Tampere University

Eetu Mäkelä

NEAR AND CLOSE

A Corpus-Based Study on Near-Synonyms

Faculty of Information Technology and Communication Sciences
Bachelor's Thesis
May 2020

TIIVISTELMÄ

Eetu Mäkelä: *Near* and *Close* – A Corpus-Based Study on Near-Synonyms
Kandidaatintutkielma
Tampereen yliopisto
Englannin kielen, kirjallisuuden ja kääntämisen tutkinto-ohjelma
Toukokuu 2020

Tämän tutkimuksen lähtökohtana oli selvittää, miten lähisynonyymit *near* ja *close* eroavat merkityksiltään ja käyttöyhteyksiltään englannin kielessä. Tutkimusaiheen valintaan vaikuttivat keskeisesti kaksi huomiota. Ensinnäkin, sanojen erot eivät välttämättä selviä sanakirjojen määritelmien perusteella, ja toiseksi, aiheesta ei ole aikaisemmin tehty laajaa korpuksiin nojaavaa tutkimusta. Vertailu tapahtui tarkastelemalla sanojen kollokaatiota eli niiden taipumusta esiintyä toistuvissa rakenteissa muiden sanojen kanssa. Sanoja voidaan käyttää useiden sanaluokkien sanoina, mutta tutkimusalue rajattiin adjektiivikäyttöön. Keskeisiä teorioita tutkimuksessa olivat kollokaatio ja synonymia sekä niiden taustalla vaikuttava semantiikka. Tutkimusmetodina toimi korpuslingvistinen kielentutkimus, ja tutkimus nojaa metodologialtaan aiempiin korpustutkimuksiin lähisynonymiasta. Lisäksi tutkimuksessa verrattiin sanakirjojen lähisynonyymeille antamia merkityksiä.

Tutkimuksen aineisto kerättiin suorittamalla hakuja nykyaikaisia englanninkielisiä tekstejä sisältävässä *Corpus of Contemporary American English* -korpuksessa. Hakuprosessissa käytettiin korpuksen vertailuominaisuutta, jonka avulla pystyy etsimään molempien sanojen kollokaatteja samanaikaisesti. Kollokaatiota tarkastelemalla voitiin havaita lähisynonyymien välisiä merkityseroavaisuuksia sekä niiden taipumusta esiintyä tiettyyn aihepiirin kuuluvien sanojen yhteydessä. Vertailu keskittyi yhteyksiin, joissa *near* ja *close* -sanat toimivat substantiivin attribuuttina. Haku suoritettiin erikseen adjektiivien perus-, komparatiivi- ja superlatiivimuodoilla.

Tutkimuksen tulosten perusteella voidaan todeta, että aineiston sisällä *near* ja *close* esiintyvät yhdessä erilaisten kollokaattien kanssa. Suuria eroja havaitaan esimerkiksi ihmissuhteisiin viittaavissa käyttöyhteyksissä sekä etäisyyksien määrittelyssä. Lisäksi tutkimuksessa ilmenee, että lähisynonyymit saavat toisistaan poikkeavia merkityksiä myös samojen kollokaattien kanssa. Tutkimuksessa kuitenkin todetaan joitakin metodologisia ongelmia, kuten virheitä korpuksen automaattisessa sanaluokkien tunnistuksessa. Tutkimustulosten perusteella voidaan myös esittää mahdollisuuksia aiheen tarkasteluun jatkossa. Jatkotutkimuksissa olisi mahdollista esimerkiksi käsitellä sanojen historiallista kehitystä tai laajentaa aihealuetta muihin sanaluokkiin.

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

1 Introduction

The adjectives *near* and *close* may at first glance seem very similar in meaning. Indeed, they are what could be classified as synonyms or near-synonyms of each other. However, upon further consideration one may realize that they are used in slightly different ways. We talk about the *near future*, but not the *close future.* We have *close friends,* not *near friends*. Of course, this is not limited to a single set of words. There seems to exist a greater tendency for *near* and *close* to co-occur with certain types of words, one that is embedded within the semantics of the near-synonyms themselves.

