Available online at www.jlls.org

JOURNAL OF LANGUAGE AND LINGUISTIC STUDIES

ISSN: 1305-578X

Journal of Language and Linguistic Studies, 16(1), 238-252; 2020

A corpus analysis on the language on TV series

Hatice Sezgin[a] *, Mustafa Serkan Öztürk[b]

[a] *Selcuk University, School of Foreign Languages, Alaaddin Keykubat Kampusu, Selcuklu , Konya, 42130, Turkey*
[b] *Necmettin Erbakan University, Ahmet Kelesoglu Faculty of Education, Department of English Language Teaching, Yeni Meram Cd. Meram, Konya, 42090, Turkey*

APA Citation:

Sezgin, H., & Ozturk, M. S. (2020). A corpus analysis on the language on TV series. *Journal of Language and Linguistic Studies, 16*(1), 238-252. Doi: 10.17263/jlls.712787

Submission Date: 14/11/2019

Acceptance Date: 10/03/2020

Abstract

The purpose of the present study is to find out the extent to which the real spoken language is reflected in TV series in terms of vocabulary. In accordance with this purpose, a corpus, named as the British TV Series Corpus (BTSC) was compiled for the present study using two British TV series, Sherlock and Doctor Who, and this corpus was compared to the spoken part of the British National Corpus (BNC), more than 40% of which was compiled from naturally occurring speech in order to find out whether there is a relationship between two corpora. The results showed that the TV series corpus covered the 98.54% of the most frequent lemmas in the spoken part of the British National Corpus, so the language used in TV series reflects the language spoken in the real life in terms of the vocabulary items and their frequency. Accordingly, it can be claimed that TV series can be used as effective in-class and extra-curricular materials for teaching vocabulary and speaking and listening skills.

Keywords: TV series; vocabulary; British National Corpus Introduction

1. Introduction

For most foreign language learners, speaking is the most difficult skill to master. Learners can experience foreign language speaking anxiety even when they are competent to some extent in other skills and areas. This problem results from various reasons, one being the shortness of active vocabulary knowledge, while another can be the problems in listening competence, which is the complementary receptive skill of the productive speaking skill. In this regard, watching movies, TV shows or series in the target language, which is a favoured activity by students, can help in developing speaking skills by contributing to the improvement of both listening skill and active vocabulary. However, what is the extent to which the language used in these TV shows corresponds to the real spoken language? The answer to this question could be found in corpus studies, which focus on collecting texts for linguistic research.

* Corresponding author. Tel.: +90-332-223-1136
E-mail address: h.sezgin@selcuk.edu.tr

The purpose of the present study is to find out the extent to which the real spoken language is reflected in TV series in terms of vocabulary. The British TV Series Corpus (BTSC) was formed for the present study and consisted of two British TV series (Doctor Who and Sherlock) to find out the extent to which TV series reflect the real spoken language and to have an opinion on the efficiency of TV series as materials for extra-curricular speaking and vocabulary activities.

1.1. Literature review

While there have been various definitions of corpus made by different linguists, one definition covering many of these in linguistic terms may be "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (Meyer, 2002). Although the first well-known study related to corpus linguistics in relation with English Language Teaching was conducted by West (1953), under the name of The General Service List (GSL), corpus studies date back to a far earlier date. According to Kennedy (1998) "first significant pieces of corpus-based research with linguistic associations involved using the Bible as a corpus". Taking this into account, Meyer (2008) classifies corpora as pre-electronic and electronic corpora and defines the first as "corpora created prior to computer era, consisting of a text or texts that served as the basis of a particular project" and the latter as "the mainstay of the modern era and the consequence of the computer revolution". Some examples of pre-electronic corpora provided by Meyer (2008) are; biblical concordances, grammars, dictionaries and SEU Corpus.

However, "the real breakthrough in corpus linguistics came with the access to machine-readable texts, which could be stored, transported, and analysed electronically" (Johansson, 2008). After the introduction of computers to corpus studies, the first computer-based corpus for linguistic purposes was developed by Brown University in 1961 under the name of Brown University Standard Corpus of Present-Day American English, which is commonly referred to as Brown Corpus (Francis & Kucera, 1964). This was followed by The Lancaster-Oslo/ Bergen (LOB) corpus (Johansson, Leech & Goodluck, 1978) of written British English compiled between 1970 and 1978 by the University of Lancaster, University of Oslo and Norwegian Computing Centre for the Humanities in Bergen; The London-Lund Corpus (LLC) by Startvik (1990) starting in 1975; along with some corpora for varieties of English, such as The Kolhapur Corpus of Indian English, Wellington Corpus of Written New Zealand English, and Australian Corpus of English (ACE), which including the Brown Corpus were defined by Kennedy (1998) as the First Generation Corpora.

