section 4.5.

In addition to overused word sequences, an examination of underused phraseology is necessary to have a complete understanding of Turkish learners written interlanguage in comparison with NS writing. Table 43 displays the underused phraseology by Turkish learners.

Table 43

Underused Word-Combinations (freq. >12/6/4/3 Occurring in TICLE in Comparison with LOCNESS, Raw Frequencies, and Relevant Log Likelihood Statistics)

Word Combinations	**LOCNESS**	**TICLE**	
Underused Phrases	**Freq.**	**Freq.**	**Log Likelihood**
would be	165	52	-72.00
united states	111	-	-163.66
would have	67	-	-98.79
American literature	64	-	-94.36
the fact that	75	29	-25.23
the amount of	23	-	-33.91
out of the	30	-	-44.23
in favour of	23	-	-33.91
the death penalty	71	-	-104.69
in public schools	36	-	-53.08
This type of	26	-	-38.34
prayer in public schools	36		-53.08
the invention of the	23		-33.91
the drinking age would	15		-22.12
the fact that the	14		-20.64
due to the fact	12	-	-17.69
in the case of	15	-	-22.12
lowering the drinking age	14	-	-20.64
the teaching of new age	13	-	-19.17
prayer in public schools	12	-	-17.69

Log Likelihood test results have shown that the underused sequences in Table 43 are statistically significant. The comparison of TICLE and LOCNESS considering the underused phrases led to interesting findings on the content level of the texts as well as on the linguistic level of the preferred sequences. The expressions such as *death penalty, prayer in public schools, lowering the drinking age, the invention of the* point to the recurring themes in LOCNESS thus clarifying the "aboutness" of the essays. On the linguistic level, the word combinations *would be, would have, the fact that, the amount of, in favour of, this type of, due to the fact, and in the case of* attract attention.

4.4. Qualitative Findings: Written Corpora

Besides carrying out quantitative analysis which has yielded important insights into written interlanguage of Turkish learners, there is also a need to examine the same data from a qualitative point of view. Such a need is highlighted in the literature as well since only a few studies of L2 written data have performed structural and functional categorization of word combinations (Y.-H. Chen & Baker, 2010). Combining quantitative analysis with qualitative analysis facilitates seeing the different aspects of same data, and it also enables to compare the results with the findings in the literature in a fruitful way as it is easier to talk about general categories rather than individual word combinations. Considering these issues, following research question which requires qualitative analysis was formulated: “what are the structural and functional features of recurrent sequences of two or more word combinations prevalent in written interlanguage of Turkish EFL learners?”

To answer this question, TICLE and LOCNESS have been re-analysed in terms of recurrent 2-, 3-, 4-, and 5-word sequences. Data from LOCNESS are included in the analysis due to the fact analysis of the learner language through a comparable native corpus would allow to have sound claims. 6-word sequences are excluded in the analysis as they are mostly contextual (see section 4.3.1.4.) limiting the possibility of reaching general conclusions. To identify the recurrent phrases for structural and functional classification, frequency thresholds adopted in quantitative analysis has been changed. Following the literature (Biber et al., 2004; Cortes, 2002b; Hyland, 2008a)and considering the size of corpora under investigation, frequency thresholds were set as 75, 25, 10 and 10 for 2-, 3-, 4- and 5-word combinations respectively. This enabled to turn the data into a manageable size. The search retrieved 171 2-gram types, 116 3-gram types, 99 4-gram types and 20 5-gram types from TICLE; 89 2-gram types, 47 3-gram types, 38 4-gram types and 6 5-gram types from LOCNESS.

4.4.1. Structures of Recurrent Phrases in TICLE in Comparison with LOCNESS

In the structural classification of the word sequences identified in both corpora, a taxonomy offered by Biber et al., (2007) has been used. Accordingly, word sequences are structurally grouped into two broad categories: Phrasal and Clausal. Phrasal expressions are further categorized as noun phrase-based, prepositional phrase-based and verb phrase-based. Clausal bundles, on the other hand, are formed for example by a that-clause fragment and a verb followed by a to-clause fragment. The taxonomy also include a third group under the category of "other expressions" which contain word sequences "that do not fit neatly into any of the other categories" (Biber et al., 2007p.1024). Figure 12 is the graphic representation of the overall distribution of main structural types of word combinations in Turkish learners' written interlanguage and NS written English.

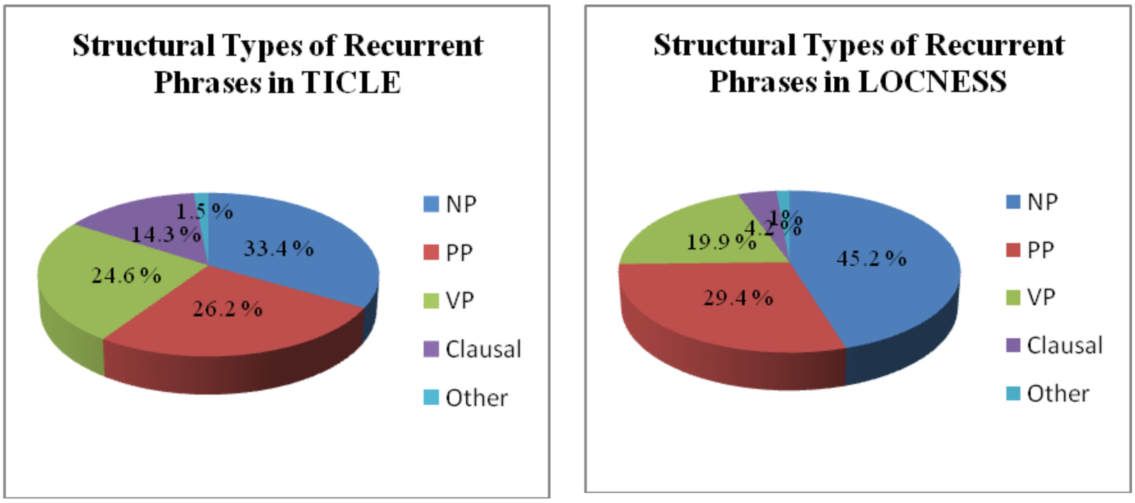


Figure 12. Distribution of major structural types in TICLE and LOCNESS

As shown in Figure 12, the largest part of the recurrent expressions in both NS and TL writing is comprised of noun phrases (NP) although noun phrases have a higher proportion in NS data. The second biggest group of word combinations in both corpora is the prepositional phrase (PP) which is followed by the verb phrase fragments with the rate of 24.6 % in Turkish written interlanguage and 19.9% in native speaker written English. The clausal expressions are observed to be relatively rare in both corpora.

This distributional pattern in terms of structures of recurrent phrases is in line with the previous findings. Based on spoken and written corpus of English (LSWE), Biber et al., (2007) note that most recurrent expressions "in academic prose are building blocks for extended noun phrases or prepositional phrases" (p.992).In their study comparing written English of Chinese learners with native expert writing (FLOB-J) and native student writing (BAWE-EN), Chen and Baker (2010) found out that noun phrases and prepositional phrases are dominant in the expert writing (FLOB-J), which contains published texts retrieved from professional journals and book sections, whereas native students and Chinese learners use noun phrases less than professional writing. A similar finding is echoed in Cortes (2002) who found that recurrent combinations in native student papers imitate the academic register in terms of structural distribution as they were mostly phrasal rather than clausal although closer examination of her data produced interesting results. In the same vein, Liu (2012),based on the investigation of the two native language corpora (COCA including professional journal articles and BNC including book chapters and journal articles) found that "noun and prepositional constructions (e.g., a/the number of and in terms of) constitute the two largest types of [multi word combinations]s in academic written English" (p.31).

The findings reported for LSWE, FLOB-J, COCA and BNC and the findings based on the analysis of LOCNESS in the present study imply that frequent use of NP and PP-based expressions are the characteristics of academic writing, and the categorical distribution in Turkish learners' data suggests that their written language is similar to academic writing to some extent. However, a detailed analysis classifying these combinations into further groups is likely to provide better understanding in terms of similarities and differences between NS and TL writing. Table 44 displays the major recurrent phrases further categorized according to their structures with examples from both TL and NS corpus. (see Appendix L for the expanded list).

Table 44

Structural categories of recurrent phrases in LINDSEI-TR and LOCNEC with example combinations

Category	Example from TICLE	Percentage TICLE	Example from LOCNESS	Percentage LOCNESS
PHRASAL				
d) Noun Phrase Based				
➢noun phrase with of-phrase fragment	*one of, kind of, first of all, point of view, the importance of, this kind of, the number of, freedom of the press*	*14.3%*	*one of, the use of, the idea of, the number of, this type of, the amount of, the case of, the majority of the,*	*27.3%*
➢other noun phrase or noun phrase fragment	*the world, the people, the same, most university degrees, sex equality, women and men*	*18.9%*	*the world, the fact, many people, the wild card, prayer in public, the drinking age*	*17.8%*
e) Prepositional Phrase Based				
➢prepositional phrase with embedded of-phrase fragment	*of the, by the help of the, as a result of, as a result of this*	*2.6%*	*of the, in favour of, as a result of, in the case of*	*2.7%*
➢other prepositional phrase fragment	*according to, in fact, in addition to, for the, in order to, around the world, on the other hand, for the real world*	*23.6%*	*as well, in order to, in the past, as long as, according to the, due to the fact that, in the long run*	*26.7%*
f) Verb Phrase Based				
➢anticipatory it/Noun/Pronoun + verb phrase/adjective phrase	*it is a, it is the, I think, they have, they do not have, they want*	*5.6%*	*it was, it is a, it is the, they have, they want to, crime does not pay*	*6.1%*
➢passive verb + prepositional phrase fragment	*are thought to be*	*0.5%*	---	---

(Table 44 Contuniued)

➢copula be + noun phrase/adjective phrase	*is one of the, are a lot of, is one of the most*	*1.5%*	*is a very, is a good, is important to, is one of the*	*1.6%*
➢Pronoun/noun phrase + be (+…)	*it is, this is, women are, people are, we are, money is, most university degrees are*	*6.6%*	*we are, they are*	*1.3%*
➢Existential "there"	*there is some, there is a, there will be, there is no need, there are many people*	*4.6%*	*there is, there are, there is a, there is no, there are many*	*3.4%*
➢(Pronoun/Noun) + Modal expressions	*they should, they will, must be, may be,* *it can be, we can see,*	*5.6%*	*would be, could be, will be, be able to, it would be,*	*7.5%*
CLAUSAL				
➢(verb phrase +) that-clause fragment	*think that, say that, believe that, think that it is, it is true that, is a fact that*	*6.6%*	*is that, I feel that, is the fact that*	*2%*
➢(verb/adjective +) to-clause fragment	*try to have, want to*	*1%*	*is important to*	*0.6%*
➢adverbial clause fragment	*if they, if there is no, when we look at, when they*	*6.6%*	*if the, when the*	*1.4*

A closer look at Table 44 reveals some noteworthy similarities as well as differences between NSs and TLs writing in terms of structures of their recurrent phrases. To begin with, even though both Turkish learners and native students perform a wide range of NP based combinations, they differ from each other in the construction of noun phrases as substantial use of noun+of phrase by native students stands out. While learners use noun phrases incorporating *of* at the rate of 14.3%, native students use them at the rate of 27.3%. The preposition *of* is a very productive unit in forming noun phrases in English, and noun phrases incorporating *of* are a feature of writing in English (Biber et al., 2007). The lesser use of NP+of fragments by Turkish learners could possibly be explained by the fact that learners' speech is less information loaded, which indicates a deviation from native-like writing. This difference might also be an indication of speech-like feature of learner writing, which is implied by the overuse of certain phrases in learner writing as touched upon in previous section. This issue will be further discussed in section 4.5. The sub-category of other noun phrases does not show much difference between NS and TL writing. Secondly, the category of prepositional phrases shows a similar distribution for both groups; however, they differ in the variety of PP based sequences. While Turkish learners mostly used PPs incorporating *in, on,* and *for*, native students made use of combinations incorporating a wide range of prepositions (consider the PPs *as well, as long as, about the, in the long run, due to the fact*). This suggests that Turkish learners use a limited set of PPs in general. High density of prepositional phrases in academic writing enables a writer to convey a great deal of information compactly. Therefore, the instructional activities should focus on the varieties of prepositional phrases, and students should be taught to recognize prepositional phrases and employ them appropriately in their own writing (Hinkel, 2004).

Thirdly, for the VP category, it has been observed that both learners and native students prefer to active sentences to passive ones, which is in line with Granger (1998) who note that learners generally prefer active discourse frames and Biber et al., (2007) who report only a few recurrent phrases incorporating *passive verb+PP* in academic prose. Next, the sub-category of modal verbs shows that although both groups of students make use of modal verbs in their writing, the native students' choice of modal verbs (mostly *would, could* and *should*) is more consistent with the academic writing (Biber et al., 2007). Turkish learners’ choices (*can, may, must*), on the other hand, are mostly used in conversation. Finally, for the category of clausal expressions, it is seen that learners use clause fragments and adverbial fragments more than native students.

In summary, structural analysis has tentatively shown that Turkish learners’ written interlanguage might possess the following features: a) speech-like features, b) preference for active verb phrases to passive verb phrases. Evaluating the data from a functional perspective would provide further evidence to explore the written English of NS and TL.

4.4.2. Functions of Recurrent Phrases in TICLE in Comparison with LOCNESS

In the analysis of the functions of the recurrent phrases in TICLE and LOCNESS, the same taxonomy used in the functional analysis of the spoken interlanguage has been used, but the category of “conversational expressions” was excluded for this part. The remaining categories are stance expressions, referential expressions and discourse markers. To have a better insight about the functions of specific combinations, these broad categories are further divided into sub-categories by Biber et al., (2007). Detailed description of the sub-categories is presented in chapter 3. However, depending on the analysis of functional properties of recurrent phrases in TICLE and LOCNESS, some sub-categories in the taxonomy have been modified in this part. For example, the sub-category of "tangible framing attributes" that refer to the abstract characteristics of the following noun is deleted as no phrase at the frequency threshold has been found to perform such a function. Similarly, the categories of "simple inquiry" and "responses" in the original taxonomy have been omitted as they are the functions generally performed in conversation. Also, the category of speech/thought reporting has been deleted. Although there were some expressions in TICLE performing the function of thought reporting such as *as I said, as I stated before*, their frequency was below the adopted threshold; therefore, they were not included in the analysis. Figure 13 presents the distribution of the major categories across learners and native speakers' writing.

