Tampere University

Niina Ikonen

BIG AND LARGE

A Corpus Study on Near-Synonyms

Faculty of Information Technology and Communication Sciences
Bachelor's Thesis
January 2022

ABSTRACT

Niina Ikonen: *Big* and *Large* – A Corpus Study on Near-Synonyms
Bachelor's Thesis
Tampere University
English Language, Literature and Translation
January 2022

The purpose of this bachelor 's thesis was to analyze how the very common English adjectives *big* and *large,* which can be considered as near-synonyms, differ in their contexts and meanings. Factors influencing the choice of the topic included interest in near-synonymy and corpus linguistics. In addition, it turned out that no comprehensive corpus-based research had been conducted on the topic in the past. The study was conducted by comparing the collocations of these adjectives, i.e., the tendency to occur repeatedly in conjunction with certain other words, in this case nouns. The initial assumption was that *big* is more common in colloquial language, while *large* occurs more often in written language.

The theoretical part of this thesis introduces the concepts of synonymy and near-synonymy and presents some previous studies on corpus linguistics and near-synonymy which this bachelor's thesis is based on. The research method utilized in the thesis was corpus linguistics. Research material was collected from the *Corpus of Contemporary American English* which is an extensive corpus of more than a billion words, containing modern English texts. The searches were made using the comparison feature of the corpus as it allows to search collocates of two words simultaneously. The study was limited to contexts where *big* and *large* are in the basic form functioning as attributive adjectives. In addition to the corpus data, two dictionaries of contemporary English were utilized in the study. The definitions of the words *big* and *large* in these dictionaries were compared with the results obtained from corpus material.

Based on the results of the study using this particular material, it can be concluded that the original hypothesis regarding the occurrence of *big* and *large* was correct. In addition, the study revealed that *big* occurs very often in a figurative sense, while *large* is most often used in a literal sense. The study also showed that dictionaries were not able to describe the differences in meaning between these two words as accurately as the corpus material. Based on the results, it is also possible to identify opportunities for further research of the same topic. In the future it could be interesting to analyze how people who speak English as a foreign or second language use these adjectives. The topic could also be studied from a more qualitative perspective by examining whether the adjectives *big* and *large* have positive or negative connotations in different contexts.

Keywords: synonymy, near-synonymy, corpus linguistics, frequency, collocation, collocate, semantics

TIIVISTELMÄ

Tämän kandidaatintutkielman tarkoituksena oli analysoida miten erittäin tavalliset englannin kielen adjektiivit *big* ja *large,* joita voidaan pitää toistensa lähisynonyymeina, eroavat toisistaan käyttöyhteyksiltään ja merkityksiltään. Aiheen valintaan vaikuttavia tekijöitä olivat mm. kiinnostus lähisynonymiaan ja korpuslingvistiseen kielentutkimukseen. Lisäksi osoittautui, ettei aiheesta ole aikaisemmin tehty kattavaa korpuspohjaista tutkimusta. Tutkimus suoritettiin vertailemalla kyseisten adjektiivien kollokaatioita eli taipumusta esiintyä toistuvasti yhdessä tiettyjen toisten sanojen, tässä tutkimuksessa substantiivien, kanssa. Alkuperäinen oletus oli, että *big* on yleisempi puhekielessä, kun taas *large* esiintyy useammin kirjakielessä.

Tutkielman teoreettisessa osuudessa perehdytään synonymian ja lähisynonymian käsitteisiin ja esitellään joitakin aiempia korpuslingvistiikkaa ja lähisynonymiaa käsitteleviä tutkimuksia, joihin tämä kandidaatintutkielma pohjautuu. Tutkimusmetodiksi valittiin korpuslingvistinen kielentutkimus. Tutkimusaineisto kerättiin *Corpus of Contemporary American English* -korpuksesta. Kyseinen korpus on laaja, yli miljardi sanaa käsittävä tietokanta, joka sisältää nykyenglannin tekstejä. Hakuihin käytettiin vertailuominaisuutta, jonka mahdollistaa kahden sanan kollokaattien haun samanaikaisesti. Tutkimus rajattiin yhteyksiin, joissa *big* ja *large* ovat perusmuodossa substantiivin attribuuttina. Lisäksi tutkimuksessa käytettiin kahta nykyenglannin sanakirjaa, joiden antamia määritelmiä sanoille *big* ja *large* verrattiin korpusaineistosta saatuihin tuloksiin.

Kyseisestä aineistosta saatujen tutkimustulosten perusteella voidaan todeta, että alkuperäinen hypoteesi koskien sanojen esiintyvyyttä piti paikkansa. Lisäksi tutkimuksessa kävi ilmi, että *big* esiintyy hyvin usein figuratiivisessa merkityksessä, kun taas *large* sanaa käytetään useimmiten kirjaimellisessa merkityksessä. Tutkimus myös osoitti, etteivät sanakirjat kyenneet kuvaamaan näiden kahden sanan merkityseroja yhtä tarkasti kuin korpusaineisto. Tulosten perusteella voidaan myös esittää mahdollisuuksia tutkielman aiheen käsittelyyn jatkossa. Jatkossa voisi esimerkiksi tutkia, miten englantia vieraana tai toisena kielenä puhuvat käyttävät kyseisiä adjektiiveja. Aihetta voisi myös tarkastella enemmän kvalitatiivisesta näkökulmasta tutkimalla onko adjektiiveilla *big* ja *large* eri käyttöyhteyksissä positiivia tai negatiivisia painotuksia.

Avainsanat: synonymia, lähisynonymia, korpuslingvistiikka, frekvenssi, kollokaatio, kollokaatti, semantiikka

Table of contents

1. Introduction

Understanding how to choose the correct word in each context is important for any language user, however, making the correct choice between words that are very close in meaning can be challenging. As Yule points out: "Whereas the word answer fits in this sentence: Karen had only one answer correct on the test, its near-synonym reply would sound odd" (Yule 1985, 95).