The purpose of the thesis is to study the variation in the meanings and use of the near-synonym adjectives *near* and *close*, comparing their dictionary definitions in the *Macmillan Dictionary* and the *Cambridge Dictionary* as well as analyzing corpus data relating to their common collocates. The topic was chosen out of interest for corpus-based research and near-synonyms. The use of corpora in linguistic analysis has become a valuable method of research, as public corpora offer an extensive, readily searchable collection of language data for analysis (Bauer 2002: 98-103). The use of corpora is well suited for the purposes of this thesis, as it allows us to observe the differences and similarities with how the words *near* and *close* are used in naturally occurring language. This will then be contrasted with the observations made from the dictionary descriptions of the use of *near* and *close*.

The words *near* and *close* were chosen due to being relatively common, meaning there is a significant quantity of data to analyze, interesting observations made during preliminary corpus searches, and the fact that there seems to be a gap in research when it comes to the relation of these two words when analyzed from the angle of corpus linguistics. Though *near* and *close* also occur as adverbs and prepositional constructions, the focus of the thesis will be placed specifically on the

attributive adjectival instances of the words, which will give deeper insight on certain aspects of their use.

The research questions of the thesis are:

- What definitions do dictionaries provide for the words *near* and *close*?
- What patterns can be observed with the collocates of the words *near* and *close* in corpus data?
- In what ways do the corpus findings correspond to the statements made by dictionaries; in what ways do they not?

2 Literature Review

In this section, we will go through some past research on the topic of near-synonyms and collocation. We will also examine some dictionary definitions of *near* and *close*.

2.1 Near-Synonyms

Near-synonyms, sometimes known as plesionyms, can be defined as words that are for the most part synonymous, but can differ slightly semantically or stylistically (DiMarco et al. 1993: 1-2). It is difficult to draw a line between synonyms and near-synonyms, and it can be argued that all synonymy is in some degree near-synonymy (Jantunen 2004: 1).

Near-synonyms have been studied for several purposes in the past, one of the reasons being their use in natural language production system development. The use of near-synonyms is a part of natural human speech, thus a system attempting to replicate language production needs to be able to know how to pick the correct near-synonym for a given context to mimic natural language generation (Inkpen & Hirst 2004: 1).

There have also been previous corpus-based studies of different near-synonym word groups, such as Liu (2010:56), who has specifically examined semantics of near-synonymous adjectives. According to Liu, the research affirms that the semantic structure of attributive adjectives can be studied by examining the nouns that they modify. These co-occurring nouns help identify the differences in meaning and usage between near-synonyms.

Despite similar meanings, near-synonymous words may not be interchangeable in various contexts due to their different collocational behavior and semantic differences (Xiao & McEnery 2006: 8; Tognini-Bonelli 2001: 34).

2.2 Collocation

Collocation refers to the co-occurrence of words in patterns, where a frequently co-occurring item is referred to as a *collocate* (Xiao & McEnery 2006: 2-3). The concept of collocation can be traced back to J.R. Firth (1957: 194) and the field of contextual semantics.

2.2.1 Definition

The concept of collocation, like many scientific terms, may vary slightly depending on the researcher. Jantunen (2004: 15-19) lists three aspects that can used to define the collocation: repetition, statistical significance and dimension.

Repetition means that for a co-occurrence to be considered collocation, it must occur recurrently (Jantunen 2004: 16). The prevailing view among many linguists (e.g. Firth 1957: 181; Hoey 1991: 6-7) is that in order to analyze the semantic relationships between two units as collocation, must they occur together repeatedly. According to Partington (1998: 17,121), collocational norms are born from a noticeable reoccurrence of collocations in language.

The second aspect, which is connected to the first, is statistical significance. Jantunen (2004: 17) explains that statistical significance is dependent on the frequency with which a certain

collocate occurs. Greenbaum (1974: 82) and Hoey (1991: 6-7) among others seem place emphasis on the statistical significance in collocation. Hoey's definition of collocation specifies that the items in question must occur "with greater than random probability in its (textual) context".