The use of the term corpus linguistics came around a decade later than the first generation corpora, in the title of a collection of papers presented at the 'Conference on the Use of Computer Corpora in English Language Research' held in Nijmegen in 1983, which was titled as Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research (Aarts & Mejis, 1984, cited in Johansson, 2008).

Following the first-generation corpora, corpus linguistics studies have undergone drastic changes in line with the technological developments. Today, there are numerous corpora of different types built for various reasons. According to Baker, Hardie and McEnery (2006), some types of corpora are reference, specialized, multilingual, parallel, learner, diachronic and monitor. Among these the first is of importance for the present study. Reference or general corpora are compiled to serve as a basis for all kinds of corpus related studies. They represent the general nature of language rather than any particular variety or domain, to be used in comparative studies. Some well-known examples of this type of corpus are; British National Corpus (BNC), the spoken part of which was utilized as a reference corpus for the present study.

The British National Corpus (BNC) is a corpus of modern British English, consisting of 100 million words. It was produced by a consortium including Oxford University Press (OUP), Longman and Chambers as dictionary publishers and Universities of Lancaster and Oxford and the Centre for Research and Development of British Library as members of academics (Burnard, 2002).

The BNC is defined as a sample corpus, being composed of text samples; a synchronic corpus, including imaginative texts from 1960 and informative texts from 1975; a general corpus, being not limited to any particular genre, register or subject field; a monolingual corpus of British English only and a mixed corpus of both spoken and written language (Burnard, 2007).

The BNC consists of around 100 million words, 90% of which makes up the written part, and 10% of which forms the spoken part. While gathering data for the written part, three criteria were taken into consideration: domain, time, and medium.

The spoken part of the BNC consists of 10 million words, and these were collected from two main sources; context-governed and demographic (Crowdy, 1993). The main concerns while selecting these data sources were representativeness and sampling. Taking these concerns into consideration, the context-governed part, which includes 6.1 million words, was categorized as; educational and informative, business, public or institutional and leisure. Each of these categories were divided into two sub-categories as monologue and dialogue, the former covering the 40% and the latter 60%. The first of these categories, educational and informative includes lectures, talks, educational demonstrations, news commentaries and classroom interactions, the second category business includes company talks and interviews, trade union talks, sales demonstrations, business meetings, and consultations. The third category, public or institutional includes political speeches, sermons, public/government talks, council meetings, religious meetings, parliamentary proceedings, and the legal proceedings. The last category leisure includes speeches, sports commentaries, talks to clubs, broadcast shows, phone-ins and club meetings (Aston & Burnard, 1998).

The trickier of the sources for the spoken part of the BNC was the demographic one, which was collected from informal encounters of the 124 volunteers, who recorded their speech for a defined period of time (at least 2 days). These individuals were selected on a balanced basis of four criteria; age, sex, social class and geographic region of origin (Aston & Burnard, 1998). The information on the recordings was also detailed including their setting, time, participants, the relationship between the speakers, etc. Consequently, a total of 700 hours of recordings, including 4.2 million words were collected from 124 adults between the ages of 15 and 60+, from 38 different parts of the United Kingdom and of four different socio-economic classes, with a balanced distribution across genders (Kennedy, 1998).

Corpus studies have been around for a long while now, and it also has contributed to language learning and teaching immensely. The relationship between corpus studies and language learning and teaching develops everyday with developing technology, and it draws more attention from every field related to language. Accordingly, more and more studies are conducted every day related to the possible contributions of corpora to Second Language Acquisition including course design (Hou, 2014), development of course materials (O'Dell & McCarthy, 2008), classroom implementations (Molino, 2018; Liu, Lanling, Jiang & Su, 2018), teacher training practices (Caliskan and Kuru Gonen, 2018; Naismith, 2016; Zareva, 2016); teaching writing skills (Yang, 2018; Staples, Biber and Reppen, 2018), vocabulary instruction (Yusu, 2014; Wang and Zeng, 2018), grammar instruction (Liu, 2011; Liu and Jiang, 2009), speaking skills (Gomez Sara, 2016), and reading skills (Brodine, 2001), etc.

1.2. Research questions

The research questions formed for the present study are as follows:

1. To what extent does the BTSC cover the items in the BNC spoken frequency lists?

2. Is there a significant relationship between the spoken part of the BNC and the BTSC in terms of frequency of the items?

3. Are there any similarities between the BNC and the BTSC in terms of the most frequent non-lemmatized and lemmatized 20 items?

2. Method

2.1. Instruments

2.1.1. The British National Corpus (BNC)

BNC is a 100 million-word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written (http://www.natcorp.ox.ac.uk/corpus/index.xml).

