T.C.
NECMETTIN ERBAKAN UNIVERSITY
GRADUATE SCHOOL OF EDUCATIONAL SCIENCES
DEPARTMENT OF FOREIGN LANGUAGES TEACHING
DIVISION OF ENGLISH LANGUAGE TEACHING

SPOKEN LANGUAGE IN TV SERIES: A COMPARATIVE CORPUS ANALYSIS

Hatice SEZGİN

M. A. THESIS

Supervisor

Assist. Prof. Dr. Mustafa Serkan ÖZTÜRK

Konya, 2019

T.C.
NECMETTIN ERBAKAN UNIVERSITY
GRADUATE SCHOOL OF EDUCATIONAL SCIENCES
DEPARTMENT OF FOREIGN LANGUAGES TEACHING
DIVISION OF ENGLISH LANGUAGE TEACHING

SPOKEN LANGUAGE IN TV SERIES: A COMPARATIVE CORPUS ANALYSIS

Hatice SEZGİN

M. A. THESIS

Supervisor

Assist. Prof. Dr. Mustafa Serkan ÖZTÜRK

Konya, 2019

T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
Eğitim Bilimleri Enstitüsü Müdürlüğü

KONYA

BİLİMSEL ETİK SAYFASI

Öğrencinin	Adı Soyadı	Hatice SEZGİN		
	Numarası	108304031007		
	Ana Bilim / Bilim Dalı	Yabancı Diller Eğitimi / İngiliz Dili Eğitimi		
	Programı	Tezli Yüksek Lisans	X	
		Doktora		
	Tezin Adı	Spoken Language in TV Series: A Comparative Corpus Analysis		

Bu tezin hazırlanmasında bilimsel etiğe ve akademik kurallara özenle riayet edildiğini, tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada başkalarının eserlerinden yararlanılması durumunda bilimsel kurallara uygun olarak atıf yapıldığını bildiririm.

31/05/2019

Hatice Sezgin

KONYA

T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
Eğitim Bilimleri Enstitüsü Müdürlüğü

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ KABUL FORMU

Öğrencinin		
	Adı Soyadı	Hatice SEZGİN
	Numarası	108304031007
	Ana Bilim Dalı	Yabancı Diller Eğitimi
	Bilim Dalı	İngiliz Dili Eğitimi
	Programı	Tezli Yüksek Lisans
	Tez Danışmanı	Dr. Öğr. Üyesi Mustafa Serkan Öztürk
	Tezin Adı	Spoken Language in TV Series: A Comparative Corpus Analysis

Yukarıda adı geçen öğrenci tarafından hazırlanan Spoken Language in TV Series: A Comparative Corpus Analysis başlıklı bu çalışma 31/05/2019 tarihinde yapılan savunma sınavı sonucunda oybirliği/oyçokluğu ile başarılı bulunarak, jürimiz tarafından yüksek lisans tezi olarak kabul edilmiştir.

	Ünvanı Adı Soyadı	İmza
Danışman	Dr. Öğr. Üyesi Mustafa Serkan Öztürk	
Jüri Üyesi	Prof. Dr. Arif Sarıçoban	
Jüri Üyesi	Dr. Öğr. Üyesi Emine Eda Ercan Demirel	

Necmettin Erbakan Üniversitesi Eğitim Bilimleri Enstitüsü Ahmet Keleşoğlu Eğitim Fak 42090 Meram Yeni Yol **Meram/KONYA**
Tel : 0 332 324 76 60
Faks : 0 332 324 55 10
Elektronik Ağ: https://www.konya.edu.tr/egitimbilimlerienstitusu
E- Posta: ebil@konya.edu.tr

ACKNOWLEDGEMENT

I would like to start by thanking my advisor Assist. Prof. Dr. Mustafa Serkan ÖZTÜRK, who made it possible for me to complete my master's studies with his guidance.

I would like to continue by expressing my gratitude to my former advisor Assist. Prof. Dr. Ece SARIGÜL, who sometimes tried harder for my thesis than myself, yet retired before I could finish it.

I am also grateful to my colleague, Assist. Prof. Dr. Mustafa DOLMACI, without whom I could not really figure out what to do during my studies. I owe him so much for being there for me with every step, for encouraging and leading me through the way.

I was very lucky during the process to have such great friends and colleagues whose support made me continue till the end.

Finally, I want to thank my family, especially my loving and caring sister Zeynep AY, who has always been more than a sister to me.

T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
Eğitim Bilimleri Enstitüsü Müdürlüğü

ÖZET

Öğrencinin	Adı Soyadı	Hatice Sezgin		
	Numarası	108304031007		
	Ana Bilim / Bilim Dalı	Yabancı Diller Eğitimi / İngiliz Dili Eğitimi		
	Programı	Tezli Yüksek Lisans	X	
		Doktora		
	Tez Danışmanı	Dr. Öğretim Üyesi Mustafa Serkan ÖZTÜRK		
	Tezin Adı	TV Dizilerinde Konuşma Dili: Karşılaştırmalı Derlem Analizi		

Bu çalışmanın amacı, öğrencilerin televizyon dizilerini izlemeye dair tercihlerini ve gerçek hayatta konuşulan dilin televizyon dizilerinde ne ölçüde yansıtıldığını ortaya çıkarmaktır. Öncelikle, öğrencilerin yaptığı İngilizce içerikli ders dışı etkinliklere dair bilgi sahibi olmak amacıyla, uzman görüşü alınarak bir anket geliştirilmiştir. Bu anket öğrencilerin İngilizce okuma, dinleme, video izleme alışkanlıklarını ve özellikle hangi dizileri izlediklerinin yanı sıra bu etkinliklerin dil becerilerinin gelişimine yaptığı katkıya ilişkin algılarını sorgulamaktadır. Daha sonra İngiliz yapımı iki televizyon dizisi kullanılarak bir derlem oluşturulmuş ve bu derlem İngiliz Ulusal Derleminin sözlü dili içeren kısmı ile karşılaştırılarak, iki derlem arasında ilişki olup olmadığı araştırılmıştır. Elde edilen sonuçlara göre; 1) öğrencilerin büyük çoğunluğu İngilizce dizi izlemektedir ve dizi izlemenin dinleme ve konuşma becerileri ile kelime bilgisi ve dil kullanımı alanlarına katkı sağladığını düşünmektedir, 2) dizilerden oluşturulan derlem, İngiliz Ulusal Derleminin sözlü dili içeren 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, televizyon dizisi, konuşma dili, kelime bilgisi, İngiliz Ulusal Derlemi

T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
Eğitim Bilimleri Enstitüsü Müdürlüğü

ABSTRACT

Author's				
	Name and Surname	Hatice SEZGİN		
	Student Number	108304031007		
	Department	Foreign Languages Teaching/English Language Teaching		
	Study Programme	Master's Degree (M.A.)	X	
		Doctoral Degree (Ph.D.)		
	Supervisor	Assist. Prof. Dr. Mustafa Serkan ÖZTÜRK		
	Title of the Thesis	Spoken Language in TV Series: A Comparative Corpus Analysis		

The purpose of the present study is to find out students' preferences regarding watching TV series and the extent to which the real spoken language is reflected in TV series in terms of vocabulary. First, a questionnaire was developed with expert opinion to have information on the English language-related extra-curricular activities of students. The items questioned students' habits of reading, listening and watching videos in English, and in specific which TV series they watched and their perceptions related to the contributions of these to the development of their linguistic skills. Then, a corpus was compiled using scripts from two British TV series, and it was compared with the spoken part of the British National Corpus in order to find out whether there is a relationship between two corpora. The results showed that 1) most of the students watch TV series in English and believe that watching TV series develops their listening & speaking skills and vocabulary knowledge and contributes to their use of English, 2) 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.

Key Words: corpus, TV series, spoken language, vocabulary, British National Corpus

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

ACE	Australian Corpus of English
ANC	The American National Corpus
AWL	Academic Word List
BASE	British Academic Spoken English
BBC	British Broadcasting Corporation
BNC	The British National Corpus
BoE	The Bank of English
BTSC	The British TV Series Corpus
CANCODE	Cambridge and Nottingham Corpus of Discourse in English
COBUILD	Collins Birmingham University International Language Database
COCA	The Contemporary Corpus of American English
DDL	Data Driven Learning
EFL	English as a Foreign Language
ELT	English Language Teaching
GSL	The General Service List
HMDC	House M.D. Pure Dialogue Corpus
IC	Interactional Competence
ICLE	International Corpus of Learner English
LLC	The London-Lund Corpus
LOB	The Lancaster-Oslo/ Bergen Corpus
MICASE	Michigan Corpus of American Spoken English
NHCC	Nottingham Health Communication Corpus
OUP	Oxford University Press
SCOTS	Scottish Corpus of Texts & Speech
SLA	Second Language Acquisition
SOFL	School of Foreign Languages
SPSS	Statistical Package for Social Sciences
STC	Spoken Turkish Corpus
TNC	Turkish National Corpus

CHAPTER 1

INTRODUCTION

1.1 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.

Many definitions have been made for the concept of corpus, such as “a collection of texts based on a set of design criteria, one of which is that the corpus aims to be representative” (Cheng, 2012), “bodies of texts assembled in a principled way” (Johansson, 1995a), and “a collection of texts, written or spoken, usually stored in a computer database” (McCarthy, 2004).

The first well-known corpus related study in the contemporary sense was conducted by West (1953), who gathered an approximate number of 2000 words, and called this body “The General Service List (GSL)”. Since then, many different corpora have been formed under different names. The most well-known of these are the British National Corpus (BNC), Corpus of Contemporary American English, and Bank of English or Australian Corpus of English.

The British National Corpus, which is included in the present study as a reference corpus to be compared to a small scale TV series corpus, “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". The BNC is a monolingual (dealing with only modern British English), synchronic (covering British English of only the late twentieth), general (including different styles and varieties, not limited to any particular subject field, genre or register, and containing examples of both spoken and written language), and a sample corpus (including samples of 45,000 words taken from various parts of single-author texts) (The British National Corpus).

1.2 Statement of the Problem

Watching TV series is a favoured activity for foreign language learners. It is believed, watching videos in the target language can develop some of the language skills and areas, such as listening, speaking and vocabulary (Díaz-Cintas, 2009). Several studies have been conducted on the language of TV series from the corpus linguistics perspective (Law, 2015; Bednarek, 2011), and some other studies focus on the use of TV series in the EFL classroom (Frumuselu, De Maeyer, Donche & Gutierrez-Colon Plana, 2015; Talavan, 2007). Therefore, the present study tries to approach TV series from a different perspective in the context of Foreign Language Teaching and corpus studies, by focusing on the language used in TV series in terms of the vocabulary used, and the extent to which vocabulary used in real-life spoken English is reflected in TV series.

1.3 Purpose of the Study

The purpose of the present study is to find out students' preferences regarding the types of materials they use for their extra-curricular activities, to form a comparatively small-scale corpus and to compare it to the spoken part of the BNC. Additionally, students' beliefs related to the contribution of watching videos in English to the development of their speaking and listening skills, vocabulary and language use are investigated. The corpus, which will be mentioned as British TV Series Corpus (BTSC) from now on, was formed for the present study and consisted of two British TV series (Doctor Who and Sherlock) that were selected based on student preferences and was

intended to find out the extent to which the students' favourite 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.

In accordance with this purpose, the research questions were formed as follows:

1. Which types of materials do the students studying English use for their extra-curricular activities?

2. What are students' favourite genres for movies and TV series?

3. What are students' beliefs related to the contribution of watching movies, TV shows and series in English to the development of their

   (a) speaking skills,
   (b) listening skills,
   (c) vocabulary,
   (d) language use?

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

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

6. Are there any similarities between the BNC and the BTSC in terms of the most frequent 20

   (a) words,
   (b) function words,
   (c) nouns,
   (d) verbs,
   (e) adjectives,
   (f) adverbs?

1.4 Importance of the Study

The present study is significant as it deals with an aspect of English that really attracts students' interest. In the first step of the study, a questionnaire was administered to English Preparatory Class students, who studied at Selcuk University School of Foreign Languages. Accordingly, almost every one of these students stated that they watched TV shows in English, and almost every one of them believed that watching these helps developing their knowledge of vocabulary and listening and speaking skills.

1.5 Assumptions

The sources of the corpus compiled for the present study were selected relying on the questionnaire conducted with participants in order to find out their preferences regarding extra-curricular materials. The assumption made while selecting these sources was that students were honest in their answers to the questions included in the questionnaire, since the present research studies the TV series, because it was found that watching TV series is a favoured extra-curricular activity related to the target language. Additionally, the TV series included in the present study were selected according to their preferences.

1.6 Limitations

The present study is limited only to the spoken part British National Corpus, which was selected based on convenience. Another limitation of the present study is data sources included, which is two British TV series. As stated above, these TV series were selected according to the results of a questionnaire administered to 132 students, who studied English Preparatory Class at Selcuk University School of Foreign Languages in 2017-2018 Academic Year. Accordingly, the present study is limited to these students, in terms of TV series preferences.

1.7 Definitions of Some Key Concepts

Content word: Words which refer to a thing, quality, state or action and have lexical meaning when used alone. Content words are mainly nouns, verbs, adjectives and adverbs (Richards & Schmidt, 2002).

Corpus: A corpus is a body of written text or transcribed speech, which can serve as a basis for linguistic analysis and description (Kennedy, 1998).

Function word: Words which have little meaning on their own and show grammatical relationships in and between sentences (grammatical meaning). Function words include conjunctions, prepositions and articles (Richards & Schmidt, 2002).

Lemma: A set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling" (Francis & Kucera, 1982).

Listening skill: The ability to pay attention to and effectively interpret what other people are saying (Oxford English Dictionary).

Script: The words of a film, play, broadcast, or speech (Cambridge English Dictionary).

Speaking skill: The ability to build and share meaning through the use of verbal and non-verbal symbols in a variety of contexts (Chaney & Burk, 1998).

Spoken corpus: A corpus consisting entirely of transcribed speech (Baker, Hardie & McEnery, 2006).

The British National Corpus (BNC): 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 (British National Corpus).

The British TV Series Corpus (BTSC): a 754378-word corpus compiled from the scripts of all aired episodes of two British TV series, Sherlock and Doctor Who, which were selected based on students' preferences.

Token: Token is a "word" within a corpus. It is used most often to talk about word count and the size of a corpus (Tang, 2015).

Type: A unique word form in a corpus. Types are placed in a word list arranged most often in order of frequency or alphabetical order, and usually shown with frequency count (Tang, 2015).

Vocabulary: The body of words known to an individual person (Oxford English Dictionary).

CHAPTER 2

REVIEW OF LITERATURE

2.1 A brief history of corpus linguistics

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 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 built for various reasons. Bonelli and Sinclair (2006) provides a historical timeline for the developmental stages of electronic corpora:

> a. The first 20 years, c. 1960–1980; learning how to build and maintain corpora of up to a million words; no material is available in electronic form, so everything has to be transliterated on a key-board.
> b. The second 20 years, 1980–2000; divisible into two decades:
> i. The 1980s, the decade of the scanner, where with even the early scanners a target of 20 million words becomes realistic.
> ii. The 1990s, the First Serendipity, when text becomes available as the by-product of computer typesetting, allowing another order of magnitude to the target size of corpora.
> c. The new millennium, and the Second Serendipity, when text that never had existence as hard copy becomes available in unlimited quantities from the Internet.

2.2 Different types of corpora

Corpus is a body of texts, compiled as a basis for linguistic analysis and description (Kennedy, 1998). Therefore, depending on the nature of the analysis to be conducted, corpora may vary. According to Baker, Hardie and McEnery (2006), some types of corpora are reference, specialized, multilingual, parallel, learner, diachronic and monitor.

Starting with the first, 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)

and Contemporary Corpus of American English (COCA), both of which consist of millions of words from almost every genre of both spoken and written English.

The second type, specialized corpora are compiled in accordance with a particular linguistic purpose unlike general corpora. The scope of such corpora is narrower than the general corpora, yet the context may vary at a wide range, from petroleum studies as in the case of Guangzhou Petroleum English Corpus (GPEC) (Zhu, 1989; cited in Kennedy, 1998) to medical studies or even more specialized as in the case of The Nottingham Health Communication Corpus (NHCC), compiled in order to document and analyse the spoken interaction between healthcare professionals and patients (Adolphs, Brown, Carter, Crawford, & Sahota, 2004).

The development of multilingual or bilingual corpora, which can also be referred to as parallel corpora, resulted from the need for mechanical translation (Mitkov, 2005). They include two or more corpora compiled similarly from different languages in a manner enabling the comparison or translation between these languages. One example of such corpora is The Arabic-English Parallel News Corpus compiled between 2001 and 2004 from news stories in Arabic and their translation to English (Evans, 2018).

Learner corpora, which can also be included in the specialized corpus type (Kennedy, 1998) and also named as non-native speaker corpora (Bonelli and Sinclair, 2006), refers to the compilation of samples of the target language uses of the foreign language learners. This is used to explore the deviance in learners in a detailed way by comparing it to the model corpora of the target language. Many learner corpora have been compiled so far, some academic and some commercial, which are listed by Nesselhauf (2004), one example being International Corpus of Learner English (ICLE), developed by University of Louvain La-Neuve in Belgium. Another learner corpus study in Turkish context was conducted by Sanal (2007).

The last two of the types of corpora listed above, diachronic and monitor corpora feature the time dimension. While the scope of diachronic corpora is limited to different periods of time to portray the characteristics of a language specific to that period of time, monitor corpora are compiled with the aim of keeping them up-to-date or

synchronic. This creates another difference between these two, as diachronic corpora are static while the monitor corpora are dynamic.

Other classifications have been made for different types of corpora, such as pedagogic corpus, which refers to the compilation of "all the language a learner has been exposed to" (Hunston, 2002) or sample-text or full-text corpora, which are designed as a "representative sample of the total population of discourse" (Kennedy, 1998). The list can go on with such examples like training, test, dialect, regional, non-standard corpora, which can also be included in the category of specialized corpora. Yet, taking the purpose of the present research into consideration, another important distinction should be made here between written and spoken corpora.

Any written text can serve as a resource to a written corpus, depending on its purpose. These could be either texts published as a hard copy, written manually or ones produced electronically. It can involve anything from books, newspapers, letters, magazines, even legal documents to e-mails, websites, and digital publications. On the other hand, gathering a spoken corpus, which can be made of transcribed speech of any context might be a little more troublesome, as stated by Weisser (2005; in Ciliz, 2010) as "written language generally tends to be far easier to process than spoken language, as it does not contain fillers, hesitations, false starts or ungrammatical constructs". Although it might seem that the resources for these two types of corpora vary dramatically, Biber (1998) reported that some of the spoken and written genres can be similar in terms of some linguistic aspects.

2.3 Various Well-Known Corpora

This section presents several influential corpora of English language gathered for several reasons in order provide a basis for comparison with British National Corpus (BNC), which serves as the reference corpus for the present study.

2.3.1 The Corpus of Contemporary American English (COCA)

COCA is "the largest, freely available corpus of English, and the only large and balanced corpus of American English. COCA is probably the most widely-used corpus

of English." (Corpus of Contemporary American English). It consists of more 560 million words, which were gathered for 27 years, by including 20 million more words from spoken fiction, popular magazines, newspapers and academic texts each year. COCA will continue to include 20 million more words each year, as it is a dynamic corpus (Xiao, 2008).

2.3.2 The American National Corpus (ANC)

The project of ANC started in 1998 in order to build a corpus comparable to The British National Corpus. Accordingly, its design was similar to BNC with some differences in sampling periods and text categories. It is "a massive electronic collection of American English, including texts of all genres and transcripts of spoken data produced from 1990 onward" (The Open American National Corpus). The second version of ANC, released in 2006, includes the total of 22 million words; 18.5 million from written, 3.9 million from spoken contexts (Xiao, 2008).

2.3.3 The Bank of English (BoE)

The BoE, which is one the most well-known monitor corpora, is a project started in 1991 and conducted by the team of Collins Birmingham University International Language Database (COBUILD) project. The written part, which makes up the 75% of the corpus, was derived from websites, newspapers, and magazines, while the spoken part (25%) was formed with conversations from television and radio, meetings, discussions, and interviews (Xiao, 2008). The full corpus contains 4.5 billion words and currently, the Bank of English covers 650 million words chosen to provide an accurate and balanced representation of contemporary English (The Collins Corpus).

2.3.4 Brown Family Corpora

The Brown University Standard Corpus of Present-day American English, also known as the Brown Corpus, is considered as the first modern corpus of English. It consists of 1,014,312 words of written American English, gathered from books published in the United States in 1961. The data collection was conducted under 15 text categories, including 500 samples of 2000+ words from each (Brown Corpus Manual).

The primary purpose for building the Brown corpus was setting a standard for compiling other corpora of variations of English and other languages for comparative studies. Realizing this primary purpose, a number of other corpora were built following this standard (Xiao, 2008). Table 1 below presents a list of such corpora, belonging to the Brown Family.

Table 1 Brown Family Corpora

Corpus	Language variety	Period	Samples	Words (million)
Brown	American English	1961	500	One
Frown	American English	1991-1992	500	One
LOB	British English	1961	500	One
Lancaster 1931	British English	1931 +/- 3 years	500	One
FLOB	British English	1991-1992	500	One
Kolhapur	Indian English	1978	500	One
ACE	Australian English	1986	500	One
WWC	New Zealand English	1986-1990	500	One
LCMC	Mandarin Chinese	1991 +/- 3 years	500	One

From, Xiao (2008)

2.3.5 Academic Word List (AWL)

The AWL, developed by Averil Coxhead, was compiled from "414 academic texts by more than 400 authors, containing 3513330 tokens (running words) and 70377 types (individual words)" from four sub-corpora of arts, commerce, law and science. The primary purpose for AWL was to be used by higher education level teachers and students for preparation to a programme (Coxhead, 2000).

2.3.6 The General Service List (GSL)

Long before the electronic corpora mentioned so for in this part, West (1953) created the General Service List of about 2000 words, which was a breakthrough in

corpus linguistics. The purpose of the list was to represent the most frequent words in English for the learners and teachers of English language. It has been more than six decades since its publication, yet the GSL is still considered as one of the best frequency-range based word lists, and the studies following it still cannot match its relevance of the subject in terms of universality, utility and usefulness (Gilner, 2011). Still, many criticisms have been raised for the GSL, related to its limitations in several terms including its being outdated (Richards, 1974). Taking these concerns into consideration, The New General Service List (NGSL) of 2800 core high frequency vocabulary words for students of English as a second language was developed by Dr Charles Browne, Dr Brent Culligan and Joseph Phillips in 2013, following West's steps (Browne, 2013).

