Eğitim ve Bilim
2013, Cilt 38, Sayı 167

Education and Science
2013, Vol. 38, No 167

Differences in English Vocabulary Use: Insights from Spoken Learner Corpus and Native Speaker Corpus

İngilizce Sözcük Kullanımında Ayrımlar: Öğrenci Bütüncesi ile Anadil Konuşmacısı Bütüncesinden Elde Edilen Bilgiler

Bilal GENÇ*

İnönü Üniversitesi

Abstract

Comparisons of native and non-native speakers' written and spoken productions enable researchers to investigate the points of divergence and parallelism between the two types of texts. Focusing on the use of two functional categories (articles and prepositions) and three content categories (nouns, adjectives and verbs), this study compares a small, spoken NNS corpus (10973 words) with a smaller, spoken NS corpus (2331 words). Participants in both groups were assigned a speaking task, the result of which produced the NNS and NS corpora and those corpora were analyzed quantitatively with the help of computer software titled *Concordance*. The results show that due to the limited number of choice, there was a large overlap in the use of articles and prepositions; close similarity between the use of nouns and verbs; and a remarkable difference in the use of adjectives.

Keywords: Vocabulary, native speaker, non-native speaker, spoken language, corpus.

Öz

Bir dili anadil olarak konuşan ve o dili anadil olarak konuşmayanların yazılı ve sözlü üretimlerinin karşılaştırılması, araştırmacılara her iki metin türü arasındaki benzerlik ve ayrım noktalarını araştırma olanağı sunmuştur. Bu çalışma, küçük bir Anadil Bütüncesi (10973 sözcük) ile daha küçük bir Anadilden Olmayan Bütünce (2331) karşılaştırmasını yapar: Her iki gruptaki katılımcılara bir konuşma ödevi verilmiş ve bunun sonucunda üretilen Anadil Bütüncesi ile Anadilden Olmayan Bütünceleri *Concordance* adlı bilgisayar yazılımı yardımıyla nicel açıdan analiz edilmiştir. Çalışma iki işlevsel kategori ile (tanımlık ve ilgeç) üç içerik kategorisine (adlar, önadlar ve fiiller) odaklanır. Sonuçlara göre konuşmacıların sözcük seçimlerindeki sınırlı tercihleri nedeniyle tanımlık ve ilgeç kullanımında büyük oranda bir örtüşme varken ad ve fiil kullanımında da yakın benzerlik görülmüştür. Önad kullanımında ise iki grup arasında kayda değer bir ayrım görülmüştür.

Anahtar Sözcükler: Sözcük, anadil konuşmacısı, anadilden olmayan konuşmacı, sözlü dil bütüncesi.

Introduction

Discourse analysis is the analysis of language in use (Brown & Yule, 1983) and corpus linguistic studies are generally considered to be a type of discourse analysis because they, too, describe the use of linguistic forms in context (Biber, Connor & Upton, 2007). Having

* Yard.Doç.Dr Bilal GENÇ, İnönü Üniversitesi Eğitim Fakültesi, Yabancı Diller Eğitimi Bölümü, billgenc@gmail.com, bilal.genc@inonu.edu.tr

explained the scope of corpus linguistics, we believe it is essential to clarify the nature of "corpus". Corpus is defined as a machine-readable collection of (spoken or written) texts that were produced in a natural communicative setting and the collection of texts is compiled with the intention to be representative (Gries, 2009a). To Gries (2009) "machine readable" means that the corpus is "stored in the form of plain ASCII or Unicode text files that can be loaded, manipulated, and processed platform-independently" (p. 7) and corpora's being produced in a natural communicative setting "means that the texts were spoken or written for some authentic communicative purpose, but not for the purpose of putting them into a corpus" (p. 8).