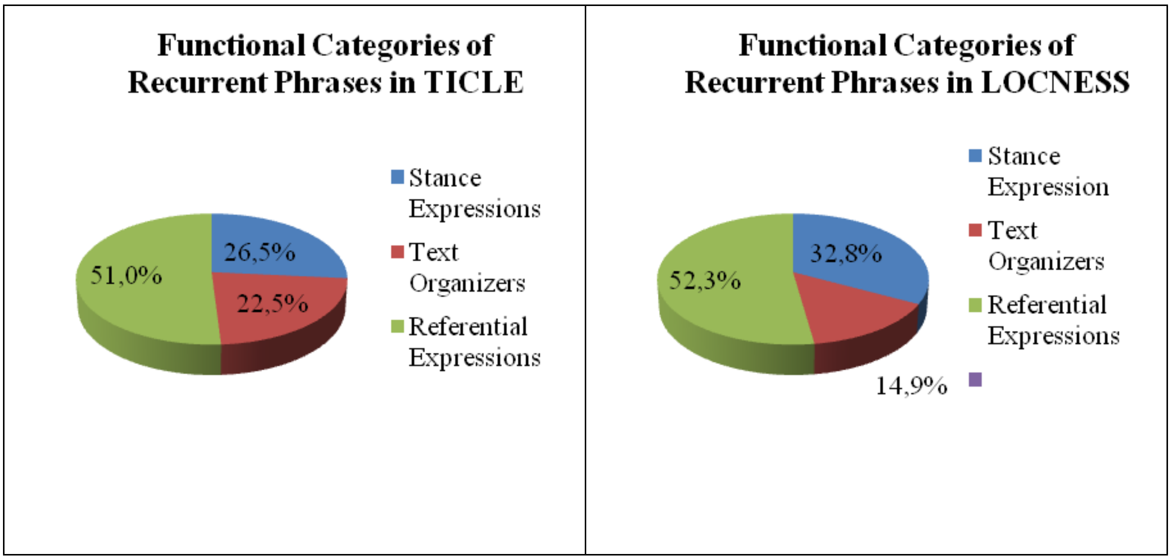


Figure 13. Distribution of major functional categories in TICLE and LOCNESS

As is obvious from Figure 9, the proportion of referential expressions accounts for the largest part of all the expressions in both native speakers and Turkish learners' writing. Stance expressions have a similar distribution in both corpora as well; however, NS and TL differ from each other in terms of dominant use of text organizers. While Turkish learners use text organizers at the rate of 26 %, native speakers employ such expressions at the rate of 14%. Prevalence of this functional category in Turkish learners' essays in comparison with NS writing can be accounted for the instructional factors as organization of text and the proper use of language items related to organization (e.g. transitions, linking words etc...) seem to be emphasized in a number of instructional materials (Gower, 2009; Savage, Mayer, Shafiei, Liss, & Davis, 2010; Zemach & Stafford-Yilmaz, 2009).

Table 45 presents the functions of common recurrent phrases in TICLE and LOCNESS in further subcategories, which could provide a detailed analysis

Table 45

Functional categories of recurrent phrases in TICLE and LOCNESS

Categories	**Sub-Categories**	**Examples from TICLE**	**Examples from LOCNESS**
Stance Expressions	**A) Epistemic Stance**		
	Personal	*I think, I think it is, in my opinion, I believe that, I think that, think that it is,*	
	Impersonal	*the fact that, in fact it is, of the fact that, is a fact that, it is true that*	*fact that, the fact that, the fact that the, due to the fact that,*
	B) Attitudinal/Modality Stance		
	B1: Desire		
	Personal	*do not want, I want to*	*want to, do not want to, would like to*
	B2: Obligation/Directive		
	Personal	*we should most important, the most important, should be, must be, should not be, it should be, the importance of, one of the most important, is the most important, the most*	

(Table 45 Contuniued)

	Impersonal	*important thing, the important thing is, there is no need to,*	*should not be, it would be, of the most important, one of the most. would have to*
	B3: Intention/Prediction		
	Personal		
	Impersonal	*it will be*	*will continue to, will not be, would be, would not be, would not have, would have been,*
	B4: Ability/Possibility		
	Personal	*we can, we cannot, can say that, we can see, we can say that, I can say that*	
	Impersonal	*can be, be able to, not be able to, to be able to, may be,*	*can be used, be able to, not be able to, could be*
Discourse/Text Organizers	**A) Topic Introduction/Focus**	*look at the, when we look at, that there are, when we look, that there is, that they are, that it is a, that there is no for instance, in fact, in addition, instead of, of course, such as, for example, because of, because of the, in addition to, to sum up, because they are, for example in, on the other hand, by the help of, in addition to this, for the sake of, in addition to these, in order to, with the help of, of course there are, are thought to be, on the other hand*	*that they are, that it is because of, such as, because they, for example, because of the, as well as, due to the fact, in order to*

(Table 45 Contuniued)

	B) Topic Elaboration /Clarification	*the, on the other hand some*	
Referential Expressions	**A) Identification/Focus**	*the students, the people, this is, one of, of this, in this, a person, the patient, the society, of these, the family, human beings, some people, the children, this is not, some of them, most of them, one of them, this is a, is the most, is one of, of the world, of the people, women and men, people who are, the people who, one of them is, of the most important, most of the people, both men and women, is one of the most*	*the government, in this, the people, of this, of these, this is, of the world, this is a, one of the, one of the most, is one of them, of the United States*
	B)Imprecision / Markers of Vagueness	*kind of, this kind of,*	*this type of*
	C) Specification of Attributes		
	C1: Quantity specification / Quantifying Sequences	*one of, all the, a lot, lot of, lots of, one of the, there are many, is one of, one of them, all of the, the number of, some of them, there are many people, have a lot of, there are a lot*	*all of, there are many, the number of, a lot of, all of the, a great deal of, the majority of the,*
	C2: Tangible Framing Attributes	---	---
	C3: Intangible Framing Attributes	*as a result of, a result of, according to, as a result of this, according to the*	*as a result of, in the case of, the only way to*

(Table 45 Contuniued)

	D) Time/Place/Text Reference		
	D1: Markers of Time	*day by day, in the past, in the future, at the end, the end of the, in a short time, at the end of, for a long time, at the same time,*	*in the past, the # century, in the #, the end of the, at the same time, in the # century, of the # century, at the end of, in the long run*
	D2: Markers of Place	*in the world, in our country, in the society, over the world, around the world, all over the world, in our society, in front of, in every part of the, all around the world*	*in public schools, in the United States*
	D3: Text deixis		
	D4: Multifunctional References	*The beginning of the, at the end,*	*The beginning of the, at the end,*

The comparative analysis of TICLE and LOCNESS from a functional perspective discloses some significant aspects of Turkish learners' written interlanguage. The first thing observed from figure 13 and table 45 is that both NS and NNS students have made use of stance expressions in their writing. However, they differ in the choice of specific combinations expressing the stance. As table 45 shows, while NS use only impersonal phrases in epistemic stance, learners have intensively made use of personal epistemic stance markers beside impersonal markers. They have used the expressions such as *I think* (♦♦♦♦),*I think it is* (◊), *I believe that* (♦♦), *I think that* (♦) and *in my opinion* (♦♦), which are also shown in the overused phrases by learners (see the section 4.3.2). Although these combinations appear in the native data as well, their frequency remains under the threshold adopted to have claims. (e.g. I think ♦, I believe ◊ in LOCNESS). The finding that learners use more personal epistemic stance expressions in their writing than NS is in agreement with the previous research (Aertselaer, 2008; Aijmer, 2002a; Neff et al., 2007). The overuse of personal epistemic stance markers suggest that writer visibility is too emphasized in learner writing, and over-emphasized writer visibility is a characteristics of non-native writing (Petch-Tyson, 1998). One possible explanation for the overuse of personal stance markers by the learners could be the register interference. The expressions *I think* and *I think that* are found to be peculiar to spoken language (Aijmer, 2002a; Huang, 2011; O'Keeffe et al., 2007). The combinations *I believe* and *in my opinion* serving the same function are more common in conversation than writing. Based on their study comparing academic prose with spoken English, Biber and Conrad (2004) state that "the most striking aspect of conversation's use of lexical bundles is the high proportion of personal stance expressions" (p 67).Therefore, it is possible to claim that Turkish learners design their writing in a more speech-like style, which is also typical of novice writers. Their use of impersonal stance together with the personal stance markers lends further support to the impression of novice writer. Consider the sentences (44) and (45) from TICLE.

(44) ***I think*** *that when two babies come to earth, one is boy and the other is girl, the girl start the game from behind with a score 0-1. Contrary* ***to this fact I believe*** *that even though a woman is weak physically and more emotional, which makes her more susceptible to what he faces in life*;

(45) *The point that* ***I want to*** *reach is the* ***fact that*** *the schematas in the universities are mostly not according to the needs in the real world*

In both examples, the student writers used both expressions belonging to spoken language and impersonal stance commonly used in written language. This juxtaposition of two strategies creates a novice writer effect regarding the tone of writing (Aijmer, 2002a).

For the second sub-category of stance expressions, Turkish learners are found to use various devices to express attitudinal/modality stance. As a cross-linguistic grammatical category, modality is related to the status of the proposition describing the event (McDouall, 2012). In English, there are a number of ways including adverbs (e.g. possibly, probably), lexical verbs (e.g. think, feel) and modal auxiliaries (Aijmer, 2002a). Table 45 shows that Turkish learners overuse obligation and directive expressions. The NSs use similar expressions with a similar function, yet the proportion is less than the Turkish learners' usage. A similar distribution of directive expressions for NS and NNS students has also been reported by Dutra (2004) who found that Brazilian learners use much more directives than NS students. In expressing the obligation, beside sequences like *it is important*, *one of the most*, Turkish learners use modal verbs *must* and *should* as well. *Should* and *must* in writing generally function to influence the readers' beliefs and attitudes, to imply that certain events are necessary or desirable by providing argument. This function of *must* and *should* is illustrated in 46, 47 and 48 from TICLE.

(46)*Yes, the fee for this operation is costly, but I believe that there* ***should be*** *an insurance policy that covers this decision.*

(47)*I think females* ***must have*** *same rights about cheating each other like on every subject.*

(48)*So, to decrease the disadvantages to minimum we* ***have to make*** *some changes or apply some regulations. Because internet if used effectively is the greatest thing that human beings have invented. We* ***must know*** *very well that life would not be the same again without internet.*

By overusing these obligation/necessity modal verbs, Turkish learners appear have more direct and emphatic style of persuasion in their writing.

For the last two sub-categories of stance expressions (intention/prediction and ability/possibility), Table 45 demonstrates that Turkish learners mostly prefer modal verbs. They generally use *will* for the prediction and *can, may* and *be able to* for the ability/possibility although there are many other modal devices serving these functions (e.g. *probably, possibly, certainly, indeed, surely*) (Aijmer, 2002a). Their overuse of modal verbs for these functions reflects the fact they are not using other modal devices expressing possibility and prediction. When the general distribution of these modal verbs are taken into consideration, it is seen that *can* is used in 1429 instances, *should* appears in 866 instances, *will* in 723, *may* in 304, *must* in 252 and *have to* appears in 163 instances, which indicates that modal auxiliaries as a whole is highly significantly overused by Turkish learners. Overuse of modal verbs is a characteristics of learner writing (Aijmer, 2002a). The search of TICLE for the other modal devices retrieved *probably* in only 30 instances, *certainly* in 25 and *indeed* in only 11 instances. The stance adverbs *possibly* and *surely* do not appear at all. One reason for this heavy reliance on only modal auxiliaries and underrepresentation of other devices for expressing stance could be the effect of instruction. As Aijmer (2002b)states, "many text books devote an unjustifiably large amount of attention to modal verbs, neglecting alternative strategies"(p.67).

As for the text organizers, Figure 13 shows that both groups employ phrases performing this function, yet the proportional weight is different. While Turkish learners use text organizers at the rate of 22,5%, native speakers use them at the rate of 14,8%. A closer look at the specific combinations indicate that the expressions *that they are, that there is a* are shared by both groups with the function of topic introduction as shown in the literature (Ädel & Erman, 2012; Bal, 2010). Turkish learners additionally use the sequence *that there is no* to introduce topic as seen in 49 and 50.

(49) *It is easily seen* ***that there is no*** *equality. If women earn money to be of help in budget, men should share*

(50) *This shows* ***that there is no*** *clear cut between the attributions of both sexes*

Although the sequence of *that there is no* does not appear in LOCNESS, it is found to be used by expert writers as well (Y.-H. Chen & Baker, 2010). Differently from NS, Turkish learners employed the sequence *when we look at* with the function of topic introduction/focus in their writing, which is discussed in section 4.3.1.4 in detail. In the category of topic clarification, learners have made use of a wide range of devices from *on the other hand* to *for instance*. Examples of some of these sequences performing the function of clarification can be found in 51, 52, 53 and 54.

(51) ***On the other hand,*** *as well as advantages of the internet there are disadvantages too. For example it makes people, especially, who spent their whole days and nights with, anti-social and from physical health of the body it is not beneficial.*

(52) *As human beings, we need love,* ***in fact****, all we need is love.* ***In addition*** *we want to be loved by all the people around us*

(53) *In some fields, generally the lecturers, by means of the system, try hard just for the graduation of the students.* ***For instance****, in the faculty of engineering in most universities of Turkey, the students learn something, study hard for the exams by memorization.*

(54) *The other perspective is financial truths. Again, think the same patient, he is alive* ***with the helps of machines*** *for years but one time the cost of his situation could not be paid* ***because of*** *the financial manner of his family*

The use of a wide range of text organizers by learners could be an indication of cohesion in discourse organization. What is more, some of these organizers such as *on the other hand, by the help of, in addition to this* are substantially overused by the Turkish learners compared to the native speakers (see section 4.3.2). This may demonstrate the effect of classroom teaching and testing as these sequences are listed and taught vigorously at schools as a short-cut to writing a “well- organized” essay (Ping, 2009).

