THE USE OF SPOKEN LEARNER CORPORA TO DETECT PROBLEMS WITH LEXICAL ACCURACY

Elif Tokdemir DEMİREL[1]
Koray ŞAHİN[2]

Öz: Bu çalışmanın amacı yabancı dil öğrencilerinin, özellikle sözcüksel öğelerin kullanımı konusundaki problemlerinin daha kolay saptanması için derlem yönteminin kullanılması yolunda bir öneri getirmektir. Her ne kadar sözlü derlem toplama aktivitesi zor ve zaman alıcı bir süreç olsa da, İngilizce'nin yabancı dil öğrencileri tarafından kullanımı veya yanlış kullanımı konusuna ışık tutma potansiyeli yüksektir. Çalışmada kullanılmak üzere, bir sözlü öğrenci İngilizcesi derlemi oluşturulmuştur. Çalışmanın katılımcıları orta ve ortanın üstü İngilizce seviyesindeki yabancı dil öğrencilerdir. Çalışmada kullanılan derlem, diğer bir deyişle 'Öğrenci Monologları Derlemi' (ÖMD) katılımcı yabancı dil öğrencilerinin iki farklı konu üzerine 35 adet konuşma kaydının çözümlemesini içermektedir. Öğrenci Monologları Derleminin oluşturulması için kullanılan konuşma konuları IELTS (Uluslararası İngilizce Dil Yeterlilik Sınavı) sınavının konuşma bölümünde kullanılan sorular arasından seçilmiştir. Derlem için toplanan ses kayıtları çözümlenmiş ve hata kategorileri yönünden elle kodlanmıştır. Öğrencilerin yaptıkları hataların kodlanmasından sonra, derlem özel bir derlem analizi programı olan AntConc 3.2.4w kullanılarak kelime kullanımı yönünden analiz edilmiştir. Çalışmanın sonuçları, öğrencilerin konuşma kayıtlarındaki en sorunlu sözcüksel grubun fiiller olduğunu ortaya çıkarmıştır. Bunun dışında, yaygın olarak saptanan yanlış kullanımlar sırasıyla zarf, isim ve sıfat kelime gruplarını içermektedir. Sonuçlar ayrıca yanlış kelime seçiminin öğrencilerin sözcüksel hatalarının en önemli sebeplerinin başında olduğuna işaret etmektedir. Çalışma, yabancı dilde konuşma becerisi öğretimi alanında öneriler içermektedir.

Anahtar Sözcükler: Derlem**,** Sözlü Öğrenci Derlemi**,** Sözcüksel Hata Analizi.

Introduction

Analysis of corpora in language and linguistics research is not a new issue. Nevertheless in the last fifty years, the fast developing technology and the use of new hardware and software in this area brought corpus analysis into

[1] Yrd. Doç. Dr., Karadeniz Teknik Üniversitesi, Edebiyat Fakültesi, Batı Dilleri ve Edebiyatı Bölümü, İngiliz Dili ve Edebiyatı Anabilim Dalı. .elif@ktu.edu.tr

[2] Okt., Giresun Üniversitesi, Yabancı Diller Yüksekokulu .koray.sahin@giresun.edu.tr

prominence in language research as it offers efficient ways of collecting and accessing genuine data. Corpus analysis and language education has a close relationship, when the data studied in the former is collected among products of the latter. In this respect, the results of corpus analysis research might contribute a lot to the field of language education.

Corpus linguistics can be defined as a method of linguistic analysis which uses different kinds of corpora. A corpus is a collection of natural language data, compiled from written texts and/or transcription of spoken language. Corpus linguistics is used to analyze and research a number of linguistic questions and offers a unique insight into the dynamics of language which has made it one of the most widely used linguistic methodologies.

It can be said that corpus linguistics has appeared in the 1950s and 1960s, with some leading studies such as compliation of The Quirk Corpus, the emergence of the Brown Corpus of American English (Francis and Kučera 1964), and the pioneering work of John Sinclair on collocation (Sinclair et al. 2004). Since the 1990s corpus linguistics has been progresively seen as an important component of the methodological toolbox by many linguists, whereas it was a specialized sub-field at first. Computer aided analysis of "a large and principled collection of natural texts" comprises the basis of corpus study methods (Biber et al. 1998).