Dimension is defined by Jantunen in four ways (2004: 18). These include the amount of words that belong in the collocation, the distance between the collocates and their base, grammatical relationship of collocates and their base as well as symmetry of the area that is being inspected. This aspect seems to contain the most variance among the interpretations of linguistics, with differing opinions regarding for example the number of items that form a collocation (Jantunen 2004: 18).

2.2.2 Collocational research

In contextual semantics, the aim is to find and study the typical usage contexts of expressions and the preferences that govern the choices of language users (Jantunen 2004: 7-9). Collocational research often takes a statistical approach (Greenbaum 1974: 82; Sinclair 1991) and some definitions only consider statistically frequent co-occurrence as collocation (e.g. Hoey 1991: 6-7). According to Sinclair (1991: 110), collocation is an example of idiomatic choice, which means that a language user chooses from a set of premade constructions that are available to him.

As explained by Jantunen (2004: 7-9), observing collocation can be used to research the differences between two near-synonyms by comparing their common collocates. The context-governed choice of a given expression is known as *contextual meaning* or *collocational meaning*, and past research has traditionally used the analysis of near-synonyms to visualize collocational distribution (Jantunen 2014: 7-9; Leech 1981: 17; Palmer 1976: 96). Examining the collocational meaning of expressions can reveal restrictions or patterns within the co-occurrence of the near-synonyms.

The use of corpus linguistics for collocational research is seen as beneficial because speakers typically cannot deduce collocational patterns well with their own intuition (Xiao & McEnery 2006: 2). Indeed, Jantunen (2004: 15-16) explains that the concept of collocation was not a popular subject of research until the advent of corpus-based technology, which gave researchers the opportunity to analyze large databases. According to Jantunen, collocative analysis has since then become one of the most important concepts in corpus linguistics.

2.2.3 Restrictions and preferences

The concept of *collocational restrictions* is presented by Cruse (1986: 277-281) as one of the two forms of semantic restrictions that restrict the co-occurrence of lexical items. The other type of semantic restrictions, which Cruse refers to as *selectional restrictions*, are based on logically necessary rules. For example, an expression denoting an inanimate object would not typically occur in a context where an animate would, which Cruse illustrates with examples such as the ill-formed "the spoon died". However, Cruse argues that collocational restrictions are not logically necessary restrictions like selectional restrictions, but rather more arbitrary rules that affect the distributional patterns and restrict the co-occurrence of words. Collocational restrictions are divided further by Cruse (1986: 281) into three groups: systematic, semi-systematic and idiosyncratic. While systematic and semi-systematic restrictions are based on the semantic characteristics of collocates (rather than logic), idiosyncratic restrictions cannot be semantically motivated.

Alongside the concept of collocational restrictions, we can talk about *co-occurrence preferences* (Jantunen 2004: 15). This is an opposite way of looking at the co-occurrence of expressions compared to the idea of restrictions. Jantunen explains that, from the perspective of preferences, we can instead interpret co-occurrence or collocational patterns as governed by certain strong preferences or tendencies to occur in specific contexts. This means possible collocates in each situation are not necessarily limited by collocational restrictions but rather the collocates have

preferences to occur or not occur in that semantic environment. Haskel (1971: 159–168) notes that deviation from these semantic preferences can occur especially when it is motivated by a desire to make the language sound particularly novel or attention-grabbing.

Yet another concept related to collocation is known as *semantic prosody*. According to Xiao & McEnery (2006: 5-7), semantic prosody can be thought of as "The collocational meaning arising from the interaction between a given node and its typical collocates". It can be argued that semantic prosody is different from the idea of semantic preference because it takes the relationship between lexical units to an even higher level of abstraction (Xiao & McEnery 2006; Sinclair 1996). Xiao & McEnery (2006: 7) suggest that semantic preferences are related to the collocates of a given word while semantic prosody is related to the node word.

2.3 Dictionary Definitions of *Near* and *Close*

The dictionaries chosen to be examined in this subsection are the online versions of the *Macmillan Dictionary* and the *Cambridge Dictionary*. The reason for their selection was their high standard and the fact that they describe contemporary English. Historical dictionaries such as the *Oxford English Dictionary* were left out for this reason (though it may be interesting to examine the etymology and historical use of the words *near* and *close*, it is not within the scope of this thesis).