Other two very common English words which most language users would consider fully synonymous are *big* and *large*. At first, they probably seem to be fully identical in meaning and would be able to substitute for each other in any sentence but when studied more closely this is in fact not the case.

The purpose of this thesis is to analyze the differences as well as similarities of meaning and use of these two very frequently occurring adjectives which many language users would consider full synonyms, when in fact more correct would be to call them near-synonyms. The analysis will be done by comparing the dictionary definitions of *big* and *large* as well as by analyzing their immediate noun collocates extracted from the Corpus of Contemporary American English (COCA) which is a comprehensive corpus containing contemporary American English. The initial assumption is that *big* is more frequently used in informal contexts while *large* occurs more often in formal contexts.

The corpus-linguistic tools which will be utilized in the analysis are frequency count and collocation. Frequency count will give a more general overview on the differences of *big* and *large* while the collocational data will be used for more detailed analysis. Concordance analysis is beyond the scope of this paper, as the objective is to analyse the adjectives *big* and *large* with their immediate noun collocates and as the collocational analysis turned out to provide enough detailed data. The concept of synonymy, with specific focus on near-synonymy, will also be discussed in more detail. Some previous research related to near-synonymous adjectives will also be introduced.

The topic was chosen due to interest in data analysis, corpus linguistics. The adjectives *big* and *large* were chosen as they are among the most frequently used adjectives in English and therefore there is a large amount of data available in COCA to be utilized for a detailed analysis. There is also very little previous research available related to these two lexemes.

2. Literature Review

This chapter will explain the concepts absolute and near-synonymy as well as discuss some previous research related to near-synonyms.

2.1 Absolute Synonymy

Most authors agree that sameness in meaning does not actually mean full synonymy. Cruse argues (1986, 268) that two lexical units would be considered absolute synonyms if and only if all their contextual relations were identical. Cruse goes as far as to say (1986, 270) that "natural languages abhor absolute synonyms just as nature abhors a vacuum". According to Lyons (1995, 61) expressions are absolute synonyms if, and only if, the following three conditions are met: (i) all their meanings must be identical; (ii) they must be synonymous in all contexts; and (iii) they must be semantically equivalent (i.e., their meaning or meanings must be identical) on all dimensions of meaning, descriptive and non-descriptive. Lyons also states (1981, 148) that absolute synonyms may co-exist only for a short time until one of them begins to be used as the standard term. Palmer argues (1976, 60) similarly that there are no real synonyms as two words with exactly the same meaning would not survive in language.

When it comes to finding examples of absolute synonyms only a few authors name some of them. They are mostly nouns which refer to the same object. Lyons (1981, 148) states that candidates for absolute synonyms can be found in highly specialized vocabulary that is purely descriptive. Lyons names (ibid.) the words *caecitic* and *typhilitis*, which both refer to inflammation

of the blind gut, as possible examples of absolute synonyms. Similarly, Murphy argues (2010, 111) that candidates for perfect synonymy can be found among technical names for plants, animals, and chemicals such as *furze/ gorse/whin* which all refer to the same plant and *groundhog/woodchuck* both of which denote the same animal.

2.2 Near-Synonymy

While most authors define absolute synonymy with some clarity as a rare phenomenon, defining near-synonymy or plesionymy is more complex as the definitions vary from author to author.

Inkpen & Hirst state (2010: 1) that "near-synonyms are words that are almost synonyms, but not quite. They are not fully inter-substitutable, but rather vary in their shades of denotation or connotation, or in the components of meaning they emphasize". Murphy defines (2010, 111) near-synonyms simply as words that are intersubstitutable in some contexts, but not in every context. Lyons defines (1995, 60) near synonyms in brief as expressions that are similar but not identical in meaning. He lists (ibid.) the following words as typical examples of near-synonyms: *mist/fog, stream/brook, dive/plunge.*

Probably the most comprehensive research on near-synonymy has been conducted by Cruse. According to Cruse (1986, 285) near-synonyms or plesionyms are not mutually entailing and they can even contrast in certain contexts: 'It wasn't *foggy* today, just *misty'*. However, Cruse stresses in his later work (2000, 159, 160) that differences between near-synonyms must be either minor, or backgrounded, or both. The following can be considered as minor differences: (i) adjacent position on scale of 'degree': *fog/mist, laugh/chuckle, hot/scorching, big/huge, disaster/catastrophe, pull/heave, weep/sob,* etc.; (ii) certain adverbial specializations of verbs: *amble/stroll, chuckle/giggle, drink/quaff;* (iii) aspectual distinctions: *calm/placid* (state vs. disposition); (iv) difference of prototype centre: *brave* (prototypically physical): *courageous* (prototypically involves intellectual and moral factors) (ibid.).

Other definitions for near-synonymy exist but for the purposes of this paper the term will be used to refer to words that can substitute each other in most contexts but not in all.

2.3 Near-Synonyms in Previous Research

Natural language processing is one of the areas where research on near-synonyms has proven useful. The aim of one of the recent studies by Inkpen & Hirst (2010, 14) was to develop a lexical knowledge base which would help to automatically select the correct word from a set of near-synonyms as making the wrong word choice in machine translation and natural language generation can be imprecise or awkward or convey unwanted implication (ibid.). Inkpen & Hirst (2010, 14) argue further that when making the choice between near-synonyms which share the same core meaning, but differ in their nuances, the collocational properties of those words can help to choose the best word in each context.

When it comes to previous corpus studies on near-synonymous adjectives Liu has used data from COCA to study the following five near-synonymous adjectives: chief, *main, major, primary,* and *principal*. In his study he focused specifically on their distributional patterns, especially the typical types of nouns that they each modify. According to Liu (2010, 56) by examining the nouns the adjectives modify it was possible to identify significant granulated differences in meaning and use among the five near-synonyms. This approach will be followed in this thesis as well.