The "representativeness" of the corpus implies that the bigger the corpus the better it is for linguistic analysis. Ragan (2001) puts forward two reservations against this belief:

- The notion that bigger is better with regard to corpus size derives from the mistaken belief of much early corpus linguistics that all or somehow enough of the sentences of natural language could be collected to give a full picture of language use.
- The increasing reliance by language researchers on quantification and significance testing of findings based on language data contributes to a misplaced anxiety over the validity and reliability of intuitive deductions based on small samples of language.

As mentioned above, corpus might consist of either spoken or written texts; both types of the texts have their advantages and disadvantages over the other. Regarding the advantages of spoken corpus, Vizcaíno (2007) suggests that main advantages of using oral corpora for research are:

- First, they can represent a wide range of genres. Therefore, they are suitable for studying spoken language in diverse communicative contexts. For instance, in this study, the participants were assigned a task which involved watching an animated movie and then commenting on various aspects of it.
- Second, these corpora contain prosodic information. This information could further be investigated through some software specifically designed for sound analysis.

Regarding the challenges of spoken corpus, Meyer (2004) argues that speech is the primary mode of human communication. However, the logistical difficulties involved in recording speech and collecting data for the spoken part of a corpus is much more difficult and involved than collecting written samples. Corpus linguistics is often associated with frequencies, quantification which implies that corpus linguistics has a very narrow application and the data gathered from corpus studies are just figures. Against this misevaluation, Gries (2009b) argues that from the data gathered from corpus studies, one can investigate a wide array of issues such as one's first language acquisition, second/foreign language acquisition, language and culture, historical developments, phonology, morphology, syntax, semantics, and pragmatics.

In a study on first language acquisition in native English-speaking children, Souter (2002) examined the Polytechnic of Wales Corpus collected in the late 1970s and investigated such issues as the rate of vocabulary growth with age in that corpus; the extent to which vocabulary is sex-specific; differences between sexes in the use of affirmatives and negatives, and in the use of male and female personal pronouns; the extent to which vocabulary size is related to socio-economic class; persistence of errors in applying regular verb endings to irregular verbs.

A most recent study by Grant (2010) investigates the use of *I don't know*/ *I dunno* phrases between the native speakers of English from Britain and New Zealand. The researcher found that with regard to both the full and the truncated phrase, British speakers use the phrase with different frequency and for different reasons than New Zealand speakers. Thus, Grant's study reveals how corpus studies help us to observe the differences between two nations having the same mother tongue.

When one is interested in the relationship between second/foreign language acquisition and corpus studies, his/her attention is focused on what is termed as learner corpora (LC). Learner corpora are electronic collections of foreign or second language learner texts assembled according to explicit design criteria. The fact that they contain data from language

learners makes them a very special type of corpus, requiring from the analyst a wider range of expertise than is necessary for native corpora (Granger, 2009). The comparison of learner corpus with NS corpus which serves as a baseline data enables the researchers in assessing second language learners' lexical proficiency and determining in linguistic elements to focus on in instruction.

In a study, for instance, comparing a learner corpus with and established NS corpus, Shirato and Stapleton (2007) investigated the extent of Japanese learners' of English deviance from the target language norms. They found that Japanese speakers of English differed from the NSs in many areas such as in their underuse of discourse markers, model items and interactive words, delexical verbs, terms for marking vagueness and hedges, and in overuse of high frequency and auxiliary verbs and common adjectives.

In addition to broadening our insight in the above mentioned areas, corpus linguistics also gives insight about cognitive development of the learners. In a study focusing on the cognition-corpus relationship, by investigating the case of verb "give" in the corpus, Mukherjee (2002) suggested that although "give" triggers a ditransitive structure and although in some examples exemplifying the use of verb "give" indirect object is omitted, "give" is still to be regarded as a "ditransitive verb in all its occurrences from a cognitive-semantic point of view because it is bound to evoke an event type which includes three argument roles, even though some 'implicit' argument roles may not be explicitised" (p. 93).

The basic motive for this study is that we agree with the statement made by Beaugrande (2001), who argues that "the uses of corpora are surely most urgent for non-native speakers who have not had extensive exposure to fluent English. Our major problem is so much not *bad English* or *incorrect English*, as is often lamented, but rather *insufficient English*" (p. 10).