Finally, for the category of referential expressions, it has been observed that Turkish learners mainly use noun phrases in the functional sub-category of identification/focus. Most of these noun phrases are structurally grouped under the title of “other NP” than those embedding the preposition *of*. As discussed in section 4.4.1. the use of word sequences with embedded *of* is an indication of native-like writing (Ping, 2009), and Turkish learners’ less use of such noun phrases contributes to the impression of non-nativeness in their writing. Though relatively less, Turkish learners are found to use phrases incorporating *of* such as *one of them, is one of the most, of the most important* as referential expressions in their writing. Expressions beginning with “one of usually softens the tone of judgement and make the statement sound objective, relative and concrete” (Ping, 2009 p.38), which is a desired feature of academic writing. The fact that Turkish learners have made use of such sequences implies that they are aware of the significance of such expressions. Regarding the markers of imprecision/vagueness in writing, learners use of the expressions *kind of* and *this kind of* is worth mentioning. While *kind of* appears in 119 instances in learner data, it has been only 14 times by native students. In the relevant literature, *kind of* has been identified as a discourse item that introduces vagueness and fuzziness in discourse and primarily used in spoken language (De Cock, 2004; Simpson-Vlach & Ellis, 2010). Its use in writing is a characteristics of learner language (Dutra, 2004; Ping, 2009). Turkish learners’ use of this combination verifies the claim that *kind of* in writing is a sign of non-nativeness. However, a closer look at the learner data where kind of is used reveals interesting outcomes. Consider the following examples:

(55) *I said 'killing a baby' as I believe that abortion is* ***a kind of*** *murder and, as we know, murder is a crime of killing a person on purpose.*

(56) *but many of the people or the student who live in our country or the other country do not give any important for cheating and plagiarism on the other hand, the student, who gives his/her knowledge as a copy, he/she does not think that this situation is* ***a kind of*** *theft*

(57) *Therefore, I don't think that assisting suicide should be approved. Because, I see it as* ***a kind of*** *suicide but in this one help comes from another person, which makes it seem easier on surface*

As seen in the examples above, learners in most cases have used *kind of* in its literal meaning as the sequence *kind of* can be paraphrased as 'type of' (if X is a sort of/kind of Y it means that X can be a hyponym of Y) (Aijmer, 2002b). However, native students have simply used the combination *type of* for such a function. This difference between NS and NNS may point to the learners' register unawareness. With regard to the specification of attributes, table 45 shows that learners have intensively used quantifying sequences and time/place references in their writing, which is in line with the literature (Cortes, 2004).

As an interim summary, the comparison of NS and NNS writing has revealed some significant differences and similarities between native students and learner academic writing. The structural analysis has shown that structural distribution of recurrent phrases in NS and NNS student essays is surprisingly similar. They both contain NPs and PPs to a great extent and clausal expressions are less used in both corpora. However, they differ from each other in the construction of NPs (e.g. NPs incorporating *of* are less preferred by learners), and variety of prepositions used by learners is limited compared with native students. Functional analysis demonstrated that both NS and Turkish learners employ word combinations recurring in their writing with certain functions. Most of the phrases they use have been identified in the literature as the building blocks of academic prose (e.g. *the fact that, one of the most, the importance of, be able to, that it is, on the other hand, one of them, the number of, as a result of, at the same time* among many). However, some expressions in the learner data point to different tendencies probably typical of learner language. The use of personal stance markers, a different tendency in the selection of specific modal verbs, use of directive expressions and heavy reliance on time and place markers are among the areas of differences in NS and NNS discourse design. As discussed above, classroom instruction, L1 influence and register interference could be possible reasons causing difference. Among them, register interference seems to be especially conspicuous in learner writing. This suggests that there is a need for further analysis of written interlanguage of Turkish learners based on the expressions found in the foregoing analysis giving the impression of speech-like design of their writing. Following section presents the results of such an analysis.

4.5. Recurrent Phrases in TICLE: Further Remarks

The last research question of the present study is "to what extent do the recurrent word combinations employed by both learners and native speakers in the written corpora overlap with or differ from those in the spoken corpora in syntactic and functional terms?"

The study has found that a number of recurrent phrases characterizing the written English of Turkish learners reflect the features of spoken language. The first thing that needs further analysis is the high writer visibility in Turkish learners' written English. When the written language of NS and Turkish learner was compared, it has been observed that, unlike NS, Turkish learners have overused the expressions such as *I think, in my opinion, I want to talk* (see section 4.3.2) in their writing. A further analysis TLs' written language in functional terms has also shown that learners have made much use of personal epistemic markers (e.g. *I believe, I think that, I think it is*) (see section 4.4.2). Overuse of these expressions creates a personal tone and makes the writer's presence strongly felt in the design of writing. Overt writer visibility, however, is not among the characteristics of the academic writing (Gilquin & Paquot, 2008). Although giving personal opinions is possible in academic writing, the way doing this is different as professional writers generally chose to use impersonal structures (e.g. *it is significant to, it is notable that*) in presenting their attitudes towards the message (Gilquin& Paquot, 2008).

One reason for the overuse of these expressions could stem from the very nature of the tasks used to collect data in TICLE, which contained argumentative essays. Recski (2004) notes that in writing argumentative essays, "personal references and subjective attitudes are certainly hard to avoid" since learners are explicitly encouraged to give their personal opinions; though, they could do this by using impersonal structures as well. Another reason for the overt writer visibility could be the learners' confusion of register. The expressions inclusive *I think, I want to, in my opinion,* and *I* *believe* are reported to be the characteristic statements of spoken language (Gilquin & Paquot, 2008). To test this claim and to see if these expressions are really common in spoken language, a frequency search for these expressions in the spoken corpus of native students (LOCNEC) was carried out. Frequencies found were adjusted to per million words (pmw) to make the findings comparable. The results showed that the expressions *I think* (3000 times pmw), *I think it* (313 pmw), *I think it's* (438 pmw), *I think that* (157 pmw), *I want to* (215 pmw), *I would like to* (339 pmw), *I want* (282 pmw) are very frequently employed by the native speakers in their spoken language. However, the expressions *in my opinion* and *I believe* do not appear in the corpus and the native students seem to prefer the sequence *I suppose* (626 pmw) in expressing their opinion. The forgoing frequencies confirm that the word combinations including *I think, I want, I would like to,* and *I suppose* are building blocks of spoken language. The fact that Turkish learners overuse these expressions in their writing (I *think*=893 pmw, *I believe*=252 pmw, *I think it is*=55 pmw, *I think that*=159 pmw, *in my opinion*=329 pmw, *I believe that*=208 pmw, *I think it*=71 pmw and *I want to*=268 pmw) suggest that they have problems in conforming register peculiarities and they design their written discourse in a speech-like manner.

The overuse of the combinations *kind of* and *of course* by Turkish learners in their writing lends further support to the claim above. The n-gram searches of TICLE have shown that the frequency of *kind of* is 652 pmw while its frequency is 83 pmw in NS writing (LOCNESS), which suggests that *kind of* is not among the preferred expressions in written language of native speakers. In fact, the combination *kind of* is reported to be one of the building blocks of spoken language. De Cock (2004) notes that *kind of* together with *sort of* as vagueness tags are commonly used in native speech and they contribute to the informality of the interaction. When the spoken corpus of native students is examined, it is seen that both expressions are frequent in spoken language (*kind of*= 658 pmw and *sort of* 4309 pmw) though *sort of* is much more frequent than *kind of*. These findings verify that *kind of* and *sort of* are the expressions of spoken language, and overuse of *kind of* by Turkish learners in their writing is further evidence about their confusion of register. Finally, the overuse of *of course* is worth mentioning here as it emerged as an overused sequence in TL writing (see section 4.3.2) and it bears the implications about speech-like design of TL writing. According to Biber et al., (2007) *of course* is used primarily for two effects: a) it indicates the certainty of a proposition, b) it implies that the audience already knows-or will readily accept the information. Being overtly certain about the proposed idea is a feature of learner writing and overuse *of course* is reported to be a feature of spoken language (Gilquin& Paquot, 2008).While the frequency of *of course* is 690 pmw in TICLE, its frequency in LOCNESS is 155. When the spoken language of NS is considered, it has been found that the frequency of *of course* is 266 pmw. This confirms that NS use *of course* in speech more commonly than in writing. Therefore, the fact that Turkish learners intensively use *of course* in their writing suggest that their written language is similar to spoken English. All in all, the analysis of written language of Turkish learners compared with native language pointed to learners' register confusion. Although the analysis here is restricted to several word sequences that emerged from the corpus-driven search, it could still make a fruitful base for future research.

To sum up, the findings and discussions about written corpora presented above have formed the second step in realizing the objectives of the present study which set out to explore Turkish EFL learners' interlanguage characteristics and compare and contrast NNS and NS use of recurrent phrases across spoken and written corpora in terms of both quantitative and qualitative variations. The overall results regarding the written language point to some overlaps between NNS and NS while there are also important discrepancies as well. 1) n-gram searches of written interlanguage of Turkish learners have shown that most of the combinations in learner language are consistent with the previous research on interlanguage. Grammatical units are most prevalent 2-word combinations, and such word sequences as *for example, because of, the same* are among the characteristic lexical choice of Turkish learners' written interlanguage. Their 3- and 4-word combinations are mostly in line with the previous findings, and 5- and 6-word sequences are mostly contextual. 2) The combinations *by the help of* and *when we look at* are apparently peculiar to Turkish learners. 3) Comparison between NS and TL writing shows that while there are overlaps in the use of bigrams, the proportion of similarity declines in other lengths. 4) The analysis based on overused/underused items indicates that Turkish learners intensively use combinations including *can* suggesting a degree of uncertainty in arguing in English. 5) Overused sequences have additionally revealed that Turkish learners' written interlanguage is characterized by high writer visibility, which is a feature of learner writing. 6) Several sequences frequently used by learners in their writing are inherent in spoken language, which suggests that register interference is influential in Turkish learners' written discourse. 7) Structural comparison of NS and TL writing demonstrated that NP-based and PP-based expressions are frequently used by both Turkish learners and native speakers; however, they differ from each other in the construction and variety of phrases. Register interference and preference of active discourse frames are the emergent themes from the structural analysis. 8) Functional comparison of NS and TL writing revealed that beside the similarities in terms of functions of the frequently used expressions, there are some deviations from the NS writing. The use of personal stance markers, a different tendency in the selection of specific modal verbs, use of directive expressions and heavy reliance on time and place markers make up the major differences between NS and Turkish learners discourse design. In addition to register interference and effect of L1, teaching-induced factors are discussed in relation with the deviations. All the findings discussed so far bear significant pedagogical implications, which are presented in the following chapter.

4.6. Chapter Summary

This chapter explains the findings obtained through the quantitative and qualitative analysis of Turkish learners' spoken and written language by using the corpora LINDSEI-TR and TICLE in comparison with the native speech and writing using the corpora LOCNEC and LONCNESS. The findings are presented in exactly the same order as the research questions posited earlier. Accordingly, in the first section, the recurrent phrases in Turkish learners' spoken interlanguage are identified and discussed in relation with the literature as an answer to the first research question. Then, to answer the second research question, a comparison between learners and native speakers' speech is provided. The second section contains the qualitative findings attained through the structural and functional analysis of the learner and native data answering the third research question. Section 3attempts to answer the fourth and fifth research questions as it presents the recurrent phrases in written language of Turkish EFL learners and then compares it with the NS writing. Section four provides a detailed description of the structures and functions of the word combinations found in learners and NS writing with the intention of answering the sixth research question of the present study. Depending on emergent themes from the analysis, section 5 provides further remarks about the written language. The results for each question are addressed with reference to relevant literature.

CHAPTER V

CONCLUSION

5.0. Introduction

This concluding chapter aims to summarize the general conclusions drawn from the study. First, it presents an overview of the study and the findings are presented briefly. Next, the pedagogical implications of the findings are discussed. Finally, this chapter ends with the suggestions for further research.

5.1. Summary of the Present Study

Within the framework of CIA adopting a corpus-driven recurrent word combination method, the present study has focused on the 2-to 6- word combinations in both spoken and written language to investigate Turkish learners' tendencies in designing their discourse in English. The main objectives were to explore the use of recurrent phrases in Turkish EFL learners' interlanguage and to compare and contrast them with native speakers' use of recurrent phrases across written and spoken language in terms of both quantitative and qualitative variation. To this end, four corpora have been analysed: LINDSEI-TR and LOCNEC for spoken language, TICLE and LOCNESS for written language investigation. Six main steps have been followed in the analysis: 1) identification of most frequent recurrent phrases in Turkish learners' spoken interlanguage 2) comparison of the word combinations with those in native speech 3) structural and functional analysis of recurrent phrases in Turkish learners' spoken interlanguage with reference to those in native speakers' 4) identification of recurrent phrases in written interlanguage of Turkish learners 5) comparison of the word combinations with those in native writing 6) structural and functional analysis of recurrent phrases written interlanguage of Turkish learners' with reference to those in native speakers. Each of these stages is briefly discussed below.

5.2. General Conclusions: Spoken Language

The n-gram searches of LINDSEI-TR for 2-word combinations have demonstrated that Turkish learners' speech is dominated by fragmentary sequences, and filled pauses have a large proportion. Their less use of such phrases as *I mean, you know* and their preference for the filled pauses instead is interpreted as markers of non-nativeness of their speech. It also points disfluency and encoding problems in learners spoken discourse (Kjellmer, 2003). The list of top-twenty 3-word combinations reinforces the claim about the disfluent nature Turkish learners' speech, which could be regarded as an indicator of lower level proficiency (Kjellmer, 2003; Lauttamus et al., 2008; Tottie, 2010). The analysis of 4-word sequences show that the phrases including the verbs *want*, *like* and *know* (e.g. *I want to be, I want to talk, and I want, eh I like*) are the commonest combinations, which is in line with the literature (Biber et al., 2007). The expression *how can I say* which is used as a communication strategy to appeal for assistance and initiate repair is found to be an idiosyncratic feature of Turkish learners' speech. 5- and 6-word combinations are found to be mostly contextual making it difficult to have extended claims.

When the speech of Turkish EFL learners is compared with native speech in terms of most frequent word combinations, striking differences beside similarities have been observed. While native speakers frequently use word sequences such as *you know, sort of*, *I mean* which are reported to be the characteristics of informal spoken English (Aijmer, 2002b, 2004; Romero-Trillo, 2002; Shirato & Stapleton, 2007; Stenström, 2006), their use in learner speech is rather rare. The combinations inclusive *yeah* which is also a distinctive feature of spoken English are frequently used by NS whereas Turkish learners hardly ever employ such expressions in their speech. These findings could make a base for the claim that Turkish learners are unfamiliar with the spoken English. The comparison of LINDSEI-TR and LOCNEC has additionally pointed to a complex picture of overuse, underuse and misuse of recurrent phrases in learner speech. The overused list of recurrent phrases shows that Turkish learners tend to use active discourse frames (e.g. *I say, I believe*) in their speech, which is reported to be a feature of learner English (Granger, 1998c); they use idiomatic expressions more intensively than NS, which could be linked to the fact that learners consider these expressions as safe points to continue their speech (Hasselgren, 2002). Some of the overused word sequences are found to realize functions seemingly idiosyncratic for Turkish learners (e.g. *and eh*), and several of them point to inappropriate use in context by learners (e.g. *of course*). The list of underused phrases lends further support to the claim about Turkish learners’ unfamiliarity with the spoken English as the sequences like *that’s right, well I mean, things like that* which are distinctive features of spoken English (Aijmer, 2004; McCarthy & Carter, 2002; Shirato & Stapleton, 2007) are very rare in NNS speech.

