A Contrastive Interlanguage Analysis of the Highest-Frequency Vocabulary in Advanced and Native English

Yunus Emre Akbana, Kahramanmaraş Sütçü İmam University, Turkey
Gülten Koşar, Social Sciences University of Ankara, Turkey

The European Conference on Language Learning 2015
Official Conference Proceedings

Abstract
Corpus linguistics has vastly developed and been addressed to mirror the frequencies of naturally occurring lexical items not only in English but also in many other languages. Learner corpora represent the written interlanguage performance of L2 or foreign language users coming from different mother tongue backgrounds. International Corpus of Learner English (ICLE) is the first computer learner corpus comprised of the argumentative essays written by advanced learners of English representing 16 different mother tongue backgrounds. In this study Turkish subcorpus of ICLE (TICLE) which represents the written performances of Turkish users of English has been analyzed and ten most frequent words have been listed. TICLE is preferred to be compared with a comparable reference corpus, Louvain Corpus of Native English Essays (LOCNESS).

The results of this data-driven study have been discussed on the basis of Contrastive Interlanguage Analysis (CIA); firstly, the top ten high-frequency words have been illustrated; secondly, the top ten high-frequency words in L2 usage have been compared with that of the native performance and then the overuse, underuse and statistically significant difference tests have been conducted to reveal any properties of interlanguage and native use. The findings revealed that the top ten words are linguistically functional words rather than content words. In addition, seven of the top ten words are commonly used by Turkish learners and American university students. Finally, the linguistic properties of those top ten words have been discussed in detail with implications to ELT in Turkey.

Keywords: ICLE, TICLE, Learner Corpus, Contrastive Interlanguage Analysis

Introduction

Corpus studies have been representative for reflecting the real usage by native or non-native users of a particular language. Regarding English language, there have been many corpus-based investigations in an attempt to mirror the language use in both written and spoken registers. Tognini Bonelli (2001) defines corpus as a collection of real life texts coming from different genres and established on systematic procedures which provide authentic written or spoken language use. In this study the written register has been analyzed by retrieving the data from two corpora in order to reflect the written productions of second/foreign language and native users of English. As corpus provides opportunities for researchers to study any language feature with a high load of data, this has paved the way for building a branch of language investigation; corpus linguistics.

Corpus linguistics has been much popularized and a field of interest with a vast number of studies on English language by researchers most particularly in the last few decades. McEnery and Wilson (2001) state that corpus linguistics is an analysis of linguistics in addition to providing a large number of examples of language use by various groups, individuals or studies on any branch of linguistics. It can be obviously stated that corpus linguistics is an approach that can be employed in any linguistic investigation.

As English has gained importance globally, learners of English seek ways to learn the authentic use of language much more than ever. Corpus assists language learners to easily access to the real language usage via a vast number of software programs. In the 21st century, learners of English have become autonomous in finding ways to access to the authentic language use by the assistance of technology. The ultimate aim of using learner corpus in language teaching programs is to assist language learners to approximate their interlanguage performance to the target language as much as possible. One of the most important learner corpora has been regarded as International Corpus of Learners English (ICLE) by Pravec (2002). It has been accepted in the corpora studies by being a representative of written interlanguage performance of L2 users of English. In addition, Louvain Corpus of Native English Essays (LOCNESS) is a native corpus which represents the authentic native language use by American speakers of English. Therefore, these two corpora have been preferred in this study so as to reveal the language use frequencies both in learner and native English usage. Lozano & Mendikoetxea (2013) state that large-scale learner data assist researchers to explore and describe the investigated language patterns more easily and add that more than 400 articles on L2 studies have made use of ICLE. Granger (2004) suggests that the learner output has been collected by Second Language (SLA) and Foreign Language Teaching (FLT) researchers "for descriptive and/or theory-building purposes". That is to say, listing the frequency of words in corpora is of high importance to SLA and FLT researchers who investigate the over-/underuse of language patterns in interlanguage phraseology or other aspects.

The ultimate purpose of the study is to investigate the top ten words used in both learner and native English and compare them quantitatively. In order to do so, the following questions have been posed to seek answers to:

1) What are the top ten high frequency words used by Turkish-speaking learners of English in TICLE and by American native speakers of English in LOCNESS?

2) Is there any over-/underuse of the top ten words in the written essays of Turkish-speaking learners of English in comparison with that of American native users of English?

Review of Literature

There have been a plethora of corpus based studies on revealing the use of any language pattern that is the concern of researchers for different purposes. There is a list of more than a thousand bibliographical references related to learner corpora research on https://www.uclouvain.be/en-cecl-lcbiblio.html. Granger (2003) points out some of the topics from this list analyzed by using ICLE: "high-frequency words, romance words, recurrent combinations, collocations and formulae, prefabricated language, lexical profiling, lexical variation, adjective intensification, the verb make, progressives, passives, modality, noun phrase complexity, demonstratives, contractions, logical connectors, causal links, conjunctions, participle clauses, direct questions, tense errors, lexical errors, part-of-speech tagging, and parsing".