2.3.1 General Observations

What is immediately noticeable about the definitions of *near* and *close* found in the *Macmillan Dictionary* and the *Cambridge Dictionary* is that they can readily substitute one for the other to define them. Take for example the definitions 1 and 2 of *near* in the *Macmillan Dictionary*:

1 close to someone or something

2 getting close to a particular state or situation

Or vice versa, from the US definition of *close* in the *Cambridge Dictionary*:

> *near in position, time, or condition:*

In some of these instances *close* or *near* are paired with a preposition such as *to*, however. The *Cambridge Dictionary* also uses "NEAR" as a guide word for distinguishing between the various meaning of *close* that are listed.

In addition, *near* and *close* are naturally listed as each other's synonyms. Something also worth mentioning is that the *Macmillan Dictionary* has no separate entries for the adverb, adjective and preposition occurrences of *near*, but rather all three share the same list of definitions. On the *Cambridge Dictionary* not only does the adjective *near* have its own entry, it has two as the *Cambridge Dictionary* has separate definitions for (British) English and American.

2.3.2 Denotations

What meanings are attached to *near* and *close*? We can first construct a list of the significant common definition aspects of the two words in either dictionary:

(1) Spatial proximity

(2) Temporal proximity

(3) Relationship (family, friends etc.)

(4) Similarity, 'almost'

(5) Careful attention

Next, let us go through the *Macmillan Dictionary* definitions for both words. For *near*, we can observe the denotations (1), (2) and (4). For *close*, (1), (2), (3), (4) and (5), in addition to several more minor definitions.

In the *Cambridge Dictionary*, *close* also has (1), (2), (3), (4) and (5). *Near* has (1), (2), (3) and (4). Note that some definitions are only found labelled UK or US.

Close seems to have a greater number of minor definitions attached to it compared to *near* in both the *Cambridge Dictionary* and *Macmillan Dictionary.* The dictionaries list various situational uses for *close,* such as these in *the Macmillan Dictionary* (s.v. *close*):

> *not willing to spend your money or give any to anyone*
>
> *not willing to share information about yourself or your emotions*

The relationship meaning of *near* is only mentioned on the *Cambridge Dictionary*. In addition, aspect (5) is not mentioned for *near* on either dictionary. Some of the definitions, especially those of the *Macmillan Dictionary*, were also rather vague. Overall during the process, the *Cambridge Dictionary* seemed to be the more comprehensive of the two dictionaries.

2.4 Other Research

There has also been previous research carried out on the nearness relations of words such as *near* and *close* within geographic space (Derungs & Purves 2016; Xu & Klippel 2012). Wallgrün, Klippel & Baldwin (2014) have used corpus-based research to compare the real world distances between locations in situations where a language has used the expressions *near*, *close* or *next to*. This is an interesting subject of research and could also be considered as characterizing semantic differences between near-synonyms. Spatial proximity is one of the denotations of the word *near* examined within this thesis, but otherwise it is quite different from the aim of this study despite sharing some thematic and methodic elements.

3 Methods and Materials

This section will detail the methods involved in corpus linguistic study and the Corpus of Contemporary American English which was utilized in this thesis.

3.1 Corpus Linguistics

A corpus is "a systematic collection of naturally occurring written or spoken language samples in context, stored on a computer readily available for qualitative and quantitative analysis" (Kayaoglu 2013: 128). Depending on their function, corpora can contain different types of texts in differing amounts. Meyer (2002: xii) argues that for the purpose of linguistic analysis, balanced corpora, which contain equal amounts of data from various genes, are best suited. As Meyer explains, this is because the inclusion of various genres allows for comparison across the data of various genres as well as the analysis of individual genres.