The structural analysis of learners’ speech in comparison with native speech has demonstrated that Turkish learner and native speech are similar to each other in terms of major structural types. That’s, both in TLs and NS speech, Verb Phrase fragments (e.g. *I think, I'm trying to*) make up the biggest proportion which is followed by Noun/Prepositional Phrase fragments (e.g. *my friends, of them, in my life*). Clausal fragments (e.g. *I don't know why, want to talk*) have the least proportion in the speech of both groups. This structural distribution of recurrent phrases in speech is consistent with the literature (Biber et al., 2007; Hernández, 2013), which reinforces the previous claim that conversation is fundamentally phrasal rather than clausal. Despite the similarities in major categories of structural taxonomy, analysis of recurrent phrases in terms of specific sub-categories has revealed several differences between NSs and TLs speech. That NSs use third person pronoun + VP fragments more than TLs implies that, unlike NS, Turkish learners rely too much on their personal experiences, which is also reported for the learner group in Hernández's (2013) study. Another difference is observed in the use of clausal expressions: while TLs' speech is dominated by Verb + to Clause fragments, native speakers use such structures relatively less. Instead, NSs employ Wh-Clause constructions in clausal fragments. Biber et al., (2007) note that Wh-Clause fragments are the frequent sequences of conversational English. Suggesting a deviation from NS speech, this finding leads to a possible claim that Turkish learners have a small repertoire of automatically retrievable Wh-clause fragments, and they most probably process such fragments on the basis of grammatical rules, which is relatively difficult in real-time production. Educational background based on separate grammatical rules rather than colloquial sequences could make a possible explanation for this finding.

Functional analysis of LINDSEI-TR in comparison with LOCNEC shows that (as in the case of structural comparison) while TL and NS are similar in terms of general functional categories of their recurrent phrases, some deviations in TLs' speech from NS speech are observed when the data are analysed in detail considering the sub-categories of the functional taxonomy. Accordingly, stance expressions have the largest proportion in both corpora, which is similar to the findings reported by Biber and Barbieri (2007) and Biber et al., (2007) and Hernández (2013). Within this category, however, epistemic stance expressions are frequently used by NS while TLs employ stance expressions related to personal desire to a great extent. Turkish learners differ from NS in the subcategories of attitudinal stance as well. While NS use combinations including *going to* to express their intention/plans, TLs use *will* with the same function. Effect of L1 is offered as a possible explanation for the difference. With regard to recurrent phrases functioning as discourse organizers in both corpora, Turkish learners are observed to mostly use expressions similar to those found in NS corpus; however, they differ in the subcategory of topic elaboration/clarification. While NSs prefer to use *I mean* and its combinations to clarify the previously stated idea, Turkish learners employ *for example* to perform the same function. Additionally, in the category of discourse organizers, it has been found that both learners and NS use the expression *I don't know* with the function of turn yielding/topic closing. Contributing to the relevant literature, this finding enabled to add a new subcategory to the original taxonomy. As for the category of referential expressions, recurrent phrases with this function show a similar distribution in both corpora; however, analysis considering the specific word combinations points to differences between NS and TLs speech. Vagueness tags are one area of difference: while NS use a wide variety of vagueness tags in their speech, in TLs' speech only one expression (i.e. *I don't know*) has been found to perform this function. That Turkish learners do not employ vagueness tags in their speech supports the claim about Turkish learners' unfamiliarity with the spoken English, and perceived foreign-soundingness of their speech. Influence of L1 and instructional factors are discussed as possible reasons of deviations.

5.3. General Conclusions: Written Language

N-gram searches of TICLE for 2-word sequences have produced findings similar to previous research as most of the bigrams in the written language are grammatical units (Granger, 2008). Some of the most frequently used bigrams out of grammatical units are *for example, because of, the same, the other, of course* and *I think* which could be interpreted as characteristic trait of academic writing with regard to lexical choice (Ebeling, 2011). Most of the three word combinations used by Turkish learners except *in the presence* are consistent with the word combinations identified in the previous research (Biber et al., 2007; Ebeling, 2011). This suggests that Turkish learners' essays bear similarities to academic prose in terms of preferred word combinations. Similar consistency with the previous research findings has been observed in the scope of 4-word combinations as most of the sequences used by Turkish learners have previously reported as the building blocks of academic writing in the literature (Bal, 2010; Biber et al., 2007; Cortes, 2002b; Ebeling, 2011) with the exception of *in the presence of* and *by the help of. In the presence of* is reported to be the most frequent combination of academic writing (Biber et al., 2007) but it is underrepresented in Turkish student written interlanguage. The combination *by the help of* appears to be an idiosyncratic use by Turkish learners as it has not been identified in the literature. Intralingual overgeneralization could make a possible explanation for this finding. Longer combinations (e.g. 5-and 6-word sequences) are mostly contextual. In other words, they recur in the data as they are required by the topics that the learners chose to write on. Other combinations that are not necessarily contextual are in line with the literature (e.g. *one of the most important, as a result of this*) with the exception of *when we look at the* which is found to be peculiar to Turkish learners.

Comparison of learner writing with native writing shows that while there is an overlap to a great extent in the 2-word combinations, the proportion of similarity declines in other lengths. The reason is that bigrams in both corpora comprise of function words. Analysis based on overused/underused phrases has pointed to several important tendencies related to Turkish learners' use of phraseology in their written discourse: Turkish learners overuse expressions including *can* which reflects a degree of uncertainty in arguing in English. And, being uncertain is a feature of learner writing (Aijmer, 2002a). The overuse of combinations including *can* is also an indication that Turkish learners infrequently use other modal devices expressing possibility preferred by native speakers. Instructional factors are proposed as the reason for this finding. The overuse of idiomatic combinations is another area of difference between NS and TL writing. Particularly, the overuse of *of course* and *kind of* is significant as they bear implications about speech-like design of learner written discourse. The intensive use of *I think, in my opinion, I want to* etc… is discussed in connection with high writer visibility, which makes another point of deviation from NS writing. In the list of underused phrases, the sequences *would be, would have, due to the fact that, this type of, in the case of* are among the combinations leading to discrepancies between NS and TL writing.

The structural analysis has revealed that Noun Phrases are the commonest combinations in both corpora. This is followed by Prepositional Phrases. Verb Phrase fragments make the third largest group of word sequences in both corpora. Although learners use Verb Phrases more than native speakers, diversity in verb choices is limited in TL writing. Clausal expressions are found to relatively rare in both data sources. This distributional pattern in terms of structures of recurrent phrases is in agreement with the literature (Biber et al., 2007; Y.-H. Chen & Baker, 2010; Cortes, 2004; Liu, 2012). Despite the similarities in the general categories, further structural analysis of recurrent phrases has revealed significant insights. Accordingly, Turkish learners differ from NS in the construction of Noun Phrases. While NSs mostly make NPs by using the preposition *of*, NPs incorporating *of* are less employed by Turkish learners. For the category of Prepositional Phrases, it has been found that Turkish learners use a limited set of prepositions in PPs compared with NS. Instructional factors are proposed as a possible reason of the difference. For the Verb Phrase category, the use of modal verbs points to a discrepancy between NS and TL. While NS choices of modal verbs are mostly consistent with academic writing, Turkish learners' choices mostly reflect register confusion.

Functional analysis of recurrent phrases has indicated that referential expressions (e.g. *most of the people, this is*) have the largest place in both NS and TLs writing. This is followed by stance expressions (e.g. *I think, the fact that*), and text organizers (e.g. *that it is, in addition to*) have relatively smaller proportion in the writings of both groups. However, further classification has produced a different pattern in NS and TL writing. For the category of stance expressions, it has been found that while NS use only impersonal phrases in epistemic stance, Turkish learners have intensively used personal epistemic stance markers along with impersonal markers. The intensive use of personal epistemic stance expressions is a characteristic of learner writing (Aertselaer, 2008; Aijmer, 2002b; Neff et al., 2007) suggesting over-emphasized writer visibility in learner writing (Petch-Tyson, 1998). Register interference could make a possible explanation for this finding. Functional analysis has also shown that the use of directive expressions, heavy reliance on time and place markers, a different tendency in the selection of specific modal devices are among the areas of deviations of learner writing from NS writing. As discussed earlier, classroom instruction, L1 influence and register confusion could be probable reasons causing difference. Finally, as register interference is continually observed as an emergent theme in the analysis of phraseology of Turkish learners, the combinations giving the impression of speech–like design in writing have been further investigated in the final part of the study which encouraged further research directions to be noted.

5.4. Pedagogical Implications

It is a fact that learner corpus research lies at the crossroads of between corpus linguistics, linguistic theory, second language acquisition and foreign language teaching (Granger, 2009). Therefore, a study based on learner corpus such as the present one is likely to yield information that is useful to learners and teachers of English as a foreign language. Although phraseological approaches should not be "the be-all and end-all of language teaching" (Granger & Meunier, 2008b p.251), it is believed that "awareness of phraseology in the wide sense should be promoted" among learners (ibid. p. 251). In the same vein, Granger (2009) states that findings attained from learner corpus research should be presented "with a view to providing a better description of one specific interlanguage and/or designing tailor-made pedagogical tools which will benefit similar-type learners" (p.20).

Following a corpus driven recurrent word combinations method, this study investigated the phraseology of Turkish learners' interlanguage in comparison with native language, and the findings have led to interpretations that could be linked pedagogical implications. One of the major findings of this study is that native and learner language does not only consist of individual building blocks, rather, their language is made up of word combinations put together more frequently than expected by chance. From a pedagogical perspective, this finding means that instead of single words, word combinations should be highlighted in language teaching as "an over-emphasis in language teaching on single words out of context may leave second language learners ill-prepared both in terms of the processing of heavily-chunked input such as casual conversation, as well as in terms of productive fluency" (McCarthy & Carter, 2002 p.38). As suggested by Gilquin (2011), materials presenting learners typical examples and drawing their attention to the most frequent phrases would make a starting point for a phraseologically enriched teaching environment.

Apart from these general suggestions, the present study bears implications for the design of language-skills-related instruction; specifically for writing and speaking skills as the analysis of these skills has been the main focus of this research.

The findings of the study could inform the design of speaking instruction in several ways. To begin with, drawn from the findings regarding learners' unfamiliarity with the spoken English, this study suggests course designs exposing learners to the features of spoken English through authentic materials to increase the their awareness of linguistic properties of spoken English. Secondly, as underlined by Shirato and Stapleton (2007) word combinations specific to oral communication need to be incorporated into the syllabus of speaking course. Creating a pedagogically useful list of recurrent phrases of spoken English is recommended as a starting point, and the language teachers could arrange the instructional activities based on such a list. Thirdly, depending on the finding regarding learners' overuse of filled pauses, it is advised that language teaching activities in speaking courses should contain communicative strategies. Explicit instruction on the recurrent phrases used by native speaker in various communicative events (e.g. topic introduction, clarification, turn yielding etc…) would be helpful for learners in terms of foreign-soundingness and in speaking more fluently. Next, this study has found that the use of vagueness markers and hedging devices (e.g. *things like that, sort of, you know, sort of like* etc…) is among the differences between learners and native speakers. Vagueness is central to informal communication and of great use when interlocutors cannot find the right words, and hedging is a characteristic of casual speech softening the tone of conversation (McCarten, 2010). Therefore, in designing speaking courses, these devices should also be incorporated into the syllabus. It is recommended that language teaching activities should place more emphasis on word combinations that enhance strategic competence "even before the acquisition of any grammatical competences" (Shirato & Stapleton, 2007 p.408). Last but not the least, this study has important implications about the functions of recurrent phrases regarding the inappropriate use of some word combinations (e.g. *of course, for example*). Thus, it is necessary that while teaching specific combinations, not only their meaning but also their function in context should be underlined.

As for the teaching of writing skill, several implications have aroused from the findings of the present study. First of all, it is observed that learners have a tendency to use the combinations with modal *can* very frequently. This causes an uncertain tone in their writing, which is particularly discouraged in academic writing. Their overuse of the clusters including *can* implies that they do not frequently use the other modal devices (e.g. *it is necessary, is likely to* etc…) which are the common combinations of academic prose. Therefore, it is recommended that the writing instruction should not cover grammatical information but also the tone of the writing. Secondly, based on the finding that learners intensively use idiomatic expressions (e.g. *for example, in my opinion* etc…) in their writings, it is suggested that learners should be taught more productive frames such as "*X is a (adj.) example of Y' and 'a (adj.) example of Y is X"* to avoid monotonous language use (Paquot, 2008). When compared with native language, in some instances of learner English, it has been observed that L1 influence is at work causing deviations and foreign-soundingness. Hence, it is advised to point out contrasts in word combinations with learners' mother tongue while teaching the phrases to learners (Nesselhauf, 2003). Finally, the findings of the study have pointed that learners in their writing tend to use features that are more typical of speech than academic writing, which suggests they are, to some extent, unaware of register differences. Therefore, while designing language courses and teaching materials, learners attention should be drawn to the register differences and to the notions of formality/informality depending on the prevalence of the word combinations that are aimed to teach.

What is more important than all these is raising learners' consciousness regarding the word combinations. The starting point should be to make learners aware of the significance of word combinations. As stated by Nesselhauf (2005) "it is essential that learners recognize that there are combinations that are neither freely combinable nor largely opaque and fixed (such as idioms) but that are nevertheless arbitrary to some degree and therefore have to be learnt" (p. 252).

5.5. Limitations of the Study and Suggestions for Future Research

Although the results of the current study provides considerable findings about Turkish EFL learners' spoken and written English, there are two notable limitations that go beyond the most basic concerns about if the L2 learners should, in fact, have the native speaker norms as target (Hunston, 2002).

First, considering the vastness of the data, phrase lengths covered and different registers included, this research is preliminary providing an overview of interlanguage characteristics of Turkish learners in terms of recurrent phrases, which leads to some probabilities. Firmer conclusions on particular issues discussed here could ultimately be reached if the quantitative scope is narrowed, or if the certain categories such as epistemic tags or markers of vagueness are predominantly considered. Second, since the structural and functional analyses of recurrent phrases were qualitatively conducted manually, it is likely that there might be some possible inconsistencies.

Despite the limitations, this study has contributed to the existing knowledge of word combinations. There are, however, still room for further research to provide better understanding of the use of word sequences in learner language. First, a research project focusing on recurrent phrases across different L1 groups would result in a richer understanding of learner language. Comparison with the other sub-corpora of LINDSEI and TICLE would shed further light on the effects of transfer, and on general tendencies across learner populations. Moreover, functional patterns of recurrent word combinations that emerged from the present study may provide useful onset for further research. In addition, depending on the findings about the heavy use of filled pauses in learner speech, more detailed studies could be conducted on the appearance of filled and unfilled pauses in spoken language, in relation to the identification and functions of word sequences. Last but not the least, future research could focus on practical applications of the recurrent phrases in educational setting. Since integration of the recurrent phrases into language teaching activities and helping students to develop a larger repertoire of spoken and written language is likely to lead to greater confidence for language learners, which in turn should greatly facilitate both spoken and written language production and communication.