2.4 The British National Corpus (BNC)

The British National Corpus (BNC), which serves as the reference corpus for the present study, 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).

When starting with the project of creating the BNC, there were several purposes to make it differ from the corpora thitherto established. The BNC would be the largest freely available corpus ever, that was synchronic, contemporary and covering both written and spoken British English with a non-opportunistic design. The commercial partners had some goals as well while investing such amount of money, like gaining a competitive advantage in publishing ELT dictionaries. For the academic partners, the goal was developing a new corpora-establishing model within the area of corpus linguistics. But the common goal for the consortium was "to build a really big corpus" (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.

For the domain criterion, texts were classified as imaginative and informative. Imaginative texts covered less than 25% of publications according to collected data, so the larger amount was allocated for informative texts. While planning the distribution between these two, the purpose was reflecting the role the literal and creative writing played on the culture. Accordingly, eight sub-domains were selected for the informative texts to be included in the BNC (Aston & Burnard, 1998). Table 2 below presents the distribution of texts included in the BNC by domain.

Table 2 Distribution of texts included in the BNC by domain

Domain	**texts**	**%**	**words**	**%**
Imaginative	625	19.47	19664309	21.91
Informative: Arts	259	8.07	7253846	8.08
Informative: Belief and thought	146	4.54	3053672	3.40
Informative: Commerce and finance	284	8.85	7118321	7.93
Informative: Leisure	374	11.65	9990080	11.13
Informative: Natural and pure science	144	4.48	3752659	4.18
Informative: Applied science	364	11.34	7369290	8.21
Informative: Social science	510	15.89	13290441	14.80
Informative: World affairs	453	14.11	16507399	18.39
Unclassified	50	1.55	1740527	1.93

For the time criterion, informative texts included were published after 1975 and, imaginative texts included were published after 1960. Table 3 below presents the distribution of texts included in the BNC by time criterion.

Table 3 Distribution of texts included in the BNC by time

Time	texts	%	words	%
1960-1974	53	1.65	2036939	2.26
1975-1993	2596	80.89	80077473	89.23
Unclassified	560	17.45	7626132	8.49

The last criterion was medium, referring to the type of the publication of the text. While defining categories for medium criterion, the creators tried to keep label categories as comprehensive as possible. The label 'Miscellaneous published' covers brochures, leaflets, manuals, advertisements. 'Miscellaneous unpublished' label refers to letters, memos, reports, minutes, essays, and the label 'Written-to-be-spoken' includes scripted television material, play scripts etc. (Aston & Burnard, 1998). Table 4 below presents the distribution of texts included in the BNC by medium.

Table 4 Distribution of texts included in the BNC by medium

Medium	texts	%	words	%
Book	1488	46.36	52574506	58.58
Periodical	1167	36.36	27897931	31.08
Miscellaneous published	181	5.64	3936637	4.38
Miscellaneous unpublished	245	7.63	3595620	4.00
Written-to-be-spoken	49	1.52	1370870	1.52
Unclassified	79	2.46	364980	0.40

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). Table 5 below presents the distribution of the context-governed sources included in the BNC by these categories.

Table 5 Distribution of the context-governed sources included in the BNC by categories

Category	texts	%	words	%
Educational and informative	144	18.89	1265318	20.56
Business	136	17.84	1321844	21.47
Institutional	241	31.62	1345694	21.86
Leisure	187	24.54	1459419	23.71
Unclassified	54	7.08	761973	12.38

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).

2.5 Corpus and ELT

2.5.1 Corpus Studies in English Language Teaching

Corpus studies have been around for a long while now, and even there have been debates about the purpose it serves among the linguists, it is an undeniable fact that corpora have contributed to both linguistics and Language Teaching immensely. According to Granger (2002), the relationship between corpora and Second Language Acquisition (SLA) started in 1980s, and SLA utilized corpus linguistics in order to

gather information on the way speakers used the language, in other words the “learner corpora”, which was defined above.

However, the relationship between corpus linguistics and Foreign Language Teaching is not limited to learner corpora. It was Johns (1986), who first suggested the positive effects of corpora on the way foreign language learners and teachers describe language. However, it took some time before researchers started to acknowledge these effects, and the relationship between corpora and foreign language learning couldn’t fully develop until the 1980s (Chambers, 2007). According to Meunier (2011), there were several reasons for the lack of corpus related studies in language learning environment, one of which is the lack of dialogue between the linguists and language teachers. Similarly, Römer (2006) claimed that the advances in corpus studies couldn’t really affect ELT (English Language Teaching) studies, although both fields made their progress separately.

Despite all these controversies, some language teachers defend the benefits of corpora for the learners strongly and argue that corpora help learners in understanding the descriptions of language by bringing the authentic language into their classrooms (Hunston, 2002). Römer (2008) classified the corpus applications in language teaching as: “indirect applications: hands-on for researchers and material writers” and “direct applications: hands on for teachers and learners (DDL-Data Definition Language)”.

Indirect effects of corpora in language teaching include the effects on syllabi and teaching materials. Accordingly, any foreign language learner has been exposed to the outputs of corpus studies (McEnery, Xiao and Tono, 2006). Dictionaries, coursebooks, course designs somehow utilize products of corpus linguistics when it comes to the field of language teaching (Hunston, 2002). Furthermore, existing pedagogical descriptions can be re-evaluated with the evidence obtained through corpus studies, even their emergence didn’t relate to corpus studies at all (Sinclair, 2004).

Direct effects of corpus linguistics focus on the “teacher-corpus interaction” and “learner-corpus interaction” (Römer, 2008), which are more aimed at teachers and learners of foreign language, who study corpora to find out about particular patterns and words in the language (Bernardini, 2002). First of the direct applications of corpora

in language teaching is concordances, which refer to lists of words used in particular texts along with their contexts (Richards and Schmidt, 2002). Concordances can be used to study word frequencies, grammar, discourse or stylistics, and concordancing can help language learners analysing the language, studying structures or lexical patterns (Gaskell and Cobb, 2004). Another direct application is Data Driven Learning (DDL), which was developed by Johns (1991). In DDL, also known as discovery learning, students take an active role in their own learning process through studying and analysing concordance lines, which helps increasing student motivation (Baker, Hardie and McEnery, 2006). Last direct application of corpus in ELT to be mentioned in the present study is the corpus-based approach, which uses corpora as source to study the language in a smaller set of data. Additionally, corpus-based approach can be utilized to test existing ideas, assumptions or knowledge about the language (Tognini-Bonelli, 2001).

The effects corpus studies on language learning and teaching are of course not limited to these and, as they develop everyday with developing technology, they draw 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 SLA 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.

2.5.2 Required percentage of vocabulary for written and spoken comprehension

Defining the number of vocabulary items in a language is an impossible task to accomplish even with the broadest corpora, and it is also impossible for any speaker of any language to know every word in a language even they are native speakers. Yet, with a certain extent of vocabulary knowledge, it is possible to accomplish some tasks, both for comprehension and production.

The issue of adequate comprehension has been studied by linguistics, and the concept has been defined as the lowest score from a comprehension test (Laufer, 1992). To put it more clearly, we can define the concept of adequate comprehension as the minimum level of vocabulary required for comprehension (Nation, 2006; Webb and Rogers, 2009a).

The question of "What is the number of words required for certain tasks?" arises here. First known study questioning this was conducted by Schonell, Meddleton & Shaw (1956), who reported that 2000 word-families made up the 99% of the spoken discourse of Australian English by studying the oral interaction among Australian workers. Taking their finding into account, some other researchers (Nation and Meara, 2002; Schmidt, 2000) acknowledged that knowledge of 2000 word families was enough for accomplishing tasks in daily spoken English until more studies were conducted on the subject.

The inclusion of computers in the corpus linguistics enabled more reliable studies on the spoken language. Another study on the required vocabulary knowledge for spoken comprehension was conducted by Adolphs and Schmidt (2003), almost 50 years later than Schonell et al. (1956). According to them, the study conducted by Schonell et al. (1956) held a very important place in the area, yet it was conducted in a much more different era of corpus linguistics, when computers hadn't come to stage. It also had several limitations in terms of the subjects included in the study (only Australian workers), whose speeches were recorded only in a certain context. Taken these limitations into consideration, they wanted the test their theory making use of two contemporary corpora; the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) and the spoken part of the British National Corpus (BNC), both of which were compiled benefiting the current technologies in corpus linguistics from a variety of subjects and settings. The CANCODE consists of 5 million words, while the spoken part of the BNC consists of 10 million words of transcribed conversations. Adolphs and Schmidt (2003) compared these two huge corpora with the findings of the study conducted by Schonell et al. (1956). Accordingly, they reported that the findings of Schonell et al. (1956) that 1623 word-families covered 98.31% of spoken discourse and 2279 word-families covered the 99.17%, didn't comply with CANCODE and BNC.

According to the frequency lists obtained from CANCODE, 2000 word-families only covered 94.76% and 3000 word-families covered 95.91% of the spoken discourse. Similarly, 2000 words covered 93.30%, and 5000 words covered 96.93% of the spoken discourse, based on the frequency lists obtained using the spoken part of the BNC.

With the findings of this study, the general opinion about the size of vocabulary knowledge for comprehension underwent some changes. In a similar attempt, Nation (2006) developed BNC sub-lists, to make estimations on the required size of vocabulary knowledge to accomplish several tasks. These sub-lists simply included the most frequent words used in English Language based on the BNC. That is, the list 1K referred to the most frequent 1000 words in the BNC, while the list 5K included the most frequent 5000 words. These lists enable researchers to compare any text or speech with the BNC to find out the extent to which the words included in the selected source are covered by the BNC for desired level of comprehension.

Utilizing this method, Nation (2006) reported that the number of words required to understand a novel or newspaper was 8000 to 9000 words, while 6000 to 7000 words were enough to understand a children's movie or an unscripted spoken interaction. Webb and Rogers conducted the same methodology with movies (2009a) and TV programs (2009b). They reported that 6000 to 10000 words provided 98% of coverage for movies and 5000 to 9000 words covered 98% of the vocabulary items in TV programs, depending on the genre.

Suggesting an exact number for the size of vocabulary required for all tasks is not possible of course as this number varies by many factors including the participants or the context for the spoken language and the genre or period for written texts. Yet, we could offer an estimate for the percentage of the required knowledge of the vocabulary. Laufer (1989) was among the first who studied the issue. According to her findings, knowing 95% of the vocabulary included in a text was enough for reading comprehension. Hu and Nation (2000) later studied the same issue from a different perspective and reported that adequate level of reading comprehension required knowledge of 98% of vocabulary included in the text.

2.5.3 The concepts of word, lemma, word family

Words are the most fundamental elements of any language. The answer to the question "What is a word?" seems obvious to many people. Many dictionaries define it as the smallest meaningful unit of language, in the most general sense. However, the concept of word is not always that simple to define or comprehend. This part provides a more detailed definition for the concepts of word, lemma and word family.

Oxford Dictionary of English defines the concept of 'word' as "*a single distinct meaningful element of speech or writing, used with others (or sometimes alone) to form a sentence and typically shown with a space on either side when written or printed*" (Oxford English Dictionary). However, it is not always to possible to apply this definition to any written unit separated with a space in English language. Function words like the article "the" or contracted forms, such as "can't" can be problematic to be defined as single words. Therefore, more distinct concepts, such as lexeme (or lexical item), lemma and word family are used in linguistics.

Lexeme refers to the smallest unit in the meaning system (Richards and Schmidt, 2002) that can consist of one or more words. A lexeme can have different forms, and any inflected form of a lexeme is considered as the same lexeme.

Lemma on the other hand refers to the headword in a dictionary or in a word list. The concept was defined by Francis and Kucera (1982) as "set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and /or spelling". Therefore, like lexemes, inflected forms of a lemma are regarded as the same lemma, yet unlike lexemes, lemmas can only consist of one word regardless of the meaning. In corpus studies, lemmas are sometimes preferred to types (any form of word separately), and similarly the frequencies of the items in the frequency lists for the BTSC formed for the present study were also calculated on lemmatized forms of the types.

The last and most comprehensive of the concepts to be defined here is the word family. A word family is the base form or root of any word covering its all forms with both inflectional or derivational prefixes and suffixes.

2.6 Related studies

As presented above, language teaching has benefited from corpus studies at a large extent. Corpus studies are included almost in every area of language teaching, from course design to practical implementations. Yet, this part mostly focuses on the studies related to teaching of spoken language, and how this area has benefited from spoken corpora.

2.6.1 Studies abroad

Knoch (2004) studied the BNC in order to investigate the structures used by native speakers for comparisons and the frequency of the comparisons in the spoken language. According to the findings, native speakers use other ways of comparisons more frequently than the adjective comparatives, and this is only the case for native speakers. Knoch (2004) also argued that textbooks mostly focus on structures of adjective comparative for comparing and contrasting, which results in non-native speakers of English preferring the adjective comparative structures more frequently, unlike native speakers.

Grant (2011) studied the use of "just" in British academic spoken English in terms of frequency and functions. She utilized Michigan Corpus of American Spoken English (MICASE) and British Academic Spoken English (BASE) to define the occurrences of "just". According to her findings, "just" is used as a minimizer by lecturers, and there are some minor differences in the usages of "just" across disciplines. She also reported a difference between the uses by the lecturers and students, suggesting that the usages of "just" should take place in the teaching of English for academic purposes. Zahra and Abbas (2018) also used MICASE as reference corpus to investigate the online corpora practices in ELT in Pakistani context. They identified the usages of lexical items in different contexts from MICASE. They found that lexical items can be used with different parts of speech depending on the context, and their positions within the sentences (right and left collocates) provide significant information on deducing the meanings in different contexts. Accordingly, they suggested the teaching of this

technique for different usages of certain lexical items and their various meanings in different contexts.

Anderson and Corbett (2010) investigated the spoken part of Scottish Corpus of Texts & Speech (SCOTS) for 'friendly' language in order to present a model for learners of English. They also intended to raise awareness on local speech varieties, which they believed to be neglected in foreign language teaching environments. On the other hand, Strik, Hulsbosch and Cucchiarini (2009) studied the multiword expressions in spoken language and ways of identifying these in a speech corpus. They reported that multiword expressions varied significantly in pronunciation.

The related literature also includes some studies relating corpora with scripted materials, such as TV shows and movies, as does the present study. One of these was conducted by Csomay and Petrovic (2012), who studied the effects of watching TV series and movies on learning technical vocabulary through a corpus-based approach. They compiled a 130.000-word corpus from TV shows and movies with legal content and studied the occurrence of legal vocabulary in terms of frequency and distribution. They found that most of the technical vocabulary was repeated more than ten times and suggested that the use of such content-specific materials can contribute to learning of English for Specific Purposes.

In another study, Liu et al. (2018) designed a Japanese films and TV series corpus to contribute to teaching of Japanese as a foreign language with real context and new teaching materials created through their corpus. They reported that teaching through a video corpus-based technique had significantly positive effects on learning the language and new vocabulary items.

Bednarek (2011), studied the language in a TV show in terms of word frequency in comparison with several corpora. According to her findings, scripted dialogues in the TV show she studied, Gilmore Girls, was more emotional, but more direct and clearer than the natural unscripted dialogues occurring naturally in real life. Law (2015) also compared a corpus formed from a TV show, House M.D. (House M.D. Pure Dialogue Corpus-HMDC) with the COCA and the spoken part of the COCA (COCA Spoken) using frequency lists. He found that the HMDC was more similar to the COCA

spoken than the COCA. It was also reported that HMDC was more negative, interpersonal and involved more disagreement than the real-life English.

2.6.1 Studies in Turkey

One of the most important studies related to corpus linguistics in Turkey was a project conducted by Aksan, Aksan, Özel, Yılmazer, Demirhan, Mersinli, Bektaş, & Altunay (2016), who constructed a corpus of Turkish language, which they called Web-Based Turkish National Corpus (TNC). TNC is defined as a balanced and general corpus of contemporary Turkish language consisting of 50 million words, following the framework of BNC (Aksan, Aksan, Koltuksuz, Sezer, Mersinli, Demirhan, Yılmazer, Kurtoğlu, Atasoy, Öz & Yıldız, 2012).

Like the TNC, a spoken corpus of Turkish language was compiled modelling the BNC under the name of Spoken Turkish Corpus (STC) (Cokal Karadas & Ruhi, 2009). Spoken Turkish Corpus Project has been conducted by the Department of Foreign Language Education of Middle East Technical University since 2008 and it was supported by TUBITAK (Spoken Turkish Corpus). The purpose of the STC was compiling a large-scale corpus of spoken Turkish language (Ruhi, Eroz-Tuga, Hatipoglu, Isik-Guler, Acar, Eryilmaz, Can, Karakas, & Cokal Karadas, 2010).

In order to study the use of pragmatic markers “hayır and cık” in Turkish language, Bal-Gezegin (2003), analysed recorded conversations of native speakers of Turkish, which were obtained from the STC. Her findings revealed differences and similarities in the syntactic and pragmatic features of hayır and cık.

Asik and Cephe (2013) compared the native and non-native speakers of English in terms of their use of discourse markers in spoken English. They compiled two corpora from native and non-native speakers of English using transcripts of student presentations. They defined the frequencies of discourse markers in these two corpora and reported that non-native speakers used a limited number of discourse markers with less variety than native speakers. They suggested that awareness should be raised on the use of discourse markers during the practice of English Language Teaching.

In another study, Peksoy and Harmaoglu (2017) used the BNC as reference corpus to study the similarity of the language used in textbooks to the language spoken by native speakers of English. They scanned all coursebooks used at high schools in Turkey and compared these with the spoken part of the BNC. They reported that coursebooks didn't have adequate resemblance with the authentic language in terms of some grammatical structures and their collocations. Based on these findings, they suggested the use of corpus-based techniques to revise existing materials or to write new materials.

Sert (2009), compiled a 90.000-word corpus from a British TV show in order to investigate potential effects of using TV series in language classroom on the Interactional Competence (IC) of language students. His findings indicate that using TV series in the language classroom can contribute to learners at a great extent by exposing them to multi-model texts.

CHAPTER 3

METHODOLOGY

The purpose of the present study is to find out students' preferences regarding watching TV series, to form a comparatively small-scale TV series corpus and compare it to the spoken part of the BNC. The British TV Series Corpus (BTSC) was formed for the present study and consisted of two British TV series (Doctor Who and Sherlock) that were selected based on student preferences and was intended to find out the extent to which the students' favourite TV series reflect the real spoken language in terms of vocabulary used and to identify whether they can be used as effective materials that might be used for extra-curricular speaking and vocabulary activities.

This part presents information about the setting, participants, instruments of the present study along with data collection and data analysis procedures.

3.1. Setting

The student questionnaires developed to find out about the types of extra-curricular activities students did was conducted at Selcuk University School of Foreign Languages in 2017-2018 Academic Year, on English Preparatory Class students.

3.2. Participants

The participants of the student questionnaire were English Preparatory Class students, who were registered to English Translation and Interpretation, English Language and Literature, International Relations and Business Administration Departments of Selcuk University. The total of 132 students participated voluntarily in the study; 72 (55%) of which were female, while 60 (45%) were male. Almost half (n=69, 52%) of the participants were enrolled in English Language related departments (English Language and Literature and English Translation and Interpretation) while the majors of the rest (n=63, 48%) weren't directly related with language studies (International Relations and Business Administration).

3.3 Instruments

3.3.1 Student Questionnaires: The first instrument used to collect data was the student questionnaires, and it was formed by the researcher to have information on the English language-related extra-curricular activities of students. Two experts of English Language Teaching were consulted for their opinions on the questionnaire. The main purpose of the questionnaire was selecting sources for the corpus to be compiled for the present study. Accordingly, the items questioned students' habits of reading, listening and watching materials in English, and in specific which movies, TV series or shows they watched, and their beliefs related to the contributions of these to the development of their linguistic skills. The first part of the questionnaire related to the personal information of the participants was included only to obtain and provide information about the gender and departments participants, and data collected from this part weren't included in the analyses. The student questionnaire form can be found in Appendix 1 and Appendix 2 in both Turkish (original) and English languages.

3.3.2 Comparison lists: As mentioned before, the purpose of the present study is to determine the extent to which the language used in TV series reflect the real-life spoken English. In order to do so, a corpus was compiled using scripts from two British TV series (BTSC) to be compared with the spoken part of the BNC, which is considered as one of the most reliable sources of the British English.

3.3.3 The British National Corpus (BNC): 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).

3.3.4 The British Television Series Corpus (BTSC): a 754378-word corpus compiled from the scripts of all aired episodes of two British TV series, Sherlock and Doctor Who, which were selected based on students' preferences.

3.4. Data Collection

3.4.1 Student Questionnaires

The student questionnaires were printed on paper, and distributed to English Preparatory Class students, who studied at Selçuk University School of Foreign Languages in the 2017-2018 academic year. Before they started filling in the forms, students were informed about the purpose of the questionnaire orally, which was also provided on the print-outs in written format. After the informing procedure, the total of 132 students voluntarily participated in the study. While they filled in the forms, the researcher was present in their classrooms and answered their questions about any points they had difficulty in understanding. Accordingly, all of the questionnaire forms were filled in properly and could be included in the research.

3.4.2 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 6 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 6 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 6, 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.

3.4.3 Developing The British Television Series Corpus

The TV series included in the British Television Series Corpus were selected based on student preferences, which were defined through the student questionnaires. In the second part of the questionnaire, the participants were asked about the extra-curricular activities they did. The findings obtained from this part revealed that watching TV series in English was the most favoured activity by the participants. In the fourth part of the questionnaire, the participants were requested to list the specific shows, movies, etc. they watched. Every response was listed and the mostly preferred two British TV shows were selected to be included in the corpus. These 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.

Table 7 Number of words and percentage for each season of Sherlock

SEASON	N OF EPISODES	N OF WORDS	%BTSC
1	3	24338	3,2262
2	3	25818	3,4224
3	3	27984	3,7095
4	3+1 SPECIAL EPISODE	42629	5,6508
	TOTAL	**120769**	**16,0090**

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. A more detailed Table presenting the word counts and percentages for each episode can be found in Appendix 3.