Corpus linguistics comprises of the descriptive and theoretical studies of language through linguistic corpora (Meyer 2002: xi). As Meyer describes it, corpus linguistics is not a linguistic paradigm but rather a type of methodology for linguistic research. The advantages of using corpora for linguistic research include the ability to freely search extensive amounts of data as well analyze it numerically (Bauer 2002: 102-103). According to Bauer, the main advantage of a public corpus specifically is that the research will be replicable. This replicability means that any researcher can verify the results themselves by accessing the corpus and checking the distributional patterns by using the same search string.

As explained by Crawford & Csomay (2015: 6), corpus linguistics can be employed in analysis of near-synonyms in two main ways. The first method is that by searching a corpus for the contexts in which these words occur, it is possible to gain information about their collocates and semantic differences. As Crawford & Csomay note, this is possible only due to the large amount of data contained in the corpus. The second method is to analyze the frequencies of the words in terms of how many times their collocates occur in in corpus.

Bauer (2002: 103-104) also raises issues and difficulties that occur with corpus-based research. The main issue is that despite their size corpora may present a false sense of accuracy by

quantitively analyzing data. Though corpora may contain large amounts of data, Bauer argues that corpora are still only samples of text and do not exhaustively describe the language. Bauer also notes the size of corpora may cause issues for the researcher by either containing too little data for a proper analysis to be done or too much of it, which results in unnecessary, extraneous data for the researcher.

It is the responsibility of the researcher to be aware of these issues and consider them when deciding the corpus or corpora that are used to study the subject in question. The researcher should also avoid making generalizations or significant claims about language based on the limited data offered by the corpus. Nevertheless, corpora are useful tools for making observations about various aspects of language use.

3.2 Corpus of Contemporary American English

The Corpus of Contemporary American English (COCA) is "the first large, genre-balanced corpus of any language, which has been designed and constructed from the ground up as a 'monitor corpus'" (Davies 2010: 447). According to Davies, one of the benefits of COCA is that it maintains a balance between the amount words collected from the genres of spoken, fiction, popular magazines, newspapers, and academic journals. This means that COCA accurately reflects developments in the real world by adding an equal amount of text from each genre with each update made to the database.

As of the year 2020, COCA contains over one billion words of text, which consists of around 20 million words per year collected during years 1990-2019. Thus, as the name suggests, the corpus contains particularly contemporary texts. The modern aspect of COCA means that it will provide an

authentic view of how the words under examination are used in contemporary English rather than their historical developments.

COCA was chosen for this thesis in part because it offers a large amount of data consisting mostly of standard American English. It also allows for a wide variety of specific search options, which suit the purpose of this thesis. Some preliminary searches were made before proceeding on with the subject and COCA as the corpus of choice.

4 Analysis

The search was performed using the Compare function of COCA. This allows for the collocates of two words to be compared in terms of their frequencies. It was decided to run separate searches for the basic forms *near*/*close*, the comparative forms *nearer*/*closer* and the superlative forms *nearest*/*closest*.

To restrict the results to instances where the words are used as adjectives, the search query was formatted to near_j and close_j. The collocates were then restricted to nouns directly on the right of the words. This method picks out the attributive adjective uses, such as "a close friend", where *friend* is a collocative directly to right of *close*. Predicative adjective uses will not be picked up, but this method ensures that the collocates are actually referring to the word in question.

4.1 Collocates

We can observe quite clearly (see Appendix A) that the collocates most strongly associated with *close* include *friend* at position 1, *relationship* at position 3, *ties* at position 9, and further down the line *relationships*, *friendship*, *friends*, *associates*, *associate, allies*, *ally*. Another strong tendency that favors *close* is the meaning 'careful attention' as seen with *attention* at position 2, *look* at position 5 and further down *scrutiny*, *watch* and *inspection*.

Near, on the other hand, makes temporal reference with collocates such as *future* and *term*. It also collocates with the sense of 'almost', such with the collocate *panic*. "Near panic" thus refers to a state that is close to a panic, but not quite the same.

(1) Everyone in the park seemed to jump and look around, in a near panic.
The ball game above their heads stopped abruptly. (2008 FIC Analog)

The comparative form results (see Appendix B) show *nearer* to be very rarely used in general. *Closer* interestingly has four quite similar words leading the list, being *look*, *inspection*, *examination* and *attention*. However, three collocates of words denoting relationship can be found within the top 10 of the listing as well: *ties*, *relationship*, *relationships*.