Some of the main analyses that can be done via computer are: (a) creating a frequency word list which provides counts of words or word groups in the corpora and (b) doing a search for all instances of a linguistic item, mostly a phrase or a word, in the corpora. Related software programs such as AntConc 3.2.4w display the results of such searches in a window where each instance of the items can be seen as concordance lines. Corpus-based approaches mostly use statistical analysis- generally in handling frequencies in a corpus- however, corpus linguistics utilizes quantitative and qualitative methods while doing the analysis (Biber et al. 1998, p. 4). For example, concordances and frequency lists of the words can be produced via a computer software, however, to make a relevant discussion and draw conclusions out of these lists definitely requires qualitative analysis on the items.

Linguistics makes use of corpus-based approaches in many areas. Lexicography may be taken as a good exampe of one the most important fields in which corpus techniques are now used very effectively. Recently, corpus methodologies have started to aid analyses of the lexical dimension of language such as lexical richness (Zhang, 2014), lexical bundles in various kinds of language (Csomay, 2013; B. Jablonkai, 2010; Gray & Biber, 2013) or Lexical associations (Rizzo, 2009). Corpora have also been used for following types of studies: text- type variation, dialect variation, the methapor studies, literary stylistics and Critical Discourse Analysis.

Currently, two main approaches exist in the status of corpus linguistics as a discipline (Hardie and McEnery 2010). One approach sees corpus linguistics as

an independent sub-field or theory of language (Teubert 2005, p. 2); according to the other perspective, corpus linguistics is considered primarily as a methodology that can be applied in various analytical and theoretical frameworks.

Resarch in corpora has a wide range of use in language teaching and learning as it has much to contribute to these fields. Barlow (2002) suggests that corpus linguistics can be applied in syllabus design, materials development, and classroom activities. Especially research in learner corpus aims to find out the links between L2 learning and aspects of learner lexis, grammar and discourse using the methods and tools of corpus linguistics, with the theoretical knowledge gained from corpus linguistics and second language acquisition (SLA) research (Tan 2005). The main objective of learner corpora studies has mostly been to get better understanding into the language used by the learners. Learner data research generally illustrates examples of 'overuses', 'underuses', or 'misuses' of the items in target language.

Leech (1997, p. 20) recognized the importance of this type of research for SLA studies, and highlighted its value in providing 'authoritative answers' for language errors that are commonly repeated by the learners. He also proposed that the outcomes of such studies might be applied to the language classrooms so that overused, underused or misused structures and lexical items of the target language by non-native learners may be taken in the actual teaching process.

The efficacy of the research on spoken corpora for the English Language Teaching has been recognized by many reserarchers (Bennett 2010; Reppen 2014). Certain differences of spoken language have been revealed by the earlier studies (McCarthy 1998; Koester 2001). Speech corpora chiefly shows authentic, natural and real language use in different contexts.

Lexical errors have been a certain part of process of the second language vocabulary learning. There has been various studies on the lexical errors in the literature (Hemchua and Schmitt 2006; Llach 2007). Analyzing a spoken corpora, this study aims to reveal the most common misuses of language items of four word categories –nouns, adjectives, verbs, adverbs- by language learners.

1. Method

1.1. Research Questions

The research questions addressed by the study are:

1- How can spoken corpora analysis be put into use in detection of the most problematic word groups and most common word misuses in students' speech?

2- What are the most problematic lexical items in students' speech regarding the four basic word categories: nouns, verbs, adjectives, and adverbs?

1.2. Participants

In total, 35 preparatory students studying English Language and Literature participated in the study. They were at intermediate to upper-intermediate

proficiency level of English at the time of the research. Each participant was informed about the study and recording process beforehand.

1.3. Materials

For the purposes of this study, a spoken corpora of learner English (Corpus of Learner Monologues) consisting 35 monologues by language learners was compiled. The monologues were either of two different tasks, chosen among IELTS speaking tasks. For the first task, learners were asked to give two minutes of speech about one of their 'childhood memories'. Learners who chose the second task made their two minute speech on 'a big public event' they had participated. This corpus consists of 6151 word tokens and 1062 different word types.

SPSS V 16.0 software was used to sort words according to their word categories after the tagging process which was manually done.