The variety of the quantitative studies is in the cross roads of low-/middle-/high frequency of the investigated language pattern or linguistic enquiry. To illustrate with an example from many in the literature, Chuang (1993) conducted a quantitative corpus based study on vocabulary and the effects of word frequency and part of speech on vocabulary acquisition. The researcher examined 83 textbooks used in the curriculum of Taiwan as a reference corpus and the exam papers of students which show their interlanguage performance. One of the milestone findings of the study points out the idea that "word frequency is far from a good predictor for students' vocabulary acquisition" (p.102). By the assistance of such studies, there has occurred a bridge between the input and output on the basis of effects of frequency findings on the vocabulary acquisition of the language learners for corpus investigation, SLA and FLT as well. One of the many other corpus based studies on word frequency is the research of Li & Fang (2011) who focus on the grammatical composition of child language with a comparison to maternal language in terms of word classes. In order to reveal a correlation between the input and output frequency, they investigated the maternal language as a source of input and child language as a source of output. The word class frequencies of both input and output based corpora showed less similar word class patterns between the child and maternal language due to the children's mental development. As an implication for SLA or FLT studies, they claimed that "a principle comprehensible input should be highlighted in adults' speech to children in order to make them achieve larger vocabulary" (p.95) In addition, it is also of high importance to present L2 learners of English the most common mistakes committed by L2 learners. As the learners of a foreign language experience the learning process and make similar mistakes, there occurs a common share of mistakes. That is to say, learner corpora can reflect the common mistakes of a community of learners, and other learners can be more conscious about their interlanguage development.

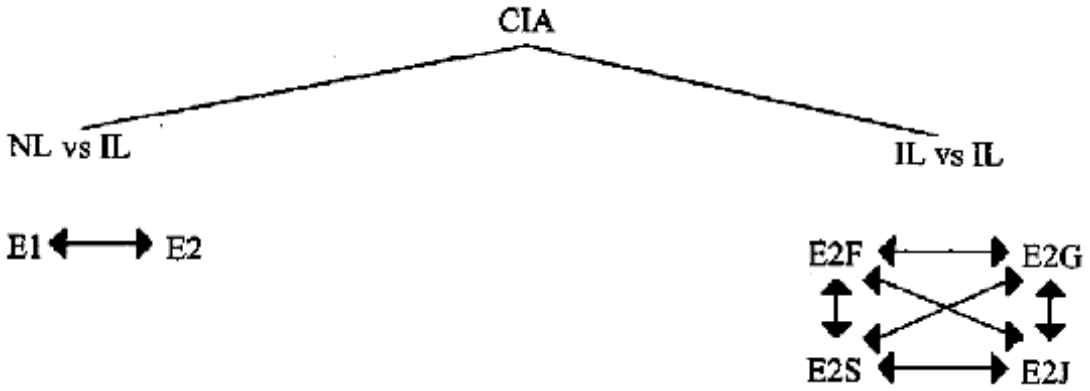
Gilquin and Granger (2010) state that learner corpora assist second language or foreign language learners to find the grammatically correct use of language functions by means of paying attention to the mostly committed grammar mistakes. For this

reason, the modern dictionaries are corpus based; specifically, learner corpus based for L2 English learners. It is of high significance to state that learners can learn with a particular attention to the grammar mistakes, the most frequent lexicogrammar functions of English, and foreign language teachers can be more conscious about the types of difficulties that learners may face in their development process of interlanguage.

In order to reveal the features of interlanguage, Contrastive Interlanguage Analysis (CIA) is an approach which has been used in carrying out corpus studies to compare the retrieved data from one corpus to another. CIA consists of two types of comparison: the native and learner language (L1 vs L2) and /or different varieties of interlanguages (L2 vs L2). (Granger, 1996, 2003, 2004). Granger (2004) suggests a bulk of studies which adopt CIA in comparing the learner and native language concerning; "high frequency vocabulary, (Ringbom 1998, 1999; Källkvist 1999; Altenberg 2002), modals (Aijmer 2002; McEnery and Kifle 2002; connectors (Milton and Tsang 1993; Field 1993; Granger and Tyson 1996; Altenberg and Tapper 1998; L. Flowerdew 1998b), collocations and prefabs (Chi Man-Lai et al. 1994; De Cock 1998, 2000; De Cock et al. 1998; Howarth 1996; Granger 1998; Nesselhauf 2003)".

Figure 1 below depicts the method of CIA. CIA paves the way for revealing any similarities or differences between a native language and mainly its second language use. The left side of the figure represents clearly the core of the present study by investigating the high-frequency words. NL in the figure stands for English, that the data is provided from American native users of English in LOCNESS and IL stands for Interlanguage for which the data is provided from Turkish-speaking L2 users of English in TICLE.

Figure 1: Contrastive Interlanguage Analysis by Granger (1996)



In addition, CIA gives a path for researchers to compare different varieties of interlanguages; however, it is not within the scope of this study. This type of comparison would provide a large opportunity for revealing any interlanguage properties that exist amongst different interlanguages. In this regard, the property of target language and the varieties of interlanguages would be clearer and the learner language would set up its own system of language. In this study, CIA is much of use to depict the over-/underuse of high frequency words excerpted from TICLE in comparison with that of LOCNESS.