2.4 Previous Research on *Big* and *Large*

It seems that *big* and *large* have not been studied in more detail until now, but they have been mentioned by some authors (Cruse, 2011. 157; Lyons, 1995, 61-6; Saeed, 2016, 62; Taylor, 2002, 26), mostly very briefly though. Murphy, however, presents (2003, 39) a slightly more detailed analysis and mentions one interesting difference of meaning between *big* and *large. Big* may communicate ‘importance’ in a way that *large* does not (ibid.). According to Murphy (ibid.) *big*

appears often in nick names (*Big John* or the *Big Apple*) and indicates not only physical size but also is used in more affective sense (ibid.).

All the authors mentioned above, apart from Taylor, base their collocational analysis on their own intuitions while linguists today more and more use text corpora for identifying the co-occurrence patterns of a given word. The use of corpora as a tool of linguistic analysis will be described in more detail in the next section.

3. Methods and Materials

This section will provide an introduction to corpus linguistics and describe the research methods used.

3.1 Introduction to Corpus Linguistics

Corpus linguistics is a field within linguistics which has become more and more the area of interest for the linguists in recent decades. According to Bennett (2010, 2) the aim of corpus linguistics is to answer the following two questions: (i) which patterns are associated with lexical or grammatical features and (ii) how these patterns differ within varieties and registers.

Tognini-Bonelli (2001, 65) makes a distinction between *corpus-based* and *corpus-driven* linguistics. The term *corpus-based* refers to a methodology which is used to test or exemplify statements or hypotheses that were made before large corpora were available (ibid.). *Corpus-driven* refers to an approach in which a corpus itself should be used as the only source for validating linguistic statements and hypotheses (ibid.). The analysis in this thesis follows the corpus-driven approach.

While the corpus-driven approach is well suited for lexical semantic interpretations based on collocation it has some limitations when used for analyzing syntactic data. Chomsky (in Oliviéri 1, 2010) states that ‘it is obvious that the set of grammatical sentences cannot be identified with any corpus of utterances obtained by the linguist in his field work’. According to the Chomskyan

formalist view language should be analyzed in isolation and examples can in principle be invented by the linguist (ibid.). Quite contrarily, British linguists such as Firth, Halliday and Sinclair argue that language should be studied in actual and authentic contexts such as the corpus (Stubbs 8, 1993). Meurers and Müller (2014, 1) state that as corpora offer a wide variation of lexical, syntactic, semantic, and contextual properties they can help the researcher to get a better overview on which of these properties are relevant for the syntactic pattern they are analyzing. According to a recent study conducted by Meurers and Müller (2014, 11) searching relevant corpus examples in a syntactically annotated newspaper corpus offers a significant contribution to empirically grounded syntactic research. However, Meurers and Müller (2014, 1) also discuss some issues and limitations related to using the corpora for syntactic research. One should be aware that 'even the largest corpora can only represent a finite subset of a language's infinite potential' (ibid.). It is also worth noting that annotated corpora exist only for a few languages: the Linguistic Data Consortium (LDC, http://www.ldc.upenn.edu) lists corpora only for 39 out of the 6000 of the world's living languages (ibid.). Meurers and Müller (2014, 3) also stress that annotated corpora contain annotation errors which can influence the results of the research.

3.2 Advantages and Limitations of Corpus DataBauer argues (2002, 102) that the main benefit of using the data from public corpora is replicability which is a sign of good science. According to McEnery and Hardy, (2012, 16) "a result is considered replicable if a replication of the method that led to it consistently produces the same result". Bauer states (2002, 103) that another benefit of corpus research is the possibility of approaching various linguistic phenomena numerically making corpora well-suited tools for quantitative research.

Even though corpora are very useful there are some limitations related to both quality and quantity of the corpus data which the researcher should take some caution in order to avoid producing inaccurate results or making overgeneralized conclusions. Bauer states (2002, 103) that by allowing quantitative treatment of various linguistic phenomena, corpus research may create a

false sense of accuracy. Comparing data from two or more different corpora can also be an issue as it can be difficult to deduce what is generating the differences (ibid.). Bauer also notes (2002, 103) that the corpus size may cause issues as it might not be large enough. On the other hand, too large a corpus can result in analyzing unnecessary data (ibid.). Chomsky (in Tognini-Bonelli, 2001, 51) advises caution regarding the quality of the data: 'Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list'.

3.3 The Corpora

The main tool within corpus linguistics is the corpus. Kennedy in Bauer (2002, 98) defines a corpus as a "body of written text or transcribed speech which can serve a basis of linguistic analysis and description".

According to Podesva et al. (2013, 269) the earliest corpora were collected in the 1960s. The Brown Corpus has a unique place in the history of corpus linguistics (ibid.). It represents the first systematic and large-scale attempt to sample written American English containing material which first appeared in print in the year 1961 (ibid.).

The Brown Corpus can be classified as a *sample corpus* (McEnery and Hardy, 2012, 6) based on its data collection regime. According to McEnery and Hardy (2012, 6), there are basically two methods which are utilized when collecting the data for corpora: the *sample* or *balanced corpus* approach which means that the data is collected according to a specific sampling frame and thus represents language at a given point in time and the *monitor corpus* approach which means that corpus expands and includes more and more data over time. In addition to the Brown Corpus, the Lancaster-Oslo-Bergen corpus (LOB) is an example of a *sample corpus* (ibid.). Examples of *monitor corpora* include the Bank of English (BoE), which is probably the best-known monitor corpus, and the Corpus of Contemporary American English (COCA) (ibid.).

3.3.1 Corpus of Contemporary American English

The Corpus of Contemporary American English is one of the most comprehensive corpora containing contemporary language.

The COCA was compiled by Professor Mark Davies at Brigham Young University. According to Davies (2010, 453) COCA is "the first reliable monitor corpus of English, and the first balanced monitor corpus of any language". An important feature in the design of the COCA is that the corpus is divided almost equally between spoken, fiction, popular magazines, newspapers, and academic journals - 20% in each genre (ibid.). The significant aspect is that the genre balance stays almost exactly the same from year to year (ibid.).