In the analysis of the corpus –determining the most frequent items used in the corpus and the sentences containing these item– AntConc 3.2.4w concordancing software was used.

1.4. Procedure

In the compilation of the corpus, 35 students were asked to give a 2 minute speech on one of the two topics given beforehand. Participants were given five minutes of time to organize their thoughts before making their speech. Each monologue was tape recorded and then transcribed into "text files" that can be processed by the concordancing software used.

After finishing the transcription of the sound files, text files were analyzed via AntConc 3.2.4w concordancing software in order to find out the most frequent 20 words in four word categories: nouns, adjectives, verbs, and adverbs. First, "word list tool" was used to create a list in which frequency of each item in the corpora can be seen clearly. Second, all the words on the list were hand-tagged using numbers according to word categories they belong to as follows: 1- noun, 2- adjective, 3-verb, 4- adverb.

Using SPSS 16.0 software, words in the list were sorted according to the category numbers and the most frequent 20 words in each category were noted down for sentence level analysis. Using AntConc 3.2.4w software, all sentences containing the words chosen in the previous stage were listed and the use of these words was analyzed concerning speakers' "wrong word choice", "use of redundant a word" and also learners' correct uses of "tense" and ''agreement'' were examined for the '20 most common verbs'.

2. Findings

The results of the analysis of the most commonly used 80 words -20 from each four lexical group- are shown in Table 1. It is found out that learners misused 22 different lexical items for 61 times in total. Analysis revealed that verbs, forming 72 percent of total errors, is the most misused word group among the four lexical groups examined in the corpus. 10 different verbs were misused for

45 times. The second lexical category which was mostly misused in learner speech is adverbs. Seven different adverbs were misused by learners in nine different sentences. Nouns were the third group that learners had difficulty in the correct usage. Three items were misused in six different occurrences. The last group items of which were found to be misused is adjectives. Two single adjectives were used in two sentences incorrectly in the learners' monologues.

Table 1: *Number and Percentages of Misused word types*

	Number of Misused word types	**Total misuses**	**Percentages**
Nouns	3	6	9,8
Adjectives	2	2	3,2
Verbs	10	44	72,1
Adverbs	7	9	14,7
Total	22	61	100

The most used 20 verbs that were included in the analysis were: be, do, have, go, feel, come, see, dance, like, take, show, start, want, run, remember, tell, know, attend and get. Participants made mistakes in the use of half of these verbs. The verb 'be', different forms of which '-am, is, are, was, were, be, been- were grouped and considered as one word in the analysis, showed up to be one of two most misused verbs (Table 2). Of all nine wrong uses, five of the misuses of 'be' were about subject-verb agreement, three about 'tense' and one about 'redundant use' of the word. The verb 'feel' having the equal number of misuses with the verb 'be' is also on the top of the list of misused verbs. Five of the mistakes with the verb 'feel' were about 'wrong verb choice', and other mistakes are about 'redundant' use of the verb. It was uncovered that learners who misused the verb 'feel', mostly used the past form of 'fell' which is 'felt' instead of the past form verb 'fall'. These misuses can be seen in Figure 1 below (e.g. lines 10, 11 and 13) which shows a concordance line search screen from AntConc 3.2.4w.