2.1.2. The British Television Series Corpus (BTSC)

The BTSC is a 754378-word corpus compiled from the scripts of all aired episodes of two British TV series, Sherlock and Doctor Who.

2.2. Data Collection

2.2.1. Spoken part of the British National Corpus (BNC)

To compare the BTSC with the BNC, frequency lists developed by Leech, Rayson and Wilson (2001) in their book Word Frequencies in Written and Spoken English: based on the British National Corpus were utilised. Table 1 below presents the number of words and their total frequency in the BNC. The frequencies presented are per million, so to provide a full representation of these lists within the BNC, the number for the whole (both written and spoken) corpus was multiplied by 100, since the whole BNC comprises of 100 million words. The total frequency for the spoken part was multiplied by 10, to show the full representation of the frequency of the words included in the lists in the 10 million words of the spoken part of the BNC.

Additionally, the frequency list created for the spoken part of the BNC includes items that have a minimum frequency of 10 per million words, which means that words with fewer frequency are not included in the lists utilized for the present study. This case applies to the whole corpus list, with a minimum frequency of 160 per million words.

Table 1. The Number of words and frequencies in the BNC Frequency Lists

BNC		**n of words**	**Total frequency (per million)**	**Total frequency in the BNC**
whole	no lemma	7726	105779	10577900
	lemmatized	6670	64049	6404900
spoken	no lemma	4841	19454	194540
	lemmatized	827	845117	8451170

As presented in Table 1, the frequency list formed for the spoken part of the BNC include 4841 words, which made up the total of 827 lemmas. The total frequency for these lemmas is 845117 per million, which was multiplied by 10 as 8451170 to have an estimation of the representation of these lemmas in the whole spoken BNC, which is around 10 million words.

2.2.2. Developing the British Television Series Corpus

The TV series included in the British Television Series Corpus were Sherlock and Doctor Who. Sherlock, based on the famous works of Sir Arthur Conan Doyle, The Adventures of Sherlock Holmes, has been broadcast on BBC since 2010. It is a modern detective story about the famous Sherlock Holmes and his friend Dr John Watson in the 20th century London. BBC has aired 13 episodes of Sherlock so far, each of which is around 90 minutes long.

Doctor Who is a science fiction show about a time traveller known as the "Doctor" and his adventures in time and space with his friends from the planet earth. The show started in 1963 and aired 847 episodes in 26 seasons until 1989 on BBC. This old version wasn't included in the BTSC. It started again in 2005, and BBC has aired 146 episodes in 11 seasons so far. Yet, the 11th season wasn't included in the BTSC, since it hadn't been released by the time the process of compiling the corpus started. Therefore, 136 episodes in 10 seasons, each one of which is around 45 minutes long, were included in the BTSC.

The scripts of the episodes were obtained in .pdf format from the official website of BBC (BBC Writers Room, Script Library, Sherlock & BBC Writers Room, Script Library, Doctor Who), then converted to .docx (Microsoft Office Word) format to exclude the non-textual parts (unspoken parts included in the scripts to provide information about the setting) from the scripts.

The total of 120769 words spoken in 13 episodes of Sherlock were included in the BTSC, which make up around 16% of the whole corpus. The total of 633609 words spoken in 136 episodes of Doctor Who included in the BTSC make up around 84% of the whole corpus. Table 2 below presents the percentage of two TV series in the BTSC.

Table 2. The Distribution of the BTSC by Series

Series	**N of Words**	**%**
Sherlock	120769	16
Doctor Who	633609	84
Total	**754378**	**100**

Then scripts of all the episodes of the two series merged into one .txt file, and Textworks 1.5.6 software rewritten by Selahattin Cilek for the current study was used to produce frequency lists for the BTSC.

Textworks produces a list of every word included in the source files along with their frequencies in the Microsoft Office Excel (.xlsx) format. Using the output file generated by Textworks, the next step was grouping misspelt words, proper names, contractions, exclamations, and abbreviations. This process was done manually, and these group of words were excluded from the BTSC. Contractions formed with apostrophe ('ve, 's, and 'd) listed under "*other*" category were also excluded from the BTSC. Table 3 below shows the distribution of the words included in and excluded from BTSC.

Table 3. Distribution of Exclusion List by Category

	TYPE	**TOKEN**
MISSPELT	457	3019
PROPER	2285	16310
CONTRACTION	27	381

EXCLAMATION	231	8961
ABBREVIATION	147	1790
OTHER	3	21967
EXCLUDED	3150	52428
INCLUDED	16625	701950
TOTAL	**19775**	**754378**

Once the above-mentioned items were excluded from the lists, the remaining 701950 items (tokens), which were formed of the total of 16625 different words (types) were lemmatized. The lemmatization process was conducted on the basis of inflectional suffixes, which covered the singular-plural forms of the nouns, basic, comparative and superlative forms of the adjectives, and tense suffixes of the verbs. For instance, the types "thing" and "things" were combined under one type as "thing", or the types "good", "better" and "best" were combined as "good", also types "go", "goes", "went" "gone" and "going" were combined as "go", also by adding up their frequencies. After the lemmatization process, the number of types included in the BTSC was reduced to 11070.

