The comparison of collocation use by Turkish and Asian learners of English: the case of TCSE corpus and icnale corpus

Elif Tokdemir Demirle, Semin Kazazoğlu

Abstract

This purpose of this study is to compare the use of collocations by Turkish learners of English and Asian learners of English. Two spoken learner corpora, namely the TC-SLE corpus (Turkish Corpus of Spoken Learner English) and ICNALE corpus (International Corpus Network of Asian Learners of English) are used in order to make comparisons. The TC-SLE consists of Turkish learners’ spoken English in the form of classroom language, monologues and language used during group work and has been compiled by the researcher. The comparisons are based on the frequency counts of collocations, types of collocations commonly used by the two different groups of learners and the types of inaccuracies done by the two different groups of learners. Based on the comparisons, the study investigates whether first language background has an important effect on the use of collocations and whether different groups of learners have similar problems in the use of English collocations. It was inferred that collocation use is among the most problematic aspects of spoken English for language learners and that factors not limited to but including first language significantly affect the use of collocations.

1. Introduction

Collocations are defined as “The occurrence of two or more words within a short space of each other in a text” (Sinclair, 1991, p. 170). The accurate use of collocations carry importance for spoken language since they make speech more fluent and native-like. Therefore, mastery of collocations is an important aspect of spoken proficiency. Cowie (1981) distinguishes four types of combinations: free combinations (e.g. drink tea), restricted collocations (e.g. perform a task) and figurative idioms (e.g. do a U-turn).
Learner speech which lacks collocation use and other aspects common in native speech usually sounds “blunt, unnatural or monotonous” (Shirato & Stapleton, 2007). As language learners are not fully capable of using the target language culturally, socially, and situationally appropriate ways (Fung & Carter, 2007). It is also commonly accepted in the Turkish EFL context that there is lack of proficiency and fluency in speaking skills of learners. One reason for this lack of proficiency is that pragmatic sides of a language are always disregarded in pedagogic settings. There is a need for more thorough investigation of learner speech in Turkey which makes use of current technology and innovations in research. One of the approaches which provide opportunities over traditional research methodologies is corpus linguistics. In Turkey currently, a spoken corpus of learner English, which would provide a valuable resource for research into spoken performances of Turkish EFL learners, does not exist.

Spoken corpora studies, however are not very recent in other countries. For example, starting in the mid-1970s, spoken language corpora came into existence in increasing quantity and variety (Leech, 2000).

Considering that it is an important first step to compile a spoken corpus of Turkish EFL learners, for the current study a corpus of EFL learner English was compiled. The purpose of this study is to compare the use of collocations by Turkish learners of English and Asian learners of English. Two spoken learner corpora, namely the TC-SLE corpus (Turkish Corpus of Spoken Learner English) and ICNALE corpus (International Corpus Network of Asian Learners of English) are used in order to make comparisons. Based on the comparisons, the study investigates whether first language background has an important effect on the use of collocations and whether different groups of learners have similar problems in the use of English collocations.

A corpus can be defined as a large, systematic and structured collection of natural texts in an electronic database (Biber, Conrad & Reppen, 1998). A corpus is comprised with written texts and transcribed speech samples, and they are core elements for linguistic analysis (Kennedy, 1998). The design, size and the nature of a corpus are influenced by the purpose for compiling it. Leech (1991) suggests that a corpus is generally designed for a particular representative function and that he defined corpus linguistics as the study of language and a method of linguistic analysis which relies heavily on the use of computerized texts of authentic language.

Compilation of a corpus for language research and using corpus linguistics methods for analyzing language provides several benefits over other traditional forms of language analysis. With the use of corpus linguistics methods it is possible to analyze complex association between samples of language use and to store huge database of language; furthermore, it provides reliable and unchanging analyses. Corpus linguistics also allows researchers to make empirical studies in many areas of linguistics such as, lexicography, sociolinguistics, studies of style and educational linguistics (Biber, Conrad & Reppen, 1998).

First generation corpora included The Brown Corpus (approximately 1,014,300 words), The Lancaster-Oslo/Bergen(LOB) Corpus (one-million words) and The London-Lund Corpus(LLC) (500,000 words of spoken British English). Other types of corpora compiled for studying spoken English include Lancaster/IBM Spoken English Corpus (SEC) (52,600 words of spoken standard British English), Corpus of Spoken American English (CSAE) (up to one million words) and Wellington Corpus of Spoken New Zealand English (one million words).
Until mid-1990s, the most widely known spoken corpus was LLC (London-Lund Corpus) with half million words. In his article “Spoken Corpora Design”, Čermák (2009) suggested basic design criteria for spoken corpora. His design consisting of three types of categories which also has a number of parameters is illustrated in Figure 1 above.