The superlative forms *nearest* and *closest* exhibit the same patterns as the basic forms (see Appendix C). The collocates of the word *closest* are overwhelmingly words relating to people and relationships. *Nearest* on the other hand has collocates that can clearly be interpreted as location-based. The superlative forms seem to have an even stronger preference for certain types of collocates than the basic forms, as the top 13 collocates most associated with *closest* are all words denoting relationships. This may be partly caused by the fact that meaning of 'careful attention' does not often occur with the superlative forms. Phrases such as "closest look" or "closest scrutiny" do not seem to be natural.

4.2 Observations

Here we will detail some notable observations made during closer analysis of the collocate listings gained from searching the corpus, as well as some limitations that were caused by the search algorithm of the corpus.

Analyzing the contexts of some collocate instances revealed another interesting thing of note. There are cases where *near* and *close* have the same collocate, but the resulting phrase differs in meaning between the two words. Take for example the collocate *majority*. Compare these sentences:

(2) …support for impeachment grew even more, from 35 percent to a **near majority**, 47 percent (2018 MAG Salon)

(3) Among Democrats, Bush received a **near majority** of the votes, 47 percent, among the eight choices presented. (2012 WEB http://hotlineblog.nationaljournal.com)

(4) …have to ask you, you say, whoever wins, it will be a **close majority** in the House. It'd be a narrow majority (2000 SPOK CNN_Novak)

(5) In 96 they barely won by a very **close majority** in a three way race (with very sleazy and illegal campaign funding, (2012 BLOG http://althouse.blogspot.com/2012/03/bush-v-gore-is-case-of-century-because.html)

The first two sentences use the phrase "near majority" to denote something that almost happened or is near to happening. However, the phrase "close majority" does not have the same meaning. Instead, it means something like 'barely'. A "close majority" is a majority, but just barely. A "near majority", on the other hand, is not quite a majority. Perhaps this variation is the result of an overlapping denotation from the definition of *near* as 'almost', while *close* denotes abstract proximity. It is also possible there is pattern to be found regarding the variation of *near* and *close* when it comes to the spatial or temporal proximities of an event.

We can also contrast the meaning of "close majority" with another collocate, *agreement*. Here, the meaning of *close* is noticeably different.

(6) The findings of the present study are in **close agreement** with the findings of Thiruvenkadan (2005) who reported that…" (2014 ACAD JAnimalPlantSci)

(7) …have on various occasions given expression to feelings and opinions with which I am in **close agreement**. (2005 ACAD Humanist)

Instead of the 'barely in agreement' meaning we might expect based on the previous examples, the phrase "close agreement" seems to have almost the opposite meaning. *Close* acts more like emphasis, as in 'very much in agreement'. This is an interesting case as it shows that there can be variation even within the meaning of just one of the words, and within two seemingly related collocates such as *agreement* and *majority*.

As another example we may also examine a case where the two words appear to be entirely interchangeable. These are instances with the collocate *thing*.

(8) It had been a **close thing**. What might have happened if the Americans had not succeeded in making lodgment (2002 MAG MilitaryHist)

(9) Brown was shaken. He had won, of course, but it was a **near thing**. (1996 MAG AmHeritage)

Here the use of either phrase, "near thing" or "close thing" appears to produce the same meaning, 'just barely'. But interestingly, when we compare the collocate *thing* with the superlative forms *nearest* and *closest*, we get an entirely new use.

(10) It is five in the morning and Sven is the **nearest thing** Frankfurt has to a pop star. (2011 FIC Mixmag)

(11) He worked with Midnight Basketball for kids at risk and was the **closest thing** some of them had to a father. (2017 FIC Bk:UndoingSaintSilvanus)

Here the phrase “nearest thing” or “closest thing” seems to denote something that most resembles something else. This appears different from the possible meaning of ‘almost’ discussed earlier, as it seems the thing in question does not have to resemble the other thing in any significant amount, as long it has some basic elements in common to be considered the closest thing possible.