Following the lemmatization process, a list of function words was created, and remaining 11070 types were categorized again as content and function words before comparing the BTSC with the BNC frequency lists. To exclude function words, Textworks was used in this step. The numbers of function and content words are presented in Table 4 below.

Table 4. Number of content & function words

WORD LIST	**TOKENS/%**	**TYPES/%**
CONTENT	331942	10917
FUNCTION	370008	153
LEMMATIZED	701950	11070

The words included in the content words list were then tagged for their parts of speech manually. Frequency lists were created for each part of speech. These were then compared with the similar lists formed for the spoken part of the BNC.

2.3. Data analysis

2.3.1. Corpora comparison

In order to find out whether there was a relationship between lemmatized form of BTSC and the spoken language, it was compared with the spoken part of the BNC. The first comparison between the BTSC and the BNC was conducted in terms of coverage. The AntWordProfiler 1.4.0w for Windows developed by Anthony (2013) was utilized for this purpose. The AntWordProfiler enables the comparison of two or more texts in terms of the words included, and it provides information about the words existing in all texts, the ones existing only in the reference text and the percentage of coverage of the other texts by the reference text.

The second step of the comparison between two corpora was conducted in terms of frequency. The common words included in both the BTSC and the BNC lists were compared in terms of frequency using paired samples T-test on the Statistical Package for Social Sciences (SPSS 20.0).

The third and last step of comparison was conducted using word lists formed using the most frequent 20 words in the whole corpus. This step provides a more detailed comparison at word level, where it becomes possible to study individual words that are common in both lists, and that are not.

3. Results

The first procedure conducted to compare the BTSC with the spoken part of the BNC was done in terms of coverage. This was done on the Ant Word Profiler 1.4.0w software. The findings are presented in Table 5 below.

Table 5. The coverage of the spoken part of the BNC by the BTSC

FILE	**TOKEN**	**TOKEN%**
BTSC/BNC SPOKEN	675	98.54
ONLY BNC SPOKEN	10	1.46
TOTAL	**685**	**100**

As presented in Table 5, the BTSC covers 98.54% of the 685 items included in the spoken part of the BNC. Only 10 lemmas included in the spoken BNC are not included in the BTSC. Below is a table of words included in spoken BNC but not in the BTSC.

Table 6. Words included in the spoken part of the BNC but not in the BTSC

1	better	6	less
2	concerned	7	mine
3	county	8	our
4	economic	9	seventy
5	eighty	10	training

These 10 items are included in the spoken part of the BNC but not in the BTSC. However, some of these items are actually included in the BTSC, but not in the lemmatized version. Starting with the first, the lemmatization process of BTSC included the superlative and comparative forms of the adjectives, which means "better" was combined with the adjective "good" as its comparative form. The same case applies to the 6th item on the list "less", which was taken as the comparative form of the adjective "little" in the lemmatization of the BTSC. But the comparative forms of adjectives are limited to these two only for the spoken part of the BNC, which suggests that the lemmatization process of the BNC also included the superlative and comparative forms of adjectives, but the irregular ones were not involved in the process. The second item on the list is also included in the non-lemmatized form of the BTSC. However, as described above, tense suffixes were also lemmatized, which means that the item "concerned" was combined with the infinitive form of the verb "concern". The same case also applies with the 10th item on the list "training", which was combined with the infinitive form of the verb "train". The 7th and 8th items also were included in the non-lemmatized form of the BTSC, but like all other pronouns and their variations, they were combined with "we" in the lemmatization process. Accordingly, we can claim that

only four of the items in the list compiled from the spoken part of the BNC are not included in the BTSC, which are "county, economic, eighty, and seventy".

As mentioned above, the frequency list for the spoken part of the BNC, utilized for the present study includes words with a minimum frequency of 10 per million words. Accordingly, the coverage of the spoken part of the BNC by the BTSC was re-tested after applying the same ratio for the BTSC. That is, the items with a frequency of lower than 10 per million were excluded from the list, and the coverage was re-calculated using the software Ant Word Profiler 1.4.0w. The findings are presented in Table 7 below.