Research Design

This data driven study has been carried out with a descriptive and quantitative research design to depict the most frequently used words and their overuse and underuse by the materials, LOCNESS and TICLE. The data retrieved from LOCNESS and TICLE have been analyzed on the basis of CIA and the Log-likelihood values of the data in each corpus material have been examined. This data driven study has collected the data from LOCNESS and TICLE which were developed by a set of certain criteria and are detailed below.

Log Likelihood Statistics (henceforth; LL), which has been previously used and suggested in many studies conducted in corpus linguistics to reflect any overuse or underuse of the investigated linguistic enquiries (e.g. Can, 2011; Granger and Rayson, 1998), has been utilized for the same purpose in this study as well. The top ten words in TICLE data have been preferred to be compared with that of the statistical values in LOCNESS.

Materials

For the current study, a learner (ICLE) and a reference/ native corpora (LOCNESS) have been preferred which are comparable to each other in many aspects and chosen in such studies. There are many issues and criteria raised when to compare a learner and a native corpus. Within the project of ICLE, which allows to compare varieties of L2 English, LOCNESS was established in order to be a comparable corpus with the same variables of age, gender, written register, writing conditions, genre et cetera. LOCNESS is the best matched comparable corpus to ICLE (Hasselgård & Johansson , 2011).

The learner corpus dimension is carried out by utilizing ICLE *version 2* data. ICLE consists of the argumentative essays written by advanced learners of English representing 16 different mother tongue backgrounds. As stated in Granger, Dagneaux, Meunier, and Paquot (2009), there are 16 subcorpora of ICLE which represent the written interlanguage productions of Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish and Tswana users of English as L2 or foreign language. In this study, Turkish subcorpus of ICLE (TICLE) has been preferred in order to create a list of highest frequency words performed in the argumentative essays. TICLE has been compared with a comparable reference corpus; LOCNESS, in that it reveals another list of high-frequency words in native usage.

TICLE is made up of 199,532 words from 280 argumentative essays produced by B2 to C2 (according to CEFR experts' ratings) level EFL learners aged between 21 to 23. The average word length of essays produced by Turkish users of English is 712 words on argumentative topics from education to environment and society. In order to compare the TICLE data with the closest token number to LOCNESS, 208 essays totaling about 149,304 word tokens have been preferred from TICLE data. Table 1 represents the features of TICLE.

Table 1. The features of TICLE

Task Variables		**Learner Variables**	
Medium	Written	**Mother tongue**	Turkish
Genre	argumentative essay	**Age**	21-23
Topic	education, society	**Gender**	Female 81% -Male 19%
Technicality	academic essay	**Language Proficiency**	B2 – C2
Task setting	untimed essay	**Learning context**	EFL classroom setting

Can (2011)

LOCNESS is made up of three sub-corpora totaling with a number of 324,304 word tokens. That is to say, it includes British pupils' A level essays: 60,209 words; British university students' essays: 95,695 words; and, American university students' essays (comprised of literary and argumentative essays): 168,400 words. In order to carry out a comparison between TICLE and LOCNESS, LOCNESS has been preferred to be as a control corpus and it represents similar topics that are examined in TICLE. To approximate the data of LOCNESS to TICLE which consists of argumentative essays, a sample of 175 written argumentative essays totaling a number of 149,574 word tokens produced by 17 to 23-year-old American university students has been preferred. Granger et al. (2009, p. 42) point out that "to ensure comparability with the ICLE data, the Louvain team has collected a corpus of essays written by native English students, the Lovain Corpus of English Essays (LOCNESS), which is the mirror of the ICLE". In addition, LOCNESS is suggested and available to researchers who conduct learner corpus studies involving a comparison of learner and native usage as a control native corpus in many studies. The following studies have made use of LOCNESS and/or suggested it as a control comparable corpus to leaner corpus: (Aarts and Granger, 1998), (Abdullah and Noor, 2013), (Aijmer, 2002), (Altenberg and Tapper, 1998), (Can, 2011), (Granger and Petch-Tyson , 1996), (Guo, 2002; 2003), (Lin, 2002), (Lorenz, 1998), (Narita, Sato, and Sugiura, 2004), (Ringbom, 1998; 1999), (Tapper, 2005), (Tono, 2004) and (Virtanen, 1998)

The data analysis procedure has been carried out firstly by investigating the data. Rayson (2008) has developed Wmatrix to retrieve the data from corpora. Wmatrix is the web interface of USAS and Claws tools and has been used in more than 60 studies and applications up to now. (Please see a full list of the Publications and Applications using Wmatrix on http://ucrel.lancs.ac.uk/wmatrix/). Particularly, Wmatrix has been used in this study in order to retrieve the data from TICLE and LOCNESS. As an initial step, the frequency analysis of the target data has been carried out by Wmatrix. Then, the occurrence of the patterns in concern has been surveyed and the following table has been constructed. Table 2 represents the list of high frequency words by Turkish and American participants.