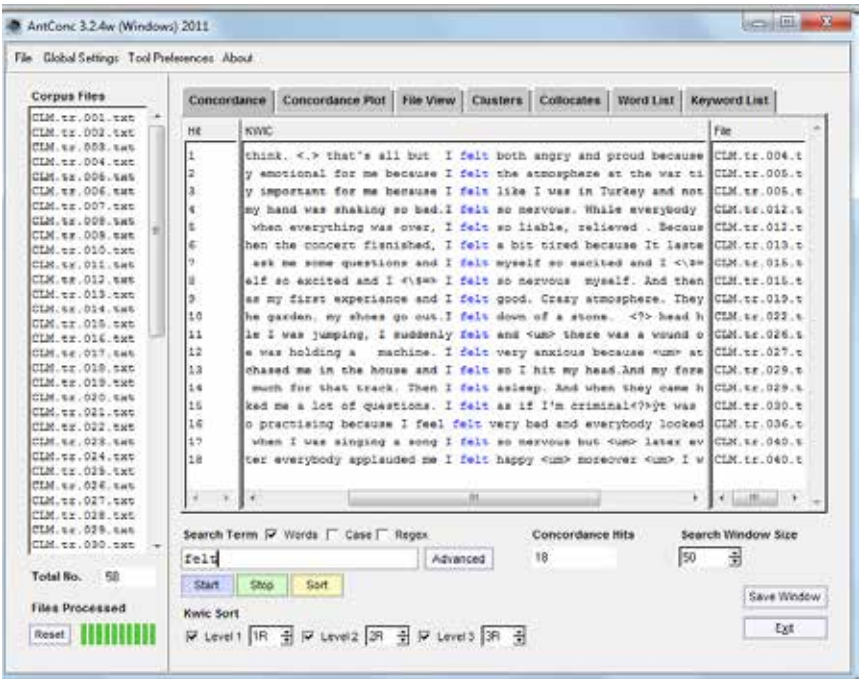

Figure 1. Concordance lines for past tense form of 'feel' and misuses.

Most of the misuses with the other verbs included 'wrong verb choice' and 'tense' errors. The analysis also revealed that there are collocational misuses of the verbs stemming from L1 influence such as the use of the verb 'feel' along with 'myself' which is a collocation of 'feel' in Turkish but not in English. This kind of misuse can be seen in Figure 1 above in lines 7 and 8.

Table 2: *Number and Percentages of Misused Verbs*

Verbs	Misuses	Percentages
To be	9	20
Feel	9	20
Have	7	15,5
Say	6	13,3
Take	5	11,1
Go	4	8,8
See	2	4,4
Do	1	2,2
Come	1	2,2
Start	1	2,2
Total	45	100

The nouns that were included in the analysis are: day(s), school, festival, time(s), friend(s), year(s), people, children, home, song(s), class, match, anniversary, father, mother, experience, teacher(s), arm, music, and sister. Table

3 shows the misused nouns and number of mistakes. It is found out, of 20 most frequent nouns analyzed there made six mistakes in the use of the nouns day, time and home. 66% of these mistakes are related with the 'redundant use of the word' and the rest is related with 'wrong word choice'.

Table 3: *Number and Percentages of Misused Nouns*

Nouns	Misuses	Percentage
day	2	33,3
time	2	33,3
home	2	33,3
Total	6	100

The analysis of the sentences including the most commonly used 20 adverbs in the CLM showed that the students misused seven adverbs in nine different sentences. The adverbs that were included in the analysis are: here, very, then, after, really, much, just, also, such, now, only, too, here, still, suddenly, together, actually, ago, away, especially. Table 4 shows the misused adverbs and the number of misuses of each item. Nine incorrect uses of adverbs include five 'word order', three 'redundant' word use and one 'wrong word choice' related errors.

Table 4: *Number and Percentages of Misused Adverbs*

Adverb	Misuses	Percentage
only	2	22,2
still	2	22,2
there	1	11,1
such	1	11,1
just	1	11,1
together	1	11,1
especially	1	11,1
Total	9	100

The following concordance lines from the CLM show the uses and misuses of adverb 'only' in the CLM corpus. For example in example 1b, there is redundant use of 'only' alongside with 'just' which has a similar meaning.

1a. fileCLM.tr.002: …yeah, that's all. This is **only** children festival and <um> <.> this day has a chi

1b. fileCLM.tr.002: … It's change changable because you <um> you not **only** just traditional but also western <$=> western…

1c. fileCLM.tr.006:…$=>he came and bring me to the emergency. I can't **only** remember I was crying. Then <um> doctor said, mad…

1d. fileCLM.tr.022:…of the day, I was at home. And I can't sleep <um> **only** <um> While playing in the garden. my shoes go out…

1e. fileCLM.tr.022:…buked him. I thought I didn't do bad thing. I just **only** spend time until the school bus came. And when I…

1f. fileCLM.tr.030:…er and I <um> went swimming to the pool <um> not **only** I my <um> cousins and go to there and <um> one d…

1g. fileCLM.tr.50:…and that day we don't <um> gathering mushroom <?> **only** <uhh> found there earing and <um> at the end we c…

Adjectives group contains the least 'misuse errors' among the four lexical categories analyzed in this study. The twenty most frequently used adjectives in the CLM were "all, excited, some, big, good, one, old, afraid, two, every, folk, other, angry, bad, different, first, happy, important, last, right, and crowded". Table 5 shows the descriptive information about the incorrectly used adjectives in the learner speech. One of the errors with the adjectives was related to 'redundant use of the word' while the other was a 'word order' error (e.g. fileCLM.tr.34:…My some relatives and I went to forest for get woods…).