Table 7. The coverage of the spoken part of the BNC by the BTSC (words with frequency lower than 10 per million excluded)

FILE	**TOKEN (n)**	**TOKEN%**
BTSC/BNC-FREQ-10perMIL	648	94.60
ONLY BNC SPOKEN	37	5.40
TOTAL	**685**	**100**

As presented in Table 7 above, after the words with frequency of lower than 10 per million were excluded, the BTSC covers the spoken part of the 94.60% of the spoken part of the BNC frequency list involving words of only with a frequency of 10 per million words or higher. Only 37 words in the frequency list for the spoken part of the BNC are not included in the BTSC frequency list.

The spoken part of the BNC and the BTSC were also compared in terms of the word frequencies within the corpora. In order to find out whether there was a statistically significant difference in terms of the frequency of the words in these two lists, paired samples T-test was conducted using the SPSS software. The results are presented in Table 8 below.

Table 8. Results of the paired samples statistics

	Mean	**Std. Deviation**	**Std. Error Mean**	**t**	**df**	**Sig. (2-tailed)**
BTSC_LEMMATIZED - BNC_SPOKEN_LEMMATIZED	.00487511179	.12673110713	.00489240013	-.996	670	.319

As presented in Table 8, there is no statistically significant difference at 5% significance level between the BTSC and the spoken part of the BNC in terms of the frequency of lemmas that are included in both lists, as the p value is lower than 0.05 ($0.319 > 0.05$).

In order to compare the BTSC with the spoken part of the BNC further, lists were formed for the most frequent 20 items. First of these was formed using the 20 most frequent non-lemmatized words in each corpus. Table 9 below presents the 20 most frequent non-lemmatized words in the BTSC and the spoken part of the BNC.

Table 9. Comparison of the 20 most frequent non-lemmatized words in the BTSC and the spoken part of the BNC

	BTSC		**BNC**	
RANK	**WORD**	**FREQ***	**WORD**	**FREQ***
1	YOU	39130	THE	39605
2	THE	35782	I	29448
3	I	34086	YOU	25957
4	IT	23098	AND	25210
5	'S	22766	IT	24508
6	TO	19594	THAT	21498
7	A	19159	A	18637
8	NOT	19090	'S	17677
9	ARE	14103	TO	14912
10	THAT	13892	OF	14550
11	AND	13142	N'T	12212
12	OF	13065	IN	11609
13	WHAT	12149	WE	10448
14	IS	11535	IS	10164
15	DO	11344	DO	9594
16	WE	9983	THEY	9333
17	IN	9641	ER	8542
18	ME	9136	WAS	8097
19	THIS	8058	YEAH	7890
20	NO	7593	HAVE	7488

*frequency per million

As presented in Table 9, 15 out of 20 words are included in both lists. These 15 words also show similarities in terms of frequency. As can be observed in Table 9 above, the first three words, “you, the and I” are the same in both lists, even their ranks are different.

The ones that are in the BTSC but not in the BNC list are “are, what, me, this and no”. On the other hand, the ones in the BNC but not in the BTSC are listed as “they, er, was, yeah and have”. The most striking difference here is the use of filler words “er and yeah”, which provide natural speech with fluency in cases, such as pauses or hesitations. It is worth mentioning here again that the BTSC was compiled from scripted speech, while the spoken part of the BNC is 40% naturally occurring dialogues. These filler words can also be found in the BTSC, yet their frequency obviously doesn’t reflect the naturally occurring speech.

Another point worth mentioning here is that most of the words in both lists are function words. That is, there is only one content word in the BTSC list, which is the verb “do” and two content words in the BNC list, verbs “do and have”.

The second list was formed using the 20 most frequent lemmatized words in each corpus. Table 10 below presents the 20 most frequent lemmatized words in the BTSC and the spoken part of the BNC. As presented in Table 10, 18 out of 20 words are common in both lists. Additionally, the items in the lists show similarity in terms of frequency. The two words that are in the BTSC list but not in the BNC list are “what and this”. Even these two words are not in the 20 most frequent lemmatized words list for

the spoken part of the BNC, they are still ranked high in the whole list. That is, "what" is the 21st and "this" is ranked 31st in the whole lemmatized list. The items in the BNC list but not in the BTSC list are "er and yeah". The reason for this finding can be explained as the BTSC being scripted and the BNC being mostly natural again. Yet again, it can be observed that most words in two lists are function words.