COCA will be utilized in this paper as it contains over one billion words of texts from various genres. COCA was chosen over another genre-balanced corpus, the British National Corpus (BNC), as it contains more recent text samples. The BNC contains texts from late 1980s till early 1990s while the COCA was sampled from 1990 until 2019(Podesva et al. 2013, 269). As this thesis focuses on analyzing the use of *big* and *large* in contemporary English the most recent data is available in the COCA. Although there are many well documented differences in grammar, vocabulary and pronunciation between British and American English a cursory glance at the top collocates of *big* (day, brother, names, race, money, mistake, hit, bang, break, match) and large (quantities, extent, sums, proportion, quantity, amounts, bowel, degree, sections, volume) in the BNC indicates that a collocate analysis utilizing the BNC would yield results similar to the COCA.

3. 4 Corpus Functionalities Used

This section will introduce the concepts of frequency count and collocation as well as describe the data extraction methods used in this paper.

3.4.1 Frequency Count

Frequency count is probably the most basic tool within the corpus linguistics, and it is used to measure how often each word, or an *n-gram* occurs in a corpus (Podesva et al., 2013, 275). Tognini-Bonelli clarifies (2001, 57) the significance of the frequency-based analysis within linguistics: scholars such as Firth, Halliday and Sinclair share the view that "each single act of communication shows the language system in operation". Also, the Firthian notion of 'repeated events' is a crucial concept when formulating generalizations of the language (ibid.).

The frequency in corpora is usually informed as *normalized* (or *relative*) frequency. According to McEnery and Hardy, (2012, 50) the most common way of informing normalized frequency is 'occurrence per thousand words' or 'occurrence per million words. Many corpora including the COCA generate the frequency automatically. The frequencies of *big* and *large* in different genres will be analyzed in chapter four of this paper.

3.4.2 Collocation

The concept of collocation in its present meaning was introduced by J.R. Firth (1957, 194). According to Firth (1957, 196) collocation is an "abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaninig of a word". Most linguists today agree that collocates of a given word can be reliably identified only by studying its co-occurrence patterns in a text corpus (McEnery and Hardy, 2012, 123). This differs from the approach used by Firth who only used some simple examples to illustrate collocation between given words, and Greenbaum who used native speakers' intuitions to identify collocates (ibid.).

When it comes to defining collocation McEnery and Hardy state (2012, 123) that collocation can be regarded as the idea that words themselves in isolation do not contain all their meanings, instead, the meanings can be found in the characteristic associations the words participate in with other words or structures that frequently co-occur with them. The most common approach today is

to define collocation as a type of co-occurrence pattern between two items which frequently occur close to each other but not necessarily next to each other (ibid.). In this sense collocation can be seen as a methodological elaboration on the concordance (ibid.). Sinclair (2004, 10) describes this approach to collocation by stating that "if word A is a node and word B one of its collocates, when word B is studied as a node, word A will become one of its collocates". McEnery and Hardy note (2012, 123) that there are also stricter approaches for defining collocation according to which collocation can be used to refer to recurring sequence of minimum two words or *n-grams* (ibid.).

The collocates analysed in this thesis were extracted using the Compare function in COCA which allows comparison of two words in terms of their frequencies. This is a very useful feature when comparing two near-synonyms. Additional reason for using the Compare function is the fact that it was not possible to perform the search using Collocates function as this search option was causing an error when running a search for noun collocates of *big* located immediately on the right side of it. As *big* is such a high frequency word, the COCA recommends including four collocates from left and four from the right side of the node in the search, but this would yield results where the collocate is too far from the node.

In order to extract the immediate noun collocates of *big* and *large* with the Compare function, the search was restricted to nouns directly on the right of the words and search string *_nn** was inserted into the 'collocates' field.

4. Analysis

This chapter will provide results of the analysis done using the research methods described above. The first part of this chapter will be dedicated to examining the dictionary definitions of *big* and *large.*

4.1. Dictionary Definitions of *big* and *large*

The dictionaries chosen for the lexicographical analysis are the online versions of the *Collins English Dictionary* and the *Macmillan Dictionary.* They are both widely used and well-established dictionaries of contemporary English.

When checking the first dictionary definitions of *big* and *large* in the *Macmillan Dictionary* we can notice that *big* has been defined as *large* and vice versa:

big: large in size

large: bigger than usual in size

The *Collins English Dictionary* also defines *big* as *large* while *large* has been defined as *great*:

large: a big person or thing is large in physical size.

big: large thing or person is greater in size than usual or average.

Both the *Macmillan Dictionary* and the *Collins English Dictionary* also mention that *large* is often used to express amount or quantity which is greater than the average. According to both dictionaries *large* should also be used when describing a company which employs a lot of people and has many activities. According to the *Collins English Dictionary* this definition applies to *big* as well.

When analyzing the definitions of *large* it is noticeable that there are only four of them in each dictionary and they are identical apart from the following entries: the *Macmillan Dictionary* points out, slightly unnecessarily, that *large* is used in clothing sizes while the *Collins English Dictionary* includes a more relevant definition which indicates that *large* can be used in more figurative sense to express a problem or issue that is very serious or important. According to both dictionaries also *big* has this meaning.