Table 2: The list of top ten words used by Turkish (TICLE) and American (LOCNESS) participants

	TICLE		LOCNESS	
	Word	*f*	Word	*f*
1	the	7605	the	9060
2	and	3517	Of	4114
3	of	3354	and	3523
4	is	3118	A	3048
5	to	2818	To	2951
6	a	2777	Is	2651
7	in	2663	In	2352
8	they	2288	*that*	2042
9	are	2035	*It*	1431
10	not	1869	*Be*	1404

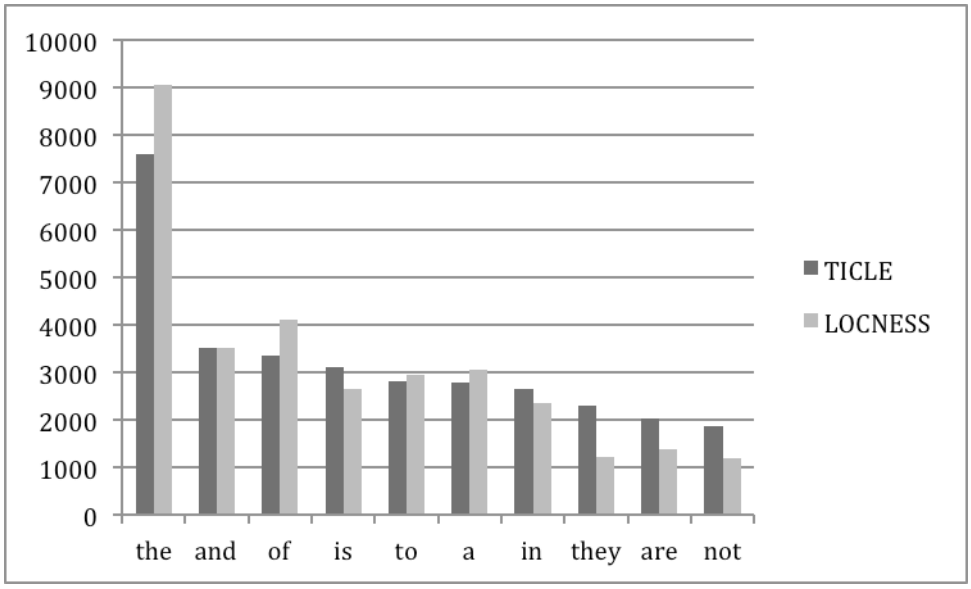
Table 2 shows the most frequently used words in nonnative (TICLE) and native (LOCNESS) usage. Seven out of ten words are determined the same in each corpus: "the, and, of, is, to, a, in", but the order changes. The top three words have been determined the same in each corpus and the highest frequency belongs to "the" in each corpus. "The" has the highest frequency of 7605 in nonnative usage. Besides, "the" is the most frequent word in native usage with 9605 frequency and in LOCNESS data. The second high-frequency word is detected as "and" with a frequency of 3517 in nonnative usage though it is in the third rank in native usage showing frequency of 3523. On the other hand, the second high-frequency word is detected as "of" with a frequency of 4114 in native usage though it is in the third rank manifesting a frequency of 3517 in nonnative usage. Another common share among the first three high-frequency words is that nonnative usage shows underuse of "the, and, of" in comparison with native usage. The fourth high frequency word identified in the nonnative usage is "is" with a frequency of 3118 though the word "is" is observed as the sixth high frequency word with frequency of 2651 in native usage of LOCNESS data. "To" has been illustrated in Table 2 as the fifth high frequency word in both nonnative and native usage. Analysing the data with Wmatrix3, "to" as the infinitive marker has been observed with a frequency of 2818 by nonnatives less than to those in native language use with a frequency of 2951. The sixth high frequency word observed in nonnative performed essays is the singular article "a" with a frequency of 2777 in TICLE. The singular article "a" is observed as the fourth high frequency word in native usage showing frequency of 3048 in LOCNESS. The component of Wmatrix 3, USAS CLAWS7 tagged "in" as general preposition with the same rank order in the frequency list. The seventh high frequency word "in" is used by nonnatives more than natives showing a frequency of 2663 in TICLE thought it is used natives with a frequency of 2352 in LOCNESS.

The eighth high frequency word observed in the TICLE data is "they" which is tagged as 3rd person plural subjective personal pronoun by CLAWS7 POS tagger of Wmatrix3. This particular word "they" showed a usage frequency of 2288 in TICLE though it does not take place in the top ten words in LOCNESS. It is used by natives as the fourteenth high frequency word with a frequency of 1215 in LOCNESS. In

addition to the 3rd person plural subjective personal pronoun "they" used by nonnatives, the linking verb "are" and the adverb "not" have been observed in nonnative usage among the top ten word list though these words have not been analysed in the top ten word list of native usage. Instead, "that" as conjunction, "it" as the 3rd person singular neuter personal pronoun, and "be" as the infinitive be are observed in native usage. The ninth high frequency word in TICLE list is "are" with a frequency of 2035. Though this word is used in LOCNESS list as the eleventh high frequency word with a frequency of 1391. The tenth high frequency word used in TICLE list is "not" with a frequency of 1869 though it is used as the fifteenth high frequency word showing frequency of 1202 in LOCNESS.