REFERENCES

Aarts, B. (1992). *Small Clauses in English: The Nonverbal Types*. New York: Mouton de Gruyter.

Aas, H. L. (2011). *Recurrent word-combinations in spoken learner English:A study of corpus data from Swedish and Norwegian advanced learners*. (Unpublished Doctoral Dissertation) University of Oslo, Sweden.

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes, 31*(2), 81–92.

Ädel, A., & Römer, U. (2012). Research on advanced student writing across disciplines and levels: Introducing the Michigan corpus of upper-level student papers. *International Journal of Corpus Linguistics, 17*(1), 3–34.

Adolphs, S., & Durow, V. (2004). Social-cultural integration and the development of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences acquisition, processing and use* (pp. 107–126). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–438.

Aertselaer, J. N. van. (2008). Contrasting English-Spanish interpersonal discourse phrases: A corpus study. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 85–100). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Aijmer, K. (2002a). 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). Amsterdam: John Benjamins Publishing Company.

Aijmer, K. (2002b). *English discourse particles: Evidence from a corpus*. Amsterdam: John Benjamins Publishing Company.

Aijmer, K. (2004). Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies*, *3*(1), 173–190.

Aijmer, K. (2009). "So er I just sort I dunno I think it's just because…": A corpus study of I don't know and dunno in learners' spoken English. In A. H. Jucker, D.

Schreier, & M. Hundt (Eds.), *Corpora: Pragmatics and discourse* (pp. 151–168). The Netherlands: Rodopi.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A.P.Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications* (pp. 101–122). Oxford: Oxford University Press.

Altenberg, B. (2002). Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 37–54). Amsterdam: John Benjamins Publishing Company.

Altenberg, B., & Granger, S. (2002a). The grammatical and lexical patterning of make in native and non-native student writing. *Applied Linguistics, 22*, 173–194.

Altenberg, B., & Granger, S. (2002b). Recent trends in cross-linguistic lexical studies. In B. Altenberg & S. Granger (Eds.), *Lexis in contrast* (pp. 1–48). Amsterdam: John Benjamins.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In Sylviane Granger (Ed.), *Learner English on computer* (pp. 81–94). U.S.A.: Addison Wesley Longman Limited.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

Axelsson, M. W., & Hahn, A. (2001). The use of the progressive in Swedish and German advanced learner English- a corpus-based study. *ICAME Journal*, *25*, 5–30.

Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of corpus linguistics*. Edinburgh: Edinburgh University Press Ltd.

Bal, B. (2010). *Analysis of four-word lexical bundles in published resesarch articles written by Turkish scholars*. Unpublished doctoral dissertation, Georgia State University, USA.

Barlow, M. (2005). Computer based analysis of learner language. In R. Ellis & G. Barkhuizen (Eds.), *Analysing learner language* (pp. 1–21). Oxford: Oxford University Press.

Barnbrook, G. (2007). Sinclair on collocation. *International Journal of Corpus Linguistics*, *12* (2), 183–199.

Bayrakci, D. (2004). *The written performance of the Turkish advanced university students of English with reference to the use of conjuncts from the quantitative* *and functional perspectives*. Unpublished master thesis, Mustafa Kemal University, Turkey.

Beckner, C., Ellis, N. C., Blythe, R., Holland, J., Bybee, J., Christiansen, M. H. & Schoenemann, T. (2009). Language Is a complex adaptive system : Position paper. *Language Learning*, *59*(1), 1–26.

Beebe, L., & Cummings, M. (1996). Natural speech act data versus written questionnaire data: How data collection method affects speech act performance. In S. M. Gass & J. Neu (Eds.), *Speech acts across cultures* (pp. 65–88). Berlin: Mouton de Gruyter.

Behrens, H. (2008). *Corpora in language acquisition research*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Béjoint, H. (2000). *Modern lexicograph: An introduction*. Oxford: Oxford University Press.

Biber, D. (2006). *University Language A corpus-based study of spoken and written registers*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Biber, D. (2010). What can a corpus tell us about registers and genres? In A. O'Keeffe & M. Mccarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 241–255). Abingdon, UK: Routledge.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286.

Biber, D., & Conrad, S. (2004). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, *20*, 56–71.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at . . .: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2007). *Longman grammar of spoken and written English*. England: Pearson Education Limited.

Biber, D., & Reppen, R. (1998). Comparing native and learner perspectives on English grammar: A study of complement clauses. In S. Granger (Ed.), *Learner English on computer* (pp. 145–158). U.S.A.: Addison Wesley Longman Limited.

Brouwer, C. E. (2003). Word searches in NNS-NS interaction: Opportunities for language learning. *Modern Language Journal*, *87*, 534–545.

Can, C. (2009). İkinci dil edinimi çalışmalarında bilgisayar destekli bir Türk öğrenici İngilizcesi derlemi: ICLE'nin bir altderlemi olarak TICLE. *Dil Dergisi, 144* (Nisan-Mayıs-Haziran), 16–34.

Can, C. (2012). Uluslararası Türk öğrenici İngilizcesi derleminde tutum belirteçleri. *Dilbilim Araştırmaları, 1,* 39–53.

Chen, H. (2010). Contrastive learner corpus analysis of epistemic modality and interlanguage pragmatic competence in L2 writing. *Arizona Working Papers in SLA & Teaching, 17,* 27–51.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30–49.

Cheng, W., Greaves, C., Sinclair, J., & Warren, M. (2008). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics, 30*(2), 236–252.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge MA: MIT Press.

Cobb, T. (2003). Analyzing late interlanguage with learner corpora : Québec replications of three European studies. *The Canadian Modern Language Review, 59*(3), 393–423.

Colson, J.-P. (2008). Cross-linguistic phraseological studies: An overview. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 191–206). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Corder, S. P. (1981). *Error analysis and interlanguage*. Oxford: Oxford University Press.

Cortes, V. (2002a). Lexical bundles in academic writing in history and biology. (Unpublished Doctoral Dissertation), Norhern Arizona University, USA.

Cortes, V. (2002b). Lexical bundles in freshman composition. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using Corpora to Explore Linguistic Variation* (pp. 131–145). Amsterdam: John Benjamins Publishing Company.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397–423.

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*. Oxford: Clarendon Press.

Coxhead, A. (2008). Phraseology and English for academic purposes: Challenges and opportunities. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 149–162). Amsterdam/Philadelphia: John Benjamins Publishing Company.

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (Eds.), *Corpus Linguistics and linguistic theory: Papers from the twentieth international conference on English language research on computerized corpora (ICAME 20)* (pp. 51–68). Amsterdam: Rodopi.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL), New Series* (2), 225–246.

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 68–80). U.S.A.: Addison Wesley Longman Limited.

Diani, G. (2004). The discourse functions of I don't know in English conversation. In K. Aijmer & A.B. Stenström (Eds.), *Discourse patterns in spoken and written corpora* (pp. 157–171). Amsterdam & Philadelphia: John Benjamins.

Diez-Bedmar, M. B., & Papp, S. (2008). The use of English article system by Chinese and Spanish learners. In G. Gilquin, M. B. Diez-Bedmar, & S. Papp (Eds.), *Linking up contrastive and learner corpus research* (pp. 147–176). Amsterdam: Rodopi.

Diniz, L. S. (2007). *Highly frequent function words in the light of the idiom principle: The case of "the."* (Unpublished Doctoral Dissertation), Georgia State University, USA.

Dutra, D. P. (2004). Bundles in learner corpora: What a type and token analysis can reveal? Retrieved on May, 2012 from www.nilc.icmc.usp.br/elc.../sessao1_2.pdf.

Ebeling, S. O. (2011). Recurrent word-combinations in English student essays. *Nordic Journal of English Studies, 10*(1), 49–76.

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33–68). Cambridge: Cambridge University Press.

Ellis, N. C. (2002). Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188.

Ellis, N. C. (2008). Phraseology: The periphery and the heart of language. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 1–14). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Ellis, N. C., Frey, E., & Jalkanen, I. (2009). The psycholinguistic reality of collocation and semantic prosody (1): Lexical access. In U. Römer & R. Schulze (Eds.), *Exploring the lexis-grammar interface* (pp. 89–116). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008a). The processing of formulas in native and second language speakers:Psycholinguistic and corpus determinants. *TESOL Quarterly*, *42*(3), 375–396.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008b). Formulaic language in native and second language speakers : Psycholinguistics , corpus Linguistics and TESOL. *TESOL Quarterly*, *42*(3), 375–396.

Ellis, R. (2003). *Second language acquisition* (9th ed.). Oxford: Oxford University Press.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford: Oxford University Press.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Erman, B. (2007). Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics*, *12*(1), 25–53.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*, 29–62.

Forchini, P., & Murphy, A. (2008). N-grams in comparable specialized corpora: Perspectives on phraseology, translation, and pedagogy. *International Journal of Corpus Linguistics*, *13*(3), 351–367.

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course* (3rd editio.). New York: Routledge.

Gilquin, G. (2011). Lexical infelicity in causative constructions. Comparing native and learner collostructions. In J. Leino & R. von Waldenfels (Eds.), *Analytical causatives* (pp. 1–23). München: Lincom Europa.

Gilquin, G. (2012). LINDSEI. Retrieved on March, 2013 from www.uclovain.be/en-cecl-lindsei.html

Gilquin, G., & De Cock, S. (2011). Errors and disfluencies in spoken corpora: Setting the scene. *International Journal of Corpus Linguistics*, *16*(2), 141–172.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, *6*(4), 319–335.

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1–26.

Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction*, *1*(1), 41–61.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Gower, R. (2009). *Real writing*. Cambridge: Cambridge University Press.

Granger, Sylviane. (1998a). The computer learner corpus: A versatile new source of data for SLA research. In Sylviane Granger (Ed.), *Learner English on computer* (pp. 1–19). U.S.A.: Addison Wesley Longman Limited.

Granger, Sylviane. (1998b). *Learner English on computer*. London & New York: Addison Wesley Longman Limited.

Granger, Sylviane. (1998c). Prefabricated patterns in advanced EFL writing: collocations and formulae. In A.P.Cowie (Ed.), *Phraseology: theory, analysis, and applications* (pp. 145–160). Oxford: Clarendon Press.

Granger, Sylviane. (2002). A bird’s eye view of learner corpus research. In Sylviane Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–36). Amsterdam & Philadelphia: John Benjamins.

Granger, Sylviane. (2004). Computer learner corpus research : Current status and future prospects. In U. Connor & T. A. Uptom (Eds.), *Applied Corpus Linguistics. A multidimensional perspective* (pp. 123–146). Amsterdam: Rodopi.

Granger, Sylviane. (2005). Pushing back the limits of phraseology. How far can we go. In C. Cosme, G. Céline, M. Fanny, & P. Magali (Eds.), *Proceedings of the phraseology conference* (pp. 165–168). Louvain-la-Neuve: Louvain-la-Neuve.

Granger, Sylviane. (2008). Learner corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An international handbook* (pp. 259–275). Berlin & New York: Walter de Gruyter.

Granger, Sylviane. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13–32). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Granger, Sylviane, Dagneaux, E., Meunier, F., & Paquot, M. (2009). *The international corpus of learner English.Version 2. Handbook and CD-ROM.* Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, Sylviane, & Meunier, F. (2008a). *Phraseology: An interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Granger, Sylviane, & Meunier, F. (2008b). Phraseology in language learning and teaching: Where to from here? In Sylviane Granger & F. Meunier (Eds.), *Phraseology in foreign language learning and teaching* (pp. 247–252). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Granger, Sylviane, & Paquot, M. (2008). Disentangling the phraseological web. In Sylviane Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27–50). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Granger, Sylviane, & Rayson, P. (1998). Automatic profiling of learner texts. In Sylviane Granger (Ed.), *Learner English on computer* (pp. 108–119). U.S.A.: Addison Wesley Longman Limited.

Granger, Sylviane, & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, *15*, 19–29.

Grant, L. E. (2010). A corpus comparison of the use of I don't know by British and New Zealand speakers. *Journal of Pragmatics*, *42*(8), 2282–2296.

Greenbaum, S. (1996). *The Oxford English Grammar*. Oxford: Oxford University Press.

Gries, S. T. (2006). Introduction. In S. T. Gries & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics* (pp. 1–18). Berlin: Mouton de Gruyter.

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. In Sylviane Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 3–26). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Gries, S. T. (2010a). Useful statistics for Corpus Linguistics. In A. Sánchez & M. Almela (Eds.), *A mosaic of corpus linguistics: selected approaches* (pp. 269–291). Frankfurt: Peter Lang.

Gries, S. T. (2010b). Corpus linguistics and theoretical linguistics A love-hate relationship ? Not necessarily …. (A. Sánchez & M. Almela, Eds.) *International Journal of Corpus Linguistics*, *15*(3), 327–343.

Grigaliūnienė, J., & Juknevičienė, R. (2011). Formulaic language, learner speech and the spoken corpus of learner English LINDSEI-LITH. *Kalbotyra*, *3*(3), 12–18.

Guan, B., & Zheng, S. (2005). A Corpus-based contrastive study of recurrent word combinations in English essays of Chinese college students and native speakers. *CELEA Journal*, *28*(1), 37–48.

Hasko, V. (2011). Qualitative corpus analysis. *The Encyclopaedia of Applied Linguistics*, *13*, 1–10.

Hasselgren, A. (2002). Learner corpora and language testing: Small words as markers of learner fluency. In Sylviane Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 149–174). Amsterdam & Philadelphia: John Benjamins.

Hernández, P. S. (2013). Lexical bundles in three oral corpora of university students. *Nordic Journal of English Studies*, *13* (2004), 187–209.

Hinkel, E. (2004). *Teaching academic ESL writing: Practical techniques in vocabulary* and grammar. Mahwah, NJ: Lawrance Erlbaum

Housen, A. (2002). A corpus-based study of the L2-acquisition of the English verb system. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 77–116). Amsterdam & Philadelphia: John Benjamins.

Howarth, P. (1998). The Phraseology of learners' academic writing. In A.P.Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications* (pp. 161–186). Oxford: Clarendon Press.

Huang, L. (2011). Discourse markers in spoken English: A corpus study of native speakers and Chinese non-native speakers. (Unpublished Doctoral Dissertation), The University of Brimingham, England.

Hunston, S. (2002). *Corpora in Applied Linguistics* (3rd ed.). Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar:A corpus driven approach to the lexical grammar of English*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hyland, K. (2008a). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21.