Although the first definition of *big* in both the *Collins English Dictionary* and the *Macmillan Dictionary* is related to expressing physical size both dictionaries indicate that, in

addition to this literal sense, *big* has a vast array of metaphorical meanings and it is also often being used in idioms and slogans. The definitions of *big* vary slightly between the two dictionaries, however both dictionaries mention that *big* is being used to express something major, important, or significant. *Big* can also be used to describe a person who is powerful of successful in particular area or activity. According to both dictionaries *big* is also used as a synonym for *older* in expressions such as *big brother/sister.*

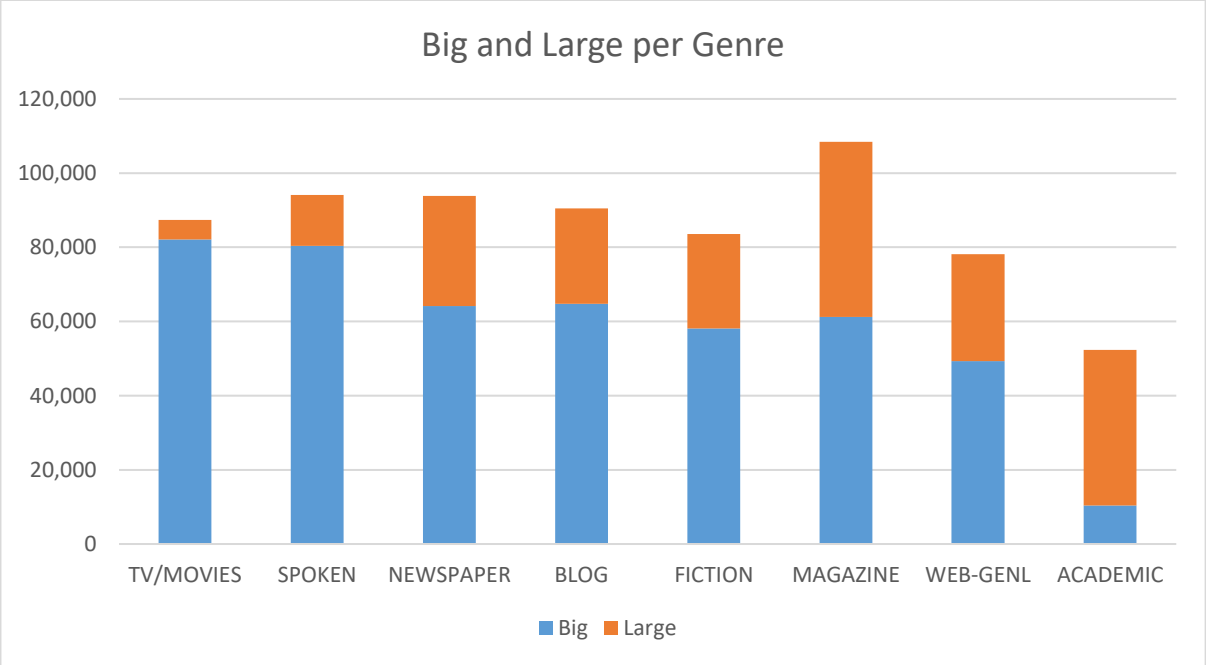
According to the above dictionary definitions, *big* and *large* are intersubstitutable when expressing physical size of a person or a thing. Both can also be used to describe a major company or business as well as a problem which is serious or important. The definitions of *large* are more consistent between the two dictionaries and indicate that *large* is mostly used in literal sense to express size or quantity. However, the *Collins English Dictionary* does advise that *large* has a more figurative sense when expressing a serious problem or issue. *Big* on the other hand has several metaphorical meanings and is often being used in idioms and slogans. There is variation in the definitions of *big* between the two dictionaries which might indicate that giving a comprehensive account of its metaphorical meanings is not straightforward.

Overall, the dictionary definitions already give some insight into the differences of use and meaning of *big* and *large*, however, in order to gain a more comprehensive understanding, this thesis will next focus on the analysis of corpus data.

4.2. Frequency Count

A cursory overview of the frequencies of *big* and *large* shows that *big* is a very high frequency word which occurs 470,291 times in the COCA compared to 218,120 occurrences of *large* – ratio of 2:1 indicates a clear preference for using *big* in data contained in the COCA.

When comparing the number of occurrences per genre shown in the two graphs below, it is noticeable that *big* is more frequently used on TV/movies and in spoken language i.e., in informal contexts, while *large* is used in formal and written contexts such as in magazines and in academic writing.



4.3. Collocates

When examining the immediate noun collocates of *big* and *large* listed in Appendix A, it is noticeable that *big* collocates most frequently with words such as *leagues* at position 1, *plays* at position 3, and *league* at position 6. *Big league/leagues* refer mostly to major leagues in sports, but they are also used metaphorically to describe state of great achievement. *Plays* on the other hand seems to be related to sports only and it seems to refer to a significant performance during a game.

1. Hudson is poised to enter the Hollywood **big leagues** with her portrayal of Nomzamo Winfreda Madikizela. (2012 WEB http://www.nextmovie.com/blog/new-movie-trailers/winnie-trailer/)
2. For a **big league** philosopher, the author writes in a concise and engaging way that we commoners can digest. (2012 BLOG http://checkwithchip.blogspot.com/2012/11/health-care-is-not-right.html)

3. They say Stewart had spent the regular season in the "Slash" role, making the same sort of **big plays** he'd contributed as a rookie in 1995, when the Steelers advanced to the Super Bowl. (1997 NEWS: Washington Post)

Another set of collocates which, when paired with *big,* are used in a metaphorical sense to express something important, major, or serious include *trouble* at position 4, *surprise* at position 5, *deal* at position 8 and *secret* at position 9. When paired with *hug* at position 7, *big* is being used in an affective sense. *Big* is also quite frequently being used in the meaning *older* with the nouns *sister* at position 2 and *brother* at position 11.