Table 8 Number of words and percentage for each season of Doctor Who

SEASON	N OF EPISODES	N OF WORDS	%
1	13	59185	7,8455
2	15	64076	8,4938
3	13	61541	8,1578
4	13+3 SPECIAL EPISODES	77305	10,2475
5	13	61540	8,1577
6	14	65507	8,6835
7	13+1 SPECIAL EPISODE	61108	8,1004
8	12+1 SPECIAL EPISODE	70318	9,3213
9	12	53770	7,1277
10	12+1 SPECIAL EPISODE	59259	7,8553
	TOTAL	**633609**	**83,9909**

The total of 633609 words spoken in 136 episodes of Doctor Who were included in the BTSC, which make up around 84% of the whole corpus. A more detailed Table presenting the word counts and percentages for each episode can be found in Appendix 4. Table 9 below presents the percentage of two TV series in the BTSC.

Table 9 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 has three main functions:

a) Deleting some of the unwanted words automatically

b) Excluding the list of words prepared beforehand.

c) Organizing words as word per line format (Dolmaci, 2015).

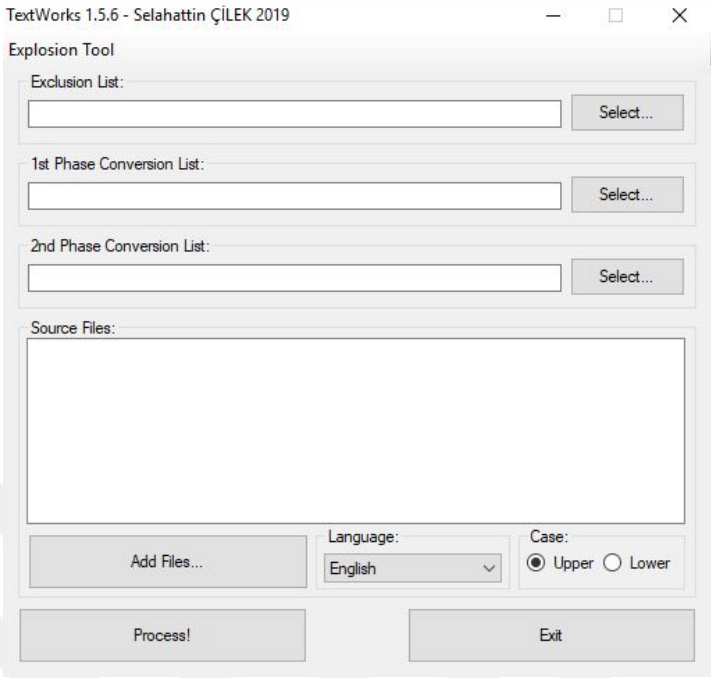

Figure 1 The interface of Textworks 1.5.6

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. The list of misspelt words, proper names, contractions, exclamations and abbreviations could be found in Appendix 5, 6, 7, 8, and 9. Table 10 below shows the distribution of the words included in and excluded from BTSC.

Table 10 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**

Contractions formed with apostrophe (‘ve, ‘s, and ‘d) listed under “*other*” category were also excluded from the BTSC, and Table 11 below presents the frequencies of these contracted forms in a more detailed form, which were presented as “Other” in Table 10 above.

Table 11 Frequencies of the Contracted forms ‘ve, ‘s, and ‘d

Contracted form	**Frequency**
‘ve	3673
‘s	17174
‘d	1120
TOTAL	**21967**

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 12 below. The whole list of function words can be found in Appendix 10.

Table 12 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.

3.5 Data analysis

3.5.1 Student questionnaires

The primary purpose of the student questionnaire was to define the most favoured extra-curricular activities done by the students. Therefore, the items included in the questionnaire simply questioned whether the students did or didn't do the activities included in the questionnaire. The questionnaire allowed multiple selections if students did more than one of the activities included in the questionnaire form. After the application, the responses of the students were input to Microsoft Excel software, and percentages of students doing each activity were calculated on the same software.

3.5.2 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, and different parts of speech. 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.

CHAPTER 4

RESULTS AND DISCUSSION

The purpose of the present study is to find out students' preferences regarding the types of materials they use for their extra-curricular activities, to form a comparatively small-scale corpus and to compare it to the spoken part of the BNC. Additionally, students' beliefs related to the contribution of watching videos in English to the development of their speaking and listening skills, vocabulary and language use are investigated. The British TV Series Corpus (BTSC) was formed for the present study and consisted of two British TV series (Doctor Who and Sherlock), which were selected based on student preferences, and it was intended to find out the extent to which the students' favourite 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.

In accordance with this purpose, the research questions were formed as follows:

1) Which types of materials do the students studying English use for their extra-curricular activities?
2) What are students' favourite genres for movies and TV series?
3) What are students' beliefs related to the contribution of watching movies, TV shows and series in English to the development of their
    (a) speaking skills,
    (b) listening skills,
    (c) vocabulary,
    (d) language use?
4) To what extent does the BTSC cover the items in the BNC spoken frequency lists?
5) Is there a significant relationship between the spoken part of the BNC and the BTSC in terms of frequency of the items?

6) Are there any similarities between the BNC and the BTSC in terms of the most frequent 20

   (a) words,
   (b) function words,
   (c) nouns,
   (d) verbs,
   (e) adjectives,
   (f) adverbs?

4.1 Findings on the Student Questionnaires

In order to answer the first research question, English preparatory class students were provided with a questionnaire questioning the extra-curricular activities they did in order to develop their language skills. The questionnaire consists of seven items. The first is about the types of extra-curricular activities the students do, the second is about the genre of the videos, the third is on the specific videos (TV series or movies) they watch, and finally fourth, fifth, sixth, and seventh questions are about students' beliefs related to the contribution of the extra-curricular activities to the development of their language skills and areas.

Table 13 Extra-curricular activities done by participants

Activity		**n**	**%**
Reading	Magazine	21	15.9
	Newspaper	22	16.6
	Short story	67	50.75
	Novel	30	22.7
Listening	Radio show	16	12.12
	Songs	117	88.6
Watching	TV series	96	72.7
	Documentary	57	43.2
	Movie	94	71.2
	TV show	56	42.4

Table 13 above shows the percentages of the students doing the extra-curricular activities mentioned in the questionnaire. Accordingly, 21 (15.9%) of the participants reported that they read magazines, 22 (16.6%) read newspapers, 67 (50.75%) read short stories and 30 (22.7%) read novels in English. As for listening, 16 (12.12%) of the participants reported that they listened to radio-shows and 117 (88.6%) listened to songs in English. As for the videos, which were chosen as the target data resource for the present study, 96 (72.7%) of the participants reported that they watched TV series, 57 (43.2%) watched documentaries, 94 (71.2%) watched movies and 56 (42.4%) watched TV shows in English. Based on the findings presented in the Table 13 above, TV series were chosen as the resource of the data for the corpus compiled for the present study, as watching TV series was among the most favoured activities for the students. Listening to songs in English is the activity that was reported to be done by the highest number of students. However, since songs utilize a special language and limited in terms of vocabulary items compared to TV series, TV series were preferred to songs.

Table 14 The genres of the videos (TV series and movies) preferred by participants

Genre	**n**	**%**
Action	71	53.78
Animation	33	25
Documentary	49	37.12
Drama	34	25.75
Fantasy	71	53.78
Horror	23	17.42
Adventure	73	55.3
Romance	24	18.18

As presented in Table 14 above, the most preferred genres for movies and TV series by the participants are adventure (n=73, 55.3%), action (n=71, 53.78%), and fantasy (n=71, 53.78%). Accordingly, two British TV series selected as the sources of the British TV Series Corpus compiled for the present study represent these three genres. Sherlock, which is a crime related drama can also be included in the genre of adventure, while Doctor Who, which is about time-travelling is included in the genres

of fantasy, action and also adventure. Additionally, these two shows were the two mostly preferred British TV series by the participants.

Table 15 Participants' beliefs related to the contribution of watching videos to the development of language skills and areas

Perception	n	%
1. Speaking	121	91.6
2. Listening	124	93.9
3. Vocabulary	121	91.6
4. Use	104	78.78

As mentioned above, the last part of the questionnaire was about participants' beliefs related to the contribution of these activities to their linguistic skills. 93.9% of the students stated that watching videos in the target language contributed to the development of their listening skill, 91.6% reported improvement in speaking skill, 91.6% reported improvement in their vocabulary knowledge, and finally 78.78% of students stated that they believed watching videos in the target language contributed to their use of language.

4.2 Findings on the Comparison of the BTSC with the BNC

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 16 below.

Table 16 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 16, 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 17 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 presented in Chapter 3, 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 18 below.

Table 18 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 18 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. These 37 words included in the spoken part of the BNC but not in the BTSC are presented in Appendix 11.

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 19 below.

Table 19 Results of the Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
BTSC_LEMMATIZED	,1143964232	671	,40866956121	,01577651343
BNC_SPOKEN_LEMMATIZED	,1192715350	671	,41252055024	,01592517921

As presented in Table 19, there are the total of 671 lemmas included in both lists formed with the BTSC and the spoken part of the BNC. The average frequency of the words in the BTSC list is 0.1144, while the average frequency of the words in the spoken BNC list is 0.1192. These averages for two corpora are similar, which indicates a similarity between two corpora in terms of frequency.

Table 20 Results of the Paired Samples Test

	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 20, 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 words at different levels. First of these was formed using the 20 most frequent non-lemmatized words in each corpus. Table 21 below presents the 20 most frequent non-lemmatized words in the BTSC and the spoken part of the BNC.

Table 21 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 21, 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 21 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. These filler words, excluded from the lemmatized version of the BTSC, can be found in the exclusion list provided in the Appendix 8.

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 22 below presents the 20 most frequent lemmatized words in the BTSC and the spoken part of the BNC. As presented in Table 22 below, 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 22 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

The third list was formed using the 20 most frequent function words in each corpus. Table 23 below presents the 20 most frequent function words in the BTSC and the spoken part of the BNC.

As presented in Table 23 below, 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 “no and can”. Instead of these two, “but and for” are among the 20 most frequent function words in the spoken part of the BNC. Still, it should be included here that “no” is ranked 21st and “can” is ranked 25th in the BNC function word list, while “but” is ranked 23rd and “for” is ranked 25th in the

BTSC function word list. These findings emphasize the similarity at a further level between two lists.

Table 23 Comparison of the 20 most frequent function 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	28937
5	IT	23606	AND	25210
6	A	21284	IT	24846
7	TO	19594	TO	23565
8	NOT	19090	THAT	21498
9	THAT	13892	A	20483
10	WE	13269	NOT	17272
11	AND	13142	OF	15093
12	OF	13065	THEY	13801
13	WHAT	12152	IN	12916
14	HE	10476	WE	11507
15	THEY	9795	HE	8715
16	IN	9641	ON	7508
17	THIS	8058	WHAT	7313
18	NO	7593	BUT	6366
19	ON	6902	FOR	6346
20	CAN	6775	THIS	5627

*frequency per million

As mentioned above, most of the words in the 20 most frequent words in the whole lists are mostly function words. Therefore, the lists for just function words don't vary much from those lists.

In addition to these, comparison lists were formed in terms of parts of speech. The fourth list was formed using the 20 most frequent nouns in each corpus. Table 24 below presents the 20 most frequent nouns in the BTSC and the spoken part of the BNC.

As presented in Table 24 below, seven nouns are common in both lists. These are; "time, thing, way, people, man, day and year". Number one item in the BTSC list

is “doctor”, which is not included in the BNC list. The reason for the high frequency of the item is that both TV series used to compile the BTSC feature main characters who are doctors. In the show Sherlock, main character Sherlock Holmes’s fellow in his adventures is Doctor Watson, and he is frequently addressed with his title as “Doctor”. However, the show contributing to this finding more is Doctor Who, in which the main character doesn’t have a name since it changes every season, sometimes a man sometimes a woman. For this reason, the main character in the show Doctor Who is always addressed to as “Doctor”. Moreover, as explained in Chapter 3, Doctor Who has been aired for a longer period of time, and the scripts of this show form the 84% of the BTSC. Taken this information into consideration, the item ranked first on the BTSC list being “Doctor” is not surprising.

Table 24 Comparison of the 20 most frequent nouns in the BTSC and the spoken part of the BNC

	BTSC		BNC	
RANK	**WORD**	**FREQ***	**WORD**	**FREQ***
1	DOCTOR	4797	TIME	2331
2	TIME	3473	THING	2240
3	THING	2263	PEOPLE	2071
4	WAY	1542	YEAR	1507
5	PEOPLE	1513	WAY	1358
6	MAN	1291	DAY	1064
7	SIR	1090	LOT	992
8	DAY	1086	POUND	899
9	YEAR	1054	WEEK	829
10	WORLD	1027	SORT	800
11	HUMAN	1021	BIT	747
12	LIFE	1009	POINT	747
13	GOD	976	NUMBER	638
14	NAME	843	MONEY	637
15	FRIEND	742	MAN	629
16	MINUTE	725	WORK	601
17	DOOR	709	MR	570
18	PLACE	707	PROBLEM	568
19	PLANET	705	HOUSE	557
20	EARTH	680	SCHOOL	549

*frequency per million

The show Doctor Who is also considered as the reason for the variations in other items. As mentioned before, the show is about a time traveller known as the “Doctor”, and his adventures in time and space with his friends from the planet earth. Therefore, the high frequency of the items “world, place, human, planet or earth” results from the context of this show.

The fifth list was formed using the 20 most frequent verbs in each corpus. Table 25 below presents the 20 most frequent verbs in the BTSC and the spoken part of the BNC.

Table 25 Comparison of the 20 most frequent verbs in the BTSC and the spoken part of the BNC

	BTSC		BNC	
VERB	**WORD**	**FREQ***	**WORD**	**FREQ***
1	BE	48652	BE	57016
2	DO	16642	HAVE	19689
3	HAVE	8862	DO	16621
4	GET	7680	GET	9230
5	GO	7386	KNOW	6119
6	KNOW	5603	SAY	6119
7	COME	4059	GO	5986
8	THINK	3667	THINK	5063
9	LIKE	3586	SEE	3162
10	SEE	3411	COME	3061
11	SAY	3311	WANT	2548
12	LOOK	3123	MEAN	2525
13	TELL	2471	TAKE	2018
14	NEED	2243	LOOK	1986
15	WANT	2183	MAKE	1902
16	TAKE	2049	PUT	1827
17	MAKE	1953	GIVE	1428
18	MEAN	1603	TELL	1180
19	STOP	1550	LIKE	1170
20	FIND	1520	NEED	937

*frequency per million

As presented in Table 25 above, 18 out of 20 words are included in both lists, which also present similarity in terms of their ranks in the list. Two items that are in the

BTSC but not in the BNC list are “stop and find”. The verb “stop” ranks 50 in the most common verbs list for the BNC, and the verb “find” is the 24th most frequent verb in the BNC. On the other hand, two verbs that are in the BNC list but not in the BTSC list are “put and give”, the first of which ranks 50th while the latter ranks 27th in the most frequent verbs list of the BTSC.

The sixth list was formed using the 20 most frequent adjectives in each corpus. Table 26 below presents the 20 most frequent adjectives in the BTSC and the spoken part of the BNC.

Table 26 Comparison of the 20 most frequent adjectives in the BTSC and the spoken part of the BNC

	BTSC		BNC	
	WORD	**FREQ***	**WORD**	**FREQ***
1	RIGHT	3040	RIGHT	3229
2	GOOD	2948	GOOD	2091
3	BACK	2305	OTHER	1499
4	SORRY	1829	ALRIGHT	774
5	LITTLE	1238	LITTLE	700
6	SOME	1090	SMALL	664
7	LONG	925	BIG	659
8	OLD	916	NICE	615
9	DEAD	908	NEW	606
10	LAST	858	OLD	603
11	EVERY	840	LONG	562
12	BIG	832	DIFFERENT	500
13	NEW	813	SURE	479
14	WRONG	737	SORRY	431
15	OTHER	734	GREAT	430
16	FINE	640	FAR	405
17	BAD	632	BAD	397
18	SAME	582	ABLE	339
19	WHOLE	573	LATE	337
20	ELSE	570	LOCAL	301

*frequency per million

As presented in Table 26 above, half of the adjectives, (n=10) are common in two lists, while ten items are different. The ones included in the BTSC list but not in the

BNC list are “back, some, dead, last, every, fine, wrong, same, whole and else”. The ones included in the BNC but not in the BTSC list of 20 most frequent adjectives are “alright, small, nice, different, sure, great, far, able, late and local”. Ten adjectives included in both are listed as “right, good, sorry, little, long, old, big, new, other, and bad”. It can also be observed from Table 26 that, the ranks of these adjectives are similar, especially the first two in both lists are the same and their frequencies are similar.

The seventh and last comparison list was formed using the 20 most frequent adverbs in each corpus. Table 27 below presents the 20 most frequent adverbs in the BTSC and the spoken part of the BNC.

Table 27 Comparison of the 20 most frequent adverbs in the BTSC and the spoken part of the BNC

	BTSC		**BNC**	
	WORD	**FREQ***	**WORD**	**FREQ***
1	JUST	5457	WELL	5881
2	NOW	3879	JUST	3820
3	HERE	3633	THEN	3474
4	WELL	2995	NOW	2864
5	THEN	2609	ONCE	2646
6	NEVER	1829	VERY	2373
7	VERY	1440	REALLY	1727
8	MORE	1356	HERE	1640
9	REALLY	1334	MORE	1573
10	ONLY	1311	ONLY	1323
11	STILL	1177	ACTUALLY	1239
12	ALWAYS	1022	MUCH	1196
13	AGAIN	893	QUITE	1050
14	EVER	884	AGAIN	836
15	MUCH	720	OBVIOUSLY	768
16	MAYBE	628	STILL	739
17	EXACTLY	494	NEVER	700
18	ACTUALLY	445	AROUND	660
19	AROUND	432	TOO	629
20	ONCE	411	ALWAYS	597

*frequency per million

As presented in Table 27 above, most of the adverbs, (n=17) are common in two lists, while three items are not. The ones included in the BTSC list but not in the BNC list are “ever, maybe and exactly”. The ones included in the BNC but not in the BTSC list of 20 most frequent adverbs are “quite, obviously and two”. Seventeen adverbs included in both are listed as “just, now, here, well, then, never, very, only, more, really, still, always, again, much, actually, once and around”.

CHAPTER 5

CONCLUSION

This chapter summarizes and discusses the findings of the present study and provides pedagogical implications along with limitations of the study and suggestions for further research.

5.1 Discussions

The present study was conducted in order to see what types of extra-curricular activities are preferred by the students of English as a foreign language and the extent to which these activities can reflect the real-life spoken language. In order to do so, a questionnaire was formed by the researcher to question the types of extra-curricular activities performed by students in the target language, which was the first of research questions. According to the findings obtained through the analysis of the data collected with the student questionnaires, most of the students love watching TV series and movies in English as an extra-curricular activity. 72.7% of the students watch TV series in English, while 71.2% watch movies. Accordingly, it can be claimed that using TV series in the classroom as well can provide students with an extra element of motivation by attracting their interest.

The second research question was about the students' favourite genres for movies and TV series. According to the findings obtained from the student questionnaire, the most preferred genres for movies and TV series by the participants are adventure (n=73, 55.3%), action (n=71, 53.78%), and fantasy (n=71, 53.78%). TV series as the sources of the BTSC were selected taking this finding into consideration. Additionally, it can be suggested that including video materials, such as TV series and movies of these genres as in-class materials can attract students' interest more than any other genre and can be more effective by including students more actively.

The third research question was about students' beliefs about the contribution of watching videos in English to the development of the related language skills and areas. The analysis of the data obtained from the student questionnaires showed that 93.9%

of the students believed that watching TV series and other videos in English developed their listening skills, while 91.6% reported that they believed watching TV series contributed to the development their speaking skill and vocabulary knowledge, and 78.78% stated that watching such videos in the target language contributed to their use of English. These findings also give information about the efficiency of videos as extra-curricular and also in-class materials, because students can only pay attention to any input as long as they believe they work. Students need to experience success to be motivated (Williams & Williams, 2011).

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 fourth 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 fifth 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 the average of the frequency of the lemmas was ,1143964232 for the BTSC and ,1192715350 for the BNC. Additionally, 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 words from different parts of speech were formed for both corpora. First of these was the 20 most common words 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.

The greatest difference was found in the fourth list, which was formed using the 20 most frequent nouns. Only seven of the items were common in the BTSC and the BNC. These were “time, thing, way, people, man, day and year”. Some of these words like “way” can be used with different meanings, and some like “time or day” are frequently used in common phrases, such as “next day” or “last time”. On the other hand, the difference between the most frequent nouns in the BTSC and the BNC is believed to have resulted from the specific contents of the TV series used to compile the BTSC.

The fifth, which is the list of most frequent 20 verbs, also presented similarity between the BTSC and the BNC, with 18 out of 20 common words. It is not surprising that a context-resulted difference like the difference between nouns did not occur here, as almost every verb on the list can be used with different meanings in different contexts or with different objects. For instance, the verb “take” can be used with different meanings as in “take a walk”, “take a photograph”, “take someone to somewhere”, or “take shower”. The list can go on with the verb “take”, and it can even be longer with other verbs, such as “have, do, get or make”.

The second greatest difference was observed in the sixth list, which included the 20 most frequent adjectives in the BNC and the BTSC. Only half (10 out of 20) of the adjectives were common. These 10 adjectives also have a very broad usage, as it is possible define numerous concepts as “good and bad, old and new, or big and little”.

The last of the comparison lists was formed of the most frequent 20 adverbs in the BNC and the BTSC, which presented a 17 out of 20 similarity between two corpora. According to the findings of these comparisons, it can be concluded that the BTSC and the BNC present similarities in terms of the most frequent vocabulary items, which suggests a similarity between scripted language of TV series and real-life spoken language.

5.2 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. As presented by the findings obtained from the student questionnaires, 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. As the differences found especially between the most frequent nouns show that vocabulary used in specific contexts can vary at a great extent, series or movies with specific contexts, such as law or medicine, can be used as materials for teaching of English for Specific Purposes.

5.3 Limitations of the Study

The corpus compiled for the present study is limited to two TV series. Additionally, these two series selected based on student preferences 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.

5.4 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.

REFERENCES

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425-438.

Adolphs, S., Brown, B., Carter, R., Crawford, P., & Sahota, O. (2004). Applying corpus linguistics in a health care context. *Journal of Applied Linguistics, 1*(1), 9–28.

Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., Demirhan, U., Yılmazer, H., Kurtoğlu, Ö., Atasoy, G., Öz, S. &Yıldız, İ. (2012). Construction of the Turkish National Corpus (TNC). *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012).* İstanbul.

Aksan, Y., Aksan, M., Özel, S., Yılmazer, H., Demirhan, U., Mersinli, Ü., Bektaş, Y. & Altunay, S. (2016). Web Tabanlı Türkçe Ulusal Derlemi (TUD). [Web-based Turkish National Corpus (TNC)]. *Akademik Bilişim'14- XVI. Akademik Bilişim Konferansı Bildirileri*, 723-730, 5-7 Şubat, 2014, Mersin Üniversitesi.