Hyland, K. (2008b). Academic clusters : Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*(1), 41–62.

Ishikawa, S. (2009). Phraseology overused and underused by Japanese learners of English : A Contrastive interlanguage analysis. In K. Yagi & T. Kanzaki (Eds.), *Phraseology, Corpus Linguistics and Lexicography: Papers from phraseology 2009* (pp. 87–100). Nishinomiya, Japan: Kwansei Gakuin University Press.

Johansson, S. (2006). How well can well be translated? On the English discourse particle well and its correspondences in Norwegian and German. In K. Aijmer & A. M. S. Vandenbergen (Eds.), *Pragmatic markers in contrast* (pp. 115–138). Oxford: Elsevier Inc.

Juknevičienė, R. (2009). Lexical bundles in learner language : Lithuanian learners vs . native speakers. *Kalbotyra*, *61*(3), 61–72.

Jung, K. K. (2004). L2 Vocabulary development through conversation: A conversation analysis. *Second Language Studies*, *23*(1), 27–66.

Kaltenbock, G. (2004). Using non-extraposition in spoken and written texts. In K. Aijmer & A.-B. Stenström (Eds.), *Discourse patterns in spoken and written corpora* (pp. 219–242). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Kilimci, A. (2000). *The Computer corpus-based lexical and syntactic contrastive interlanguage analysis on the written performance of the English language learners at Çukurova University with reference to the native speaker performance*. (Unpublished Doctoral Dissertation) Çukurova University, Turkey.

Kilimci, A. (2001). Automatic extraction of the lexical profile of EFL learners through corpus query techniques. *Çukurova Üniversitesi Eğitim Fakültesi Dergisi*, *21*(2), 37–47.

Kilimci, A. (2002). Constructional and functional tendencies of prepositions in the written discourse of advanced Turkish learners of English. In *11th*

International Conference on Turkish Linguistics-ICTL. Cyprus: The Eastern Mediterranean University.

Kilimci, A. (2003). Stance and attitude in advanced Turkish learners' written discourse. In *The 38th Linguistics Colloquium: Language and Language-Processing*. Piliscsaba, Hungary.

Kilimci, A. (2009). *Negotiation of meaning in L2 academic writing***.** In G. Socarras (Ed) Proceedings of 1st International Conference on Literature, Languages and Linguistics. Athens: ATINER

Kilimci, A., & Can, C. (2009). Uluslararası Türk öğrenici İngilizcesi derlemi / TICLE: Turkish ınternational corpus of learner English. In M. Sarıca, N. Sarıca, & A. Karaca (Eds.), *XXII. Ulusal Dilbilim kurultayı bildirileri* (pp. 1–11). Ankara: Yüzüncü Yıl Üniversitesi Yayınları.

Kırkgöz, Y. (2010). An analysis of written errors of Turkish adult learners of English. *Procedia - Social and Behavioral Sciences*, *2*(2), 4352–4358.

Kjellmer, G. (1991). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 111–127). London: Longman.

Kjellmer, G. (2003). Hesitation: In defence of ER and ERM. *English Studies*, *84*(2), 170–198.

Lakshmanan, U., & Selinker, L. (2001). Analysing interlanguage : How do we know what learners know ? *Second Language Research*, *17*(4), 393–420.

Lauttamus, T., Nerbonne, J., & Wiersma, W. (2008). Filled pauses as an evidence of L2 proficiency: Finnish Australians speaking English. In B. Heselwood & C. Upton (Eds.), *Proceedings of Methods XIII: Papers from the Thirteenth International Conference on Methods in Dialectology* (pp. 240–251). Frankfurt am Main: Peter Lang.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics* (pp. 105–122). Berlin: Mouton de Gruyter.

Leech, G. (2000). Grammars of spoken English : New outcomes of corpus-oriented research. *Language Learning*, *50*(4), 675–724.

Léon, J. (2005). Claimed and unclaimed sources of Corpus Linguistics. *Henry Sweet Society Bulletin*, (44), 36–50.

Lightbown, P. M., & Spada, N. (2006). *How languages are learned* (3rd ed.). Oxford: Oxford University Press.

Lindstromberg, S., & Boers, F. (2008). Phonemic repetition and the learning of lexical chunks: The power of assonance. *System*, *36*(3), 423–436.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes*, *31*(1), 25–35.

LOCNESS. (2010). *CECL Team*. Retrieved December 05, 2012, from http://www.uclouvain.be/en-cecl-locness.html

Lombard, R. J. (1997). Non-native speaker collocations : A corpus-driven characterization from the writing of native speakers of Mandarin.

Luzón, M. J. (2009). The use of we in a learner corpus of reports written by EFL Engineering students. *Journal of English for Academic Purposes*, *8*(3), 192–206.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Mahwah,NJ: Lawrence Erlbaum Associates.

Mason, O. (2008). Stringing together a sentence: Linearity and the lexis-syntax interface. In A. Gerbig & O. Mason (Eds.), *Language, People, Numbers-Corpus Linguistics and Society* (pp. 231–248). Amsterdam: Rodopi.

McCarten, J. (2010). Corpus-informed course book design. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of Corpus Linguistics* (pp. 413–427). Abingdon, UK: Routledge.

McCarthy, M. (2006). *Explorations in Corpus Linguistics*. Cambridge: Cambridge University Press.

McCarthy, M., & Carter, R. (2002). This, that and the other: Multi-word clusters in spoken English as visible patterns of interaction. *The Iris Association for Applied Linguistics*, *21*, 30–52.

McCarthy, M., & O'Keeffe, A. (2010). Historical Perspective: What are corpora and how they evolved? In M. McCarthy & A. O'Keeffe (Eds.), *The Routledge handbook of Corpus Linguistics* (pp. 3–13). Abingdon, UK: Routledge.

McDouall, A. (2012). A corpus based investigation into the use of English modal auxiliaries by adult Korean L2 learners. *KLING*, *6*, 33–44.

McEnery, T., & Gabrielatos, C. (2006). English Corpus Linguistics. In B. Aarts & A. McMahon (Eds.), *The Handbook of English Linguistics* (pp. 33–71). Oxford: Blackwell Publishing.

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., & Wilson, A. (2001). *Corpus Linguistics: An introduction* (2nd Ed.). Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies An advanced resource book*. London: Routledge.

McKenny, J. A. (2006). *A corpus-based investigation of the phraseology in various genres of written English with applications to the teaching of English for academic purposes*. (Unpublished Doctoral Dissertation),The University of Leeds, USA.

Meyer, C. F. (2004). *English Corpus Linguistics: An introduction*. Cambridge: Cambridge University Press.

Moon, R. (1998). Frequencies and forms of phrasal lexemes in English. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 78–100). Oxford: Clarendon Press.

Mukherjee, J. (2009). The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 203–230). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Myles, F. (2005). Review article Interlanguage corpora and second language acquisition research. *Second Language Research*, *21*(4), 373–391.

Narita, M., Sato, C., & Sugiura, M. (2004). Connector usage in the English essay writing of Japanese EFL learners. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)* (pp. 1171–1174). Lisbon/Portugal.

Narita, M., & Sugiura, M. (2006). The use of adverbial connectors in argumentative essays by Japanese EFL college students. *English Corpus Studies*, *13*, 23–42.

Nascimento, M. F. B. do, Mendes, A., & Antunes, S. (2006). Typologies of multiword expressions revisited-A corpus driven approach. In Y. Kawaguchi, S. Zaima & T. Takagaki (Eds.), *Spoken language corpus and linguistic ınformatics* (pp. 227–244). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Neff, J., Ballesteros, F., Dafouz, E., Martínez, F., & Rica, J.-P. (2007). A contrastive functional analysis of errors in Spanish EFL university writers' argumentative texts: corpus-based study. In E. Fitzpatrick (Ed.), *Corpus Linguistics beyond the word: Corpus research from phrase to discourse* (pp. 203–225). Amsterdam: Rodopi.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, *24,* 223–242.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125–157). Amsterdam: John Benjamins Publishing Company.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins Publishing Company.

O'Donnell, M., & Römer, U. (2009). Proficiency development and the phraseology of learner language. *ICAME 30,* 1–23.

O'Grady, G. (2010). *A grammar of spoken English discourse*. London: Continuum International Publishing Group.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. New York: Cambridge University Press.

Oostdijk, N. (1991). *Corpus Linguistics and the automatic analysis of English*. Amsterdam: Rodopi.

Oxford Advanced Learner's Dictionary. (2000). *Oxford advanced learner's dictionary*. Oxford: Oxford University Press.

Özhan, D. (2012). *A comparative analysis on the use of but, however and although in the university students' argumentative essays: A corpus-based study on Turkish learners of English and American native speakers*. (Unpublished Master's Thesis) Middle East Technical University, Turkey.

Paananen-Porkka, M. (2007). *Speech rhythm in an interlanguage perspective: Finnish adolescents speaking English. Pragmatics, ideology and contact monographs*. (Unpublished Doctoral Dissertation) University of Helsinki, Finland.

Paquot, M. (2010). *Academic vocabulary in learner writing from extraction to analysis*. London: Continuum.

Paquot, M., & Bestgen, Y. (2009). Distinctive words: A comparison of three statistical tests. In M. Hundt, D. Schreier, & A. H. Jucker (Eds.), *Corpora: Pragmatics and discourse* (pp. 243–265). Amsterdam: Rodopi.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 107–118). London: Longman.

Ping, P. (2009). A study on the use of four-word lexical bundles in argumentative essays by Chinese English: A comparative study based on WECCL and LOCNESS. *CELEA Journal*, *32*(3), 25–45.

Polat, B. (2011). Investigating acquisition of discourse markers through a developmental learner corpus. *Journal of Pragmatics*, *43*(15), 3745–3756.

Pravec, N. A. (2002). Survey of learner corpora. *ICAME Journal*, *26*, 81–114.

Qi, Y., & Ding, Y. (2011). Use of formulaic sequences in monologues of Chinese EFL learners. *System*, *39*(2), 164–174.

Rafiee, M. (2011). Structural analysis of lexical bundles across two types of English newspapers edited by native and non-native speakers, *Modern Journal of Applied Linguistics*, 3 (2)136–155.

Recski, L. J. (2004). Expressing standpoints in EFL written discourse. *Revista Virtual de Estudos da Linguagem*. Retrieved on November 20, 2012, from www.revelhp.cjb.net

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In Sylviane Granger (Ed.), *Learner English on computer* (pp. 40–53). U.S.A.: Addison Wesley Longman Limited.

Rodriguez, L. (2005). Exploring recurrent word combinations in a business English learner corpus : A parallel corpus analysis and its curricular implications. (Unpublished Doctoral Dissertation) University of Puerto Rico, USA.

Romero-Trillo, J. (2002). The pragmatic fossilization of discourse markers in non-native speakers of English. *Journal of Pragmatics*, *34*, 769–784.

Savage, A., Mayer, P., Shafiei, M., Liss, R., & Davis, J. (2010). *Effecitve academic writing*. Oxford: Oxford University Press.

Schmitt, N. (2004). *Formulaic sequences acquisition, processing and use*. (N. Schmitt, Ed.). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic sequences* (pp. 127–152). Philadelphia: John Benjamins Publishing Company.

Scott, M. (2010). *WordSmith tools Version 5.0.* (Vol. 5). Liverpool: Lexical Analysis Software.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *10*, 209–240.

Shirato, J., & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research*, *11*(4), 393–412.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (1996). EAGLES. Preliminary recommendations on Corpus Typology. Retrieved April 12, 2012, from www.ilc.pi.it/EAGLES96/corpustyp/corpustyp.html

Sinclair, J. (2004a). Developing linguistic corpora: A guide to good practice corpus and text basic principles. *Tuscan Word Centre*. Retrieved June 12, 2011, from http://ahds.ac.uk/creating/guides/linguistic-corpora/

Sinclair, J. (2004b). *Trust the text: Language, corpus and discourse*. London: Routledge.

Sinclair, J. (2008). The phrase, the whole phrase, and nothing but the phrase. In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 407–410). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Stenström, A.-B. (2006). The Spanish discourse markers o sea and pues and their English correspondences. In K. Aijmer & A. M. S. Vandenbergen (Eds.), *Pragmatic markers in contrast* (pp. 155–172). Oxford: Elsevier Inc.

Stubbs, M. (2001). Texts, corpora and problem of ınterpretation: A response to Widdowson. *Applied Linguistics*, *22*(2), 149–172.

Stubbs, M. (2002). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell Publishing.

Stubbs, M. (2007). An example of frequent English phraseology: Distribution, structures and functions. In R. Facchinetti (Ed.), *Corpus Linguistics 25 years on* (pp. 89–105). Amsterdam: Rodopi.

Şanal, F. (2007). *A learner corpus based study on second language lexicology of Turkish students of English*. (Unpublished Doctoral Dissertation), Çukurova University, Turkey.

Taylor, C. (2008). What is corpus linguistics ? What the data says. *ICAME Journal*, 32, 179–200.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, *10* (1), 1–13.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at work*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Tognini-Bonelli, E. (2010). Theoretical overview of the evolution of Corpus Linguistics. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of Corpus Linguistics* (pp. 14–27). Abingdon, UK: Routledge.

Tono, Y. (2000). A Computer learner corpus based analysis of the acquisition order of English grammatical morphemes. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the third international conference teaching and language corpora* (pp. 123–132). Frankfurt: Peter lang.

Tono, Y. (2003). Learner corpora : Design , development and applications. In D.Archer, P. Rayson, A. Wilson, & A. McEnery (Eds.), Proceedings of Corpus Linguistics 2003 (pp. 800–809). England: Lancaster University.

Tottie, G. (2010). Uh and Um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*, *16*(2), 173–197.

Waibel, B. (2007). *Phrasal verbs in learner English: A corpus-based study of German and Italian students*. (Unpublished Doctoral Dissertation), Albert-Ludwigs-Universität Freiburg, Germany.

Wei, N. (2009). On the phraseology of Chinese learners spoken English: Evidence of lexical chunks from COLSEC. In A. Jucker, D. Schreier, & M. Hundt (Eds.), *Corpora: Pragmatics and discourse. Papers from the 29th international conference on English language research on computerized corpora (ICAME 29)* (pp. 271–296). Ascona, Switzerland: Rodopi.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A., & Fitzpatrick, T. (2008). Why can't you just leave it alone? Deviations from memorized language as a gauge of nativelike competence. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 123–148). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, *20*(1), 1–28.

Xiao, R. Z. (2007). What can SLA learn from contrastive corpus linguistics? The case of passive constructions in Chinese learner English. *Indonesian Journal of English Language Teaching, 3*(1), 1–19.