Flowerdew (2001) suggests that in the field of corpus linguistics in 1990s, two researchers either investigated such large corpora as Cobuild Corpus with 300 million words and British National Corpus with 100 million words or they focused on smaller corpora which were compiled for various academic purposes. With a small learner corpus and a smaller native corpus, this study aims to find out similarities and differences in the speech of non-native speakers and native speakers who were assigned the same speaking task.

Research Questions

In this study we tried to find answers to the following questions:

- To what extent do the speeches of Turkish speakers of English exhibit native-like qualities in terms using function words?
- To what extent do the speeches of Turkish speakers of English exhibit native-like qualities in terms using content words?
- What differences occur in the frequency and type of function and content word use between native speaker group and Turkish speakers of English?

In order to display the differences in the use of function words, we focused on the most used prepositions and articles. Regarding the use of content words, we focused on the use of three categories of content words: nouns, adjectives and verbs.

Data Collection

This study was carried out with the participation of both Turkish students majoring in English at the ELT Department, University of Cukurova, and four native speakers of English. In order to elicit oral narratives, the participants were first asked to watch a movie titled *Mickey's Christmas Carol* based on the novel *Christmas Carol* by the famous English novelist Charles Dickens, then they were asked to provide an oral comment of the film.

Being interviewed individually, the Turkish participants, treated as the non-native speaker group (NNS), were invited to the researcher's office, where their narratives were

audio-recorded and then transcribed on a personal computer. The four native speakers of English and the native speaker group (NS), were also asked to present their narratives in the researchers' office, which were also audio-recorded and then transcribed on the personal computer. Each interview session lasted 5-7 minutes.

Participants

The participants in this study are categorized into two groups: (1) Turkish 1st-year students of English studying at the ELT Department, University of Cukurova (NNS), all being trained to be prospective English teachers (n=30; 22 females, 8 males; 21-25 years of age range). Since they already are students of the ELT Department, they were all enrolled in an obligatory two-semester speaking class. The other group consists of four native speakers of English (NS), all females, of 20-23 age range, all had an undergraduate degree.

Instrument

The main motive behind the selection of *Mickey's Christmas Carol* was the relatively small number of characters in the film (Ebenezer Scrooge, Bob Cratchit –Scrooge's overworked employee- and the ghosts, being the major persona), and the relatively less complex plot of the novel made it easy for the participants, especially for the NNSs, to refer to the persons, things, and events in the film. The cartoon is a twenty-four minute animated short film produced by Walt Disney Productions and originally released in the United Kingdom in the October of 1983.

Data Analysis

The recordings of NNS and NS were grouped as separate entities; therefore, two units of recordings were investigated in this study. All the units were transcribed into standard orthography for analysis. We included very short isolated utterances within longer conversation units and omitted unintelligible utterances from the transcription. Backchannels such as *hmm* and *er*, were also included into the data and treated as full words. Contracted forms of auxiliaries were counted as a single word. When all the units were transcribed, the teachers' utterances put in brackets were extracted from the data. As a result, only utterances spoken by students and four native speakers were analyzed through analytical software titled *Concordance*. The term concordance means the lists of the occurrences of a given word or phrase in a corpus.

As the title of the software suggests, *Concordance* is capable of making indexes and word lists, count word frequencies, compare different usage of a word, analyse key words, find phrases and idioms[*]. In order to extract recurring words, we generated the rank-order frequency lists of single word and lemmatised single word sequences. The next step was to determine the most commonly employed prepositions, articles, nouns, adjectives and verbs and to compare the use of those items in the NNS corpus and NS corpus.

Results and Discussion

As the first step of our analyses, we summarized the counts for the NNS and NS corpora in Table 1. Using the *Concordance* software, we found 2331 token and 521 types in NS corpus and 10973 token and 1209 types in NNS corpus. The software also provided us the counts of lemmata –lemma is defined by Francis and Kucera (1982: 1) as a 'set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling'.