We can also note that although *close* exhibits a strong tendency for denoting relationships, there are a few common collocates particularly where it denotes proximity, such as *range*, *contact* and *quarters*. These collocates seem to be semantically restricted to co-occur with *close*. *Near* on the other hand seems to lack denotations of relationship almost entirely, suggesting that it has stronger tendencies for certain types of collocates compared to *close.*

It should be noted that despite the search being restricted to *near* and *close* as adjectives, there were instances of other parts of speech slipping through. For example, collocates of the word *close* included *lid* and *eye* at positions 40 and 95 respectively. Closer examination of the context revealed that they were instances of the imperative verb form of *close,* found in instructions.

(12) Spoon batter onto waffle iron. **Close lid**, and cook 3 minutes, or until waffle is crisp. (2010 MAG VegTimes)

(13) Breathe deep, **close eyes**, take another breath. (2002 FIC NewEnglandRev)

Other instances of mistagging occurred with the word *agreement. Agreement* was earlier discussed as a collocate of *close*, but it also appeared as a collocate of *near* with nine instances in the results. However, unlike the legitimate cases of “close agreement” where the use was clearly adjectival, eight out of the nine instances of “near agreement” are erroneously tagged. In these cases, *near* seems to function as a preposition.

(14) Negotiators for players and owners slowed their pace but were said to be **near agreement** on all the major issues as they close in on a deal to end… (1996 NEWS AssocPress)

(15) The Congress and I are **near agreement** on sweeping welfare reform

(1996 SPOK CBS Special)

However, these types of corpus errors should not cause any major distortion in the results as they are insignificant with regard to the collocate ratios being compared in the corpus analysis.

4.3 Comparison

Based on the results it can be stated that *close* has a strong preference for words relating to relationships as well as words denoting careful attention. *Near* on the other hand is more associated with temporal and spatial references, as well as the sense of 'almost'. The dictionary definitions examined earlier do not make any clear statements regarding *near* and *close* having such preferences, though the *Macmillan Dictionary* does not even list any meaning related to relationships for *near* in the first place. *Cambridge Dictionary*, however, does, and while there are instances of collocates that denote relationships with *near*, based on the results it is quite rare. Even in the instances where *near* is used with a collocate denoting a person, it often occurs with meaning other than relationship, such as in the sense of 'almost' in "near genius" and "near stranger". With the superlative form we also get phrases such as "nearest doctor".

The results indicate that there are several collocates that are restricted to either *near* or *close*. The dictionary definitions, however, simply list the fact that both *near* and *close* can be used for various contexts without explaining these connotations or collocative restrictions. Perhaps this works for a native speaker of English, but further research on non-native learners of English and their use of these words could bring more insight into the subject and whether the dictionaries are lacking or not.

Patterns observed with the usage of *near* and *close* have shown an interesting tendency where *near* seems to denote actions or situations have not reached their conclusion, while *close* denotes a

situation has been brought to that point. This tendency could be seen with the collocate *majority*. It can be assumed that the use of *near* follows the meaning of 'almost', but the use of *close* is rather unusual.

It has been noted that *close* has a set of collocates with a preference for spatial meaning. This use seems to be restricted to the particular set of words (*range*, *quarters*, *contact*) and could possibly be considered as an idiosyncratic restriction on the words in question, as there does not seem to be semantic motivation for the deviation.

It is also interesting to compare the differences between the collocation of the near-synonyms to an instance where they do in fact seem to serve the same function. This was noted with the expression "near/close thing", where the collocation does interestingly hold one of the meanings ('almost') that seems less strongly in preference of one of the words over the other, compared to the more consistently distributed meanings of proximity and relationships.

The inclusion of comparative forms did not yield much data for analysis, but the superlative forms showed strong tendencies for the same type of distribution as the base forms.

5 Conclusion

This study was focused solely on the adjectival instances of the near-synonyms *near* and *close*. However, in future research it may be beneficial to account for instances of prepositions and adverbs as well, as observations during the research process indicated that there is not much semantic difference within these different instances. This can be seen, for example, in the fact that *Macmillan* did not list separate entries for the different forms of *near*. One observation that was made is that *near* would in some cases be used in the same way as the combination *close to,* while in other instances the combinations *near to* and *close to* would have similar uses for denoting spatial proximity.