Yong, W., Jingli, W., & Zhou, C. (2010). The use of I think by Chinese EFL learners : A study revisited. *Chinese Journal of Applied Linguistis,* 33(1), 3–23.

Zemach, D., & Stafford-Yilmaz, L. (2009). *Writers at work*. Cambridge: Cambridge University Press.

APPENDICES

APPENDIX A

LINDSEI TRANSCRIPTION GUIDELINES

1. Interview identification

Each interview is preceded by a code of this type: <h nt="TR" nr="FR+*three-figure number*">

All interviews should end with the following tag (on a separate line):</h>

2. Speaker turns

Speaker turns are displayed in vertical format, i.e. one below the other. Whilst the letter "A" enclosed between angle brackets always signifies the interviewer's turn, the letter "B" between angle brackets indicates the interviewee's (learner's) turn. The end of each turn is indicated by either </B> or </B>.

3. Overlapping speech

The tag <overlap /> (with a space between "overlap" and the slash) is used to indicate the beginning of overlapping speech. It should be indicated in both turns. The end of overlapping speech is not indicated.

4. Punctuation

No punctuation marks are used to indicate sentence or clause boundaries.

5. Empty pauses

Empty pauses are defined as a blank on the tape, i.e. no sound, or when someone is just breathing. The following three-tier system is used: one dot for a "short" pause (< 1 second), two dots for a "medium" pause (1-3 seconds) and three dots for "long" pauses (> 3 seconds).

6. Filled pauses and backchanneling

Filled pauses and backchanneling are marked as (eh) [brief], (er), (em), (erm), (mm), (uhu) and (mhm). No other fillers should be used.

7. Unclear passages

A three-tier system is used to indicate the length of unclear passages: <X> represents an unclear syllable or sound up to one word, <XX> represents two unclear words, and <XXX>represents more than two words.

If transcribers are not entirely sure of a word or word ending, they should indicate this by having the word directly followed by the symbol <?>.

Unclear names of towns or titles of films for example may be indicated as <name of city> or <title of film>.

8. Anonymisation

Data should be anonymised (names of famous people like singers or actors can be kept). Transcribers can use tags like <first name of interviewee>, <first name and full name of interviewer> or <name of professor> to replace names.

9. Truncated words

Truncated words are immediately followed by an equals sign.

10. Spelling and capitalisation

British spelling conventions should be followed. Capital letters are only kept when required by spelling conventions on certain specific words (proper names, I, Mrs, etc) – not at the beginning of turns.

11. Contracted forms

All standard contracted forms are retained as they are typical features of speech.

12. Non-standard forms

Non-standard forms that appear in the dictionary are transcribed orthographically in their dictionary accepted way:*cos, dunno, gonna, gotta, kinda*, *wanna* and *yeah*.

13. Acronyms

If acronyms are pronounced as sequences of letters, they are transcribed as a series of upper-case letters separated by spaces.

If, on the other hand, acronyms are pronounced as words, they are transcribed as a series of upper-case letters not separated by spaces.

14. Dates and numbers

Figures have to be written out in words. This avoids the ambiguity of, for example, "1901", which could be spoken in a number of different ways.

15. Foreign words and pronunciation

Foreign words are indicated by <foreign> (before the word) and </foreign> (after the word).

As a rule, foreign pronunciation is not noted, except in the case where the foreign word and the English word are identical. If in this case the word is pronounced as a foreign word, this is also marked using the <foreign> tag.

16. Phonetic features

(a) Syllable lengthening

A colon is added at the end of a word to indicate that the last syllable is lengthened. It is typically used with small words like *to*, *so* or *or*. Colons should not be inserted within words.

(b) Articles

-when pronounced as [ei], the article *a* is transcribed as a[ei];

-when pronounced as [i:], the article *the* is transcribed as the[i:].

17. Prosodic information: voice quality

If a particular stretch of text is said laughing or whispering for instance, this is marked by inserting <starts laughing> or <starts whispering> immediately before the specific stretch of speech and <stops laughing> or <stops whispering> at the end of it.

18. Nonverbal vocal sounds

Nonverbal vocal sounds are enclosed between angle brackets.

19. Contextual comments

Non-linguistic events are indicated between angle brackets only if they are deemed relevant to the interaction (if one of the participants reacts to it, for example).

20. Tasks

The three tasks making up the interview (set topic, free discussion and picture description) should be separated from each other. This is done using the following tags: <S> (before the set topic), </S> (after the set topic), <F> (before the free discussion), </F> (after the free discussion), <P> (before the picture description), </P> (after the picture description). These tags should occupy a separate line and should not interrupt a turn.

APPENDIX B
LINDSEI TASKS

LINDSEI

I'd like to interview you informally on things of interest in your life for fifteen minutes. To get the conversation started could you please choose one of the following topics and think about what you are going to say. You should aim to be able to talk for 3-5 minutes. The conversation will then continue informally.

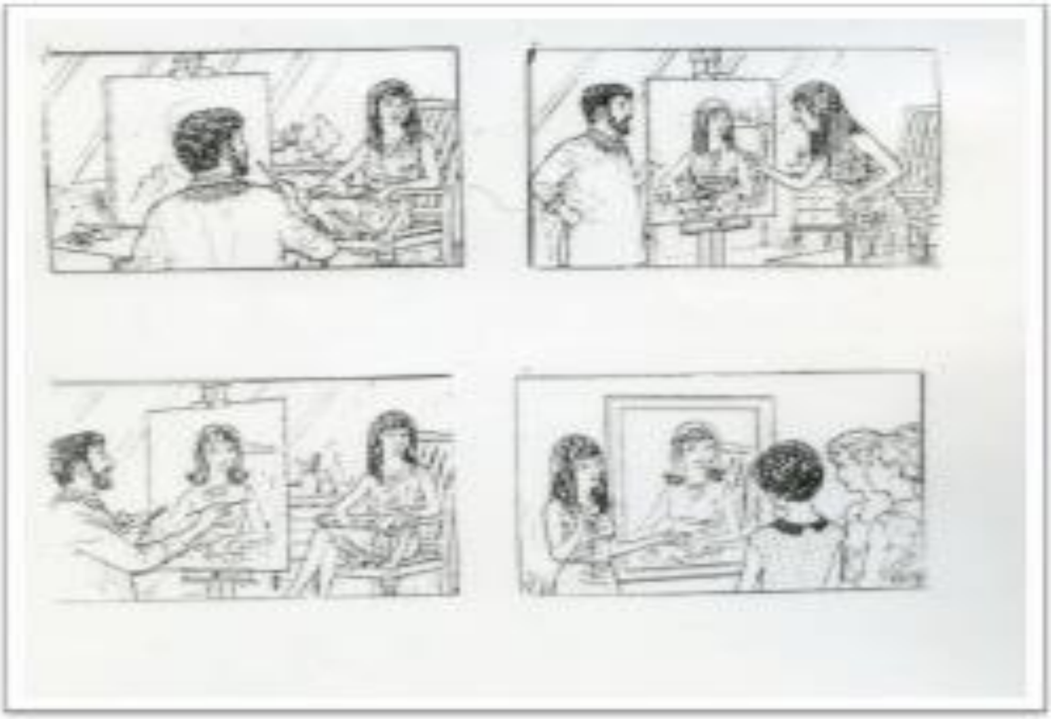
Topic 1: An experience you've had which has taught you an important lesson.Youshould describe the experience and say what you have learnt fromit.

Topic 2: A country you have visited which has impressed you. Describe yourvisit and say why you found the country particularly impressive.

Topic 3: A film/play you've seen which you thought was particularly good/bad.Describe the film/play and say why you thought it was good/bad.

Please don't take any notes as I would like it to be a spontaneous talk.

Story for retelling: The four pictures below tell a story. Study the pictures and then make up a story around them.



APPENDIX C

EXPANDED LIST OF BIGRAMS IN LINDSEI-TR

Word sequence	Frequency	Word sequence	Frequency	Word sequence	Frequency
the woman	143	very much	81	woman eh	63
to be	140	was eh	80	yes yes	62
I was	138	when I	79	and em	61
I can	134	her friends	77	and they	61
I eh	131	so I	76	like eh	61
you know	126	was a	76	she eh	60
of the	125	was very	76	be a	59
at the	120	that I	75	the painter	59
eh we	119	eh because	74	wants to	58
that eh	119	so eh	74	I had	57
for example	115	eh to	73	my eh	57
and then	114	the first	73	of them	57
she is	113	and he	71	she wants	57
eh they	112	is very	71	don’t like	57
the eh	111	then eh	71	there are	57
is a	110	eh my	70	to go	57
and the	108	eh for	69	yes I	57
I am	108	the film	69	eh there	56
I I	108	the i	69	but the	54
I will	105	in my	68	if I	54
in a	102	picture eh	68	don’t eh	54
it is	100	the man	68	the same	53
I don	98	there is	68	this is	53
eh I’ m	98	because i	65	to her	53
to the	98	like that	65	go to	52
and she	96	he is	64	like the	52
eh a	96	of course	64	me eh	52
for me	94	a lot	63	a very	51

APPENDIX D

EXPANDED LIST OF 3 GRAMS IN LINDSEI-TR

Word sequence	Frequency	Word sequence	Frequency	Word sequence	Frequency
it was a	33	eh it is	27	like the picture	21
can I say	32	don t like	26	the painter eh	21
eh for example	32	the woman eh	26	to talk about	21
eh I have	32	be a teacher	25	eh and I	20
and eh the	31	eh I can	25	eh in a	20
eh the woman	31	for example eh	25	first of all	20
I can say	31	I couldn’t eh	25	her picture eh	20
I like eh	31	the I woman	25	I was eh	20
picture and eh	31	to her friends	25	eh eh I	20
the picture and	30	eh at the	24	in my life	19
eh I I	29	I think I	24	she is not	19
and eh he	28	it is not	24	there was a	19
at the end	28	she wants to	24	they don t	19
eh in the	28	draw her picture	23	and eh we	19
how can I	28	eh the man	23	eh the picture	18
I eh I	28	I can t	23	eh there are	18
in the picture	28	it was eh	23	eh they are	18
want to eh	28	the picture is	23	end of the	18
eh but eh	27	eh she is	22		

APPENDIX E

EXPANDED LIST OF 4 GRAMS IN LINDSEI-TR

Word sequence	Frequency	Word sequence	Frequency
I would like to	12	I don t eh	7
at the same time	11	I m trying to	7
eh first of all	11	I will be a	7
end of the film	11	in a different way	7
in the first picture	11	it was a good	7
a lot of things	10	looks at the picture	7
as I said before	10	once upon a time	7
her picture and eh	10	picture to her friends	7
I think it is	10	she wants to be	7
it is very good	10	shows the picture to	7
the picture and eh	10	so I want to	7
there are lots of	10	the I woman eh	7
there is a woman	10	the picture and the	7
eh it was a	9	the picture the woman	7
eh it was eh	9	time with my friends	7
eh it was very	9	to be a good	7
eh there is a	9	and at the end	6
eh there was a	9	and eh I I	6
for me because i	9	and eh I think	6
I want to go	9	and eh I was	6
the woman in the	9	and eh the man	6
woman in the picture	9	and the painter eh	6
a very beautiful woman	8	be an English teacher	6
doesn t like the	8	but I don t	6
don t want to	8	don t know but	6
draw her picture and	8	eh at the same	6
eh at the end	8	eh I can eh	6
eh how can I	8	eh it is not	6
eh Icouldn t	8	eh she didn t	6
first of all I	8	eh she wants to	6

I can say that	8	eh the woman in	6
I don t have	8	eh they don t	6
I don t want	8	firstly I want to	6
in the picture is	8	have a lot of	6
like it very much	8	I don't have any	6
picture and eh she	8	I Iwant to	6
she doesn t like	8	I like it very	6
doesn’t like the picture	8	I said before I	6
with my friends eh	8	I want to have	6
and it was a	7	like the picture and	6
draw her picture eh	7	one day a woman	6
eh I can say	7	plans for the future	6
eh Ididn t	7	she wants him to	6
eh in the first	7	she wants to eh	6
eh she eh she	7	the I woman is	6
eh the man eh	7	the picture is not	6
eh the painter eh	7	to paint her picture	6
eh the woman eh	7		

APPENDIX F

EXPANDED LIST OF 5- AND 6-GRAMS IN LINDSEI-TR

5-Grams		6-Grams	
Word sequence	**Frequency**	**Word sequence**	**Frequency**
to draw her picture eh	5	at the end of the film	9
want to talk about a	5	I want to be a teacher	9
woman in the picture is	5	eh I want to talk about	6
and eh I want to	4	I want to talk about a	5
as I said before I	4	the woman in the picture is	5
as much as I can	4	eh at the end of the	4
but I don t know	4	eh the name of the film	4
eh firstly I want to	4	I have a lot of friends	4
eh the name of the	4	okay I want to talk about	4
em how can I say	4	the end of the film eh	4
end of the film eh	4	and at the end of the	3
have a lot of friends	4	doesn t like the picture and	3
I don t know why	4	don't have any chance I think	3
I had a chance to	4	eh she wants him to draw	3
I have a lot of	4	I don't have any chance I	3
I want to eh I	4	I want to be a good	3
I would like to be	4	I want to talk about eh	3
okay I want to talk	4	I will be an English teacher	3
the name of the film	4	k p s s exam eh	3
to be a good teacher	4	one day a woman comes to	3
to be a teacher eh	4	that there is no one in	3
will be an English teacher	4	the picture to his to her	3
a student in Çukurova university	3	the woman on the portrait was	3
a very beautiful eh woman	3	want to talk about a film	3
and at the end of	3	when I was in high school	3
and eh I think eh	3		
at the same time eh	3		

APPENDIX G
STRUCTURAL CLASSIFICATION OF THE RECURRENT PHRASES IN LINDSEI-TR AND LOCNEC (EXPANDED LIST)