Table 10. Comparison of the 20 most frequent lemmatized words in the BTSC and the spoken part of the BNC

	BTSC		**BNC**	
RANK	**WORD**	**FREQ***	**WORD**	**FREQ***
1	I	48964	BE	57016
2	BE	48652	THE	39605
3	YOU	45467	I	31893
4	THE	35782	YOU	26077
5	IT	23606	AND	25210
6	A	21284	IT	24508
7	TO	19594	THAT	21498
8	NOT	19090	HAVE	19689
9	DO	16642	A	18637
10	THAT	13892	NOT	17272
11	WE	13269	DO	16621
12	AND	13142	TO	16615
13	OF	13065	OF	14550
14	WHAT	12152	THEY	12517
15	HE	10476	IN	11609
16	THEY	9795	WE	11507
17	IN	9641	GET	9230
18	HAVE	8862	HE	8628
19	THIS	8058	ER	8542
20	GET	7680	YEAH	7890

*frequency per million

4. Discussions

The positive effects of the use of videos in the target language have been discussed and proven in many ways by numerous studies so far. The related literature shows that watching videos, such as films, TV series and shows in English develops reading comprehension (Saricoban & Yuruk, 2016); contributes learning vocabulary and use of language (Ariogul & Uzun, 2008); improves communicative competence (Yang & Fleming, 2013); listening skills (Tekin & Parmaksiz, 2016) and speaking skills (Leopold, 2016). Acknowledging these positive effects, the present study approaches the subject from a different perspective. Watching videos in target language, such as movies and TV series develops language skills and areas, including the speaking skill. Nevertheless, the extent to which these reflect the naturally occurring speech has not been investigated by these studies.

In order to find an answer to this question, the present study utilizes corpus linguistics. A corpus, named as the British TV Series Corpus (BTSC) was compiled for the present study using two British

TV series, Sherlock and Doctor Who, and this corpus was compared to the spoken part of the British National Corpus (BNC), more than 40% of which was compiled from naturally occurring speech.

The BTSC and the BNC were first compared in terms of coverage in order to answer the first research question. The BNC frequency lists formed by Leech, Rayson and Wilson (2001) including the words with minimum frequency of 10 per million were compared with the BTSC lists. It was found that the BTSC covered the 98.54% of the most frequent 685 lemmas in the spoken part of the BNC. Moreover, lemmas with less frequency than 10 per million were excluded from the BTSC list, and it was compared with the BNC list again. This analysis revealed that 94.60% of the lemmas in the BNC list were covered by the lemmas with minimum frequency of 10 per million in the BTSC. These findings suggest that the language used in TV series reflect the language used in real life at a great extent in terms of the vocabulary items used.

The BNC and the BTSC were also compared in terms of frequency of the items to answer the second research question. With this purpose, common lemmas in the frequency lists of both corpora were compared on SPSS using paired samples t-test. The results of the SPSS analyses revealed that there was no statistically significant difference between the BTSC and the BNC lists in terms of frequency. These findings also indicate a similarity between the BTSC and the BNC, in other words the language used in TV series and the language used in the real life.

In order to answer the last research question, the last comparison between two corpora was conducted in terms of the most common individual items. Lists of 20 most frequent non-lemmatized and lemmatized items were formed for both corpora. First of these was the 20 most common items in the non-lemmatized versions of the BTSC and the BNC. Not surprisingly, all 20 items in both lists were function words, which are a must for sentence building in English language. Out of the 20 items, 15 items were common in both lists. The ones that were not common were also function words. One big difference in the first list was the filler words in the BNC. These filler words, such as “er and yeah” were in the BTSC as well, yet the two corpora presented difference in terms of the frequency of these items. The reason for this difference is considered as the BTSC being scripted and the BNC not being scripted. It can be concluded from this finding that scripted language of the TV series falls short in reflecting these natural elements of the language spoken in real-life.

The second list was formed with the 20 most frequent lemmas in both corpora. The similarity was higher in this list with 18 out 20 items. This list was also mostly formed of function words, except for very common verbs, such as have, do and be. Yet again, these verbs serve also as function words most of the time. The third list consisted of only function words, and again 18 out of 20 items were common in the BTSC and the BNC.

5. Conclusions

5.1. Pedagogical Implications of the Study

As stated above, the findings of the present study show that the naturally occurring speech is reflected in the TV series at a great extent in terms of the vocabulary used. Additionally, audio-visual materials, such as TV series can reflect other elements of a language, such as mimes, gestures, pauses, or hesitations. Accordingly, it can be claimed that TV series can be reliable sources for teaching of general speaking skills and listening skills as well. Students love watching these in their free time. Therefore, adding the motivation factor into equation, watching TV series in the target language can be considered as an efficient extra-curricular activity for language learners. Moreover, through a structured course

plan, they can even be used as in-class materials to teach not only vocabulary, but also pronunciation, language use, the culture and more broadly the speaking skills.

5.2. Limitations of the Study

The corpus compiled for the present study is limited to two TV series. Additionally, these two series are limited in terms of context. Sherlock is about the adventures of an extraordinary detective, while Doctor Who is the fantastic story about a traveller, who travels through time and space. Accordingly, the context of these two series is different than the context of everyday spoken language. Another limitation of the corpus compiled for the present study is that a significant amount of it (84%) was formed of one of these TV series. Finally, part of speech tagging and therefore the lemmatization process of the BTSC was done based on individual words, independent from their contexts and their uses within the sentences, since part of speech tagging is a very troublesome process for such great size of texts and requires serious labour and time.