[*] For more information the reader is referred to http://www.concordancesoftware.co.uk/

Table 1.

Tokens, types and lemmas of the NS and NNS corpora

	NS	NNS
Tokens	2331	10973
Types	521	1209
Lemmas	272	312

As can be inferred from Table 1, while the average token count in the NS group was around 580, the members of NNS group's token count was 360 on average. Given that it took 5-7 minutes for each participant to complete the task, we see that NSs are far more fluent speakers than the NNSs. When we have a look at the types, we see that the NNSs used almost twice the types NSs used. However, there was only 40 lemmata difference between the two groups' speeches. In the following table we analyzed the use of 36 lemmata (2 articles, 9 prepositions, 8 nouns, 8 adjectives and 11 verbs) in the native and non-native corpora.

Results regarding the use of articles and prepositions by NNS and NS are given in Table 2 in which the use of the definite article *the*, indefinite articles *a*, *an*, and 7 most frequently appearing prepositions are presented.

Table 2.

The most frequent determiners/prepositions in NS and NNS corpora

	NS		NNS	
Preposition	Frequency	%	Frequency	%
the	111	4.76	477	4.34
a/an	35	1.50	299	2.73
to	83	3.56	308	2.80
in	29	1.24	211	1.92
of	49	2.10	140	1.28
for	14	0.60	95	0.87
with	32	1.37	60	0.55
about	19	0.82	48	0.44
from	8	0.34	44	0.40
Total	380	16.29	1682	15.33

As shown in Table 2, the use of the definite article, indefinite articles and seven of the most used prepositions make up the over 15% of the tokens in each corpus. In terms of the use of the definite article, NNSs and NSs reveal very close patterns, which is a strong indicator of foreign language proficiency of NNS whose mother tongue does not have any definite article. In the same vein, the parallelism between the use of the indefinite articles and seven most widely used prepositions implies that NNSs reveal native like qualities in their speech at least in terms of the use of function words.

The use of prepositions, unlike the use of nouns, adjectives and verbs -which are the other categories analyzed in this study- displays peculiar characteristics as given in Table 2. It is the only table in which all the items listed in the table enjoy the same ranking in native and non-native corpora.

Regarding the use of the definite article and indefinite articles, it was observed that in the NNS corpus there are relatively more repetitions which involve the use of in/definite article and the noun following it. When have a an overview of the narratives and dwell on the closeness of figures related with the use other preposition, it is seen that both native and native speakers' narratives focussed on the same characters, locations, and events, which was expected due to the simple plot of the animation movie.

In the next phase of our analyses, the use of content words including nouns, adjectives and verbs were investigated separately. To see the extent of use of nouns, adjective and verbs, the items displayed in the tables were compared with the 3000 words used to write the definitions in the *Oxford Advanced Learner's Dictionary 7th Edition*.

In Table 3, the results of the analyses on the use of nouns in NNS and NS corpora are presented.

Table 3.

The most frequently used nouns in NS and NNS corpora

	NS			NNS	
Noun	Frequency	%	Noun	Frequency	%
money	25	1.07	*money*	195	1.78
Christmas	23	0.99	*people*	74	0.67
people	20	0.86	*Christmas*	71	0.65
ghost	13	0.56	*house*	53	0.48
family	13	0.56	*day*	53	0.48
story	12	0.51	*film*	52	0.47
future	11	0.47	*life*	33	0.30
things	9	0.39	*assistant*	33	0.30
Total	126	5.41	Total	564	5.13

When we have look at Table 3, we see that both groups of speakers mostly used *money*, *people*, and *Christmas* in their speech. Regarding other nouns, we see that there are no common nouns among the mostly preferred ones in the speeches of both groups. Although the speeches of the NSs focused on *ghost*, *family*, *story*, *future* and *things*, NNSs focussed on *house*, *day*, *film*, *life* and *assistant*. It is quite probable that the speakers of both groups referred to the movie when they were talking about the *film* or the *story*.