Linguistics studies using a spoken corpus have investigated various research questions such as the factors creating individual differences in the utterances of native and non-native speakers (e.g. Fuller, 2003; Fung and Carter, 2007; Aijmer, 2011) or regional differences in informal speech. (e.g. Torgersen, Gabriellatos, Hoffman and Fox, 2010). Authentic data provided by a carefully compiled spoken corpus provides important insights into the speech of native and non-native speakers since spoken language is different from written language and is more difficult and time-consuming to capture.

2. Methods

1.2. Corpus Compilation

The corpus compiled for the present study consists of, 58 spoken learner speech samples on two subjects chosen from among the IELTS spoken test subjects. The students whose voices were recorded were preparatory class students at the English Language and Literature department of a public university in Turkey. Their English level is between intermediate and advanced level. The corpus was compiled with the help of a digital voice recorder. Each student’s recording was made he/she wanted to talk about, was allotted some time to take notes and prepare themselves for the talk. When he/she was ready, the student was instructed to talk about their selected subject for 2 minutes. After the recordings were made, they were transcribed and coded for spoken discourse issues such as hesitations and false starts and given an identification number. The topics given to the students were one of the following:

- Talk about a memorable childhood experience of yours
- Talk about a big public event that you have attended

The Turkish Corpus was labeled TC-SLE (Turkish Corpus of Spoken learner English). This corpus consists of approximately 1500 words at present but work on compiling is ongoing. In order to make comparisons between Turkish EFL learners and other groups of learners and native speakers, two corpora which were available online were selected as comparison corpora. For learner comparisons the ICNALE corpus (Ishikawa, 2010) was selected. In order to make the ICNALE and TC-SLE comparable, a section of the ICNALE was randomly selected for the study. The size of the selected section is 1533 words and consists of a blend of Asian learners: Japanese, Indonesian, Chinese. Similar to the TC-SLE corpus, the ICNALE corpus selected consists of 2 minute short recorded and

Figure 1. Design criteria for spoken corpora
transcribed talks. In order to make comparisons with native-speaker speech the spoken section of the BNC (British national Corpus) was used. Table 1 provides information on the corpora used in the study.

| Corpus size (words) | TC-SLE | ICNALE | BNC          |
|---------------------|--------|--------|--------------|
| content             | 1500   | 1533   | 100,000,000  |
|                     | 2-minute recorded EFL learner speech samples | 2-minute recorded EFL learner speech samples | Spoken native speaker English from various contexts |
| reference           | Demirel (2014), Karadeniz Technical University, Turkey | Ishikawa (2010), Kobe University, Japan | Brigham Young University (1980) |

1.3. Data Analysis

For the analysis of the corpus data for TC-SLE corpus and the ICNALE corpus, AntConc concordance program was used. The steps of the data analysis is listed below:

- Step 1: Most frequently used to words in the following categories were chosen from the TC-SLE by using AntConc 3.2.4.w concordancing program (Anthony, 2011). Figure 2 shows the lists of most frequently used words in three part of speech categories of nouns, adjectives and verbs.

![Figure 2. Most frequent nouns, adjectives and verbs in the TC-SLE Corpus](image)

- Step 2: Collocations of the words in the three word categories were found in the three corpora: TC-SLE, ICNALE and BNC

- Step 3: The collocations and their MI (Mutual Information) values through BNC and AntConc and compared across the three corpora. The MI value is a measure of collocational strength. The higher the MI score, the stronger the link is between two items. An MI score of 3.0 or higher to be taken as evidence that two items are collocates. The closer to 0 the MI score gets, the more likely it is that the two items co-occur by chance. A negative MI score indicates that the two items tend to shun each other. The MI scores were recorded for the analyzed collocations to see whether there are differences between the three corpora in terms of collocational strength.
Step 4: The similarities and differences of the collocations were explored.

3. Results and Discussion

The purpose of the study was to compare the collocations of the most frequent nouns, adjectives and verbs across three corpora; TC-SLE (the main non-native EFL corpus), the ICNALE (the EFL corpus for comparison) and the BNC (the native-speaker corpus for comparison) to see whether there is a variation in collocation use between non-native EFL learners and native speakers.

As a result of the analyses differences were observed between the preferred collocates by learners and native speakers. For example, for the selected ten most frequent verbs, a total of 57 collocates were found. Out of these 57 collocates, 23 collocations were preferred only by native speakers and the remaining 24 were preferred by either only Turkish learners or by only Asian learners and sometimes by both. Table 2 shows the preferences for collocations with the past tense verbs went, started and attended.

| Table 2. Preferences for verb collocations |
|------------------------------------------|
| Native speaker preferences | Asian learner preferences | Turkish learner preferences |
| went into | went to | went to |
| went out | went into |
| went down | |
| went back | |
| went round | |
| first started | started to | started to |
| started talking | I started | I started |
| started laughing | |
| people attended | apparently attended | attended conference |
| attended by members attended | | attended |