Table 5: *Misuse of adjectives 'some' and 'big'*

Adjectives	Misuses	Percentage
some	1	50
big	1	50
Total	2	100

Figure 2 demonstrates the reasons lying behind the misuses. It is clear that 'wrong word choice' was the most important reason of the lexical errors in learner speech (~%30). 'Redundant word use' (%24) and 'subject verb agreement related errors' (%21) get the second and third places in the list. Tense related errors are responsible for 19 percent of the misuses while 6.5 percent of the misuses were the result of 'word order' related problems.

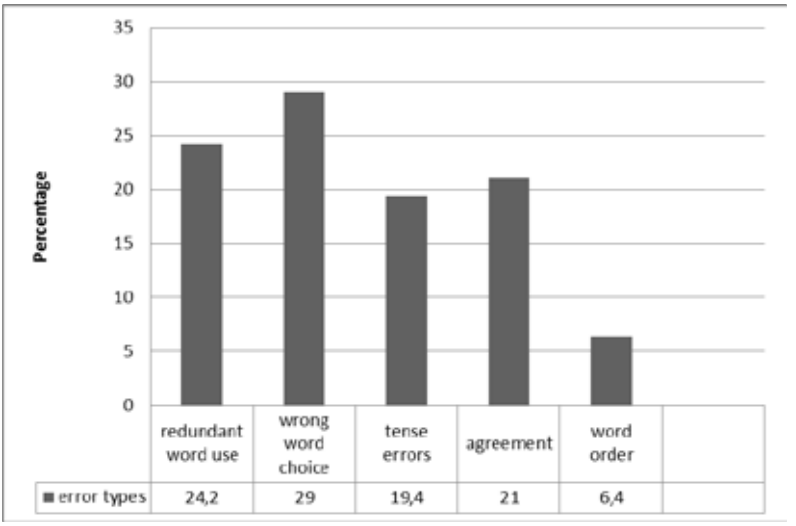

Figure 2. Misuse Categories (%)

Discussion and Results

The major aim of this research was to identify the most common lexical misuses that occur in Turkish elementary and pre-intermediate learners' speech, and

detect and categorize the main reasons underlying these misuses. The findings of the study suggested that students who participated the study made errors mostly in correct use of the verbs. Apart from the verbs, participants made many errors in adverbs' and adjectives' proper use. First crucial reason for these types of errors, especially errors in verbs' use, might be the L1 influence. It was understood that in many verb use- word choice based errors, learners used translation equivalents of the verbs' in their L1 which is Turkish. There are some studies in the literature with similar findings (Hemchua and Schmitt 2006). In their study about lexical errors, Hemchua and Schmitt pointed out that L1 influence is among the reasons of the errors made by Thai learners. Another important reason for these misuses might be learners' inadequate knowledge of the semantic and syntactic properties of verbs. The low proficiency levels of participants might also be added among the reasons listed.

Findings of this research showed that word choice errors came up to be the most important reason for the misuses. One of the main reasons behind word choice errors might be learners' lack of lexical knowledge. It seems that learners had problems with choosing the correct lexical options all the time while speaking. Another reason behind this might be the anxiety factor that might have occurred in the data collection process for the corpus compilation. As the students were called in teacher's room to give their speeches one by one, they might have felt some discomfort and anxiety, which might be accounted as one of the reasons that caused word choice errors.

One of the aims of this research was to answer the question whether a spoken corpora analysis could be put into use in detection of the most problematic word groups and most common word misuses in students' speech. It is clear that providing concordance lines and frequency lists, a corpus and methodology based on corpus make it easy to get to the point in detecting lexical errors.

One of the implications of this study can be on teaching practices in Turkey. As it was shown that students made mistakes while using almost one third of most frequent lexical items, there is a need for improvement in teaching of vocabulary at undergraduate level, especially in teaching of verbs.