Language learners are expected to employ vague language use more than native speakers do. Having a look at Table 3, however, we see that "thing" is among the most frequently used nouns in the NNs' corpus. When we look up the words appearing in Table 3 in Oxford Dictionary's list we see that *Christmas*, and *ghost* do not exist in the Oxford's 3000 core vocabulary list.

Following the analysis of the nouns, the use of adjectives in the speeches of NSs and NNSs were investigated. The descriptive statistics values of the most frequently occurring adjectives in the speech of both groups are presented in Table 4.

Table 4.

The most frequent adjectives in NS and NNS corpora

	NS			NNS	
Adjective	Frequency	%	Adjective	Frequency	%
tiny	6	0.26	*happy*	62	0.57
main	6	0.26	*poor*	58	0.53
important	5	0.21	*important*	30	0.27
little	4	0.17	*mean*	22	0.20
cute	4	0.17	*rich*	21	0.19
small	3	0.13	*good*	16	0.15
poor	3	0.13	*little*	13	0.12
obsessed	3	0.13	*small*	12	0.11
Total	34	1.46	Total	234	2.14

Compared with functional categories and other content categories analyzed in this study, adjectives are the least employed word types. With percentages 1.46% by NSs and 2.14% by NNSs, the use of the most frequently occurring adjectives occupies a small portion among other items.

Besides this, the most remarkable discrepancy between the speech of NSs and NNSs is the

adjective they employed. There is no overlap in the most used two adjectives by the two groups; *important*, *poor*, *small* and *little* are the four adjectives used by both groups. Thus, out of the eight most employed adjectives, *tiny*, *main*, *cute*, and *obsessed* occurred only in the speech NS group and *happy*, *mean*, *rich*, *good* were observed only in the speech of NNS group. While the NNS group described Donald Duck's love for money with the adjective "mean", the NSs described his character with the adjective "obsessed". Furthermore, these two adjectives and *cute* are not found in Oxford's 3000 words list –*mean* is listed only a as verb.

The last content category analyzed in the study was verbs: the most occurring 11 verbs and the frequencies and percentage values are given in Table 5.

Table 5.

The most frequent verbs in NS and NNS corpora

	NS			NNS	
Verb	Frequency	%	Verb	Frequency	%
be	69	2.96	*be*	502	4.57
see	29	1.24	*have*	110	1.00
have	16	0.69	*want*	36	0.33
go	8	0.34	*think*	36	0.33
love	7	0.30	*love*	28	0.26
show	7	0.30	*talk*	27	0.25
come	6	0.26	*take*	25	0.23
change	6	0.26	*come*	25	0.23
give	6	0.26	*see*	21	0.19
say	5	0.21	*know*	20	0.18
laugh	4	0.17	*help*	20	0.18
Total	163	6.99	Total	850	7.75

When we have a look at Table 5, we see that most frequently used verbs make up almost 7% of the native corpus and 8% of the non-native corpus. The use of verbs occupies a larger portion in the corpora than the use of noun or adjectives. While NNSs relied heavily upon *be* and *have* in their account of the movie, NSs preferred *be* and *see* in their comments on the cartoon movie. Of the 11 verbs, *be*, *see*, *have*, *love*, *come* were used by both groups of speakers. In addition to these verbs the comments of the NSs were conveyed thorough *go*, *show*, *change*, *give*, *say*, *laugh* and those of the NNSs were conveyed through *want*, *think*, *talk*, *take*, *know* and *help*.

When we have a look at sentences displaying the use of *be*, we see that both groups employed *be* as auxiliary only few times; as seen from the table the use of "be" accounted for the 2.96 % of the total types in NS corpus and 4.57% in the NNS corpus. Only a few of the nouns and adjectives are not found on Oxford Dictionary's list; however, all the verbs appearing in the table could be found in the list.