The top collocates of *large* include nouns such as *measure* at position 2, *proportion* at position 6 and *quantities* at position 9 and *cohort* at position 11 which are mostly related to expressing quantity:

4. ...you can retain a **large measure** of self-control and self-confidence by understanding and practicing emotional intelligence. (2012 WEB http://www.helpguide.org/mental/work_stress_management.htm)
5. RECENT STUDIES SUGGEST that a **large proportion** of the population are frequently lonely. (2002 ACAD Journal of Psychology)
6. …it is better to have the money supply controlled by gold miners and entities hoarding **large quantities** of gold. (2012 BLOG http://thedailyknell.wordpress.com/2012/11/09/daily-bell-review-on-banking-the-nature-of-money-and-dubious-affiliations/)
7. Given this relatively **large cohort** of exposed persons and a good assessment of exposure. (1998 ACAD Journal of Environmental Health)

When paired with *capacity* at position 5, *large* is often used to describe the capacity of ammunition:

8. Possession of a firearm or **large capacity** ammunition feeding device by a police officer or sworn peace officer of another state. (2012 WEB http://wings.buffalo.edu/law/bclc/web/NewYork/ny3(b).html)

Large capacity is also used in more figurative sense to express person's great mental or emotional capacity:

9. But he won't flip on Joyce Dempsey. Well, I guess he has a **large capacity** for love. (TV 2011 CSI: Crime Scene Investigation)

Large is also used to modify nouns related to scientific and medical terms and occurs with collocates such as *hadron (collider)* at position 1, *intestine* at position 7 and with *bowel* further down on the list. *Large* occurs also in recipes with nouns referring to cooking utensils and food items such as *saucepan* at position 3, *garlic* (clove) at position 8 and *skillet* at position 10.

The corpus search also erroneously picked up one adjective which was excluded from the analysis. Although the search string _nn* should only pick up nouns, the search also returned *ol'* at position 10 as an immediate collocate of *big*. This type of mistagging may have occurred because the spelling of the collocate (*ol'*) is not standard. The exclusion does not have any impact on the results of the analysis.

10. Even a **big ol'** house has got 4 walls that can bear down like a summer squall Green grass surrounds a hidden moat.. (2012 WEB http://www.lyricsmania.com/it_has_not_been_a_good_day_lyrics_tom_flannery.html)

4.4 Other Obervations

This chapter will introduce some additional findings made when further analyzing the noun collocates of *big* and *large*.

When reviewing the COCA data in more detail there seem to be cases where *big* and *large* are interchangeable. It seems that when modifying nouns such as *difference, problem,* and *question* either one can be used. In these contexts, l*arge* is often preceded by the adverb *very.*

1. Handling, usability, and features can make a **big difference**. (BLOG 2012 http://blog.chasejarvis.com/blog/2012/09/nikon-d600-camera-is-here/)
2. …and those external factors like diet and lifestyle can make a **large difference**. (MAG 2019 Engadget)
3. That's not a **big problem** for a person with one tumor on the back of a hand. (MAG 2019 Prevention)
4. this creates a **large problem** for those of us with legitimate disabilities and service dogs. (BLOG 2012 http://www.patriciamcconnell.com/theotherendoftheleash/dog-laws-around-the-world)
5. So the **big question** here should be how much better is the AF on the K-5 II over K-5 (BLOG 2012 http://www.dpreview.com/news/2012/11/09/Just-posted-studio-test-images-from-pentax-k-5-II-k-5-IIs-DSLRs)
6. the new Bagger had one **very large question**: Are the Oscars really worth all the money (NEWS 2014 New York Times)

Similarly, *big,* and *large* seem to be intersubstitutable when describing a large company or corporation.

7. The important thing is that you are asking questions that the **big companies** are not. (BLOG 2012 http://www.techdirt.com/blog/casestudies/articles/20111119/01564816843/value-is-relationship-not-mp3-file.shtml)

8. I've heard some **large companies** have internal day care facilities and I really think that should be more common. (WEB 2012 http://offbeatmama.com/2012/10/going-back-to-work-judgement)
9. Why would we still need **big corporations** to provide the copy and distribution of cultural goods? (WEB 2012 http://www.techdirt.com/articles/20120620/03552119398/business-model-failure-is-not-moral-issue.shtml)
10. Today, most **large corporations** have structures in place, like public policy departments, (NEWS 2018 Washington Post)

In addition to strong tendency for metaphorical usage the COCA data indicates that *big* is being used in literal sense as well. Hence it can replace *large* in most cases, however, these usages seem to occur in informal contexts:

11. While the baby sleeps Lore will cook green things in a **big skillet,** (FIC 2016 The Virginia Quarterly Review)
12. and a **big proportion** of those children available for adoption are black or of other minority groups. (NEWS 1994 Denver Post)

The COCA data indicates that *big* is also being used in slogans and brand names. The examples include *Big Mac* and *Big Blue*:

13. they will see in bright lights that a **Big Mac** and large fries weighs in at 1,050 calories (WEB 2012 http://www.businessweek.com/articles/2012-09-18/mcdonald-s-enters-the-age-of-transparency)
14. as **Big Blue** continues to focus on the cloud, analytics, software, and research and development. (WEB 2012 http://www.zdnet.com/ibm-ceo-virginia-m-rometty-elected-chairman-of-the-board-7000004793/)

The popularity of *big* in tradenames and slogans is also supported by data collected in Trademark Electronic Search System (TESS). This system allows you to search the United States Patent and Trademark Office's (USPTO) database of registered trademarks and prior pending applications. It returns 24 254 hits for *big* compared to only 1658 for *large.*

4.5 Comparison

Based on the above analysis it can be concluded that for the pair of words covered by this thesis, the dictionaries tend to describe only the literal sense of the words, but they do not always succeed in explaining all the connotations or stylistic differences which are of importance when making the correct choice between them. The analysis of the corpus data has helped to reveal some interesting differences between *big* and *large* which were not captured by either of the dictionaries used.

When reviewing the frequencies per genre it is notable that *big* is being used in informal contexts while *large* more often appears in formal contexts.

The collocational analysis reveals further that *big* has a strong tendency to be used in metaphorical sense while *large* is often used in literal sense with very little idiomatic usage.