Anderson, W. & Corbett, J. (2010). Teaching English as a friendly language: lessons from the SCOTS corpus, *ELT Journal: English Language Teaching Journal, 64* (4), 414-423.

Anthony, L. (2013). AntWordProfiler (Version 1.4.0) [Computer Software]. Tokyo, Japan: Waseda University. Available 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.

Asik, A. & Cephe, P. T. (2013). Discourse Markers and Spoken English: Nonnative Use in the Turkish EFL Setting. *English Language Teaching, 6* (12) (144).

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.

Bal Gezegin, B. (2013). How do We Say No in Turkish?: A Corpus-Based Analysis of Hayır and Cık in Turkish. *Dil ve Edebiyat Dergisi / Journal of Linguistics and Literature, 10:2*, 53-73.

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

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

Bednarek, M. (2011). The language of fictional television: a case study of the 'dramedy' Gilmore Girls. *English Text Construction 4/1,* 54-83.

Bernardini, S. (2002). Exploring new directions for discovery learning. (Edited by: Kettemann, B., & Marko, G.). *Teaching and learning by doing corpus analysis*. Amsterdam: Rodopi, 165-182.

Biber, D. (1998). *Corpus linguistics: Investigating language structure and use*, Cambridge: Cambridge University Press.

Bonelli, E., & Sinclair, J. (2006). Corpora. (Edited by: Brown, K.). *Encyclopedia of Language and Linguistics* (2nd Edition) Oxford, United Kingdom: Elsevier, 206-219.

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

Brown Corpus Manual, Retrieved from http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM on March, 1, 2019.

Browne, C. (2013). The New General Service List: Celebrating 60 years of Vocabulary Learning. *The Language Teacher. 4. 37*, 13–16.

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 from http://www.natcorp. ox.ac.uk /docs/Burnage93a.htm#4.4 on January, 4, 2019.

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.

Cambridge English Dictionary, retrieved from https://dictionary.cambridge.org /dictionary/english/script on June, 1, 2018.

Chambers, A. (2007). Integrating corpora in language learning and teaching. *ReCALL, 19(3),* 249-251.

Chaney, A. L. & Burk, T.L. (1998). *Teaching Oral Communication in Grades K-8.* Boston: Allyn & Bacon.

Cheng, W. (2012). *Exploring corpus linguistics: language in action*. New York: Routledge.

Ciliz, M. (2010). *Building Up A Learner Corpus Through Creative Nonfiction Prose: An Experimental Research*, MA Thesis, University of Gaziantep Graduate School of Social Sciences, Gaziantep.

Cokal Karadass, D. & Ruhi, S. (2009). Features for an internet accessible corpus of spoken Turkish discourse. *Working Papers in Corpus-based Linguistics and Language Education 3*, pp. 311--320.

Corpus of Contemporary American English, Retrieved from https://www.english-corpora.org/coca/ on March, 1, 2019.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly, 34(2),* 213-238.

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

Csomay, E. & Petrovic, M. (2012). "Yes, Your Honor!": A Corpus-Based Study of Technical Vocabulary in Discipline-Related Movies and TV Shows. *System: An International Journal of Educational Technology and Applied Linguistics, 40,* 305-315.

Díaz-Cintas, J. (2009). Introduction. (Edited by: Díaz-Cintas, J., & Anderman, G.) *Audiovisual translation: Language transfer on screen.* England: Palgrave Macmillan, 1-17.

Dolmaci, M. (2015). *A Corpus Study Of Academic Vocabulary In Turkish As A Foreign Language*, PhD Thesis, Gazi University Graduate School of Educational Sciences, Ankara.

Evans, D. (2018). Unit 1: Introduction. (Edited by: de Koster, R.). *Corpus building and investigation for the Humanities: An online information pack about corpus investigation techniques for the Humanities.* Retrieved from: https://www.birmingham.ac.uk/Documents/collegeartslaw/corpus/Intro/Unit1.pdf, on March, 12, 2019.

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.

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

Frumuselu, A., De Maeyer, S., Donche, V., & Gutierrez-Colon Plana, M. (2015). Television series inside the EFL classroom: Bridging the gap between teaching and learning informal language through subtitles. *Linguistics and Education,* 2015, 1-11. http://dx.doi.org/10.1016/j.linged.2015.10.001

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System, 32*(3), 301-319.

Gilner, L. (2011). A primer on the General Service List. *Reading in a Foreign Language 23*(1), 65-83.

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.

Granger, S. (2002). A bird's-eye view of learner corpus research. (Edited by: Granger, S., Hung, J., & Petch-Tyson, S.). *Computer learner corpora, second language acquisition and foreign language teaching* Amsterdam: John Benjamins, 3-33.

Grant, L. (2011). The frequency and function of 'just' in British academic spoken English. *Journal of English for Academic Purposes 10* (3), 183-197.

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.

Hu, M. & Nation, I.S.P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language 13*(1), 403–430.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Johansson, S. (1995). Mens sana in corpora sano: On the role of corpora in linguistic research. *The European English Messenger, 4*(2), 19-25

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.

Johns, T. (1986). Micro-concord: A language learner's research tool. *System, 4*(2), 151-162.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *ELR Journal, 4*, 1-16.

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

Knoch, U. (2004). A new look at teaching comparisons: A corpus-based study. *Journal of Language and Learning, 2*(2), 171-185.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? (Edited by: Lauren, C., & Nordman, M.). *Special language: from humans thinking to thinking machines* Bristol, UK: Multilingual Matters, 316-323.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? (Edited by: Bejoint, H., & Arnaud, P.). *Vocabulary and Applied Linguistics*. New York: Macmillan, 126-132.

Law, L. (2015). House MD Corpus Analysis: A linguistic intervention of contemporary American English. *Proceedings of AsiaLex 2015*. 2015. Hong Kong: Hong Kong Polytechnic University, 230-249.

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

McCarthy, M. (2004). *From corpus to classroom*. Cambridge: Cambridge University.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies. An advanced resource book.* London: Routledge.

Meunier, F. (2011). Corpus linguistics and second / foreign language learning: exploring multiple paths. *RBLA, Belo Horizante, 11*(2), 459–477.

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.

Mitkov, R. (2005). Introduction. (Edited by: Mitkov, R.) *The Oxford handbook of computational linguistics.* Oxford: Oxford University Press, xvii-xx.

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

Nation, I. S. P. (2006). Second language vocabulary. (Edited by: Brown, K.). *Encyclopedia of Language and Linguistics* (2nd Edition), Oxford: Elsevier, 448-454.

Nation, P., & Meara, P. (2002). Vocabulary (Edited by: Schmitt, N.). *An Introduction to Applied Linguistics.* London: Arnold, 35-54.

Nesselhauf N. (2004). Learner corpora and their potential for language teaching. (Edited by: Sinclair, J.). *How to use corpora for language teaching.* Amsterdam & Philadelphia: John Benjamins, 125-152.

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

Oxford English Dictionary, Retrieved from https://en.oxforddictionaries.com/definition/word on March, 28, 2019.

Peksoy, E. & Harmaoglu, O. (2017). Corpus Based Authenticity Analysis of Language Teaching Course Books. *International Journal of Languages' Education and Teaching, 5* (4), 287-307.

Richards, J. C. (1974). Word lists: Problems and prospects. *RELC*, *5*(2), 69–84.

Richards, J. C., & Schmidt, R. (2002). *Longman Dictionary of Language Teaching and Applied Linguistics* (Third Edition). Pearson Education Limited. Essex: England.

Römer, U. (2006). Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift für Anglistik und Amerikanistik, 54*(2), 121-134.

Römer, U. (2008). Corpora and language teaching. (Edited by: Lüdeling, A., & Kytö, M.). *Corpus Linguistics: An International Handbook.* Berlin, Germany: De Gruyter, Volume 1, 112-130.

Ruhi, S., Eroz-Tuga, B., Hatipoglu, C., Isik-Guler, H., Acar, M. G. C., Eryilmaz, K., Can, H., Karakas, O., & Cokal Karadas, D. (2010). Sustaining a corpus for spoken Turkish discourse: Accessibility and corpus management issues. *In Proceedings of the LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management.* Paris: ELRA.

Sanal, F. (2007). *A learner corpus based study on second language lexicology of Turkish students of English*. Phd Dissertation, Çukurova University Institute of Social Sciences, Adana.

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.

Schmitt, N. (2000). *Vocabulary in Language Teaching.* Cambridge: Cambridge University Press.

Schonell, F. J., Meddleton, I. G., & Shaw, B. A. (1956). *A Study of the oral vocabulary of adults*. Brisbane: University of Queensland Press.

Sert, O. (2009). Developing Interactional Competence by Using TV Series in "English as an Additional Language" Classrooms. *Enletawa Journal, 2,* 23-50.

Sinclair, J. (2004). *How to use corpora in language teaching.* Amsterdam: John Benjamins.

Spoken Turkish Corpus, Retrieved from https://std.metu.edu.tr/en/hakkinda/, on March, 22, 2019.

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

Strik, H., Hulsbosch, M. & Cucchiarini, C. (2010) Analyzing and identifying multiword expressions in spoken language. *Lang Resources & Evaluation, 44*, 41-58. https://doi.org/10.1007/s10579-009-9095-y

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

Talaván, N. Z. (2007). Learning vocabulary through authentic video and subtitles. *TESOL-SPAIN Newsletter, 31,* 5–8.

Tang, W. M. (2015). *A corpus linguistic glossary*. Retrieved from https://wmtang.org/corpus-linguistics/glossary/ on 22 March 2010

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 from http://www.natcorp.ox.ac.uk/corpus/index.xml, on March, 22, 2019.

The Collins Corpus, Retrieved from https://collins.co.uk/pages/elt-cobuild-reference-the-collins-corpus on March, 1, 2019.

The Open American National Corpus, Retrieved from http://www.anc.org on March, 1, 2019.

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

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

Webb, S., & Rodgers, M.P.H. (2009a). The lexical coverage of movies. *Applied Linguistics*, *30*, 407-427.

Webb, S., & Rodgers, M.P.H. (2009b). The vocabulary demands of television programs. *Language Learning*, *59*, 335-366.

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

Williams, K., & Williams, C. (2011). Five key ingredients for improving motivation. *Research in Higher Education Journal,* 11.

Xiao, Richard. (2008). Well-known and influential corpora. (Edited by: Lüdeling, A., & Kytö, M.). *Corpus Linguistics: An International Handbook.* Berlin, Germany: De Gruyter, Volume 1, 383-457.

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.

Zahra, T. & Abbas, A. (2018). Pedagogical Implications of Corpus-Based Approaches to ELT in Pakistan. *Journal of Education and Educational Development*, *5* (2), 259-275.

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

APPENDICES

Appendix 1-Student Questionnaire Form (Original Version in Turkish)

Sevgili öğrenciler;

Bu anket İngilizceyi yabancı dil olarak öğrenenlerin yabancı dil becerilerini geliştirmek adına yaptığı ders dışı etkinleri değerlendirmek üzere yapılacak çalışmada veri toplamak üzere uygulanmaktadır. Toplanan veriler ve kişisel bilgileriniz yalnızca bilimsel amaçlarla kullanılacak, hiçbir şekilde ticari kuruluşlarla paylaşılmayacaktır.

Katkılarınız için teşekkür ederim.

Öğr. Gör. Hatice SEZGİN

Cinsiyet:
- ☐ Erkek
- ☐ Kadın

Bölüm:
- ☐ İngiliz Dili ve Edebiyatı
- ☐ İşletme
- ☐ Mütercim-Tercümanlık İngilizce
- ☐ Uluslararası İlişkiler
- ☐ Diğer (Lütfen belirtiniz) _____

1. **Hedef dilde (İngilizce) becerilerinizi geliştirmek için yaptığınız ders dışı etkinlikleri işaretleyiniz.**

☐ **İngilizce orijinal metinler okuma**

- ☐ Dergi
- ☐ Hikaye
- ☐ Gazete
- ☐ Roman
- ☐ Diğer (Lütfen belirtiniz) _____

☐ **Dinleme**

- ☐ İngilizce radyo programı
- ☐ İngilizce şarkı
- ☐ Diğer (Lütfen belirtiniz) _____

☐ **Video**

- ☐ İngilizce dizi
- ☐ İngilizce belgesel
- ☐ İngilizce film
- ☐ İngilizce TV programı (Talk Show)
- ☐ Diğer (Lütfen belirtiniz) _____

2. **İngilizce olarak izlediğiniz dizi ve filmler arasından en sevdiğiniz türler nelerdir?**

☐ Aksiyon	☐ Fantastik
☐ Animasyon	☐ Korku
☐ Belgesel	☐ Macera
☐ Dram	☐ Romantik

☐ Diğer (Lütfen belirtiniz) _____

3. **Lütfen şu ana kadar izlemiş olduğunuz veya halen izledikleriniz içinden en sevdiğiniz İngilizce dizi, film ve programları belirtiniz.**

Dizi: _____
_____
Film: _____
_____
TV programı: _____
_____

4. İngilizce olarak izlediğiniz dizi, film ve programların konuşma becerilerinizin gelişmesinde etkili olduğunu düşünüyor musunuz?
   ☐ Evet ☐ Hayır
5. İngilizce olarak izlediğiniz dizi, film ve programların dinleme becerilerinizin gelişmesinde etkili olduğunu düşünüyor musunuz?
   ☐ Evet ☐ Hayır
6. İngilizce olarak izlediğiniz dizi, film ve programların İngilizce kelime bilginizin gelişmesinde etkili olduğunu düşünüyor musunuz?
   ☐ Evet ☐ Hayır
7. İngilizce olarak izlediğiniz dizi, film ve programların İngilizceyi kullanımınızın gelişmesinde etkili olduğunu düşünüyor musunuz?
   ☐ Evet ☐ Hayır

Appendix 2- Student Questionnaire Form (English Version)

Dear students;

The present questionnaire is implemented in order to collect data for the research to be conducted in order to evaluate extra-curricular activities done by learners of English as a foreign language to develop language skills. Collected data and personal information will only be used for scientific purposes and won't be shared with businesses by any means.

Thank you for your contribution.

Inst. Hatice SEZGİN

Gender:
- ☐ Male
- ☐ Female

Department:
- ☐ English Language and Literature
- ☐ Business Administration
- ☐ Translation and Interpretation-English
- ☐ International Relations
- ☐ Other (Please specify) _____

1. **Please mark the extra-curricular activities you do to develop your skills in the target language (English).**

☐ **Reading authentic texts in English**
- ☐ Magazine
- ☐ Short story
- ☐ Newspaper
- ☐ Novel
- ☐ Other (Please specify) _____

☐ **Listening**
- ☐ Radio programs in English
- ☐ Songs in English
- ☐ Other (Please specify) _____

☐ **Video**
- ☐ TV series
- ☐ Documentary
- ☐ Movie
- ☐ TV programs (Talk Show)
- ☐ Other (Please specify) _____

2. **What are your favourite genres for the TV series and movies you watch in English?**

☐ Action	☐ Fantastic
☐ Animation	☐ Horror
☐ Documentary	☐ Adventure
☐ Drama	☐ Romance

☐ Other (Please specify) _____

3. **Please specify your favourites among the TV series, movies and shows you have watched or are currently watching.**

Series: _____
_____
Movies:
_____
_____
TV show: _____
_____

4. Do you think the TV series, movies and other TV shows you watch in English contribute to the development of your speaking skills?
   ☐ Yes ☐ No
5. Do you think the TV series, movies and other TV shows you watch in English contribute to the development of your listening skills?
   ☐ Yes ☐ No
6. Do you think the TV series, movies and other TV shows you watch in English contribute to the development of your vocabulary?
   ☐ Yes ☐ No
7. Do you think the TV series, movies and other TV shows you watch in English contribute to the development of your use of English?
   ☐ Yes ☐ No

Appendix 3- Number of types and tokens for each episode of Sherlock

SEASON	EPISODE	TYPE	TOKEN	TTR	%BTSC	SEASON TOTAL	%
1	1	1183	6881	5,81656805	0,91214219		
	2	1470	7579	5,15578231	1,00466875		
	3	1785	9878	5,53389356	1,30942313	24338	3,22623406
2	1	1560	9037	5,79294872	1,19794056		
	2	1480	8506	5,7472973	1,12755144		
	3	1578	8275	5,24397972	1,09693019	25818	3,42242218
3	1	1527	8059	5,27766863	1,06829733		
	2	1811	9947	5,49254555	1,31856974		
	3	1678	9978	5,94636472	1,32267908	27984	3,70954614
4	S	1849	10405	5,62736614	1,379282		
	1	1812	10407	5,74337748	1,37954712		
	2	1717	11069	6,44670938	1,46730154		
	3	1714	10748	6,27071179	1,42474993	42629	5,65088059
	TOTAL		**120769**		**16,009083**	**120769**	**16,009083**

Appendix 4- Number of types and tokens for each episode of Doctor Who

SEASON	EPISODE	TYPE	TOKEN	TTR	%BTSC	SEASON TOTAL	%
1	1	820	3316	4,04390244	0,43956743	59185	7,84553632
	2	1051	4281	4,07326356	0,56748739		
	3	1045	5049	4,83157895	0,66929311		
	4	1025	4527	4,41658537	0,60009703		
	5	1091	4766	4,36846929	0,63177876		
	6	949	4397	4,63329821	0,58286429		
	7	1090	4836	4,43669725	0,64105793		
	8	781	4108	5,25992318	0,54455459		
	9	865	4017	4,64393064	0,53249167		
	10	1007	5142	5,10625621	0,68162115		
	11	1195	5280	4,41841004	0,69991437		
	12	1117	5291	4,73679499	0,70137252		
	13	859	4175	4,86030268	0,55343608		
2	1	1049	4754	4,53193518	0,63018805	64076	8,49388503
	2	1088	4556	4,1875	0,60394126		
	3	1046	4554	4,35372849	0,60367614		
	4	937	4744	5,06296692	0,62886245		
	5	1048	4828	4,60687023	0,63999745		
	6	918	4463	4,86165577	0,59161322		
	7	979	4452	4,54749745	0,59015507		
	8	963	4473	4,64485981	0,59293882		
	9	964	5015	5,20228216	0,66478609		
	10	1026	5136	5,00584795	0,68082579		
	11	1016	5207	5,125	0,69023752		
	12	1025	4708	4,59317073	0,62409031		
	13	916	4289	4,68231441	0,56854786		
	14	453	1433	3,16335541	0,18995782		
	15	474	1464	3,08860759	0,19406717		
3	1	1044	4694	4,49616858	0,62223448	61541	8,1578466
	2	1186	5195	4,38026981	0,68864681		
	3	1040	5524	5,31153846	0,73225889		
	4	975	4699	4,81948718	0,62289727		
	5	938	4654	4,96162047	0,6169321		
	6	934	4389	4,69914347	0,58180382		
	7	864	4098	4,74305556	0,54322899		

	8	982	4695	4,78105906	0,62236704		
	9	956	4579	4,78974895	0,60699013		
	10	794	4192	5,27959698	0,55568959		
	11	985	4861	4,93502538	0,64437192		
	12	1097	5243	4,77939836	0,69500966		
	13	996	4718	4,73694779	0,62541591		
4	1	930	4019	4,32150538	0,53275679	77305	10,2475152
	2	1082	5190	4,79667283	0,68798401		
	3	1044	4644	4,44827586	0,6156065		
	4	1206	5005	4,15008292	0,66346049		
	5	1001	4416	4,41158841	0,58538292		
	6	962	4449	4,62474012	0,58975739		
	7	1216	5411	4,44983553	0,71727967		
	8	905	4742	5,23977901	0,62859733		
	9	878	4933	5,61845103	0,65391621		
	10	989	5875	5,94034378	0,77878729		
	11	1066	5426	5,09005629	0,71926806		
	12	1047	4537	4,33333333	0,60142263		
	13	1238	5856	4,73021002	0,77626866		
	S1	1320	6809	5,15833333	0,9025979		
	S2	371	988	2,66307278	0,13096882		
	S3	1040	5005	4,8125	0,66346049		
5	1	1036	6162	5,94787645	0,81683188	61540	8,15771404
	2	1017	4389	4,31563422	0,58180382		
	3	992	4307	4,34173387	0,57093393		
	4	972	4881	5,02160494	0,64702311		
	5	878	5249	5,97835991	0,69580502		
	6	909	4026	4,4290429	0,5336847		
	7	937	4892	5,22091782	0,64848127		
	8	964	4353	4,51556017	0,57703167		
	9	997	4868	4,88264794	0,64529984		
	10	949	4407	4,64383562	0,58418989		
	11	1000	4897	4,897	0,64914406		
	12	937	4112	4,38847385	0,54508482		
	13	933	4997	5,35584137	0,66240002		
6	1	943	4551	4,82608696	0,60327846	65507	8,68357773
	2	949	4788	5,04531085	0,63469507		
	3	971	4108	4,23069001	0,54455459		
	4	948	5392	5,68776371	0,71476103		

	5	943	3939	4,17709438	0,52215202		
	6	962	4678	4,86278586	0,62011352		
	7	1040	4989	4,79711538	0,66133954		
	8	918	4608	5,01960784	0,61083436		
	9	771	3903	5,06225681	0,51737988		
	10	885	4336	4,89943503	0,57477816		
	11	975	4453	4,56717949	0,59028763		
	12	1038	5242	5,05009634	0,6948771		
	13	1030	5109	4,96019417	0,67724669		
	14	993	5411	5,44914401	0,71727967		
7	1	920	4445	4,83152174	0,58922715	61108	8,10044832
	2	1160	5003	4,31293103	0,66319537		
	3	902	3716	4,11973392	0,49259125		
	4	1120	4592	4,1	0,6087134		
	5	782	3742	4,78516624	0,4960378		
	6	816	4208	5,15686275	0,55781054		
	7	777	3445	4,43371943	0,45666761		
	8	902	3751	4,15853659	0,49723083		
	9	888	3779	4,25563063	0,5009425		
	10	897	4071	4,53846154	0,53964988		
	11	976	3695	3,78586066	0,4898075		
	12	1086	5153	4,74493554	0,68307931		
	13	893	4549	5,09406495	0,60301334		
	S	1165	6959	5,97339056	0,92248183		
8	1	1306	7376	5,64777948	0,97775916	70318	9,32132167
	2	989	4831	4,88473205	0,64039513		
	3	1114	4686	4,2064632	0,621174		
	4	853	4839	5,67291911	0,64145561		
	5	953	4417	4,63483736	0,58551548		
	6	1080	5584	5,17037037	0,74021247		
	7	956	5051	5,2834728	0,66955823		
	8	1193	6135	5,1424979	0,81325277		
	9	910	4938	5,42637363	0,654579		
	10	1072	5284	4,92910448	0,7004446		
	11	912	4717	5,17214912	0,62528335		
	12	1153	5451	4,72766696	0,72258205		
	S	1113	7009	6,29739443	0,92910981		
9	1	950	3962	4,17052632	0,52520089	53770	7,12772642
	2	969	4580	4,72652219	0,60712269		