STRUCTURE	EXAMPLES FROM LINDSEI-TR	EXAMPLES FROM LOCNEC
1. Verb Phrase Fragments		
(connector +)1st/2nd person pronoun+VP fragment	***I think*** *(and I think, I think it, I think eh, I think I, I think it is, I think that)* ***I like*** *(and I like, eh I like, I like very much)* ***I want*** *(so I want, I want to, eh I want, and I want, I want to, eh I want to, I want to be, I want to eh, and I want to, so I want to, I want to do, I want to have, I want to go, I want to talk)* ***I have*** *(eh I have, I have to)* ***I was*** *(I was eh, I was a child)* ***I can*** *(I can say, I can say that)* ***you know*** *(you know eh)* ***I will*** *(I will be)* ***I don’t*** *(I don’t know, I don’t have any, I don’t like, I don’t know but,)* ***we don’t know,*** ***I had, I have,*** *(I have to)* ***I didn’t know, I went,*** *(I went to)* ***I would like, as I said (*** *I*	***Ithink*** *(I think it’s, I think I, I think it, yeah I think)* ***you know*** *(you know it’s, you know you, and you know, you know I, you know what I mean)* ***I mean*** *(but I mean, yeah I mean, I mean I, you know the, I mean it’s,)* ***I would like to, I wanted to (*** *I want to be a, I want to talk about* ***,*** ***I don’t*** *(I don’t like, I don’t know, I don’t know I mean, I don’t know I don’t,)* ***you have to, I had to, I used to, you have to, I went to***

	said before), ***I'm trying to***	
(connector +) 3rd person pronoun+ VP fragment	***it was*** *(and it was, it was very, it was very nice,)* ***it is*** *(it is very, it is a, it is not, but it is not, it is very good,)* ***she doesn't like,*** ***she wants,****(she wants to, eh she wants, she wants to be,)* ***he is*** *(eh he is,),* ***she is, they are, this is,*** ***she didn't like***	***it was*** *(and it was, it was just, no it was, it was really, but it was, it was very, it was like, it was it, so it was, er it was, it was the, yeah it was)* ***so that was,***
➢Discourse markers + VP fragment	*I think it is, of course I want*	*sort of you know, you know it was,*
➢Verb Phrase (with non passive verb)	*wants to, like it,* *talk about, be an English teacher, had a chance, have a lot of, draw her picture, like it very much, looks at the picture,*	*paint a picture, know it,*
➢verb phrase with passive verb	------	*was very impressed*
➢yes/no question fragments	*can I say*	----------
➢Wh-Question fragments	*how can I say*	*what else did I*
2. Dependent Clause Fragments		
➢$1^{st}/2^{nd}$ person pronoun+Dependent Clause	*she wants him to, I don't know but, I don't know why*	*I know but, you know but I,*
➢Wh-clause fragments	*when I was a child, when I came to*	*I don't know what, you know what I, can't remember what, what I want to, what I wanted to, when I was*

		in, when I was there, while I was there, I did when I, which is a, which is very, when you go
➢if-clause fragments	-------	
➢verb/adjective+to-clause	*I want to, I'm trying to, I want to be, I want to do, want to go, want to have, I want to talk, I would like to, she wants to be, want to be a*	*like to go, like to see, to go to,*
➢that-clause fragment	*I can say that, I don't think that, that there is no one*	
<u>3. Noun Phrase and Prepositional Phrase Fragments</u>		
➢Noun phrase with –of phrase fragment	*of them,* *end of (the end of the,* *a lot of (a lot of things)* *of the film, (end of the film)* *the name of*	*a bit of, a couple of (couple of years), a lot of (a lot of the), a picture of, awful lot of (an awful lot of,), end of (the end of) one of, lots of, sort of (sort of a, sort of things, sort of you, sort of like, sort of the), kind of*
➢Noun phrase with other post-modifier	*plans for the future, good experience for me*	
➢Other noun phrase expressions	*the picture, my friends, the woman, her picture, her friends, the film, the painter, the man, the woman and, a very beautiful woman, the*	*a look at, two and a half, and things like that, or something like that*

	first picture, one day a woman	
➢Prepositional phrase expressions	*in high school, in my life, for me, with my friends, for four years, to her friends, in the, at the end, in a different way, from my family, in the picture, at the same time, at the picture*	*in the, at all, at the, at the moment*
➢Comparative expressions	------	
4. OTHER EXPRESSIONS	*once upon a time,*	*yeah yeah yeah,*

(Longer combinations embedding the same phrases are given in parenthesis)

APPENDIX H
EXPANDED LIST OF BIGRAMS IN TICLE

Word sequence	Frequency	Word sequence	Frequency	Word sequence	Frequency
of the	1123	and they	210	each other	147
in the	884	will be	206	of a	143
it is	794	because of	203	of their	143
they are	470	the most	203	all the	139
to the	420	that the	197	men and	139
to be	417	the same	189	if they	135
do not	375	that they	184	can t	133
is the	374	are not	182	and women	132
for the	354	for example	181	of this	132
is not	346	people who	180	such as	132
there are	338	at the	178	when they	132
is a	333	not be	171	according to	130
there is	316	have a	169	in this	128
as a	314	in our	169	you can	128
should be	307	from the	168	a lot	127
and the	281	in their	165	the real	127
the other	276	have to	163	of course	126
they can	271	i think	163	to live	124
in a	267	the women	163	try to	124
the world	263	think that	162	be a	123
don t	262	to do	161	it s	123
on the	261	most of	157	a person	122
the students	249	this is	156	kind of	119
can not	242	women are	155	real world	119
can be	238	one of	153	who are	119
they have	238	of them	152	the patient	118
we can	231	to have	151	for a	117
the people	223	by the	150	because they	115

APPENDIX I

EXPANDED LİST OF 3 GRAMS IN TICLE

Word sequence	Frequency	Word sequence	Frequency	Word sequence	Frequency
a lot of	115	people who are	54	prepare students for	38
it is not	112	in our country	53	the help of	38
men and women	112	because of the	52	to sum up	38
on the other	107	women and men	51	between men and	37
they do not	107	can not be	50	of the students	37
in order to	105	i want to	49	they can t	37
there is no	99	of the world	48	don t have	36
in the world	96	day by day	47	in the future	36
one of the	95	that they are	46	they want to	36
the other hand	95	of the people	45	it can be	35
the most important	93	they can not	45	students for the	35
they don t	91	be able to	44	point of view	34
most of the	87	of the most	44	to be a	34
the real world	75	should not be	44	at the same	33
as a result	74	to have a	44	i do not	33
there is a	73	is one of	43	look at the	33
there are some	70	university degrees are	41	the real life	33
it is a	69	in the past	40	there are a	33
it is the	67	in the society	40	if there is	32
do not have	62	is not a	40	it should be	32
that it is	61	is the most	40	a result of	31
first of all	60	that there is	40	can say that	31
in my opinion	60	the right to	40	do not prepare	31
there are many	59	according to the	39	it will be	31
they are not	59	but it is	38	most university degrees	31
the people who	55	i believe that	38	over the world	31
for the real	54	in addition to	38	they have to	31

APPENDIX J

EXPANDED LIST OF 4 GRAMS IN TICLE

Word Sequence	Frequency	Word Sequence	Frequency
on the other hand	94	all around the world	17
one of the most	41	and do not prepare	17
for the real world	40	the people who are	17
is one of the	37	in addition to this	16
between men and women	36	there is no need	16
students for the real	34	are theoretical and do	15
as a result of	31	for a long time	15
are a lot of	30	have the right to	15
there are a lot	30	in every part of	15
all over the world	29	theoretical and do not	15
do not prepare students	27	we can say that	15
university degrees are theoretical	27	men and women have	14
prepare students for the	26	the most important thing	14
not prepare students for	24	the other hand the	14
at the same time	23	they don t have	14
by the help of	23	to get rid of	14
most university degrees are	23	at the end of	13
do not want to	22	but it is not	13
to be able to	22	for the sake of	13
both men and women	21	in the real world	13
degrees are theoretical and	21	one of them is	13
most of the people	20	that there is no	13
of the most important	20	the important thing is	13
is the most important	19	think that it is	13
we look at the	19	we do not have	13
when we look at	19	do not have to	12
in addition to these	12	of course there are	11
is no need to	12	of men and women	11
men and women are	12	of the fact that	11

that it is a	12	prepare students for real	11
the help of the	12	that most university degrees	11
the other hand some	12	the end of the	11
there are many people	12	there are lots of	11
they do not have	12	they do not want	11
with the help of	12	to the real world	11
a result of this	11	women and men are	11
equality between men and	11	and they don t	10
has a right to	11	are thought to be	10
have the same rights	11	but they are not	10
I can say that	11	can not find a	10
if they do not	11	do not have the	10
in a short time	11	for the real life	10
in fact it is	11	freedom of the press	10
is a kind of	11	has the right to	10
have a lot of	12	I think it is	10

APPENDIX K

EXPANDED LIST OF 5- AND 6 GRAMS IN TICLE

5-Grams		6-Grams	
Word sequence	**Frequency**	**Word sequence**	**Frequency**
there are a lot of	30	prepare students for the real world	19
prepare students for the real	26	do not prepare students for the	18
students for the real world	26	not prepare students for the real	18
do not prepare students for	24	most university degrees are theoretical and	16
is one of the most	22	and do not prepare students for	15
university degrees are theoretical and	21	are theoretical and do not prepare	14
most university degrees are theoretical	20	degrees are theoretical and do not	14
one of the most important	20	theoretical and do not prepare students	14
not prepare students for the	18	university degrees are theoretical and do	14
and do not prepare students	16	is one of the most important	11
are theoretical and do not	15	that most university degrees are theoretical	8
degrees are theoretical and do	14	one of the most important inventions	7
on the other hand the	14	do not prepare students for real	6
theoretical and do not prepare	14	prepare students for the real life	6
on the other hand some	12	money is the root of all	5
there is no need to	12	the freedom of the press is	5
when we look at the	12	a lot of women working as	4
as a result of this	11	are a lot of women working	4
by the help of the	11	don t prepare students for the	4
equality between men and	11	every human being has a right	4

women			
they do not want to	9	everyone is supposed to be equal	4
freedom of the press is	8	human being has a right to	4
it is a fact that	8	lot of women working as a	4
prepare students for real world	8	men and women have the same	4
that most university degrees are	8	nail polish and lots of other	4
women and men are equal	8	not prepare students for real world	4
at the end of the	7	on the other hand some people	4

APPENIX L

STRUCTURAL CLASSIFICATION OF THE RECURRENT PHRASES IN TICLE AND LOCNESS (EXPANDED LIST)

Category	Example from TİCLE	Percentage TİCLE	Example from LOCNESS	Percentage LOCNESS
<u>PHRASAL</u>				
g) Noun Phrase Based				
➢noun phrase with of-phrase fragment	*one of, kind of, first of all, point of view, one of them, the importance of, this kind of, the number of, some of them, most of the people, freedom of the press*	*14.3%*	*one of, of this, part of, the use of, the idea of, invention of the, the end of, the number of, this type of, the amount of, the case of, the effects of, one of the most, the majority of the, the teaching of new age, the root of all evil*	*27.3%*
➢other noun phrase or noun phrase fragment	*the other, the world, the students, the*	*18.9%*	*the same, the world, the other,*	*17.8%*

	people, the same, the women, the real, most university degrees, some people, sex equality, women and men		*the fact, the people, the only, the death, many people, the wild card, prayer in public, the drinking age*	
h) Prepositional Phrase Based				
➢prepositional phrase with embedded of-phrase fragment	*of the, by the help of the, as a result of, as a result of this*	*2.6%*	*of the, in favour of, as a result of, in the case of*	*2.7%*
➢other prepositional phrase fragment	*according to, in fact, in addition to, for the, in the, in this, for them, in order to, in the world, in the past, in the society, around the world, in the future, at the end, in the family, on the other hand, for the real world,*	*23.6%*	*in the, with the, in a, as well, about the, in order to, in the past, as long as, according to the,in the world, at the same time, due to the fact that, in the long run*	*26.7%*

	for the sake of			
i) Verb Phrase Based				
➢anticipatory it/Noun/Pronoun + verb phrase/adjective phrase	*it is a, it is the, İ think, they have, they do not have, they want*	*5.6%*	*it was, it is a, it is the, they have, they want to, crime does not pay*	*6.1%*
➢passive verb + prepositional phrase fragment	---	---	---	---
➢copula be + noun phrase/adjective phrase	*is one of the, are a lot of, is one of the most*	*1.5%*	*is a very, is a good, is important to, is one of the*	*1.6%*
➢Pronoun/noun phrase + be (+…)	*it is, this is, women are, people are, we are, money is, most university degrees are*	*6.6%*	*we are, they are*	*1.3%*
➢Existential “there”	*there is some, there is a, there are many, there will be, there is no need, there are many people*	*4.6%*	*there is, there are, there is a, there is no, there are many*	*3.4%*
➢(Pronoun/Noun) + Modal expressions	*they should, we should, they will, must be,*	*5.6%*	*should be, can be, would be,*	*7.5%*

	may be, they cannot, it can be, it should be, we can see, should not be		*could be, will be, should not, be able to, it would be, would not be, the drinking age would*	
CLAUSAL				
➢(verb phrase +) that-clause fragment	*think that, say that, believe that, think that it is, it is true that, is a fact that*	*6.6%*	*is that, İ feel that, is the fact that*	*2%*
➢(verb/adjective +) to-clause fragment	*try to have, want to, are thought to be*	*1.5%*	*is important to*	*0.6%*
➢adverbial clause fragment	*if they, if there is no, when we look at, when they*	*6.6%*	*if the, when the*	*1.4*

CURRICULUM VITAE

PERSONAL INFORMATION

Name: Aysel ŞAHİN KIZIL
Date of Birth/Place: 1980/Elazığ
Email: ayselsahin1@gmail.com

EDUCATIONAL BACKGROUND

Date	**Degree**	**University**	**Field**
2013	Doctor of Philosophy	Çukurova University Social Sciences	English Language Teaching
2007	Master of Arts	Karadeniz Technical University Social Sciences	Applied Linguistics
2002	Bachelor of Arts	Atatürk University	English Language and Literature

JOB EXPERIENCE

2009-	Instructor	Fırat University
2002-2009	Instructor	Karadeniz Technical University

ACADEMIC WORK

Arslan, R. Ş. & Şahin-Kızıl, A. (2010). "How can the use of blog software facilitate the writing process of English language learners?", *Computer Assisted Language Learning*, 23(3), 183-197.

Arslan, R. Ş. & Şahin-Kızıl, A. Extending writing instruction beyond school walls. International Conference on Foreign Language Education, Tuning in: Learners of Language, Language of Learners 24-26 May, 2007 İstanbul TURKEY

Şahin-Kızıl, A. Weblogs: Clearing the path towards learner autonomy, II. Uluslararası Bilgisayar ve Öğretim Teknolojileri Sempozyumu Bildiriler Kitabı, 2008 Ankara: Pegem A Yayınevi

Şahin-Kızıl, A. EFL teachers attitudes towards Information and Communication technologies (ICT). 5th International Computer & Instructional Technologies Symposium 22-24 September 2011, Elazığ, TURKEY

Şahin-Kızıl, A. & Arslan, R.Ş. EFL Students' Experiences with Blog-Integrated Writing Instruction. International Conference "ICT for Language Learning" 5th Edition 15-16 November 2012, Florance, ITALY.

Şahin-Kızıl, A. Google assisted EFL reading instruction. 5th International Conference on Education and New Learning Technologies, 1-3 July,2013, Barcelona, SPAIN.