Both the *Collins English Dictionary* and the *Macmillan Dictionary* mention 'large in (physical) size' as the first definition of *big* while all the top ten collocates of *big* in the COCA are related to expressing idiomatic or metaphorical meanings. According to the COCA data, *big* occurs in the literal sense mainly in informal contexts. As *big* can be used in metaphorical as well as literal sense it is more versatile in use than *large*. Both the COCA data and the dictionaries also indicate that *big* is being used in idioms and slogans as well as in the names of trademarks. The popularity of *big* in tradenames and slogans is also supported by data in Trademark Electronic Search System (TESS).

The dictionaries as well as the COCA data indicate that *large* is mainly used to expressing size or quantity. According to the COCA data, there is also a clear preference to use *large* in

scientific and medical terminology as well as in recipes. This tendency has not been mentioned by either of the dictionaries, whereas the *Macmillan Dictionary* somewhat unnecessarily mentions that *large* is a clothing size.

Even though *large* has very little idiomatic usage the *Collins English Dictionary* includes a definition where *large* is used in a more metaphorical sense to express a problem or issue that is very serious or important. This usage is reflected in the COCA data as well.

The collocates which can be modified by both adjectives include *difference, problem,* and *question.* Both *big* and *large* can also be used to modify nouns such as *company* and *corporation.* These meanings are included in at least in one of the dictionaries.

There seems to be only one instance of mistagging in the data which lead to the exclusion of the collocate: the adjective *ol'* was tagged as a noun probably due to its non-standard spelling. However, this error is not significant and did not have any impact on the results of the analysis.

5. Conclusion

The aim of this study was to find out the differences as well as the similarities of use and meaning of the two very common adjectives, *big* and *large.* The initial assumption was that *big* and *large* are not full synonyms but rather near-synonyms and that *big* is more informal in use while large more frequently occurs in formal langauge. The corpus data analysis confirmed that these assumptions were correct. The collocational analysis revealed also that *big* has very strong tendency to be used in metaphorical sense while *large* Is mostly used in literal sense. Based on the data analysed, *big* seems to be able to replace *large* in most cases, however, these usages mainly occur in informal contexts.

Based on the above, it can be stated that, although the amount of data extracted from the COCA was relatively small, the analysis of the immediate noun collocates of *big* and *large* helped to reveal relevant aspects about their usage and about the effect they have on the words in their

immediate proximity. As Inkpen points out: "Some of the fine-grained senses are also close to each other, so they might ond toccur in similar contexts, while the coarse-grained senses are expected to occur in distinct contexts" (Inkpen, 2007: 3). It is also worth noting that the corpus examples described the usage of the two adjectives more comprehensively than the dictionaries. It can also be stated that, although the data included expressions which are specific to American culture, conducting this type of analysis using the BNC would lead to similar results.

When it comes to identifying opportunities for future research, it would be interesting to research *big* and *large* used as predicative adjectives preceded by copular verbs such as be, feel, seem, appear, look, sound, smell, taste, become, get and possibly one or more modifiers. It could be also interesting to investigate how non-native speakers use these two adjectives by analyzing the data in a learner English corpus such as the International Corpus of Learner English (ICLE).

In conclusion it can be stated that conducting a study using the methods introduced in this thesis has proven to be a useful way of analyzing the differences and similarities of use and meaning of near synonyms such as *big* and *large.* This study has also provided some interesting opportunities for further research which can help to reveal additional subtle but important diffreneces between these two adjectives.

Bibliography

Primary Sources

Collins Cobuild English Language Dictionary. Accessed from https://www.collinsdictionary.com/

Corpus of Contemporary American English (COCA). Accessed from https://corpus.byu.edu/coca/

Macmillan Dictionary online. Accessed from https://www.macmillandictionary.com/

Secondary Sources

Bauer, Laurie. 2002. “Inferring Variation and Change from Public Corpora”. *The Handbook of Language Variation and Change,* edited by J.K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, 97-114. London: Blackwell.

Bennett, Gena R. 2010. Using Corpora in the Language Learning Classroom. Corpus Linguistics for *Teachers. University of Michigan Press.*

Cruse, D. Alan, et al. 1986. *Lexical Semantics.* Cambridge University Press.

Cruse, D. Alan. 2011. *Meaning in language: an introduction to semantics and pragmatics* (3rd ed.). Oxford University Press.

Davies, Mark. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing.* 447–464.

Meurers, Detmar and Stefan Muller. 2014. “Corpora and Syntax”. *The HSK Corpus Linguistics*, article 42, section G.

Firth, J. R. 1957. Papers in Linguistics 1934–51. 1964 edition. Oxford University Press.

Inkpen, Diana and Graeme Hirst. 2006. “Building and Using a Lexical Knowledge-Base of Near-Synonym Differences”. *Computational Linguistics.*

Inkpen, Diana. 2007. A Statistical Model for Near-Synonym Choice. *ACM Transactions on Speech and Language Processing.*

Leech, Geoffrey. 1981. *Semantics: The Study of Meaning.* Penguin Books.

Liu, Dilin. 2010. "Is it a Chief, Main, Major, Primary, Or Principal Concern?: A Corpus-Based Behavioral Profile Study of the Near-Synonyms." *International Journal of Corpus Linguistics.* 15.1: 56-87.

Lyons, John. 1995. *Linguistic semantics: an introduction.* Cambridge University Press.

Lyons, John. 1981. *Language and linguistics: an introduction.* Cambridge University Press.

McEnery, Tony ., and Andrew Hardie. 2011. *Corpus Linguistics.* Cambridge University Press.

Murphy, M. Lynne. 2010. *Lexical meaning.* Cambridge University Press.

Murphy, M. Lynne. 2003. *Semantic relations and the lexicon antonymy, synonymy, and other paradigms.* Cambridge University Press.

Oliviéri, Michèle. 2010. "Syntax and Corpora". *Open Edition Corpus Journal.* 10.1: 7-20.

Palmer, Frank Robert. 1976. *Semantics.* Cambridge University Press.

Podesva, Rober J. and Devyani Sharma. 2013. *Research Methods in Linguistics.* Cambridge University Press.