	3	989	4869	4,9231547	0,6454324		
	4	841	3842	4,56837099	0,50929375		
	5	1007	4596	4,56405164	0,60924364		
	6	1114	5285	4,74416517	0,70057716		
	7	875	3765	4,30285714	0,49908666		
	8	854	4451	5,21194379	0,59002251		
	9	987	4726	4,78824721	0,62647638		
	10	982	5107	5,200611	0,67698157		
	11	689	2959	4,2946299	0,39224368		
	12	985	5628	5,71370558	0,74604509		
10	1	819	4352	5,31379731	0,57689911	59259	7,85534573
	2	1051	5190	4,93815414	0,68798401		
	3	963	4350	4,51713396	0,576634		
	4	919	4227	4,59956474	0,56032917		
	5	1047	4771	4,55682904	0,63244156		
	6	1026	4228	4,1208577	0,56046173		
	7	910	4534	4,98241758	0,60102495		
	8	994	4537	4,56438632	0,60142263		
	9	997	3952	3,96389168	0,5238753		
	10	960	4867	5,06979167	0,64516728		
	11	849	4132	4,86690224	0,54773602		
	12	929	4811	5,17868676	0,63774394		
	S	1063	5308	4,99341486	0,70362603		
		TOTAL	**633609**		**83,990917**	**633609**	**83,990917**

Appendix 5- List of Misspelt Items Excluded from the BTSC

	ITEM	FREQ		ITEM	FREQ		ITEM	FREQ
1	ADDIC	24	154	HOWIT	1	307	RUHE	1
2	AFAMILY	1	155	HUGEWAVE	1	308	S	294
3	AFRIC	1	156	HUMANSAND	1	309	SA	4
4	AGENTIN	1	157	HUMANZ	1	310	SAVEJOHN	1
5	AIIEE	1	158	ICAN	1	311	SAVINGTHE	1
6	AJELLYFISH	1	159	IDON	2	312	SAYJEALOUSY	1
7	AJOURNALIST	1	160	IES	1	313	SCAREA	1
8	ALIVEAND	1	161	IFJAMES	1	314	SCIENCEY	1
9	ALWAYSA	1	162	IGHT	1	315	SCO	2
10	AME	1	163	IJUST	18	316	SCOOTORI	1
11	AMENHOTEP	1	164	IL	4	317	SE	4
12	AMM	1	165	ILOOK	1	318	SEARCHEA	1
13	AMUNDO	1	166	INA	1	319	SHEJUST	1
14	ANDITCHANGED	1	167	INDO	1	320	SHORTSURPRISE	1
15	ANO	1	168	ING	1	321	SHUTETTY	1
16	ANORGAN	1	169	INGERPRINZ	1	322	SILVERTHING	1
17	APPROXING	1	170	INON	1	323	SJUST	10
18	AR	1	171	INTERPOLTHINK	1	324	SJUSTJEALOUS	1
19	ARRIV	1	172	INTRU	1	325	SKALA	1
20	ASSIM	1	173	INTRUD	1	326	SLARG	1
21	ATTEN	1	174	IPPING	1	327	SLEEPMEN	2
22	ATTENED	1	175	IRTING	1	328	SLICKEEN	1
23	ATYOU	1	176	ISJUST	2	329	SLIMEBAIT	1
24	AU	1	177	ISKIND	1	330	SOAS	1
25	AUVERS	1	178	ITHING	1	331	SONT	1
26	B	25	179	ITHINK	3	332	SOORY	1
27	BABA	1	180	ITJUST	2	333	SORTOF	1
28	BACKSIES	1	181	ITJUSTIFIES	1	334	SPOILTHINGS	1
29	BACKSTREET	1	182	ITSELFAT	1	335	SSERWHEN	1
30	BAKELITE	1	183	IWAS	1	336	STATTS	1
31	BANA	2	184	IWOULD	1	337	STEALINGA	1
32	BEARA	2	185	IX	1	338	STEINO	1
33	BEINCINERATED	1	186	IZ	4	339	STINKIN	1

34	BELOWTHE	1	187	J	2	340	STRA	1
35	BI	2	188	JA	2	341	SUBO	1
36	BIGTROUBLE	1	189	JAC	1	342	SUPERFICIALJOB	1
37	BLO	1	190	JAGGIT	1	343	SUPPOSESHE	1
38	BOL	2	191	JAST	1	344	SUPRE	1
39	BOM	38	192	JUSTIFYALL	1	345	SUR	1
40	BOYO	1	193	K	30	346	T	951
41	BRAINL	1	194	KABEL	1	347	TALKINGTO	1
42	BTSIX	2	195	KATAA	1	348	TER	1
43	BUTTHEN	1	196	KEPTSHOUTING	1	349	TERRABERSERKER	1
44	BUTYOU	1	197	KEPTTHE	1	350	TEXAN	1
45	BYTHE	1	198	KIDDIN	1	351	THAFS	1
46	CAESOFINE	2	199	KNO	2	352	THATES	1
47	CALLEDFOR	1	200	KNOCKINGAROUND	1	353	THATJUSTIFIES	1
48	CANBUT	1	201	KO	4	354	THATS	1
49	CATRIGAN	1	202	L	17	355	THEGANGERS	1
50	CE	1	203	LELP	1	356	THEMS	1
51	CENTREOF	1	204	LIFEB	1	357	THEREIS	1
52	CH	1	205	LIFTTHEMSELVES	1	358	THESKY	1
53	CHARING	1	206	LITTE	1	359	THINKIN	1
54	CHEAM	1	207	LLJUST	1	360	THINKL	2
55	CHIEN	1	208	LLY	1	361	THINKTHAT	1
56	CI	2	209	LONGERIF	1	362	THISJOKE	1
57	CLAAR	1	210	LUCKNOW	1	363	THOUGHTL	1
58	CLICH	1	211	LYINGTO	1	364	THWAITES	1
59	CLONEFEED	3	212	M	85	365	TIL	2
60	CLONEWORLD	1	213	MAEVE	1	366	TILLL	1
61	CN	1	214	MAKEJOKES	1	367	TJUST	3
62	CO	30	215	MALIN	1	368	TKNOW	1
63	COMPUTERDOESN	1	216	MARKETIS	1	369	TOJOHN	1
64	CONNECZ	1	217	MARTH	1	370	TOJUST	1
65	CORRU	1	218	MARTINO	1	371	TOMMYROT	1
66	COULDBE	1	219	MATIC	1	372	TOTEN	1
67	COULDN	9	220	MESELF	1	373	TOV	1
68	COULDYOU	1	221	METHROUGH	1	374	TRANSPA	1
69	COUNZ	1	222	MEUNDER	1	375	TRI	1
70	COVERCOMPROMISED	1	223	MI	11	376	TROUGHTON	1

71	CRABTREE	1	224	MIGHTJUST	2	377	TRU	3
72	CREDITJUNKIE	1	225	MILAYA	1	378	TU	3
73	CURRENCYL	1	226	MINERA	1	379	TWELVEMONTH	1
74	CUTOUT	1	227	MIVVIES	1	380	U	21
75	D	472	228	MOSTFEARED	1	381	UGM	1
76	DAEMOS	1	229	MU	1	382	UKSUBTITLES	24
77	DALEKED	1	230	MUCHAS	1	383	UNG	1
78	DAVINADROID	1	231	MULTO	1	384	UNIVERSEIS	1
79	DEEPCOLD	1	232	MUSG	1	385	UNS	1
80	DEFOO	1	233	MUTUALFRIEND	1	386	UNTILJOHN	1
81	DESEEN	1	234	MYBONES	1	387	USH	1
82	DESES	1	235	MYJOB	1	388	UTAN	1
83	DESII	1	236	N	48	389	UTIFU	1
84	DIAMONDFALL	1	237	NAR	1	390	UTTAR	1
85	DIDIT	1	238	NEC	1	391	UTTERSPLATTERING	1
86	DIFFICULZ	1	239	NEUTRALINO	1	392	UX	2
87	DIMEN	1	240	NEVERJUDGE	1	393	V	2
88	DOESA	1	241	NEXZ	1	394	VAZ	1
89	DOIN	2	242	NI	2	395	VE	22
90	DOKE	1	243	NINEPENCE	1	396	VEA	1
91	DOOBLE	1	244	NORTHOVER	1	397	VEEXTENDED	1
92	DRACONIA	1	245	NOTFUNNY	1	398	VEN	2
93	DRJOHN	1	246	NOTJUST	4	399	VENI	3
94	DUNDRA	1	247	NTED	1	400	VERRY	4
95	E	29	248	O	84	401	VI	1
96	EASYWHEN	1	249	OA	3	402	VIBRATINGTHE	1
97	EASZ	1	250	OAF	3	403	VIII	2
98	ECTION	1	251	OCCURSJUST	1	404	VISCUM	1
99	ECZEEMA	3	252	OD	1	405	VLLI	1
100	EDDIZ	2	253	OFCOURSE	2	406	VOGA	1
101	EET	2	254	OFFI	1	407	VORBEI	1
102	EETING	1	255	OFJEALOUSY	1	408	VOSTRO	1
103	EEX	1	256	OFTHE	7	409	VVHAT	1
104	EMERG	1	257	OG	1	410	W	21
105	EMSELVES	1	258	OLDESTCLIFF	1	411	WA	7
106	ENDGAME	1	259	OOT	1	412	WAKIN	1
107	ES	1	260	ORI	1	413	WARNOCK	1

108	ET	3	261	ORJUST	2	414	WASJOHN	1
109	EXPLAININGTO	1	262	OUIZMANIA	1	415	WASNEVER	1
110	EXPRESSWAY	1	263	OUZ	1	416	WATH	1
111	EXTER	1	264	P	3	417	WATSO	1
112	F	10	265	PAAR	1	418	WAZ	1
113	FAILTO	1	266	PALIN	1	419	WEEDKILLER	1
114	FAIRWELL	1	267	PARTHENOGENESISL	1	420	WENEED	1
115	FALLIN	1	268	PENTALLIANI	1	421	WENZ	1
116	FAVOURITESCHOOL	1	269	PERSIL	1	422	WEREA	1
117	FIANC	2	270	PICTURELINK	1	423	WERES	1
118	FINITOGLASS	1	271	POK	1	424	WHATSITS	1
119	FLIPPIN	3	272	POSSIB	1	425	WHENWE	1
120	FOODSTUFF	1	273	POTENCYSTATS	1	426	WIENCEY	1
121	FOTO	1	274	PRIORITIESJUST	1	427	WILLBE	1
122	FRANE	1	275	PRISONERZERO	1	428	WITHAN	1
123	FROMTHE	1	276	PRISONNIER	1	429	WKD	1
124	FUNF	1	277	PROBABLYJUST	1	430	WODY	1
125	FY	1	278	PTED	1	431	WORDSEARCH	1
126	G	14	279	PU	1	432	WORKSOP	1
127	GARDE	2	280	PUSHCHAIRS	1	433	WORL	1
128	GAVEJOHN	1	281	PUZ	2	434	WOULDBE	1
129	GAWD	1	282	PV	1	435	WRO	1
130	GB	1	283	Q	4	436	WUMANY	1
131	GETJOHN	1	284	QUI	1	437	WUMPY	1
132	GI	1	285	QUIETNER	1	438	X	33
133	GINGE	2	286	R	15	439	XII	1
134	GIRLJUST	1	287	RAVAN	1	440	XV	1
135	GOODTHING	1	288	RAZ	1	441	Y	55
136	GOTO	1	289	REALWORLD	1	442	YCROFZ	3
137	GOZ	1	290	RECORDEA	1	443	YOUJUST	1
138	GREYPACKING	1	291	REDBEARA	2	444	YOURE	1
139	GRY	1	292	REDBEARAI	1	445	YOURJACKET	1
140	H	63	293	REHEMMED	1	446	YOURJOB	1
141	HALKE	1	294	REJOKING	1	447	YOURJOKES	1
142	HAMERICA	1	295	REMEMBERERED	1	448	YOUS	1
143	HANGON	1	296	REXI	1	449	YOVER	1
144	HARDTO	1	297	RI	3	450	YOVERS	1

145	HATEAPPLES	1	298	RIGHTIO	1	451	Z	39
146	HAVEBEGUN	1	299	RIGTH	1	452	ZE	1
147	HAYLER	1	300	ROBO	1	453	ZEROWILL	1
148	HEERE	1	301	ROMANCEZ	1	454	ZETA	1
149	HEJUST	1	302	RONAY	1	455	ZIZZED	1
150	HEREJUST	1	303	ROSIEJUST	1	456	ZO	1
151	HERJUST	1	304	ROUTEMASTERS	1	457	ZZ	1
152	HERRINF	1	305	RS	2		**TOTAL FREQUENCY**	**3019**
153	HIGHWAYMAN	1	306	RUH	1			

Appendix 6- List of Proper Nouns Excluded from the BTSC

	ITEM	FREQ		ITEM	FREQ		ITEM	FREQ
1	ABARAXAS	1	763	FIBETTE	1	1525	OSTALGIE	1
2	ABBA	2	764	FIBONACCI	1	1526	OSTERHAGEN	15
3	ABBADON	1	765	FINNEGAN	4	1527	OSWALD	86
4	ABBERLINE	1	766	FLEISHMAN	1	1528	OSWIN	37
5	ABBOTT	2	767	FLEMING	2	1529	OTHELLO	1
6	ABERDEEN	5	768	FLICKR	1	1530	OWEN	1
7	ABI	4	769	FLO	2	1531	OWENS	3
8	ABIGAIL	2	770	FLORENCE	2	1532	OXFORD	2
9	ABOOT	1	771	FLORIDA	9	1533	OZZIE	1
10	ABOUTJOHN	1	772	FLORIZEL	2	1534	PAAB	1
11	ABSORBAKLON	1	773	FLYDALE	8	1535	PABLO	1
12	ABSORBALING	1	774	FLYNN	1	1536	PACEM	1
13	ABSORBATHON	1	775	FO	5	1537	PACIFICA	1
14	ABSORBATRIX	1	776	FOON	6	1538	PADDINGTON	1
15	ABTEILUNG	1	777	FOSS	2	1539	PADRA	2
16	ABZORBALOFF	3	778	FOTCH	5	1540	PADRIVOLE	1
17	ACELT	1	779	FRAID	2	1541	PAIGE	1
18	ACHILLES	1	780	FRANCE	21	1542	PAJATOS	2
19	ACR	1	781	FRANCESCO	4	1543	PAKISTAN	1
20	ADA	17	782	FRANCINE	1	1544	PAKOO	1
21	ADAM	18	783	FRANCISCO	1	1545	PALESTINE	1
22	ADAMS	5	784	FRANKLAND	9	1546	PALLIDOME	1
23	ADDAMS	1	785	FRANKLIN	1	1547	PALLUNI	1
24	ADDLESTONE	1	786	FRANZ	1	1548	PALLUSHI	1
25	ADE	2	787	FRANZETTA	1	1549	PANBABYLONIANS	1
26	ADELAIDE	2	788	FRAU	1	1550	PANCROSIS	1
27	ADEOLA	1	789	FRED	1	1551	PANDOFFI	1
28	ADI	2	790	FREEDONIA	6	1552	PANDORA	4
29	ADIEU	1	791	FRENCH	21	1553	PANDORICA	42
30	ADLER	17	792	FREUD	2	1554	PANKHURST	1
31	ADOLF	2	793	FRIDAY	21	1555	PANKY	1
32	ADRIAN	3	794	FUHRER	2	1556	PANTAPHOBIA	2
33	AEQUOREA	1	795	FUNKENSTEIN	1	1557	PANTHEON	2

34	AFGHAN	2	796	GABREAN	1	1558	PAPOOSE	3
35	AFGHANISTAN	11	797	GABRIEL	2	1559	PAPUA	1
36	AFRICA	9	798	GADDABEE	1	1560	PARIS	19
37	AFRICAN	3	799	GAFFABEQUE	1	1561	PARLEYSES	1
38	AGATHA	32	800	GAGAN	4	1562	PARTINGTON	2
39	AGINCOURT	1	801	GAIL	1	1563	PASAMEER	2
40	AGORAX	2	802	GAINSBOROUGH	1	1564	PATERNOSTER	1
41	AGRA	13	803	GALLAGHER	2	1565	PATRICK	5
42	AHMED	2	804	GALLIFREY	65	1566	PAUL	14
43	AICKMAN	1	805	GALLIFREYAN	6	1567	PAVALE	2
44	AITCHINSON	1	806	GALLIFRY	1	1568	PAVEL	9
45	AJAY	10	807	GANDALF	1	1569	PAVLOVA	1
46	AKAKO	2	808	GANDHI	1	1570	PAZ	2
47	AKHAAATEN	4	809	GANTOK	2	1571	PEDRO	7
48	AKHATEN	6	810	GANYMEDE	2	1572	PEGGY	1
49	AKHET	1	811	GARDNER	1	1573	PEMBERTON	2
50	AKIDO	1	812	GARETH	7	1574	PENHAXICO	1
51	AL	2	813	GARN	1	1575	PENTAGON	6
52	ALAMO	1	814	GARR	2	1576	PENTONVILLE	3
53	ALAN	8	815	GARRIDEB	4	1577	PERGANON	1
54	ALASKA	4	816	GARRIDEBS	1	1578	PERU	3
55	ALAYA	12	817	GARSENNON	1	1579	PERUVIAN	1
56	ALBANIAN	2	818	GARVIE	1	1580	PERVY	1
57	ALBAR	2	819	GARY	2	1581	PESHWAMI	2
58	ALBERT	4	820	GAVIN	2	1582	PESO	1
59	ALBION	2	821	GEER	2	1583	PETE	34
60	ALBTRAUM	2	822	GEFANGENER	1	1584	PETER	35
61	ALCOPOPS	1	823	GEH	1	1585	PETERSON	2
62	ALDATES	1	824	GEHEN	1	1586	PETH	3
63	ALEC	1	825	GEHLAL	1	1587	PETHERIDGE	1
64	ALEX	32	826	GEIGER	2	1588	PETRIFOLD	3
65	ALEXANDER	1	827	GEJELH	3	1589	PETRONELLA	2
66	ALEXANDRA	2	828	GELTH	13	1590	PETRUS	3
67	ALEXANDRIA	1	829	GENEVA	6	1591	PHALVITORIUS	1
68	ALEXANDRIAN	1	830	GENGHIS	3	1592	PHELAN	2
69	ALEXEI	3	831	GEOFF	1	1593	PHILLIP	6
70	ALF	1	832	GEOFFREY	1	1594	PHILLIPS	3

71	ALFAVA	1	833	GEORGE	68	1595	PHOENIX	1
72	ALFIE	34	834	GEORGIA	2	1596	PICASSO	1
73	ALI	2	835	GERALD	1	1597	PICCADILLY	2
74	ALICE	6	836	GERMAN	15	1598	PINOCCHIO	1
75	ALISON	3	837	GERMANS	9	1599	PIOTR	1
76	ALISTAIR	2	838	GERMANY	7	1600	PISA	1
77	ALLA	1	839	GHANA	1	1601	PITT	6
78	ALLES	1	840	GHOSH	1	1602	PLASMAVORE	4
79	ALLL	1	841	GIBBIS	2	1603	PLYMOUTH	2
80	ALLONS	16	842	GIDEON	3	1604	POGGIT	5
81	ALLY	1	843	GILES	1	1605	POIROT	3
82	ALONSO	6	844	GILLYFLOWERLAND	1	1606	POISSON	5
83	ALPACA	1	845	GILLYFLOWERTOWN	1	1607	POLA	1
84	ALPHAGRADE	1	846	GIOFFRE	1	1608	POLAND	1
85	ALPS	1	847	GISELLE	1	1609	POLKA	1
86	ALUN	2	848	GLAMIS	1	1610	POLYCARBIDE	2
87	AMANDA	2	849	GLASGOW	11	1611	POMPEII	21
88	AMARRA	1	850	GLASMIR	1	1612	PONCING	2
89	AMAS	1	851	GLEICH	1	1613	PONDICUS	1
90	AMAT	1	852	GLENN	1	1614	PONTEFRACT	2
91	AMBROSE	15	853	GLORIA	1	1615	PONTICUM	1
92	AMELIA	76	854	GLORIANA	1	1616	POPEMOBILE	1
93	AMERICA	43	855	GLOUCESTER	3	1617	PORLOCK	2
94	AMERICAN	23	856	GODDARD	5	1618	PORTON	3
95	AMERICANS	8	857	GODZ	1	1619	POSITRONIC	1
96	AMIE	1	858	GOFFLE	1	1620	POSSA	1
97	AMO	2	859	GOGH	19	1621	POWELL	2
98	AMONTILLADO	1	860	GOLDSMITH	1	1622	PRADESH	1
99	AMSTERDAM	3	861	GOLIGHTLY	3	1623	PRAGUE	1
100	AMY	529	862	GOLLUM	1	1624	PRENTIS	11
101	AMYS	4	863	GONNE	1	1625	PRESBURY	1
102	ANA	21	864	GOODING	1	1626	PRESLEY	1
103	ANAH	5	865	GOTTLE	5	1627	PRITCHARD	12
104	ANAHSON	1	866	GRACIAS	1	1628	PRIVYET	1
105	ANDA	1	867	GRAHAM	2	1629	PROBIC	4
106	ANDEREN	1	868	GRAYLE	5	1630	PROGENATE	1
107	ANDERSON	13	869	GRAYLING	2	1631	PROGENATION	3