5.3. Suggestions for Further Research

Similar corpus comparison studies can be conducted using different TV series or movies with various contexts for a better representation of the target language from different perspectives. Instead of using video materials with specific contexts, TV series based more on daily life, such as sitcoms can be used to find out their educational value in terms of the instruction of the general speaking skills. Additionally, other aspects of spoken language besides the vocabulary, such as language use, filler words or phrases, can be studied at a further level. Finally, other corpora of different varieties of English, such as the COCA, can be used as a second reference corpus to find out about the differences between different varieties of English.

6. Ethics Committee Approval

The authors confirm that ethical approval was not required for this study. (Date of Confirmation: 17.03.2020)

Acknowledgements

*This paper was derived from the MA Thesis of the first author.

*This paper was presented orally in the 3rd International Conference on Research in Applied Linguistics, on 24-26 October 2019, in Konya, Turkey.

References

Anthony, L. (2013). AntWordProfiler (Version 1.4.0) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved on June 3, 2018 from: http://www.laurenceanthony.net/software

Ariogul, S. & Uzun, T. (2008). Digital video technology in foreign language classes a case study with ' Lost'. *Dil Dergisi*, *142*, 61-70.

Aston, G. & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

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

BBC Writers Room, Script Library, Doctor Who, Retrieved on June 3, 2018 from: https://www.bbc.co.uk/writersroom/scripts/doctor-who-series-3.

BBC Writers Room, Script Library, Sherlock, Retrieved on June 3, 2018 from: https://www.bbc.co.uk/writersroom/scripts/sherlock.

Brodine, R. (2001). Integrating corpus work into an academic reading course. (Edited by: Aston, G.). *Learning with corpora*. Houston, TX: Athelstan, 138-176.

Burnard, L. (2002). Where did we go wrong? a retrospective look at the British National Corpus. (Edited by: Kettemann, B. & Markus, G.). Teaching and learning by doing corpus analysis. Amsterdam: Rodopi, 51-71.

Burnard, L. (2007 [2000]). *Encoding the British national Corpus, BNC Users Reference Guide.* Edited by Burnard, L., Retrieved on January 4, 2019 from: http://www.natcorp. ox.ac.uk /docs/Burnage93a.htm#4.4.

Caliskan, G., & Kuru Gonen, S. İ. (2018). Training teachers on corpus-based language pedagogy: Perceptions on vocabulary instruction. *Journal of Language and Linguistic Studies, 14(4),* 190-210.

Crowdy, S. (1993). Spoken Corpus Design and Transcription. *Literary and Linguistic Computing, 8*(4). 259–65.

Francis, W. N. & Kucera, H. (1964). *Manual of Information to Accompany 'A Standard Sample of Present-Day Edited American English, for Use with Digital Computers'.* (revised 1979) Providence, RI: Department of Linguistics, Brown University.

Gómez Sará, M. M. (2016). The influence of peer assessment and the use of corpus for the development of speaking skills in in-service teachers. *HOW, 23*(1), 103-128. http://dx.doi.org/10.19183/how.23.1.142.

Hou, H. I. (2014). Teaching specialized vocabulary by integrating a corpus-based approach: Implications for ESP course design at the university level. *English Language Teaching, 7*(5), 26-37.

Johansson, S. (2008). Some aspects of the development of corpus linguistics in 1970s and 1980s. (Edited by: Lüdeling, A., & Kytö, M.). *Corpus Linguistics: An International Handbook.* Berlin, Germany: De Gruyter, Volume 1, 33-53.

Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Oslo: Department of English, University of Oslo.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* London: Longman.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman.

Leopold, L. (2016). Honing EAP Learners' Public Speaking Skills by Analyzing TED Talks. *TESL Canada Journal, 33*(2), 46-58.

Liu D. Jiang P. (2009). Using a corpus-Based lexicogrammatical approach to grammar instruction in EFL and ESL contexts. *Modern Language Journal, 93*(1), 61–78. DOI: 10.1111/j.1540-4781.2009.00828.x

Liu, D. (2011). Making grammar instruction more empowering: An exploratory case study of corpus use in the learning/teaching of grammar. *Research in the Teaching of English, 45*(4), 353-377.

Liu, Y., Lanling, H., Jiang, B., & Su, X., (2018). The application and teaching evaluation of Japanese films and TV series corpus in JFL classroom. *The Electronic Library, 36*(4), 721-732. https://doi.org/10.1108/EL-09-2017-0193

Meyer, C. F. (2002). *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Meyer, C. F. (2008). Pre-electronic corpora, in Corpus Linguistics. (Edited by: Lüdeling, A., & Kytö, M.). *Corpus Linguistics: An International Handbook.* Berlin, Germany: De Gruyter, Volume 1, 1-14.