As this paper seeks to find out to what extent Turkish speakers of English use the top ten words in comparison with that of the usage represented by native speakers of English, Figure 2 shows an overall of usage of the words used by nonnatives. Figure 2 below stands for a clear outlook of the difference laid out by the two corpora usage.

Figure 2. The Overall Frequency Distribution of NNS Performance



It is clear in Figure 2 that there are four words used more by native speakers of English than nonnative speakers of English. Namely, "the", "of", "to" and "a" represent more usage frequency by American speakers of English than Turkish speakers of English. The most used word in both corpora is "the" at a stake and it is clear from Figure 2 that native speakers of English use it more frequently than Turkish speakers of English. The close frequency of "and" is conspicuous in Figure 2 that the usage frequency in TICLE data is 3517 and 3523 in LOCNESS data. In addition to illustrating the data in a figurative way, the over/under use of the words in both corpora seems clear at a stake; however, Table 3 below provides statistically overuse or underuse of the words examined in both corpora usage.

Table 3: The list of top ten over- and underuse of words used in nonnative writing in comparison with native writing

		NNS	NS	LL Ratio (*p< 0.05)
1	the	7605	9060	-126.02
2	and	3517	3523	-0.00
3	of	3354	4114	-76.86
4	is	3118	2651	+38.23
5	to	2818	2951	-2.96
6	a	2777	3048	-12.39
7	in	2663	2352	+19.55
8	they	2288	1215	+334.88
9	are	2035	1391	+154.61
10	not	1869	1202	+146.57

+ indicates overuse in TICLE relative to LOCNESS,
- indicates underuse in TICLE relative to LOCNESS

Tables 3 represents five of the words; “is”, “in”, “they”, “are”, “not”, in the top ten words lists as overused in TICLE relative to LOCNESS according to the observed significantly high LL values. Another significantly high LL values determine the underuse of the following three words in TICLE relative to LOCNESS; “the”, “of”, “a”. The remaining two words “and” and “to” have been observed without any significantly high different values in TICLE relative to LOCNESS. Following is the detailed LL values based on Table 3.

Using Wmatrix as a frequency profiling assessment software programme, we have utilized LL values as Rayson (2008) suggests. The higher the LL value, the more significant is the difference between two frequency scores. For this particular study, we have preferred the LL cut-off value as ($p<0.05$) which stands for a critical values of 3.84; thus the more LL critical value is, the more significantly difference is among the items. The results obtained Table 3 shows that the five overused words in TICLE relative to LOCNESS; ; “is”, “in”, “they”, “are”, “not”, have been confirmed by the LL calculation that the LL values are +38.23, +19.55, +334.88, +154.61 and +146.47 respectively. Hence, Table 3 displays the most overused word by Turkish speakers of English as “they” with a LL value of +334.88 indicating a very high statistically significant difference between two corpora though it is the eighth most frequent word in the TICLE list and the fourteenth in the LOCNESS list. The second most overused word by the Turkish speakers of English is “are” with a LL value of +154.61 though it is the ninth most frequent word in the TICLE list and the eleventh most frequent word in the LOCNESS list as confirmed also by the results of Table 2. Table 3 displays the third most overused word in the written productions of Turkish speakers of English as “not” though it takes the fifteenth rank order of frequency in the written productions of American native speakers of English. This result like the others obtained from Table 3 is confirmed in the same way with the frequency analysis of each items in both corpora in Table 2. Again, as confirmed by the results of Table 2, Table 3 suggests the fourth most overused word in the TICLE data in comparison with that of LOCNESS data is “is” with a calculated LL value of +38.23 indicating a statistically significant difference between two corpora. Finally, the least overused word used by Turkish speakers of English relative to the performances of American native speakers has been

demonstrated as “in” with a LL value of +19.55. The particular word “in” is used in both the TICLE and LOCNESS lists as the seventh most frequent word confirmed by the results of Table 2 as well.

Table 3 displays three underused words in TICLE relative to LOCNESS; “the”, “of” and “a”. Comparing the usage between the TICLE and LOCNESS corpora, “The” has been observed the most used word with the highest frequency in Tables 2, 3 and 4 in addition to Figure 1. However, Table 3 demonstrates the highest underuse LL value regarding “the”. Hence, Table 3 clearly shows that “the” has been used the most underused word in TICLE with -126.71 LL value ($p<0.05$), indicating a statistically very high difference between two corpora. The second most underused word in TICLE has been observed as “of” with -76.86 LL value ($p<0.05$) showing statistically significant difference between TICLE and LOCNESS. Finally, the least underused word indicating statistically significant difference between the two corpora shown in Table 3 is “a” with -12.39 LL value ($p<0.05$). Apart from those above mentioned underused words in TICLE data relative to LOCNESS, there remains two other items; “and” and “to”. These two items have been investigated and found that they do not indicate any statistically significant difference with respect to any kind of underuse or overuse between the two corpora. More specifically, the use of “and” between the two corpora shows no underuse or overuse as the observed LLvalue is -0.00 ($p< 0.05$) that we need at least a LL value of 3.84 to claim any underuse or overuse of the item. Regarding “to” the LL value is -2.96 which indicates no statistically significant difference between two corpora.