Another suggestion that can be made about the corpus use in teaching and error detection. English teachers that are teaching at secondary or tertiary level might benefit from corpus analysis to detect the most common problems among their students with their language use. Even it may seem as a difficult task, the results they could get after such an analysis would provide invaluable contributions in their teaching.

This study was conducted on the data collected from undergraduate students' speech whose proficiency levels of English were ranging from intermediate to upper-intermediate. Further research may be conducted with data from graduate or high school students at different levels of English proficiency.

REFERENCES

Barlow, M. (2002). Corpora, concordancing, and language teaching. Proceedings of the 2002 KAMALL International Conference, Korea:Daejon, 135-141.

Bennett,G.R. (2010*). Using corpora in the language learning classroom: Corpus linguistics for teachers*. Michigan: The University of Michigan Press.

Biber, D. C., Susan & Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Csomay, E. (2013). Lexical Bundles in Discourse Structure: A Corpus-Based Study of Classroom Discourse. Applied *Linguistics*, 34, 369-388.

Francis, N. & Kučera H. (1964). Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English for use with Digital Computers. Providence: Brown University Department of Linguistics.

Gray, B. & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics.* 18, 109-135.

Hardie, A. & McEnery T. (2010). On two traditions in corpus linguistics, and what they have in common. *International Journal of Corpus Linguistics,* 15, 384-394.

Hemchua, S. & Schmitt, N. (2006). An analysis of lexical errors in the English compositions of Thai learners, *Prospect*, 21, 3-25.

Jablonkai, R. (2010). English in the context of European integration: A corpus-driven analysis of lexical bundles in English EU documents. *English for Specific Purposes*, 29, 253-267.

Koester, A. J. (2001). The performance of speech acts in workplace conversations and the teaching of communicative functions. *System,* 30, 167-184.

Leech, G. (1997). Teaching and Language Corpora: A Convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. *Knowles (eds.).Teaching and Language Corpora* (pp. 1-23) Harlow: Addison Wesley Longman.

Llach, M., P., A., (2007). Lexical errors in young EFL learners: How do they relate to proficiency measures, *Interlingüistica*, 17, 63-73.

McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.

Reppen, R. & Richards, J. C. (2014). Towards a Pedagogy of Grammar Instruction, *RELC Journal*, 45, 5-25.

Rizzo, C. (2009). Wireless: Some Facts and Figures from a Corpus-driven Study. *International Journal of English Studies*. 2009 Supplement, 91-114.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tan, M. (2005). Authentic language or language errors? Lessons from a learner corpus. *ELT Journal*, 59, 126-134.

Teubert, W. (2005). My Version of Corpus Linguistics. *International Journal of Corpus Linguistics,* 10, 1-13.

Zhang, Y. (2014). A corpus based analysis of lexical richness of Beijing Mandarin speakers:

variable identification and model construction, *Language Sciences*, 44, 60-69.

THE USE OF SPOKEN LEARNER CORPORA TO DETECT PROBLEMS WITH LEXICAL ACCURACY

Abstract: The purpose of this study is to offer a solution for the easier and more accurate detection of speaking problems, especially with the use of lexical items, through the use of a corpus methodology. Although the compilation of spoken corpora is a difficult and time consuming process, it has a great potential to shed light on the use or misuse of English by foreign language learners. For the purposes of this study, a spoken corpus of learner English was compiled. The participants are students at the intermediate to upper-intermediate proficiency level of English. The corpus used in the study, namely 'Corpus of Learner Monologues' (CLM) consisted of transcriptions of 35 spoken accounts by participating foreign language learners on two different topics. The topics assigned for the compilation of the corpus were chosen from among IELTS (International English Language Testing System) exam speaking section topics. The recordings were transcribed and hand coded for error categories. After the coding of errors, the corpus was analyzed by using AntConc 3.2.4w, a special software for corpus analysis. The results revealed that verbs are the most problematic lexical group in students' speech. Other common errors included adverb, noun, and adjective word groups. Results also highlighted that "wrong word choice" was the most common reason for students' lexical errors. The study carries implications for the teaching of speaking skills.

Keywords: Corpus, Spoken Learner Corpus, Lexical Error Analysis.