Saeed, J. I. 2016. *Semantics* (4th ed.). John Wiley & Sons.

Sinclair, J., Jones, S., Daley, R. and Krishnamurthy, R. 2004. *English Collocational Studies: The OSTI Report.* Continuum.

Taylor, J. R. 2002. Near synonyms as co-extensive categories: 'high' and 'tall' revisited. *Language Sciences,* 25, 263-284.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work.* 6 Vol. J. Benjamins Philadelphia, Amsterdam.

Yule, George. 1885. *The Study of Language.* Cambridge University Press.

Appendix A
First 50 collocates with the Compare function big_j / large_j + _n

	WORD	W1	W2	W1/W2	SCORE
1	LEAGUES	1347	0	2,694.0	1,249.5
2	SISTER	1078	0	2,156.0	999.9
3	PLAYS	910	0	1,820.0	844.1
4	TROUBLE	1811	1	1,811.0	839.9
5	SURPRISE	1572	1	1,572.0	729.1
6	LEAGUE	760	0	1,520.0	705.0
7	HUG	659	0	1,318.0	611.3
8	DEAL	17107	14	1,221.9	566.7
9	SECRET	558	0	1,116.0	517.6
10	OL	491	0	982.0	455.4
11	BROTHER	3764	4	941.0	436.4
12	WINNER	462	0	924.0	428.5
13	SHOTS	458	0	916.0	424.8
14	DADDY	844	1	844.0	391.4
15	BELIEVER	398	0	796.0	369.2
16	DIPPER	383	0	766.0	355.3
17	MAC	362	0	724.0	335.8
18	FELLA	352	0	704.0	326.5
19	BROTHERS	344	0	688.0	319.1
20	DAY	2475	4	618.8	287.0
21	MOMENT	587	1	587.0	272.2
22	SPEECH	271	0	542.0	251.4
23	NIGHT	993	2	496.5	230.3
24	HURRY	243	0	486.0	225.4
25	RIG	225	0	450.0	208.7
26	LAUGH	222	0	444.0	205.9
27	FINISH	207	0	414.0	192.0
28	SPENDER	205	0	410.0	190.2
29	WEEK	205	0	410.0	190.2
30	DREAMS	404	1	404.0	187.4
31	TOP	202	0	404.0	187.4
32	BANG	4237	11	385.2	178.6
33	SPENDERS	192	0	384.0	178.1
34	CHILL	190	0	380.0	176.2
35	HIT	1483	4	370.8	172.0
36	NAMES	1029	3	343.0	159.1
37	BLUE	170	0	340.0	157.7
38	REASONS	165	0	330.0	153.1
39	JOKE	323	1	323.0	149.8
40	LOSER	161	0	322.0	149.3
41	BUCKS	1595	5	319.0	147.9
42	PLUS	319	1	319.0	147.9
43	LEBOWSKI	148	0	296.0	137.3
44	MAMA	272	1	272.0	126.2
45	WORRY	136	0	272.0	126.2
46	REVEAL	134	0	268.0	124.3
47	BOYS	1071	4	267.8	124.2
48	DRAW	265	1	265.0	122.9
49	SURPRISES	131	0	262.0	121.5
50	HELP	756	3	252.0	116.9

	WORD	W2	W1	W2/W1	SCORE
1	HADRON	255	0	510.0	1,099.6
2	MEASURE	719	2	359.5	775.1
3	SAUCEPAN	656	2	328.0	707.2
4	EXTENT	1152	6	192.0	414.0
5	CAPACITY	92	0	184.0	396.7
6	PROPORTION	730	4	182.5	393.5
7	INTESTINE	173	1	173.0	373.0
8	GARLIC	168	1	168.0	362.2
9	QUANTITIES	1324	8	165.5	356.8
10	SKILLET	1114	7	159.1	343.1
11	COHORT	78	0	156.0	336.4
12	EGGS	1117	8	139.6	301.1
13	SAUTÉ	68	0	136.0	293.2
14	CLOVES	128	1	128.0	276.0
15	SEGMENTS	243	2	121.5	262.0
16	AMOUNTS	2831	27	104.9	226.1
17	DIAMETER	46	0	92.0	198.4
18	SHALLOT	45	0	90.0	194.1
19	SAMPLES	88	1	88.0	189.7
20	SHALLOTS	42	0	84.0	181.1
21	QUANTITY	320	4	80.0	172.5
22	VOLUMES	311	4	77.8	167.6
23	SUBUNIT	38	0	76.0	163.9
24	MILLIMETER/SUBMILLIMETER	36	0	72.0	155.2
25	REGIONS	69	1	69.0	148.8
26	DEGREE	585	9	65.0	140.1
27	CASELOADS	32	0	64.0	138.0
28	FORMAT	127	2	63.5	136.9
29	SWATHES	62	1	62.0	133.7
30	OVEREXPRESSION	30	0	60.0	129.4
31	VARIABILITY	30	0	60.0	129.4
32	CLOVE	58	1	58.0	125.1
33	CONGREGATIONS	29	0	58.0	125.1
34	BAKING	229	4	57.3	123.4
35	ZUCCHINI	28	0	56.0	120.7
36	AMOUNT	2023	37	54.7	117.9
37	MAGNITUDE	27	0	54.0	116.4
38	ENSEMBLES	53	1	53.0	114.3
39	BOWEL	26	0	52.0	112.1
40	SCALES	103	2	51.5	111.0
41	PORTOBELLO	25	0	50.0	107.8
42	ARRAYS	24	0	48.0	103.5
43	EGGPLANT	24	0	48.0	103.5
44	INTESTINES	24	0	48.0	103.5
45	CADRE	23	0	46.0	99.2
46	LEEKS	23	0	46.0	99.2
47	MIXING	274	6	45.7	98.5
48	ARRAY	178	4	44.5	95.9
49	CELERY	22	0	44.0	94.9
50	SAUCEPOT	22	0	44.0	94.9