108	ANDES	1	870	GRAYSON	2	1632	PROGENITOR	9
109	ANDO	25	871	GRAYSTARK	1	1633	PROSEGUIRE	1
110	ANDRE	1	872	GREEK	6	1634	PRYDONIAN	2
111	ANDREA	3	873	GREENLAND	2	1635	PRYOR	2
112	ANDREAS	1	874	GREENWICH	1	1636	PSYCHOCHRONOGRAPH	1
113	ANDREW	6	875	GREG	15	1637	PURCELL	2
114	ANDROZANI	5	876	GREGOR	11	1638	PYLEEN	1
115	ANDY	4	877	GREGSON	2	1639	PYLONS	1
116	ANGELA	2	878	GRETA	2	1640	PYROVILE	11
117	ANGELES	1	879	GRETCHEN	3	1641	PYROVILLIA	8
118	ANGELINA	1	880	GRETEL	3	1642	PYROVILLIAN	2
119	ANGELO	21	881	GREXNIK	1	1643	QAEDA	1
120	ANGIE	20	882	GRIFFOTH	2	1644	QIN	2
121	ANGLICAN	1	883	GRIMM	1	1645	QOM	1
122	ANITA	8	884	GRIMPEN	3	1646	QUADROCYCLE	3
123	ANNA	1	885	GUANTANAMO	1	1647	QUAYLE	2
124	ANNABEL	4	886	GUIDO	2	1648	QUEENIE	1
125	ANNALISE	7	887	GUINEA	3	1649	QUOLDONITY	1
126	ANNE	6	888	GUINNEVERE	1	1650	RA	3
127	ANNIE	1	889	GUS	6	1651	RACNOSS	1
128	ANS	1	890	GWEN	6	1652	RADFORD	1
129	ANSIBLE	1	891	GWYNETH	15	1653	RAFFAELLA	1
130	ANTARCTICA	2	892	HABBO	1	1654	RAFFALO	2
131	ANTOINETTE	2	893	HADES	4	1655	RAJ	1
132	APALAPU	1	894	HADROJASSIC	2	1656	RAJESH	4
133	APALAPUCIA	9	895	HAERETICUM	3	1657	RALEIGH	1
134	APALAPUCIANS	2	896	HALPEN	11	1658	RALPH	1
135	APLAN	2	897	HAMILTON	2	1659	RAMBO	1
136	APLANS	6	898	HAMISH	10	1660	RAMSAY	1
137	APOLLO	16	899	HAMMILL	1	1661	RAMSDEN	3
138	APPLEDORE	9	900	HAMNET	2	1662	RANDOMER	1
139	APRIL	6	901	HAMPSTEAD	1	1663	RANJIT	1
140	ARBUCKLE	1	902	HAMPTON	2	1664	RAOUL	9
141	ARCADIA	2	903	HANNAH	3	1665	RASPUTIN	1
142	ARCATEENIAN	1	904	HANNIBAL	6	1666	RASSILON	4
143	ARCATEENIANS	2	905	HANSEL	3	1667	RASSMUSSEN	7
144	ARCHIBALD	2	906	HARCOURT	3	1668	RATTIGAN	7

145	ARCHIE	3	907	HARLOW	1	1669	RAXACORICOFALLAPATORIUS	1
146	ARDEN	8	908	HAROLD	13	1670	RAZBAHAN	1
147	ARDENNES	1	909	HARRIET	33	1671	REAGAN	1
148	AREN	5	910	HARRIS	1	1672	REDBEARD	18
149	ARGENTINA	1	911	HARRISON	1	1673	REDBEAROJ	1
150	ARGOS	1	912	HARROGATE	2	1674	REDFERN	6
151	ARIANA	1	913	HARRY	29	1675	REDMOND	5
152	ARIDIUS	1	914	HARTLEY	1	1676	REG	5
153	ARISTOTLE	9	915	HARTMAN	2	1677	REICHENBACH	6
154	ARKIPHETS	1	916	HARTNELL	1	1678	REINETTE	25
155	ARMSTRONG	5	917	HARUKA	2	1679	RELS	7
156	ARNOLD	1	918	HARUSPEX	1	1680	RENFREW	2
157	ARTHUR	6	919	HARVEY	5	1681	REPETEZ	1
158	ARTIE	12	920	HASSAN	2	1682	RESTAC	8
159	ARWELL	7	921	HAVANA	1	1683	RETROTOPES	1
160	ASCINTA	1	922	HAVERSTOCK	1	1684	RETROVIDS	1
161	ASGARD	2	923	HAWTHORNE	1	1685	REXEL	3
162	ASHILDR	47	924	HAZELDINE	1	1686	REXICOR	1
163	ASHINGTON	1	925	HAZLEHEAD	1	1687	REXICORICO	1
164	ASHTON	9	926	HAZRAN	4	1688	REXICORICOPHALVITORIAN	1
165	ASIA	3	927	HEATHER	8	1689	REXICORICOPHALVITORIUS	8
166	ASQUITH	2	928	HEATHROW	4	1690	REYKJAVIK	2
167	ASTRID	14	929	HEDGEWICK	3	1691	REYNOLDS	1
168	ATIF	4	930	HEIDI	4	1692	REZH	1
169	ATLANTIC	6	931	HEIL	2	1693	RHODRI	1
170	ATLANTIS	1	932	HEILIGE	1	1694	RHYS	1
171	ATLAS	1	933	HEINKEL	1	1695	RICHARD	23
172	ATMOS	42	934	HEISSE	1	1696	RICHARDS	3
173	ATRAXI	6	935	HELEN	5	1697	RICHES	3
174	ATTA	1	936	HELMAND	1	1698	RICKSTON	12
175	ATTABOY	4	937	HELMIC	2	1699	RICKY	24
176	ATTILA	1	938	HELTER	1	1700	RICOLETTI	32
177	ATTLEE	1	939	HENRY	52	1701	RIDDELL	4
178	AUGUST	3	940	HERALDS	1	1702	RIGSY	32
179	AUGUSTE	1	941	HERBERT	1	1703	RIO	11
180	AUGUSTUS	6	942	HERBES	1	1704	ROBERT	14
181	AULD	6	943	HERCULE	1	1705	ROBIN	31

182	AURA	1	944	HERRINGS	1	1706	ROBINA	2
183	AURELIUS	1	945	HICKMAN	2	1707	ROBINSON	3
184	AUSTEN	3	946	HIDEY	1	1708	ROBSON	1
185	AUSTRALASIA	1	947	HIER	2	1709	ROBYN	1
186	AUSTRALIA	4	948	HIESS	1	1710	ROCASTLE	1
187	AUSTRALIAN	2	949	HIGGINS	1	1711	ROCCO	1
188	AUSTRIA	1	950	HIMALAYAS	1	1712	RODRICK	7
189	AUTONS	1	951	HIMMLISCHER	2	1713	ROEDEAN	1
190	AVANTI	2	952	HIRST	1	1714	ROENTGEN	2
191	AVERY	4	953	HITCHINGSON	2	1715	ROGER	7
192	AVIV	1	954	HITCHLEY	3	1716	ROMAN	39
193	AZ	2	955	HITLER	18	1717	ROMANS	20
194	AZBANTIUM	1	956	HOCHHEILIGE	1	1718	ROMANY	1
195	BAALEN	4	957	HOCKLEY	1	1719	ROME	18
196	BABBINGTON	1	958	HOELLE	1	1720	ROMEO	1
197	BABYLON	1	959	HOFF	4	1721	ROMMEL	1
198	BACH	3	960	HOLBORN	2	1722	RONALD	1
199	BAGHDAD	3	961	HOLLAND	1	1723	RORANICUS	2
200	BAHRAIN	1	962	HOLMES	298	1724	RORY	358
201	BAINBRIDGE	16	963	HOLODECK	1	1725	RORYCAM	1
202	BAINES	19	964	HOLOVID	1	1726	ROSA	1
203	BAKER	29	965	HOLOVIDS	1	1727	ROSAMUND	3
204	BALAMORY	1	966	HONG	3	1728	ROSANNA	7
205	BALDERDASH	1	967	HOOB	1	1729	ROSIE	19
206	BALHOON	3	968	HOOLOOVOO	1	1730	ROSSETTI	1
207	BALTIMORE	2	969	HOOPER	17	1731	ROSSITER	1
208	BANNA	1	970	HOPETHORNE	2	1732	ROSWELL	2
209	BANNAKAFFALATTA	20	971	HOPKINS	1	1733	ROSY	1
210	BANTOS	1	972	HOPPLEDOM	1	1734	ROTMEISTER	3
211	BAPHIX	1	973	HOSHBIN	1	1735	ROXBORNE	2
212	BARBARELLA	1	974	HOSKINS	2	1736	ROYSTON	1
213	BARCELONA	5	975	HOSPFIAL	1	1737	RU	24
214	BARNES	1	976	HOUDINI	2	1738	RUBBISCH	1
215	BARNEY	6	977	HOUNSLOW	1	1739	RUBICON	1
216	BARNICOT	3	978	HOUSTON	3	1740	RUDI	2
217	BARROWMAN	3	979	HOWARD	6	1741	RUFUS	1
218	BARRY	2	980	HOWIE	11	1742	RUPERT	9

219	BARRYMORE	7	981	HUBBARD	1	1743	RUSHMORE	1
220	BART	7	982	HUDDERS	1	1744	RUSSELL	1
221	BARTOCK	1	983	HUDSON	57	1745	RUSSIA	5
222	BARTS	2	984	HUGH	2	1746	RUSSIAN	8
223	BASKERVILLE	16	985	HULKS	1	1747	RUSSIANS	3
224	BASKERVILLES	1	986	HUMBER	1	1748	RUTANS	1
225	BASSEY	1	987	HUMPHREY	1	1749	RUTHERFORD	1
226	BASTIC	1	988	HUTCHINSON	6	1750	RYAN	1
227	BATTERSEA	6	989	HYDROKINOMETER	1	1751	RYCBAR	1
228	BAVARIA	1	990	IAN	10	1752	RYDER	5
229	BAXTER	2	991	IANS	1	1753	RYMAN	1
230	BAXTERS	1	992	IANTO	4	1754	SABREWOLVES	1
231	BBC	9	993	ICELANDIC	2	1755	SACRAMENTO	1
232	BEATLES	4	994	ICH	5	1756	SAHARA	2
233	BEATRICE	2	995	IDA	22	1757	SAHEED	3
234	BEAUCOUP	1	996	IDEE	1	1758	SAIBRA	8
235	BEBO	2	997	IDRIS	6	1759	SAIGON	1
236	BECK	1	998	ILLYRIA	1	1760	SALL	1
237	BEETHOVEN	12	999	IMMER	1	1761	SALLY	43
238	BEIJING	3	1000	INDIA	10	1762	SALVAIN	1
239	BELARUS	1	1001	INDIAN	1	1763	SAM	16
240	BELGIAN	1	1002	INDIANA	2	1764	SAMARITAN	2
241	BELGIANS	2	1003	INDIRA	3	1765	SAMARRA	5
242	BELGIUM	6	1004	INDRA	1	1766	SAMSON	2
243	BELLISSIMO	1	1005	IPADS	3	1767	SAMUEL	2
244	BEN	14	1006	IPHONE	2	1768	SAN	10
245	BENE	8	1007	IPOD	1	1769	SANDEEP	1
246	BENEDICT	4	1008	IPODS	1	1770	SANDEFORD	3
247	BENGSDOTTER	2	1009	IPSWICH	1	1771	SANS	1
248	BENJAMIN	5	1010	IRAN	1	1772	SANTA	5
249	BENNETT	19	1011	IRAQ	3	1773	SANTINI	1
250	BENNY	2	1012	IRAXXA	6	1774	SANTORI	2
251	BERGEN	1	1013	IRELAND	3	1775	SANTOS	3
252	BERGERAC	2	1014	IRENE	11	1776	SARAH	45
253	BERING	1	1015	IRISH	1	1777	SARFF	11
254	BERLIN	5	1016	IRRA	1	1778	SATSUMA	2
255	BERMUDA	1	1017	ISAAC	27	1779	SATURDAY	11

256	BERNARD	3	1018	ISABELLA	16	1780	SATURDAYS	2
257	BERTH	1	1019	ISANDLWANA	1	1781	SATURNYNE	3
258	BESSAN	1	1020	ISLAMABAD	1	1782	SAUL	1
259	BESTEN	1	1021	ISOBEL	4	1783	SAXE	1
260	BESZ	2	1022	ISOLUS	13	1784	SCANNY	1
261	BETH	5	1023	ITALIAN	1	1785	SCHLAF	2
262	BETHLEM	1	1024	ITALIANS	2	1786	SCHLAFT	1
263	BETTENJOHN	1	1025	ITALY	1	1787	SCHUBERT	1
264	BETTY	1	1026	ITO	2	1788	SCHWARZENEGGER	1
265	BEVY	1	1027	IVAN	4	1789	SCOOBY	3
266	BEXLEY	1	1028	JABE	5	1790	SCOOTI	8
267	BEZ	1	1029	JACK	81	1791	SCOT	1
268	BEZZIE	1	1030	JACKIE	58	1792	SCOTCH	1
269	BEZZY	1	1031	JACKSON	6	1793	SCOTLAND	32
270	BIEN	1	1032	JACQUELINE	6	1794	SCOTS	4
271	BIENTOT	1	1033	JAGRA	1	1795	SCOTSMAN	2
272	BILLY	12	1034	JAGRAFESS	10	1796	SCOTT	5
273	BIRMINGHAM	1	1035	JAHANNAM	2	1797	SCOTTISH	26
274	BITTE	3	1036	JAHOO	1	1798	SEAMUS	1
275	BLACKFRIARS	1	1037	JAKE	10	1799	SEAN	2
276	BLAGGARD	1	1038	JALANDRA	1	1800	SEATTLE	1
277	BLAIDD	5	1039	JAMES	27	1801	SEBASTIAN	9
278	BLAINE	5	1040	JAMIE	7	1802	SEKHMET	1
279	BLAIR	1	1041	JAMMIE	1	1803	SELDEN	1
280	BLASCO	1	1042	JANE	33	1804	SEPTEMBER	1
281	BLEAURGH	1	1043	JANEIRO	1	1805	SER	1
282	BLESSINGTON	2	1044	JANET	1	1806	SERBIAN	5
283	BLINOVITCH	1	1045	JANINE	14	1807	SERENISSIMA	1
284	BLISH	1	1046	JANIS	1	1808	SERIESSUB	17
285	BLON	4	1047	JANUARY	2	1809	SHABOGANS	1
286	BLYTON	1	1048	JANUS	9	1810	SHADMOCH	1
287	BOB	36	1049	JAPAN	6	1811	SHAFE	10
288	BOBBY	2	1050	JAPANESE	5	1812	SHAKESPEARE	23
289	BOE	23	1051	JARIA	1	1813	SHAKRI	9
290	BOEKIND	1	1052	JATE	1	1814	SHALLACATOP	2
291	BOESHANE	1	1053	JATHAA	1	1815	SHALLANNA	1
292	BOGGONS	2	1054	JATT	1	1816	SHAMBONI	1

293	BOHEMIAN	2	1055	JEANETTE	2	1817	SHARINE	2
294	BOIS	1	1056	JEFF	11	1818	SHARON	4
295	BONAPARTE	1	1057	JEFFERSON	28	1819	SHAUN	1
296	BONJOUR	2	1058	JEFFREY	2	1820	SHEBA	1
297	BONNIE	10	1059	JEHOVAH	2	1821	SHEFFIELD	3
298	BORGIA	4	1060	JEMIMA	1	1822	SHEILA	1
299	BORGIAS	2	1061	JEN	20	1823	SHEN	1
300	BORS	1	1062	JENKINS	8	1824	SHEPPY	2
301	BOSTEEN	1	1063	JENNA	9	1825	SHERL	7
302	BOSTON	1	1064	JENNIFER	32	1826	SHERLOCK	567
303	BOSWELL	1	1065	JENNY	61	1827	SHERLY	1
304	BOULOGNE	1	1066	JENS	1	1828	SHERRINFORA	3
305	BRABBIT	2	1067	JEREMY	3	1829	SHERRINFORAJ	1
306	BRACCATOLIAN	1	1068	JERICHO	1	1830	SHERRINFORD	6
307	BRACEWELL	11	1069	JERSEY	1	1831	SHERWOOD	4
308	BRACKNELL	4	1070	JERUSALEM	1	1832	SHERYL	1
309	BRADFORD	1	1071	JESSICA	1	1833	SHETLAND	2
310	BRADLEY	10	1072	JESUS	24	1834	SHEZZA	3
311	BRAKOVITCH	1	1073	JETHRO	14	1835	SHEZZER	1
312	BRAM	6	1074	JETSON	1	1836	SHILCOTT	1
313	BRAN	1	1075	JETTISON	9	1837	SHIPTON	6
314	BRAZIL	2	1076	JETZT	2	1838	SHIREEN	7
315	BRECON	2	1077	JEX	18	1839	SHIRLEY	1
316	BREITLING	1	1078	JIM	36	1840	SHIVARATRI	1
317	BRIAN	21	1079	JIMBO	3	1841	SHO	2
318	BRIDGET	16	1080	JIMMY	13	1842	SHOLTO	14
319	BRIEN	2	1081	JIVAL	4	1843	SHONA	4
320	BRIGHTON	3	1082	JO	7	1844	SHUK	4
321	BRISBANE	1	1083	JOAN	4	1845	SHUKINA	1
322	BRISTOL	2	1084	JOCRASSA	1	1846	SIDNEY	1
323	BRITAIN	46	1085	JODHPURS	1	1847	SIE	10
324	BRITISH	46	1086	JODRELL	2	1848	SIEBEN	1
325	BRITS	1	1087	JOE	30	1849	SIERPINSKI	1
326	BRIXTON	3	1088	JOERGENSEN	1	1850	SIGMAFOLIO	1
327	BROFF	2	1089	JOH	3	1851	SILENCIO	4
328	BROOKLYN	5	1090	JOHANN	2	1852	SILFRAX	1
329	BRUCE	2	1091	JOHANNESBURG	1	1853	SILURIAN	7

330	BRUHL	1	1092	JOHH	1	1854	SILURIANS	9
331	BRUNSWICK	1	1093	JOHN	440	1855	SIMEON	5
332	BRUSSELS	1	1094	JOHNNY	4	1856	SIMMONS	3
333	BRYAN	1	1095	JOLCO	2	1857	SIMON	4
334	BRYANT	1	1096	JONATHAN	3	1858	SIMONP	2
335	BUCH	1	1097	JONES	110	1859	SIND	2
336	BUCKDEN	1	1098	JOOFIE	1	1860	SINDA	1
337	BUCKINGHAM	7	1099	JOPLIN	1	1861	SINGH	1
338	BUCKNALL	1	1100	JORDAN	1	1862	SIRIUS	1
339	BUDAPEST	1	1101	JOSEPH	1	1863	SISTINE	1
340	BURMESE	1	1102	JOSHUA	2	1864	SKALDAK	28
341	BURNE	2	1103	JUBBLY	1	1865	SKARO	24
342	BYRON	1	1104	JUDAS	1	1866	SKAROSA	3
343	BYZANTINE	1	1105	JUDOON	18	1867	SKASAS	3
344	BYZANTIUM	9	1106	JUDY	4	1868	SKORR	5
345	C	25	1107	JULIENNE	1	1869	SKOVOX	7
346	CA	2	1108	JULY	1	1870	SKYPE	2
347	CAAN	16	1109	JUMNA	1	1871	SLAVA	1
348	CAESAR	6	1110	JUNE	5	1872	SLAVIC	1
349	CAIO	1	1111	JURASSIC	2	1873	SMALLWOOD	8
350	CAIRO	1	1112	KAHLER	9	1874	SMITH	131
351	CALIBURN	5	1113	KAI	2	1875	SMITHS	1
352	CALIENTE	1	1114	KALED	1	1876	SMYTHE	1
353	CALIFORNIA	1	1115	KALOON	1	1877	SNEED	10
354	CALISTO	1	1116	KAMEN	1	1878	SNOWDON	1
355	CALLISTA	1	1117	KANDAHAR	1	1879	SOBORIAN	1
356	CALLUFRAX	1	1118	KANZO	1	1880	SOHO	1
357	CALVIERRI	7	1119	KAPUT	4	1881	SOLANA	5
358	CAMBERWELL	1	1120	KAR	13	1882	SOLLTE	1
359	CAMBODIA	1	1121	KARABRAXOS	19	1883	SOLOMON	21
360	CAMBRIDGE	2	1122	KARACHI	1	1884	SONICED	2
361	CAMDEN	2	1123	KARASS	1	1885	SONICING	1
362	CAMILLA	1	1124	KAREN	1	1886	SONICKING	2
363	CAMMINO	1	1125	KARIM	2	1887	SONST	1
364	CAMPBELL	3	1126	KARINA	1	1888	SONTAR	52
365	CAMPTOWN	2	1127	KARN	3	1889	SONTARAN	43
366	CANADA	1	1128	KASBAH	1	1890	SONTARANS	39

367	CANNAE	1	1129	KASTERBOROUS	3	1891	SONTARON	1
368	CANTERBURY	1	1130	KATE	25	1892	SONTERRUN	1
369	CANTON	23	1131	KATH	1	1893	SONTERRUNS	1
370	CANTRELL	3	1132	KATHERINE	4	1894	SONTERUNS	1
371	CAPALDI	8	1133	KATHY	14	1895	SONY	1
372	CAPITANO	1	1134	KATIE	1	1896	SOOTHSAYED	1
373	CARACAS	1	1135	KEL	1	1897	SOOTHSAYER	6
374	CARDEW	4	1136	KELLY	2	1898	SOOTHSAYERS	5
375	CARDIFF	24	1137	KELVIN	2	1899	SOPHIE	30
376	CAREW	1	1138	KEMBEL	1	1900	SOUTHAMPTON	1
377	CAREY	1	1139	KEN	2	1901	SOUTHBANK	2
378	CARL	12	1140	KENDAL	1	1902	SOUTHWARK	6
379	CARLISLE	4	1141	KENDRICK	3	1903	SOVIET	2
380	CARLO	3	1142	KENNEDY	15	1904	SPACEHOPPER	1
381	CARLYLE	1	1143	KENNY	10	1905	SPACELANE	1
382	CARMEN	3	1144	KENT	1	1906	SPAIN	2
383	CARMICHAEL	9	1145	KESS	3	1907	SPANISH	4
384	CAROL	2	1146	KEZZIA	4	1908	SPARKPLUG	1
385	CAROLS	1	1147	KHADENI	1	1909	SPARTACUS	7
386	CARRA	1	1148	KHAN	3	1910	SPEELFOX	1
387	CARRIONITE	5	1149	KILBURN	1	1911	SPENCER	1
388	CARRIONITES	6	1150	KINCADE	1	1912	SPINNIN	2
389	CARTA	1	1151	KINCAID	2	1913	SPION	1
390	CARTWRIGHT	3	1152	KINGLOUISXX	10	1914	SPIRODON	1
391	CARTWRIGHTS	1	1153	KINGSLEY	1	1915	SPOCK	7
392	CASANOVA	3	1154	KIRK	1	1916	SPONGEBOB	1
393	CASPERTINE	1	1155	KIRSTY	5	1917	SPOONHEAD	1
394	CASS	20	1156	KISSOGRAM	4	1918	SPOONHEADS	1
395	CASSADYNE	1	1157	KISTANE	7	1919	SPOTIFY	1
396	CASSANDRA	20	1158	KIZLET	3	1920	SPRINGFIELD	3
397	CASSAVALIAN	1	1159	KLARJ	1	1921	SPRITE	1
398	CATCHLOVE	8	1160	KLEIN	1	1922	SPURRINA	2
399	CATH	1	1161	KLEMPARI	1	1923	SQUELCHY	1
400	CATHERINE	5	1162	KLINGON	1	1924	STAAL	13
401	CATHICA	5	1163	KNABE	1	1925	STACY	7
402	CATHY	5	1164	KNOX	1	1926	STALIN	3
403	CATULLUS	1	1165	KODION	1	1927	STAMFORD	4