Conclusion

By exploring the most frequently used ten words in the essays of Turkish speakers of English in Turkish International Corpus of Learner English (TICLE) and American native speakers of English in Louvain Corpus of Native English Essays (LOCNESS), our study aimed to investigate the use of the top ten words in a Contrastive Interlanguage Analysis (CIA) manner and the following remarks have been drawn:

Overuse and/or underuse of the words tend not to show any native use property but interlanguage use. Hence, most properties of interlanguage use have been observed with the overuse and underuse of the words observed in the written productions of Turkish speakers of English. Out of the top ten words, only two words have been found to reveal native use “and” and “to” as there are no statistically shown overuse or underuse of the items in concern. This is also a proof of the fact that Turkish speakers of English can use these particular two items almost at the same rate with that of the American native speakers of English.

The findings revealed that the top ten words illustrated in the body of this study are linguistically functional words rather than content words. In order to see a larger list of the highest frequency words Appendix 1 has been presented with parts of speech of each item in the argumentative essays of both Turkish learners and American university students.

Our first research question investigated the top ten high frequency words used by Turkish-speaking learners of English in TICLE and by American native speakers of English in LOCNESS. Findings revealed that Turkish learners use “the, and, of, is, to,

a, in, they are, not" and American university students use the first seven words of Turkish learners but differ in the last three words with the use of "that, it, be". As Table 2 shows Turkish learners and American university students use seven common words out of the top ten words. Hence, Turkish learners can use as many highest frequency words as American university students. The argumentative essays of the two corpora examined in this study include similar frequency scores in the order of top ten words.

The second research question of the study was posed to find out whether there is any over-/underuse of the top ten words in the written essays of Turkish-speaking learners of English in comparison with that of the American native users of English. The statistically overused words by Turkish learners in comparison to American university students in both TICLE and LOCNESS data are "is, in, they, are, not". The statistically underused words by Turkish learners relative to American university students' written productions are "the, of, a". (see Table 3)

Implications to Language Teaching in Turkey

The present study has revealed that Turkish learners and American university students show a strong tendency to use similar words with highest frequencies. This leads us to conclude that Turkish learners approximate their interlanguage performance to the native use performed by the American native speakers of English in LOCNESS data. This conclusion stands as a proof for the fact that how successful advanced users of English language in Turkey. However, in order to prevent any probable difficulty that an elementary user of English can face in their interlanguage process, this list can be posed to their textbooks and the textbook authors should well be informed about corpus driven tools to show how the native usage performs. In a similar manner but similar design of a study, the similar implications were drawn in the study of Shin and Nation (2006) that they conducted a study of investigation of spoken highest frequency words and made implications for elementary speakers of English L2 learners. They investigated ten million word BNC spoken section and suggested for inclusion of the most frequent 2,000 words of English, that many of these collocations could be usefully taught in an elementary speaking course.

This study sheds light on the idea that the inclusion of the most frequently used words into the textbooks in the language teaching programs in Turkey is a good requisite for providing comprehensible input in the process of acquiring English as L2 or FL.

Suggestions for Further Research

This particular present study is limited to argumentative essays of Turkish learners and American university students; that, the further research might involve a broader research body of including literary texts as well in addition to other interlanguage productions other than Turkish learners' from ICLE or other leaner corpora. By doing this, it can lead the further research depicting the performances of other L2 users of English. That is to say, we can see how well a Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, and Tswana user of English performs the highest frequency in their own system of interlanguage and we can compare it with that of native use with LOCNESS data or any other larger data sets. Furthermore, as the American native usage has been

investigated in this study as a control corpus to depict the native usage, British native can also be included to stand for a more comprehensible native usage. In addition, this study can also be broadened to a wider setting of spoken corpus and the most frequent words in addition to collocations can be drawn to the attention of textbook authors and dictionary writers. Finally, the reasons why some certain items are used more or less frequently than likelihood items by Turkish learners might well be researched.

References

Aarts, J., & Granger, S. (1998). Tag sequences in learner corpora: a key to interlanguage grammar and discourse. In S. Granger (Ed.), Learner English on computer (pp. 132–141). London & New York: Addison Wesley Longman.

Abdullah, S., & Noor, N. M. (2013). Contrastive Analysis of the Use of Lexical Verbs and Verb-noun Collocations in Two Learner Corpora: WECMEL vs. LOCNESS. In S. Ishikawa (Ed.), Papers from LCSAW2013 (Vol. 1, pp. 139–160). Japan: Kobe University.

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), Computer learner corpora, second language acquisition and foreign language teaching (pp. 55–76). Amsterdam & Philadelphia: John Benjamins.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learner's written English. In S. Granger (Ed.), Learner English on computer (pp. 80–93). London & New York: Addison Wesley Longman.