Overall, the items selected for analysis in this study (the 36 lemmata analyzed in the study) make up nearly 30% of the total tokens in both corpora. *Money*, *Christmas* and *people* were the most used nouns and *be*, and *have* are the most commonly used verbs in both groups; only *important* is one of the most frequently used adjectives in both groups. Thus, it seems that there is a great overlap in the use of articles, prepositions, nouns and verbs by NNSs and NSs; however, the use of adjective exhibits relatively less similarity.

Consequently, Turkish speakers of English participated in this study revealed native-like qualities in terms of the use of word types analyzed in this study. As the average word counts reveals, however, their speech patterns are far from native-like in terms of fluency. The use of multi-word expressions, discourse markers and hedges also seem possible areas of discrepancy between native and non-native speakers.

Conclusion

In this study, to investigate the differences in vocabulary use between native and non-native speakers, both quantitative analyses of corpus-based data were performed. Our overall findings show that the speeches of NSs and NNSs in this study, in large part, display similarities in terms of the categories investigated in the study. The results obtained lead us to the conclusion that the current proficiency of Turkish NNSs in vocabulary does not display a remarkable deviance from that of native speakers.

From this data, it is suggested that regarding the one word level NNS and NS speech display close similarities; the use of multi-word expressions, discourse markers and hedges needs further investigation. The lack of analyses on the use of multi-word expressions, discourse markers and hedges, thus, is one of the limitations of this study.

Finally, we believe that the teachers might provide the results of analyses of NNS and NS corpora for the students for their self study, or the teacher might discuss the results with the students in the classroom to discuss the strengths and weaknesses of the students in speaking the target language.

References

Beaugrande, R. de (2001). Large corpora, small corpora, and the learning of "language". In M. Ghadessy, A. Henry, R. L. Roseberry (Eds.) *Small corpus studies and ELT* (pp.3-28). Amsterdam: John Benjamins.

Brown, G., & G. Yule. (1983). *Discourse analysis*. Cambridge: Cambridge University Press.

Biber, D., Connor, U., & Upton, T. (2007). Discourse analysis and corpus linguistics. In D. Biber, U. Connor and T. Upton (Eds.) *Discourse on the move: Using corpus analysis to describe discourse structure* (pp.1-20). Amsterdam: John Benjamins.

Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, A. Henry, R. L. Roseberry (Eds.) *Small corpus studies and ELT* (pp. 363-379). Amsterdam: John Benjamins.

Francis, W.N., & Kucera, H. (1982). *Frequency analysis of English Usage: Lexicon and grammar*. Boston, MA: Houghton Mifflin.

Granger, S. (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: John Benjamins.

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.

Gries, S. T. (2009a). *Quantitative coprus linguistics with r: A practical introduction*. London: Routledge.

Gries, S. T. (2009b). What is corpus linguistics? *Language and Linguistics Compass, 3*, 1–17.

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

Mukherjee, J. (2002). Corpus data in a usage-based cognitive grammar. In K. Aijmer & B. Altenberg (Eds.) *Advances in corpus linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23) Göteborg 22-26 May 2002*. (pp.85-100). Amsterdam: Rodopi.

Oxford Dictionary (7th ed.). (2000). Hinsdale, IL: Penguin Press.

Ragan, P. H. (2001). Classroom use of a systemic functional small learner corpus. In M. Ghadessy, A. Henry, R. L. Roseberry (Eds.) *Small corpus studies and ELT* (pp.207-236). Amsterdam: John Benjamins.

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.

Souter, C. (2002). Aspects of spoken vocabulary development in the Polytechnic of Wales Corpus of Children's English. In K. Aijmer & B. Altenberg (Eds.) *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23) Göteborg 22-26 May 2002*. (pp.279-296). Amsterdam: Rodopi

Vizcaíno, M. J. G. (2007). Using oral corpora in contrastive studies of linguistic politeness. In E. Fitzpatrick (Ed.) *Corpus linguistics beyond the word: Corpus research from phrase to discourse*. (pp. 117-142). Amsterdam: Rodopi