Molino, A. (2018). 'What I'm speaking is almost English...': A corpus-based study of metadiscourse in English- medium lectures at an Italian university. *Educational Sciences: Theory & Practice, 18,* 935–956. http://dx.doi.org/10.12738/ estp.2018.4.0330

Naismith, B. (2016). Integrating corpus tools on intensive CELTA courses. *ELT Journal, 71*(3), 273-283. doi:10.1093/elt/ccw076

O'Dell, F., & McCarthy, M. (2008). *English collocations in use: Advanced*. Cambridge, England: Cambridge University.

Saricoban, A. & Yuruk, N. (2016). The use of films as a multimodal way to improve learners' comprehension skills in reading in English language and literature department at Selçuk University. *Turkish Online Journal of English Language Teaching (TOJELT), 1*(3), 109-118.

Staples, S., Biber, D., & Reppen, R. (2018). Using Corpus-Based Register Analysis to Explore the Authenticity of High-Stakes Language Exams: A Register Comparison of TOEFL iBT and Disciplinary Writing Tasks. *The Modern Language Journal, 102*(2), 310-332. DOI: 10.1111/modl.12465

Svartvik, J. (1990). *The London Corpus of Spoken English: Description and Research.* Lund: Lund University Press.

Tekin, I. & Parmaksiz, R. S. (2016). Impact of Video Clips on the Development of the Listening Skills in English Classes: A Case Study of Turkish Students. *Journal of Education and Training Studies, 4*(9), 200-208.

The British National Corpus, Retrieved on March 22, 2019, from: http://www.natcorp.ox.ac.uk/ corpus/index.xml,.

Wang, S., & Zeng, X. F. (2018). Effect of English Corpus on Reform of College English Teaching and the Improvement of Students' Vocabulary Competence. *Educational Sciences: Theory & Practice, 18*(6), 3493-3499. http://dx.doi.org/ 10.12738/estp.2018.6.258

West, M. (1953). *A General Service List of English Words*. London: Longman, Green and Co.

Yang, L. H. & Fleming, M. (2013) How Chinese college students make sense of foreign films and TV series: implications for the development of intercultural communicative competence in ELT. *The Language Learning Journal, 41*(3), 297-310. DOI: 10.1080/09571736.2013.836347

Yang, X. (2018). A corpus-based Study of Modal Verbs in Chinese Learners' Academic Writing. *English Language Teaching, 11*(2), 122-130.

Yusu, X. (2014). On the application of corpus of contemporary American English in vocabulary instruction. *International Education Studies, 7*(8), 68-73. DOI:10.5539/ies.v7n8p68.

Zareva, A. (2016). Incorporating corpus literacy skills into TESOL teacher training. *ELT Journal, 71*(1), 69-79. doi:10.1093/elt/ccw045

TV dizilerinde kullanılan dil üzerine bir derlem incelemesi

Öz

Bu çalışmanın amacı gerçek hayatta konuşulan dilin kullanılan kelimeler açısından TV dizilerinde ne derece yansıtıldığını ortaya çıkarmaktır. Bu amaçla, iki İngiliz TV dizisi kullanılarak bir derlem oluşturulmuş ve bu derlem İngiliz Ulusal Derleminin sözlü dil kısmıyla karşılaştırılıp, aralarında ilişki olup olmadığı sorgulanmıştır. Sonuçlara göre, TV dizilerinden oluşturulan derlem İngiliz Ulusal Derlemi sözlü kısmında en sık kullanılan lemmaların %98.54'ünü kapsamaktadır, dolayısıyla dizilerde kullanılan dil, gerçek hayatta konuşulan dili kullanılan kelimeler ve bunların sıklığı açılarından yansıtmaktadır. Sonuç olarak, televizyon dizilerinin kelime bilgisi ile konuşma ve dinleme becerilerinin öğretimi için sınıf içinde ve dışında etkin materyaller olarak kullanılabileceği savunulabilir.

Anahtar Kelimeler: derlem; TV dizisi; kelime; İngiliz Ulusal Derlemi

AUTHOR BIODATA

Hatice Sezgin is a lecturer at Selcuk University School of Foreign Languages. She received her BA degree in English Language Teaching from Hacettepe University, and MA degree in English Language Teaching from Necmettin Erbakan University. She is interested in teaching of speaking skills, phonetics and linguistics.

Mustafa Serkan Ozturk is a faculty member at Necmettin Erbakan University, Faculty of Education, Department of English Language Teaching. He received his PhD degree in English Language Teaching from Gazi University. He is interested in foreign language teaching strategies.