Can, C. (2011) "Conjunctive Adverbs in Learner English: A Usage-based Approach." The Dialogue of Language, the Dialogue of Culture Ed. A. Lyda, D.G. Baker, M. Blaszak, T. Wasza, 92-105 pp., Warsaw, Poland, University of Silesia. 2011

Chuang, Y. (1993) 'A quantitative corpus analysis of word frequency and part of speech in the English textbooks used in senior high schools in Taiwan', Proceedings of PACFoCoL, p. 96–106.

Gilquin, G. & Granger, S. (2010). How can data-driven learning be used in language teaching?. In A. O'Keeffe & M. McCarthy (Eds.), The Routledge handbook of corpus linguistics (pp. 359-70). London: Routledge.

Granger, S. (1996). From CA to CIA and back: an integrated approach to comuterized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (Eds.), Languages in contrast. Text-based cross-linguistic studies (pp. 37-51). Lund: Lund University Press.

Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. TESOL Quarterly, 37(3), 538–546.

Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In U. Connor & T. Upton (Eds.), Applied corpus linguistics: a multidimensional perspective (pp. 123–145). Amsterdam & Atlanta: Rodopi.

Granger, S., 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, S., & Petch-Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. World Englishes: Journal of English as an International and Intranational Language, 15(1), 17–27.

Granger, S., & Rayson, P. (1998). Automatic lexical profiling of learner texts. In S. Granger (Ed.), Learner English on computer (pp. 119–131). London & New York: Addison Wesley Longman.

Guo, X. (2002). A comparative corpus study of modal verbs in COLEC and LOCNESS. Presented at the First Inter-Varietal Applied Corpus Studies (IVACS) conference, Limerick.

Guo, X. (2003). Between verbs and nouns and between the base form and the other forms of verbs - a contrastive study into COLEC and LOCNESS. In D. Archer, P.

Rayson, A. Wilson, & T. McEnery (Eds.), (Vol. 16, pp. 274–281). Presented at the Corpus Linguistics 2003 Conference, Lancaster University: University Centre for Computer Research on Language.

Hasselgård, H. and Johansson, S. (2011) Learner corpora and contrastive interlanguage analysis. In Meunier F., De Cock S., Gilquin G. and Paquot M. (eds), *A Taste for Corpora. In honour of Sylviane Granger*. Amsterdam: Benjamins, 33-62..

Li, H.-H. and A.C. Fang. (2011). Word Frequency of the CHILDES Corpus: Another perspective of child language features. In ICAME Journal, Vol 35. pp 95-116.

Lin, L. (2002). Overuse, underuse and misuse: using concordancing to analyse the use of "It" in the writing of Chinese learners of English. In M. Tan (Ed.), Corpus studies in language education (pp. 63–76). Bangkok (Thailand): IELE Press.

Lorenz, G. (1998). Overstatement in advanced learners' writing: stylistic aspects of adjective intensification. In S. Granger (Ed.), Learner English on computer (pp. 53–66). London & New York: Addison Wesley Longman.

Lozano, C., & Mendikoetxea, A. (2013). Corpus and experimental data: Subjects in second language research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead. Proceedings of the First Learner Corpus Research Conference (LCR 2011) (pp. 313–323). Louvain-la-Neuve: Presses Universitaires de Louvain. Retrieved from http://www.uclouvain.be/445428.html

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

Narita, M., Sato, C., & Sugiura, M. (2004). Connector Usage in the English Essay Writing of Japanese EFL Learners. In LREC. European Language Resources Association. Retrieved from http://www.lrecconf.org/proceedings/lrec2004/pdf/48.pdf

Pravec, N. (2002). Survey of Learner Corpora. ICAME Journal 26: 82-114

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*. 13:4 pp. 519-549. DOI: **10.1075/ijcl.13.4.06ray**

Ringbom H. (1998). Vocabulary frequencies in advanced learner English: a crosslinguistic approach. In S. Granger. (Ed.), Learner English on computer (pp. 41-52). London & New York: Addison Wesley Longman.

Ringbom H. (1999). High-frequency verbs in the ICLE corpus. In A. Renouf (Ed.), Explorations in Corpus Linguistics (pp. 191-200). Amsterdam and Atlanta: Rodopi, 191-200.

Shin, D. & Nation, I. S. P. (2006). Beyond single words: The most frequent collocations in spoken English. ELT Journal 62, 4, 339–48.

Tapper, M. (2005). Connectives in advanced Swedish EFL learners' written English – preliminary results. In F. Heinat & E. Klingvall (Eds.), The Department of English in Lund: Working Papers in Linguistics 5 (pp. 115–144). Lund: Department of English, Lund University.

Tognini Bonelli, E. (2001). Corpus linguistics at work. Amsterdam and Philadelphia: John Benjamins.

Tono, Y. (2004). Multiple comparisons of IL, L1 and TL corpora: the case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In G. Aston, S. Bernardini, & D. Stewart (Eds.), Corpora and language learners (pp. 45–66). Amsterdam; Philadelphia: John Benjamins.