404	CAVILL	1	1166	KOH	2	1928	STANFORD	2
405	CAWDOR	1	1167	KOMME	1	1929	STAPLETON	11
406	CENTURIAN	2	1168	KONG	3	1930	STATTEN	22
407	CHAKA	1	1169	KOP	1	1931	STELLA	3
408	CHAN	38	1170	KOREA	3	1932	STEPASHIN	5
409	CHANDRAKALA	5	1171	KOREAN	1	1933	STEPHEN	5
410	CHANEL	4	1172	KORWIN	23	1934	STETSON	3
411	CHANG	4	1173	KOVARIAN	4	1935	STETSONS	2
412	CHANTHO	4	1174	KRAFAYIS	3	1936	STEVE	3
413	CHAPLIN	3	1175	KRAKATOA	1	1937	STEVEN	6
414	CHARLEMAGNE	2	1176	KREMLIN	1	1938	STEVIE	3
415	CHARLES	18	1177	KRILLITANE	3	1939	STEWART	10
416	CHARLIE	24	1178	KRILLITANES	5	1940	STILLE	1
417	CHARLOTTE	4	1179	KRO	1	1941	STO	8
418	CHATEAUROUX	1	1180	KRONE	1	1942	STONEHENGE	3
419	CHATTERJEE	2	1181	KRONKBURGER	5	1943	STORMAGEDDON	6
420	CHAUDHRY	2	1182	KROP	2	1944	STORMCAGE	3
421	CHAVIC	1	1183	KUR	1	1945	STRADIVARIUS	1
422	CHEEM	2	1184	KWON	2	1946	STRASSE	1
423	CHEEN	1	1185	KYLIE	1	1947	STRATHCLYDE	1
424	CHELONIAN	2	1186	KYOTO	1	1948	STRAX	26
425	CHELSEA	1	1187	LAMBETH	1	1949	STREETE	5
426	CHEN	4	1188	LAMMASTEEN	1	1950	STREPSIL	1
427	CHENGHUA	1	1189	LANA	2	1951	STROOD	1
428	CHERCHEZ	2	1190	LANCASHIRE	4	1952	STUART	4
429	CHERNOBYL	2	1191	LANGDALE	2	1953	STUBBIE	10
430	CHESTER	1	1192	LANGER	1	1954	STUKAS	1
431	CHICAGO	1	1193	LAS	1	1955	SUDOKU	2
432	CHILDERS	1	1194	LASAGNE	3	1956	SUE	4
433	CHINA	29	1195	LASSAR	1	1957	SUERKY	1
434	CHINESE	22	1196	LASSEN	1	1958	SUEZ	1
435	CHINNY	1	1197	LASZLO	27	1959	SUKI	9
436	CHISWICK	9	1198	LATIMER	8	1960	SULEJMANI	2
437	CHLOE	57	1199	LATIN	13	1961	SULLIVAN	2
438	CHOPRA	19	1200	LAUCASS	12	1962	SUMATRA	5
439	CHORLE	1	1201	LAUDER	1	1963	SUMMERSON	2
440	CHOWDRY	6	1202	LAURA	3	1964	SUNDAY	12

441	CHRISSIE	1	1203	LAUREL	1	1965	SUNDAYS	3
442	CHRISTENDOM	1	1204	LAURISTON	5	1966	SUNITA	1
443	CHRISTIE	30	1205	LAUTA	2	1967	SURREY	1
444	CHRISTINA	1	1206	LAWRENCE	3	1968	SUSAN	3
445	CHRISTODOLOU	1	1207	LAZARUS	21	1969	SUSANNE	1
446	CHRISTOPHER	2	1208	LEA	2	1970	SUSIE	2
447	CHRONODYNE	6	1209	LEANDRO	2	1971	SUSSEX	3
448	CHRONOLOCK	7	1210	LECTRICKS	1	1972	SUTCLIFFE	11
449	CHULA	7	1211	LEE	3	1973	SUTTON	3
450	CHURCHILL	18	1212	LEEDS	6	1974	SUZANNE	1
451	CHUZZLEWIT	1	1213	LEEN	1	1975	SUZETTE	7
452	CIAO	1	1214	LEGENAJ	1	1976	SWEDE	1
453	CID	2	1215	LEINSTER	3	1977	SWEDEN	1
454	CINDERELLA	1	1216	LENNY	1	1978	SWEDISH	1
455	CLAIRE	19	1217	LEO	18	1979	SWEENEY	2
456	CLARA	585	1218	LEONIANS	1	1980	SWEETVILLE	10
457	CLARAS	1	1219	LEONIS	1	1981	SYDNEY	1
458	CLARE	1	1220	LERNER	2	1982	SYLVIA	5
459	CLARINS	1	1221	LESTRADE	27	1983	TABITHA	1
460	CLARK	8	1222	LETHBRIDGE	5	1984	TACORIAN	6
461	CLARKE	2	1223	LETITIA	4	1985	TAE	2
462	CLASSABINDI	1	1224	LILITH	1	1986	TAJ	1
463	CLAUDE	3	1225	LINCOLN	2	1987	TAKESIES	1
464	CLAUDETTE	4	1226	LINDA	12	1988	TALLULAH	16
465	CLAVADOE	1	1227	LIVINGSTON	1	1989	TANDOCCA	4
466	CLAYTON	3	1228	LIZ	10	1990	TANYA	3
467	CLEO	1	1229	LLOYD	3	1991	TARDIS	637
468	CLEOPATRA	6	1230	LO	1	1992	TARDISES	9
469	CLIFTON	2	1231	LOCH	3	1993	TARMACKERS	1
470	CLIVE	4	1232	LOCKIGEN	1	1994	TARMACKING	2
471	CLOM	8	1233	LOCKLEY	1	1995	TAROVIAN	1
472	CLOONEY	1	1234	LOIS	1	1996	TASTIC	3
473	CLUEDO	2	1235	LOMBARDY	1	1997	TAUNTON	1
474	CLYDE	2	1236	LONDON	242	1998	TAUREAN	1
475	COBB	7	1237	LONDONER	2	1999	TAYLOR	1
476	COFELIA	2	1238	LONGSHOE	2	2000	TBILISI	8
477	COFFA	1	1239	LONGSTAFF	1	2001	TED	2

478	COHEZIC	2	1240	LORNA	4	2002	TEDDY	2
479	COLCHESTER	2	1241	LOS	1	2003	TEGAN	2
480	COLEMAN	5	1242	LOUIS	1	2004	TEGISING	1
481	COLIN	6	1243	LOUISA	1	2005	TENA	1
482	COLOMBIA	3	1244	LOUISE	6	2006	TENNESSEE	1
483	COLOMBIAN	1	1245	LOXLEY	2	2007	TENZA	6
484	COLOSANTO	1	1246	LUCANIANS	1	2008	TERILEPTIL	1
485	COLTRANE	1	1247	LUCAS	6	2009	TERILEPTILS	1
486	CONDURRA	1	1248	LUCIE	5	2010	TESCO	1
487	CONNIE	8	1249	LUCIFER	2	2011	TESELECTA	6
488	CONNOLLY	9	1250	LUCIUS	17	2012	TESLA	1
489	CONSADINE	1	1251	LUCY	23	2013	TESSA	2
490	CONSTANTINE	4	1252	LUDENS	2	2014	TETRADECAGON	3
491	CONTRAFIBULATOR	1	1253	LUDMILA	1	2015	TEXAS	3
492	CORDOLAINE	5	1254	LUDOVIC	1	2016	THAILAND	1
493	CORFU	1	1255	LUDWIG	2	2017	THAMES	22
494	CORGI	1	1256	LUGAL	1	2018	THARSIS	1
495	CORNELIUS	3	1257	LUKE	16	2019	THARSISIAN	3
496	CORNISH	2	1258	LUKER	1	2020	THATCHER	14
497	CORNWALL	4	1259	LUKIS	12	2021	THAY	2
498	CORVIN	1	1260	LUKO	1	2022	THECO	1
499	COSTELLO	3	1261	LUMIC	35	2023	THEDION	2
500	COUP	3	1262	LUNDVIK	1	2024	THICKANIA	1
501	COURTNEY	34	1263	LUNN	24	2025	THICKETTY	1
502	COVENTRY	7	1264	LURKWORMS	1	2026	THINGAMABOB	1
503	CRAIG	81	1265	LUSITANIA	1	2027	THOMAS	10
504	CRAYHILL	1	1266	LUV	1	2028	THRACE	1
505	CRECHE	1	1267	LYDIA	1	2029	THURSDAY	14
506	CREET	3	1268	LYI	3	2030	TIBERION	2
507	CRESPALLION	2	1269	LYNDA	15	2031	TIBET	3
508	CRETINO	1	1270	LYNLEY	4	2032	TIENES	3
509	CRIMEA	1	1271	LYNN	2	2033	TIM	8
510	CRIMEWATCH	1	1272	LYNNE	1	2034	TIMELORD	1
511	CRISPIANS	1	1273	LYONS	1	2035	TIMESICK	1
512	CRISPIN	5	1274	MABEL	1	2036	TIMEWALL	1
513	CROATIA	1	1275	MACARENA	7	2037	TIMEY	21
514	CROCODILOPOLIS	1	1276	MACBETH	2	2038	TIMMY	2

515	CROMER	1	1277	MACCATEER	1	2039	TIMOTHY	4
516	CROMWELL	2	1278	MACHST	1	2040	TIS	9
517	CROOT	1	1279	MACKENZIE	3	2041	TITANIC	21
518	CROSACTIC	1	1280	MACKESON	1	2042	TIVOLEANS	1
519	CROSBIE	6	1281	MACLEISH	1	2043	TIVOLI	5
520	CROYDON	3	1282	MACRA	4	2044	TOBIAS	3
521	CRUSOE	2	1283	MACRAE	3	2045	TOBY	40
522	CUBA	1	1284	MADGE	21	2046	TOCLAFANE	12
523	CULVERTON	10	1285	MADRID	1	2047	TODD	2
524	CURBISHLEY	2	1286	MAEBH	54	2048	TOKYO	2
525	CURBISHLEYS	1	1287	MAFEKING	3	2049	TOM	19
526	CURIOSCANNER	1	1288	MAGAMBO	1	2050	TOMMY	21
527	CURTIS	1	1289	MAGELLAN	1	2051	TONTO	1
528	CURUCA	1	1290	MAGGIE	1	2052	TONY	13
529	CYBERBRAINS	1	1291	MAGNA	2	2053	TOODLE	2
530	CYBERDEARS	1	1292	MAGNO	3	2054	TORAGY	1
531	CYBERFORM	2	1293	MAGNUSSEN	40	2055	TRAFALGAR	8
532	CYBERIAD	8	1294	MAGPIE	17	2056	TRANSMAT	18
533	CYBERKINETIC	1	1295	MAHA	1	2057	TRANSMATS	1
534	CYBERKING	1	1296	MAHAL	1	2058	TRANSMATTED	2
535	CYBERLEADER	3	1297	MAISIE	6	2059	TRANSMATTING	1
536	CYBERMAT	7	1298	MAITLAND	1	2060	TREEBORGS	2
537	CYBERMATS	3	1299	MAJORCA	1	2061	TRENT	1
538	CYBERMITES	2	1300	MAJORIA	1	2062	TRENZALORE	15
539	CYBERNETIC	2	1301	MALCASSAIRO	1	2063	TREPPA	2
540	CYBERSHIP	1	1302	MALDOVAR	5	2064	TREVOR	1
541	CYBERSHIPS	1	1303	MALOHKEH	4	2065	TRIBOPHYSICAL	1
542	CYBERTHING	2	1304	MALONE	4	2066	TRICEY	6
543	CYBERTHREAT	1	1305	MALTA	1	2067	TRIPLIGHT	1
544	CYBUS	13	1306	MANCHESTER	1	2068	TRIQUETRAL	1
545	CYRANO	1	1307	MANDY	5	2069	TRISH	5
546	CYRIL	36	1308	MANGER	1	2070	TRISHA	5
547	CZECH	3	1309	MANHATTAN	13	2071	TUESDAY	11
548	DACOSTA	1	1310	MANISTA	3	2072	TUESDAYS	1
549	DAGMAR	2	1311	MANTON	5	2073	TUMBLR	4
550	DALEKANIUM	10	1312	MARA	1	2074	TUNGUSKA	1
551	DALEKENIUM	2	1313	MARBELLA	1	2075	TUPPERWARE	1

552	DALEKS	315	1314	MARCIE	1	2076	TURKEY	5
553	DALIAN	2	1315	MARCO	3	2077	TURKISH	1
554	DAMASCUS	1	1316	MARCONI	1	2078	TURMEZISTAN	5
555	DAMIEN	1	1317	MARCUS	2	2079	TWODIS	2
556	DAN	7	1318	MARGARET	16	2080	TWOSTREAMS	12
557	DANIEL	2	1319	MARIA	3	2081	TYBURN	2
558	DANIELS	2	1320	MARIE	3	2082	TYLER	87
559	DANISH	1	1321	MARINUS	1	2083	TYLERS	1
560	DANKE	1	1322	MARIO	1	2084	TYRA	1
561	DANNY	125	1323	MARION	1	2085	TYTHONIAN	1
562	DAPHNE	1	1324	MARLEY	3	2086	UBER	1
563	DARILLIUM	3	1325	MARPLE	3	2087	UGANDA	1
564	DARLIG	1	1326	MARRAKECH	1	2088	ULTRAMANCER	1
565	DARREN	1	1327	MARTHA	230	2089	ULTRAVOX	2
566	DARTMOOR	5	1328	MARTY	1	2090	ULV	2
567	DAS	1	1329	MARY	118	2091	UMBRINGEN	1
568	DAVE	29	1330	MARYLEBONE	2	2092	UMPTION	1
569	DAVENPORT	1	1331	MATILDA	1	2093	UMQRA	2
570	DAVEY	1	1332	MATT	13	2094	URSULA	12
571	DAVID	15	1333	MATTHEW	1	2095	USA	1
572	DAVIES	1	1334	MAUPERTUIS	1	2096	UTAH	5
573	DAVITCH	2	1335	MAXARODENFOE	2	2097	VADER	1
574	DAVROS	60	1336	MAYBELLINE	1	2098	VALERIE	2
575	DAWKINS	1	1337	MAYNARD	1	2099	VALEYARD	2
576	DE	36	1338	MAZDA	2	2100	VALHALLA	4
577	DEBBIE	1	1339	MCALLISTER	1	2101	VALKYRIE	4
578	DECEMBER	1	1340	MCAVOY	2	2102	VANDALEUR	4
579	DEL	1	1341	MCCRIMMON	1	2103	VASHTA	13
580	DELANEY	4	1342	MCDONNELL	5	2104	VASHTEE	2
581	DELAWARE	10	1343	MCFLY	1	2105	VASTRA	12
582	DELHI	1	1344	MCGINTY	1	2106	VATICAN	9
583	DELPHOX	1	1345	MCGRATH	3	2107	VATOR	3
584	DEMARCO	3	1346	MCLINTOCK	1	2108	VAUXHALL	1
585	DEMUS	1	1347	MECHANOID	1	2109	VAVOOM	2
586	DER	6	1348	MEDIZINICSHEN	1	2110	VEENA	4
587	DEREK	2	1349	MEDUSA	7	2111	VEGAS	3
588	DERREN	2	1350	MEEZ	2	2112	VELMA	7

589	DES	2	1351	MEG	2	2113	VELOSIANS	2
590	DESCARTES	1	1352	MEIN	1	2114	VELTINO	1
591	DEUX	1	1353	MELINA	4	2115	VENETIAN	1
592	DEVE	1	1354	MELISSA	2	2116	VENETIANS	1
593	DEWER	5	1355	MEMORYONSMELLS	27	2117	VENEZIA	2
594	DEWEY	1	1356	MEMPHIS	1	2118	VENEZUELA	2
595	DEXTRUS	6	1357	MENSCHLICHE	1	2119	VENICE	8
596	DI	4	1358	MERCHANDANI	1	2120	VENTURA	1
597	DIA	1	1359	MERCURIO	1	2121	VENUS	3
598	DIANA	2	1360	MESOPOTAMIA	1	2122	VENUSIAN	3
599	DICH	1	1361	MESSERSCHMITTS	1	2123	VERA	1
600	DICKEN	5	1362	METALTRON	3	2124	VERITAS	19
601	DIDCOT	1	1363	METEBELIS	2	2125	VERMEER	1
602	DIEGO	2	1364	METELLA	3	2126	VERNET	1
603	DIGBY	3	1365	METRAXIS	1	2127	VERONICA	1
604	DILLANE	1	1366	MEXICAN	1	2128	VERRIER	4
605	DIMMOCK	3	1367	MEXICANS	2	2129	VERRON	1
606	DIR	2	1368	MEXICO	6	2130	VERRUCKT	1
607	DIS	2	1369	MEZON	1	2131	VERSAILLES	4
608	DISILLIUM	2	1370	MIA	1	2132	VESPIFORM	6
609	DISNEYLAND	1	1371	MICH	2	2133	VESPIFORMS	1
610	DISS	2	1372	MICHAEL	4	2134	VESTAL	3
611	DIVO	1	1373	MICHELANGELO	1	2135	VESTONES	2
612	DOCH	1	1374	MICHELIN	1	2136	VESUVIUS	10
613	DOCHERTY	5	1375	MICHIGAN	1	2137	VICI	3
614	DOCTORDONNA	3	1376	MICK	3	2138	VICKY	3
615	DOKO	1	1377	MICKETTY	1	2139	VICTOR	12
616	DONALD	1	1378	MICKEY	100	2140	VICTORIA	12
617	DONCASTER	1	1379	MICKEYS	1	2141	VICTORIAN	20
618	DONNA	226	1380	MIKE	13	2142	VIDFONE	1
619	DONNELL	16	1381	MIKEY	1	2143	VIDGAMES	1
620	DONOVAN	12	1382	MILHOUS	1	2144	VIDI	3
621	DOOMFINGER	2	1383	MILLIGAN	8	2145	VIENNA	5
622	DOPPELGANGERS	1	1384	MINDHURTY	1	2146	VIENNESE	1
623	DORABELLA	5	1385	MINNEAPOLIS	1	2147	VIETNAM	2
624	DOREEN	1	1386	MIR	1	2148	VIKING	5
625	DORIUM	10	1387	MIRANDA	3	2149	VIKINGS	9

626	DORSET	1	1388	MISSOURI	1	2150	VILLE	1
627	DOS	2	1389	MITCHELL	1	2151	VILLENGARD	4
628	DOTTIE	1	1390	MITZVAH	1	2152	VINCENT	31
629	DOUGLAS	5	1391	MOFFAT	2	2153	VINCEY	3
630	DOWNING	14	1392	MOHAMMED	1	2154	VINCI	1
631	DRAHVIN	2	1393	MOHANDES	1	2155	VINNY	1
632	DREARCLIFF	1	1394	MOIRA	1	2156	VINVOCCI	2
633	DRISCOLL	1	1395	MOLEY	1	2157	VIRGINIA	1
634	DROOD	2	1396	MOLLY	53	2158	VISCOTOXINS	1
635	DRYADS	4	1397	MONDAS	2	2159	VITEX	4
636	DU	4	1398	MONDASIAN	2	2160	VITUS	1
637	DUBLIN	1	1399	MONDAY	10	2161	VIVIAN	3
638	DUCKWORTH	1	1400	MONKFORD	11	2162	VIVIEN	1
639	DUF	3	1401	MONKY	2	2163	VOILA	1
640	DUMPTY	1	1402	MONTY	2	2164	VON	2
641	DUNBAR	1	1403	MOOKY	1	2165	VONE	2
642	DUNKIRK	2	1404	MOORE	12	2166	VOSSAHEEN	1
643	DUNLOP	1	1405	MOORHOUSE	3	2167	VOT	5
644	DUPONT	1	1406	MORAN	19	2168	VOTIVIG	1
645	DURAN	2	1407	MORGENSTERN	4	2169	WACEY	2
646	DURRANTS	1	1408	MORIARTY	95	2170	WACHT	1
647	DURY	2	1409	MORITURI	1	2171	WAINWRIGHT	1
648	DUSSELDORF	1	1410	MOROCCO	2	2172	WALES	3
649	DUTCH	2	1411	MORPETH	1	2173	WALFORD	1
650	DWALLS	1	1412	MORPHEUS	30	2174	WALKER	1
651	DYACHENKO	1	1413	MORPHIC	3	2175	WALKIE	1
652	DYLAN	1	1414	MORPOTH	1	2176	WALLBANGER	3
653	DZUNDZA	2	1415	MORRISON	3	2177	WALSH	4
654	EALING	1	1416	MORSTAN	2	2178	WALTER	1
655	EASTENDERS	1	1417	MORTARIUM	1	2179	WALTONS	1
656	EAU	1	1418	MORTIMER	6	2180	WANDSWORTH	1
657	EBAY	1	1419	MORTLOCK	1	2181	WANZ	2
658	ECTOSHINE	1	1420	MORVIN	6	2182	WARPFOLD	1
659	EDDIE	15	1421	MOSCOW	4	2183	WARPSPEED	1
660	EDDISON	15	1422	MOSES	2	2184	WASHINGTON	4
661	EDDY	1	1423	MOTT	2	2185	WASTERIDGE	1
662	EDEN	3	1424	MOUNTBATTEN	1	2186	WATERGATE	1