Secondly it was observed that the collocations showed more similarity between the two groups of learners than they did between learners and native speakers. For example, if we look at the collocates of “good” in the adjectives category. We can see that three words come out as preferred collocations of good in the EFL learner corpora, whereas they are not among the preferred collocates by native speakers.

| Table 3. MI score comparison of collocates of “good” across three corpora |
|-----------------------------|
| MI scores                  |
| adjectives                 | TC-SLE | ICNALE | BNC |
| a good                     | 4,14   | 6,65   | 0   |
| good and                   | 1,15   | 2,15   | 0   |
| good for                   | 3,82   | 7,27   | 0   |
| good relations             | 0      | -2     | 0   |
| good actor                 | 0      | 0      | 4,9 |
| good afternoon             | 0      | 0      | 3,85|
| awfully good               | 0      | 0      | 3,56|
| good boy                   | 0      | 0      | 5,47|

As can be seen in Table 3, collocates with strong MI values for both EFL groups do not show up as strong collocates in the native speaker corpus. This is an indication of the fact that EFL learners do not have enough competence regarding commonly used collocations in native speaker speech and tend to use free combinations of words instead of preferred chunks of speech which make speech more native-like and fluent. Figure 3 also illustrates that there is similarity between EFL groups and difference between the native speaker group and the EFL groups in terms of collocation use.
Another important finding is about the formation of collocates by EFL learners. The collocates preferred by learners are formed by usually core words such as articles, common prepositions or connectors. However, collocations formed with these words did not show up as strong collocates in the native speaker corpus. This result is in harmony with earlier findings by other researchers such as Altenberg & Granger (2001) and De Cock, Granger, Leech, & McEnery (1998). Similarly, Hasselgren (1994) argued that infelicitous collocations resulted from overdependence on the familiar—that is, structures that learners learned early, used widely, and with which they felt comfortable. She referred to these as “lexical teddy bears.”

4. Conclusions

Three main findings were reached as a result of the study. Firstly, it was found that preferences for collocates of the most common words show variation across native speakers and EFL learners. That is, EFL learners use vocabulary differently compared to other EFL learner groups and native speakers. This could mean that collocation use is problematic for learners regardless of their country of origin and that Turkish learners and Asian learners display similar use of collocations in spoken production when compared to native speakers. Secondly, it was seen that most of the variation in collocation use is existent between native speakers and EFL learners but not between the two groups of EFL learners.

Upon finding similar problems with advanced learners Laufer and Waldman (2011) suggest awareness raising activities for the use of collocations. For example supplementing communicative, task-based teaching with preplanned Focus on Form or Focus on Forms activities. With the current availability of online resources and user-friendly interfaces of corpora, teachers can also incorporate the exploration of corpora for collocations into their teaching. In order to get accustomed with the correct use of collocations, learners should be given more opportunities to explore native-speaker language through the use of online corpora. A hands on Data Driven Learning Approach (Johns, 2002) to teaching vocabulary should be followed in order to raise students’ awareness of collocations and also to enrich their repertoire of vocabulary.
References
Altenberg, B. & Granger, S. (2001) The grammatical and lexical patterning of MAKE in Native and non-native student writing. *Applied Linguistics*, 22(2), 173-94
Anthony, L. (2011). AntConc (Version 3.2.1)[Computer Software]. Tokyo, Japan: Waseda University.
Biber, D., Conrad S., and Reppen R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
Čermák, F. (2009). Spoken corpora design: Their constitutive parameters. *International Journal of Corpus Linguistics*, 14(1), 113-123.
Cowie, A. P. (1981). The treatment of collocations and idioms in learners’ dictionaries. *Applied Linguistics*, 2(3), 223-235.
De Cock, S., Granger, S., & Leech, G. T. McEnery. 1998. *An automated approach to the phrasicon of EFL learners. learner English on Computer*. London: Longman, 67-79.
Fuller, 2003 Janet M. Fuller The influence of speaker roles on discourse marker use.*Journal of Pragmatics*, 35 (2003), pp. 23–45
Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics*, 28(3), 410-439.
Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, 4(2), 237-258.
Johns, T. (2002). Data-driven learning: The perpetual challenge. *Language and Computers*, 42(1), 107-117.
Leech, G. (1991). The state of the art in corpus linguistics, in Aijmer K. and Altenberg B. (eds.) *English Corpus Linguistics: Studies in Honor of Jan Svartvik*, pp 8-29. London: Longman
Leech, G. (2000). Grammars of Spoken English: New Outcomes of Corpus-Oriented Research. *Language Learning*, 50: 675–724.
Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Addison-Wesley-Longman
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.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
Torgersen, E. N., Gabrielatos, C., Hoffmann, S., & Fox, S. (2011). A corpus-based study of pragmatic markers in London English. *Corpus Linguistics and Linguistic Theory*, 7(1), 93-118.