Virtanen, T. (1998). Direct questions in argumentative student writing. In S. Granger (Ed.), Learner English on computer (pp. 94–118). London; New York: Addison Wesley Longman.

Contact email: gencoglugulten@gmail.com

Appendix 1. The top a hundred word frequency and part of speech list in TICLE and LOCNESS

	TICLE			LOCNESS		
	Word	**POS**	***f***	**Word**	**POS**	***f***
1	the	AT	7605	the	AT	9060
2	and	CC	3517	of	IO	4114
3	of	IO	3354	and	CC	3523
4	is	VBZ	3118	a	AT1	3048
5	to	TO	2818	to	TO	2951
6	a	AT1	2777	is	VBZ	2651
7	in	II	2663	in	II	2352
8	they	PPHS2	2288	that	CST	2042
9	are	VBR	2035	it	PPH1	1431
10	not	XX	1869	be	VBI	1404
11	it	PPH1	1663	are	VBR	1391
12	be	VBI	1547	for	IF	1303
13	for	IF	1468	this	DD1	1216
14	that	CST	1449	they	PPHS2	1215
15	can	VM	1358	not	XX	1202
16	people	NN	1315	to	II	1119
17	their	APPGE	1217	with	IW	813
18	this	DD1	1155	people	NN	791
19	we	PPIS2	1148	their	APPGE	777
20	to	II	1126	or	CC	683
21	you	PPY	996	on	II	663
22	or	CC	924	would	VM	654
23	I	PPIS1	789	was	VBDZ	627
24	do	VD0	781	i	PPIS1	623
25	but	CCB	781	have	VH0	620
26	money	NN1	726	has	VHZ	609
27	there	EX	699	by	II	605
28	with	IW	696	can	VM	572
29	have	VH0	692	an	AT1	569
30	students	NN2	666	's	GE	529
31	should	VM	649	but	CCB	513
32	life	NN1	643	if	CS	506
33	will	VM	638	that	DD1	489
34	if	CS	628	from	II	474
35	who	PNQS	606	will	VM	471
36	these	DD2	592	these	DD2	470
37	some	DD	568	we	PPIS2	432
38	from	II	543	many	DA2	431
39	them	PPHO2	537	should	VM	431
40	by	II	508	there	EX	427
41	all	DB	482	who	PNQS	424
42	our	APPGE	469	what	DDQ	417
43	also	RR	462	as	CSA	412

44	on	II	456	because	CS	399
45	when	CS	451	he	PPHS1	392
46	which	DDQ	450	also	RR	384
47	about	II	441	were	VBDR	374
48	an	AT1	433	you	PPY	371
49	because	CS	427	all	DB	367
50	have	VHI	416	have	VHI	349
51	women	NN2	413	do	VD0	345
52	as	CSA	397	our	APPGE	325
53	many	DA2	396	women	NN2	321
54	has	VHZ	390	been	VBN	314
55	that	DD1	373	could	VM	304
56	he	PPHS1	367	one	MC1	302
57	world	NN1	356	when	CS	299
58	as	II	350	them	PPHO2	298
59	what	DDQ	350	society	NN1	292
60	education	NN1	348	some	DD	290
61	so	RR	329	which	DDQ	283
62	very	RG	328	n't	XX	279
63	important	JJ	324	his	APPGE	279
64	universities	NN2	308	at	II	278
65	person	NN1	307	life	NN1	274
66	at	II	295	as	II	263
67	think	VV0	285	children	NN2	260
68	want	VV0	279	about	II	255
69	other	JJ	276	money	NN1	242
70	men	NN2	274	does	VDZ	236
71	university	NN1	274	more	RGR	236
72	nt	XX	273	she	PPHS1	232
73	children	NN2	268	more	DAR	231
74	problems	NN2	262	other	JJ	229
75	may	VM	258	being	VBG	216
76	do	VDI	253	than	CSN	207
77	she	PPHS1	251	just	RR	206
78	your	APPGE	250	students	NN2	205
79	knowledge	NN1	250	one	PN1	201
80	his	APPGE	242	how	RRQ	198
81	most	DAT	235	may	VM	197
82	being	VBG	230	time	NNT1	195
83	good	JJ	228	argument	NN1	194
84	things	NN2	227	very	RG	191
85	abortion	NN1	227	my	APPGE	191
86	one	MC1	226	had	VHD	186
87	real	JJ	222	those	DD2	185
88	euthanasia	NN1	222	any	DD	183
89	n't	XX	215	person	NN1	182
90	no	AT	208	then	RT	174
91	way	NN1	208	no	AT	172

92	was	VBDZ	204	only	RR	166
93	even	RR	202	sex	NN1	166
94	more	RGR	200	her	APPGE	165
95	must	VM	197	problem	NN1	164
96	most	RGT	197	another	DD1	163
97	family	NN1	196	even	RR	162
98	any	DD	195	however	RR	162
99	how	RRQ	193	family	NN1	161
100	just	RR	193	's	VBZ	158