663	EDINBURGH	2	1425	MOUY	1	2187	WATERLOO	4
664	EDMUND	2	1426	MOXX	3	2188	WATSON	160
665	EDWARD	1	1427	MOYA	1	2189	WATSONS	1
666	EDWARDS	2	1428	MULLIGAN	5	2190	WEBBER	10
667	EDWIN	7	1429	MURRAY	3	2191	WEBLEY	5
668	EFFIE	1	1430	MUSGRAVE	2	2192	WEDNESDAY	12
669	EGON	1	1431	MUSSOLINI	1	2193	WEDNESDAYS	2
670	EGYPT	6	1432	MUSWELL	3	2194	WEHTUN	1
671	EGYPTIAN	5	1433	MYCROFT	88	2195	WELLIE	1
672	EGYPTIANS	1	1434	MYCROFZ	4	2196	WELLIES	2
673	EILEEN	1	1435	MYRNA	1	2197	WELLINGTON	1
674	EIN	1	1436	NABEELA	1	2198	WELSBOROUGH	12
675	EINARR	3	1437	NACHT	2	2199	WELSBOROUGHIS	1
676	EINFACH	1	1438	NAGATA	13	2200	WELSH	5
677	EINS	1	1439	NAINBY	2	2201	WELSHMAN	1
678	EINSAM	1	1440	NAISMITH	2	2202	WEMBLEY	3
679	EINSTEIN	1	1441	NAKASHIMA	2	2203	WENCESLAS	3
680	EKNODINES	1	1442	NAPLES	3	2204	WERDEN	1
681	EL	2	1443	NAPOLEON	9	2205	WEREN	8
682	ELDANE	4	1444	NARDIE	1	2206	WESSEX	2
683	ELENA	1	1445	NARDOLE	54	2207	WESTIE	7
684	ELINA	1	1446	NASREEN	11	2208	WESTMINSTER	7
685	ELIOT	2	1447	NATHAN	3	2209	WESTWOOD	1
686	ELISE	1	1448	NAVARRE	1	2210	WHITEHALL	1
687	ELIZA	9	1449	NAZI	2	2211	WHITEPOINT	3
688	ELIZABETH	9	1450	NAZIS	1	2212	WHITNEY	5
689	ELIZABETHAN	1	1451	NEBULA	5	2213	WIBBLY	10
690	ELLA	1	1452	NEBULAS	1	2214	WIBBY	1
691	ELLIE	2	1453	NEBULOUS	2	2215	WICOWSKY	1
692	ELLIOT	27	1454	NEE	1	2216	WIGGINS	3
693	ELLIS	1	1455	NEFERTITI	6	2217	WIGMORE	1
694	ELLO	1	1456	NEFFY	7	2218	WII	1
695	ELO	2	1457	NEGAPACT	1	2219	WIKIPEDIA	1
696	ELSIE	1	1458	NEIL	3	2220	WILDE	1
697	ELTON	16	1459	NELL	1	2221	WILDEBEEST	1
698	ELTONS	2	1460	NELLY	1	2222	WILF	5
699	ELVIS	9	1461	NELSON	4	2223	WILFRED	7

700	EMELIA	19	1462	NEMESIS	2	2224	WILLIAM	5
701	EMIL	1	1463	NERADA	13	2225	WILLIAMS	21
702	EMILION	1	1464	NERYS	1	2226	WILLY	1
703	EMILY	1	1465	NESTENE	10	2227	WILSON	13
704	EMMA	9	1466	NESTENES	1	2228	WILTSHIRE	2
705	EN	8	1467	NETFLIX	1	2229	WIMEY	20
706	ENGLAND	44	1468	NETHERSPHERE	6	2230	WINDIBANK	1
707	ENGLISH	36	1469	NEVADA	1	2231	WINFOLD	1
708	ENGLISHMAN	1	1470	NEVILLE	4	2232	WINIFRED	1
709	ENID	1	1471	NEWINGTON	1	2233	WINSTON	15
710	ENTSCHULDIGEN	1	1472	NEWTON	1	2234	WIR	1
711	ENZOMODAN	1	1473	NIAGARA	1	2235	WIRD	1
712	ENZOMODONS	1	1474	NICHT	3	2236	WOLLEN	1
713	EPCOT	1	1475	NICOLA	1	2237	WOLLT	1
714	EREHWON	1	1476	NIEMAND	1	2238	WOODBRIDGE	7
715	ERIC	1	1477	NIGHTINGALE	6	2239	WOOLF	1
716	ERICA	4	1478	NILLY	1	2240	WOOLWICH	2
717	ERINA	2	1479	NIMMER	1	2241	WOORS	1
718	ERISA	1	1480	NIMON	1	2242	WORA	2
719	ERNIE	5	1481	NIMROD	2	2243	WORDSMITH	2
720	ERROL	1	1482	NINA	7	2244	WORLDERS	1
721	ESH	3	1483	NIXON	6	2245	WUBBLY	1
722	ESPALDA	3	1484	NOBIS	1	2246	WYCOMBE	3
723	ESSEX	1	1485	NONNY	4	2247	XIAOLIAN	1
724	ESTEE	1	1486	NOOR	2	2248	YAMUNA	1
725	ETHIOP	1	1487	NORBURY	4	2249	YANA	7
726	ETRUSCANS	1	1488	NORTHUMBERLAND	8	2250	YANGTZE	1
727	EUROPA	1	1489	NORWAY	5	2251	YAO	4
728	EUROPE	8	1490	NORWEGIAN	2	2252	YERS	2
729	EUROPEAN	2	1491	NOTTINGHAM	10	2253	YEUCH	1
730	EUROVISION	1	1492	NOVEMBER	16	2254	YORK	51
731	EURUS	34	1493	NOZ	4	2255	YORKS	1
732	EUSTACE	24	1494	NUREMBERG	1	2256	YORKSHIRE	6
733	EVA	1	1495	NYMPHS	1	2257	YOWZ	1
734	EVAC	6	1496	NYSSA	2	2258	YOWZA	2
735	EVANS	3	1497	OAKHAM	1	2259	YOWZAH	4
736	EVARIN	11	1498	OBEGO	1	2260	YURI	1

737	EVELINA	12	1499	OCTOBER	1	2261	YVONNE	11
738	EVELYN	1	1500	ODIN	13	2262	ZACHARY	2
739	EVEREST	1	1501	ODYSSEY	1	2263	ZACK	17
740	EVERETT	6	1502	OFFY	2	2264	ZAFFIC	1
741	EVIE	1	1503	OFSTED	1	2265	ZEALAND	1
742	EWAN	1	1504	OKLAHOMA	4	2266	ZED	2
743	EWART	4	1505	OLIVER	7	2267	ZEN	1
744	EXEDOR	5	1506	OLVERON	1	2268	ZENITH	1
745	EXETER	2	1507	ONEGIN	2	2269	ZEUS	2
746	EXPELLIARMUS	3	1508	OOD	126	2270	ZHI	4
747	EXXILON	1	1509	OODKIND	2	2271	ZHOU	3
748	FACEBOOK	6	1510	OODLES	2	2272	ZHU	4
749	FAIRFORD	2	1511	OODS	3	2273	ZHUKOV	1
750	FAIRWEATHER	1	1512	OOGLY	1	2274	ZIMMERMAN	1
751	FALKLANDS	1	1513	OPRAH	1	2275	ZIPPO	1
752	FANSHAWE	2	1514	OPTOGRAM	1	2276	ZOE	1
753	FAUNTLEROY	1	1515	ORRIBLE	4	2277	ZOG	1
754	FAWKES	1	1516	ORRIE	1	2278	ZORRO	1
755	FEBRUARY	5	1517	ORSAY	2	2279	ZOVIRAX	1
756	FECUNDITATIS	2	1518	ORSON	6	2280	ZUR	2
757	FEL	5	1519	ORWELLIAN	1	2281	ZYGELLA	6
758	FELMAN	3	1520	OSCAR	2	2282	ZYGON	42
759	FELSPOON	3	1521	OSGOOD	42	2283	ZYGONKIND	1
760	FERMAT	2	1522	OSGOODS	5	2284	ZYGONS	31
761	FERRARI	3	1523	OSHOGASU	1	2285	ZYTON	1
762	FERRIN	2	1524	OSKAR	1		**TOTAL FREQUENCY**	**5382**

Appendix 7- List of Contracted Forms Excluded from the BTSC

	ITEM	FREQ
1	AIN	6
2	AINT	1
3	CANNOT	7
4	CANT	1
5	DIDN	73
6	DIDNT	1
7	DOESN	21
8	HADN	5
9	HASN	4
10	HAVEN	27
11	HAVENT	1
12	HED	1
13	IM	1
14	INNIT	13
15	ISN	71
16	MUSTN	4
17	NEEDN	1
18	RE	82
19	RU	4
20	SHOULDN	7
21	WASN	29
22	WHOD	1
23	WOTCHA	1
24	WOULDA	1
25	WOULDN	18
TOTAL FREQUENCY		**381**

Appendix 8- List of Exclamations and Filler Words Excluded from the BTSC

	ITEM	FREQ		ITEM	FREQ		ITEM	FREQ
1	AAAAAAAAARGH	1	78	GOOO	2	155	PGPR	1
2	AAAAAARGH	1	79	GOSH	6	156	PHEW	6
3	AAAAARGH	2	80	GRROOSSS	2	157	PISH	1
4	AAAAGH	2	81	HA	223	158	POW	1
5	AAAAGHHH	1	82	HAH	5	159	PSST	2
6	AAAAH	1	83	HAHA	4	160	PUFF	2
7	AAAAHHH	2	84	HAI	1	161	RARGH	1
8	AAAARGH	10	85	HEEBIE	1	162	ROAAJ	1
9	AAAARRGGHH	1	86	HEY	283	163	RRRRRRRR	1
10	AAAGH	3	87	HI	108	164	RRRRUFF	1
11	AAAH	5	88	HIYA	8	165	SASHAY	1
12	AAARGH	18	89	HM	34	166	SAYONARA	1
13	AAGH	5	90	HMM	197	167	SBLOOD	1
14	AAH	16	91	HMMM	2	168	SH	19
15	AARGH	20	92	HMPH	2	169	SHH	63
16	AARRGH	1	93	HMS	1	170	SHHH	16
17	AARRRGGH	1	94	HO	21	171	SHHHH	1
18	AGGH	3	95	HOHO	1	172	SHMAVITY	1
19	AGH	29	96	HOO	15	173	SHMUCKS	1
20	AH	417	97	HOORAY	6	174	SOO	10
21	AHA	5	98	HOWDY	2	175	SQUEEEEE	1
22	AHEM	5	99	HUH	71	176	SSH	32
23	AHH	10	100	HUMAAAAANS	1	177	SSHH	3
24	AHHH	6	101	HURRAH	6	178	SSHHH	1
25	AHHHH	2	102	HURRAY	1	179	SSSH	12
26	ALAS	2	103	HUUUUUU	3	180	SSSSH	1
27	ARG	1	104	HYAH	1	181	UGH	15
28	ARGH	77	105	HYYA	7	182	UGHHH	1
29	ARGHHHHH	1	106	IA	3	183	UH	103
30	ARRGGH	4	107	II	5	184	UHH	1
31	ARRRRGH	1	108	III	6	185	UM	226
32	AW	19	109	JEEBIES	1	186	UMM	2
33	AWS	1	110	JEEZ	2	187	URGH	8
34	AWW	2	111	KABOOM	1	188	URGHHH	1
35	AWWWK	1	112	MAAAN	3	189	URRRRR	1
36	AY	12	113	MEH	3	190	UUGH	1
37	AYE	31	114	MMM	52	191	WAAARD	1
38	BA	3	115	MMMM	1	192	WAARGH	1

39	BAA	5	116	MWAH	2	193	WAH	5
40	BAH	3	117	MWUH	1	194	WAHEY	2
41	BAM	1	118	NA	2	195	WAKEY	15
42	BLAH	9	119	NAAAH	2	196	WH	2
43	BLIMEY	51	120	NADA	2	197	WHA	2
44	BO	5	121	NAH	45	198	WHAM	3
45	BOLLOCKS	2	122	NIX	1	199	WHAZ	4
46	BOO	16	123	NOOO	1	200	WHEE	1
47	BRR	1	124	NOOOO	2	201	WHEEE	1
48	BYEEEE	1	125	NOOOOOO	1	202	WHEW	1
49	BZZZ	1	126	NOOOOOOO	1	203	WHOA	108
50	CRIKEY	2	127	NOOOOOOOO	1	204	WHOO	15
51	DAH	1	128	NOPE	35	205	WHOOMPH	2
52	DAHH	2	129	NUH	1	206	WHOOO	1
53	DAMMIT	3	130	OCH	3	207	WHOOPS	5
54	DAMN	45	131	OH	3189	208	WHOOPSY	1
55	DOKEY	1	132	OHH	17	209	WI	31
56	EE	5	133	OHHH	14	210	WO	3
57	EEEEUUUURRR	1	134	OHHHHH	1	211	WOAH	4
58	EGADS	1	135	OHHHHHH	1	212	WOO	15
59	EGGGGZZZ	1	136	OI	119	213	WOW	34
60	EGGGZZZZZZZZ	1	137	OIK	1	214	YA	24
61	EGGZZ	6	138	OM	2	215	YAAAARGH	1
62	EH	121	139	OMM	2	216	YAAH	1
63	EM	59	140	OO	9	217	YARGH	2
64	ER	245	141	OOH	132	218	YAY	6
65	ERM	92	142	OOOH	12	219	YE	3
66	ERRRM	1	143	OOOOOH	1	220	YEA	21
67	EUGH	4	144	OOPS	20	221	YEAH	1735
68	EURGH	3	145	OOR	12	222	YECCH	1
69	EW	4	146	OORS	6	223	YEE	2
70	EWW	2	147	OOSH	1	224	YEP	73
71	EXTERMINAAAAATE	1	148	OUCH	1	225	YO	11
72	EY	1	149	OW	88	226	YOO	1
73	FFFF	1	150	OWWWW	1	227	YOOKAY	2
74	GAH	1	151	OY	4	228	YOOROPEE	1
75	GERONIMO	13	152	PAH	1	229	YUCKY	1
76	GEZ	2	153	PFFRRRT	1	230	YUM	3
77	GOOA	1	154	PFPH	1	231	YUP	12

TOTAL FREQUENCY 1437

Appendix 9- List of Abbreviations Excluded from the BTSC

	ITEM	FREQ		ITEM	FREQ		ITEM	FREQ
1	BEV	1	51	GPS	1	101	PCM	1
2	CCTV	7	52	GRA	2	102	PD	1
3	CDS	1	53	HADS	2	103	PE	25
4	CERN	5	54	HB	1	104	PKD	2
5	CIA	7	55	HC	7	105	PL	1
6	CK	3	56	HQ	3	106	PLC	1
7	CLA	1	57	HR	2	107	PM	9
8	COMM	3	58	HT	6	108	PO	1
9	COMMS	9	59	IFL	5	109	PR	4
10	COS	181	60	IISN	1	110	PROF	1
11	CPR	1	61	IQ	3	111	PSI	9
12	CRA	1	62	IRA	1	112	PW	1
13	CRB	1	63	ISIS	1	113	QRA	1
14	CRIT	1	64	IST	4	114	RADS	2
15	CSI	1	65	ITV	3	115	RAF	2
16	CT	1	66	IV	3	116	RCH	1
17	CV	1	67	IVF	1	117	RD	10
18	DA	24	68	JK	1	118	RDF	2
19	DC	1	69	JR	1	119	REA	4
20	DEAA	1	70	KBO	2	120	RF	1
21	DEAAJ	1	71	KM	3	121	RGA	1
22	DL	10	72	KN	6	122	RIP	19
23	DNA	38	73	LA	8	123	RSVP	1
24	DOC	32	74	LBS	2	124	RX	1
25	DR	157	75	LE	10	125	RY	8
26	DRWG	5	76	LES	2	126	SCR	1
27	DTS	1	77	LL	12	127	SM	1
28	DVD	8	78	LTD	7	128	SMT	1
29	DVDS	14	79	MAO	1	129	SOS	4
30	DX	1	80	MIT	1	130	SS	2
31	EAA	3	81	ML	1	131	ST	57
32	EC	1	82	MM	47	132	TA	15
33	ED	26	83	MOI	1	133	TD	3
34	EEK	1	84	MP	11	134	TE	1

35	EMP	2	85	MPS	1	135	TEK	3
36	EST	4	86	MR	592	136	TH	62
37	ETC	2	87	MRI	1	137	TK	1
38	EXPED	1	88	NASA	8	138	TRO	1
39	FBI	5	89	NATO	5	139	TS	1
40	FET	2	90	NAV	8	140	TV	26
41	FL	2	91	ND	10	141	UFO	5
42	FM	1	92	NHS	2	142	UK	31
43	FS	2	93	NNYPD	2	143	UN	16
44	FTL	2	94	NS	1	144	USP	1
45	FYI	1	95	NTA	1	145	VA	1
46	GCSE	2	96	NW	1	146	VR	1
47	GFP	1	97	OCD	4	147	WWW	52
48	GHT	1	98	OTA	1	TOTAL FREQUENCY		1790
49	GM	1	99	PA	13			
50	GP	1	100	PC	4			

Appendix 10- List of Function words in the BTSC

	ITEM	FREQ		ITEM	FREQ		ITEM	FREQ
1	A	16056	53	IT	17808	105	THREE	375
2	ABOUT	2064	54	MAY	198	106	THROUGH	573
3	ABOVE	71	55	MIGHT	495	107	THROUGHOUT	21
4	ACCROSS	1	56	MILLION	183	108	THY	9
5	ACROSS	142	57	NEITHER	60	109	TILL	126
6	AINT	29	58	NINE	92	110	TO	14781
7	ALL	4225	59	NINETEEN	3	111	TOWARDS	24
8	ALTHOUGH	33	60	NINETY	3	112	TRILLION	13
9	AMIDST	1	61	NO	5728	113	TWELVE	10
10	AMONG	30	62	NOBODY	172	114	TWENTY	17
11	AMONGST	11	63	NOR	16	115	TWO	819
12	AND	9914	64	NOT	14401	116	UNDER	230
13	ANY	896	65	NOTHING	755	117	UNLESS	94
14	ANYBODY	52	66	NOTWITHSTANDING	1	118	UNTIL	150
15	ANYONE	366	67	OF	9856	119	UNTO	4
16	ANYTHING	586	68	OFF	941	120	UP	2245
17	AS	1593	69	OK	1381	121	UPON	58
18	AT	2079	70	ON	5207	122	UPTO	1
19	BE	36702	71	ONE	3012	123	VERSA	2
20	BECAUSE	924	72	ONTO	51	124	VIA	4
21	BELOW	44	73	ONWARDS	6	125	WE	10010
22	BESIDE	22	74	OR	1257	126	WHAT	9167
23	BETWEEN	123	75	OTHERWISE	36	127	WHATEVER	262
24	BEYOND	79	76	OUGHT	12	128	WHATSOEVER	5
25	BILLION	109	77	OUT	2627	129	WHEN	1273
26	BY	998	78	PER	19	130	WHENEVER	14
27	CAN	5111	79	PERCENT	3	131	WHERE	1753
28	DOWN	1077	80	SEVEN	126	132	WHEREABOUTS	2
29	DURING	25	81	SEVENTEEN	3	133	WHEREAS	5
30	EIGHT	72	82	SHE	4871	134	WHEREBY	1
31	EIGHTEEN	1	83	SHOULD	879	135	WHEREVER	39
32	EITHER	73	84	SIX	167	136	WHETHER	25
33	ELEVEN	18	85	SIXTEEN	4	137	WHICH	485
34	ENOUGH	414	86	SIXTY	4	138	WHILE	205

35	EVERYBODY	131	87	SO	3516	139	WHILST	3
36	EVERYMAN	3	88	SOMEBODY	117	140	WHO	2022
37	EVERYONE	439	89	SOMEONE	521	141	WHOEVER	55
38	EVERYTHING	577	90	SOMETHING	1252	142	WHOM	18
39	FIFTEEN	7	91	SUCH	189	143	WHOMEVER	1
40	FIFTY	7	92	THAN	567	144	WHOSE	37
41	FIVE	312	93	THAT	10480	145	WHY	2115
42	FOR	4116	94	THE	26993	146	WILL	4650
43	FORTY	12	95	THEREFORE	19	147	WITH	2843
44	FOUR	201	96	THESE	594	148	WITHIN	70
45	FOURTH	17	97	THEY	7389	149	WITHOUT	230
46	FROM	1764	98	THIRD	32	150	WOULD	1480
47	HE	7903	99	THIRTEEN	3	151	YES	1820
48	HOW	2425	100	THIRTY	4	152	YOU	34299
49	HOWEVER	43	101	THIS	6079	153	ZERO	78
50	I	36937	102	THOSE	676		TOTAL FREQUENCY	148578
51	IF	2470	103	THOU	1			
52	IN	7273	104	THOUSAND	98			

Appendix 11- 37 words in the BNC but not in the BTSC (with minimum frequency of 10 per million)

	ITEM		ITEM		ITEM
1	agreement	14	eighty	27	product
2	alright	15	election	28	production
3	application	16	fifteen	29	role
4	award	17	financial	30	seventy
5	better	18	firm	31	sixty
6	chapter	19	including	32	tax
7	committee	20	less	33	thirty
8	concerned	21	manager	34	thus
9	county	22	mine	35	training
10	despite	23	nineteen	36	twelve
11	development	24	ninety	37	various
12	economic	25	our		
13	education	26	percent		

T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
Eğitim Bilimleri Enstitüsü Müdürlüğü

KONYA

Özgeçmiş

Adı Soyadı:	**Hatice Sezgin**	İmza:	
Doğum Yeri:	**Kütahya**		
Doğum Tarihi:	**04.01.1984**		
Medeni Durumu:	**Bekar**		

Öğrenim Durumu

Derece	Okulun Adı	Program	Yer	Yıl
İlköğretim	**Milli Egemenlik İ.O.**		**Tavşanlı**	**1994**
Ortaöğretim	**Anadolu İ.H.L.**		**Tavşanlı**	**1998**
Lise	**Anadolu Öğr. Lisesi**	**Y. Dil Pr.**	**Kütahya**	**2001**
Lisans	**Hacettepe Ün.**	**İng. Öğr.**	**Ankara**	**2010**
Yüksek Lisans	**Necmettin Erbakan Ü.**	**İng. Öğr.**	**Konya**	**2019**

İş Deneyimi:	**Selçuk Üni. Y.D.Y.O. (2011-)**
Hakkımda bilgi almak için önerebileceğim şahıslar:	**Prof. Dr. Arif Sarıçoban** **Dr. Öğr. Üyesi Mustafa Serkan Öztürk** **Dr. Öğr. Üyesi Mustafa Dolmacı**
Tel:	**05443988430**
Adres	**Selçuk Üni. Y.D.Y.O.**