The initial findings made at the start of the project based on preliminary searches were shown to be quite accurate. *Close* is associated more with words relating to human relationships while *near* is frequently used with collocates to denote time or distance. However, different types of collocates turned out to be rather frequent as well, such as the sense of 'almost' and 'careful attention'. It can also be said that *close* still occurs more readily with collocates of proximity than *near* would occur with collocates of relationship.

In addition, several interesting observations were made during the research process regarding the various collocative co-occurrence patterns of *near* and *close*. These included the difference in meaning between the use of either *near* or *close* in certain instances, such as *close/near majority* and the differing meanings occurring with the collocate *thing* when the adjectives are in superlative form (*nearest*/*closest thing*).

As this thesis was very limited in scope, we cannot draw many definite conclusions regarding the use of *near* and *close*. However, certain patterns have been observed when it comes to the noun collocates of the adjectives and their comparative and superlative forms. With the results of this thesis, it is possible to state that at the least, preference for certain meanings in the collocates is strong within the data collected by the Corpus of Contemporary English. There are many possibilities of expanding research into different areas within this topic, such as studying the historical development of *near* and *close* or expanding into the adverbial and prepositional uses of the words. Interestingly, *near* also occurs as a verb with sense of 'to get near', while *close* forms the phrasal verb *close in* which has similar meaning - this could also be interesting to further analyze. Yet another possible future research possibility would be additional synonyms of *near* and *close* and how they would compare with the results reached here. It could also be interesting to perform more extensive corpus searches, perhaps on another corpus with a different source of texts, so long as it can be done effectively.

In this study we have taken advantage of the possibilities offered by corpus linguistics to research the synonymy of words. We have observed how a simple pair of near-synonymous words such as *near* and *close* can follow various noteworthy patterns and analyzed how that reflects in the meanings of the words and their dictionary definitions. Perhaps most importantly, this study has given reason for further research within the various aspects of *near* and c*lose* semantics and collocation. Following this trace of thought should result in more insight regarding the characteristics of near-synonyms.

Bibliography

Primary sources

Cambridge Dictionary. Cambridge University Press 2020. Accessed from https://dictionary.cambridge.org

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

Macmillan Dictionary. Macmillan Publishers Limited 2009–2020. Accessed from https://www.macmillandictionary.com

Secondary sources

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

Crawford, William, and Eniko Csomay. 2015. *Doing Corpus Linguistics*. Routledge.

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

Derungs, Curdin, and Ross S. Purves. 2016. "Mining Nearness Relations from an N-Grams Web Corpus in Geographical Space." *Spatial Cognition & Computation* 16.4: 301-22.

DiMarco, Chrysanne, Graeme Hirst, and Manfred Stede. 1993. "The Semantic and Stylistic Differentiation of Synonyms and Near-Synonyms". *AAAI Spring Symposium on Building Lexicons for Machine Translation.*

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

Greenbaum, Sidney. 1974. "Some Verb-Intensifier Collocations in American and British English." *American speech* 49.1/2: 79-89.

Haskel, Peggy Irene. 1971. "Collocations as a Measure of Stylistic Variety.", in *The Computer in Literary and Linguistic Research: Papers from a Cambridge Symposium*, edited by Roy A. Wisbey, 159–168. Cambridge University Press.

Hoey, Michael. 1991. *Patterns of Lexis in Text*. 299 Vol. Oxford University Press.

Inkpen, Diana Zaiu, and Graeme Hirst. 2004. "Near-Synonym Choice in Natural Language Generation". *Recent Advances in Natural Language Processing.*

Jantunen, Jarmo H. 2004. "Synonymia Ja Käännössuomi." *Korpusnäkökulma samamerkityksisyyden kontekstuaalisuuteen ja käännöskielen leksikaalisiin erityispiirteisiin.Joensuun yliopiston humanistisia julkaisuja* 35.

Kayaoglu, M. Naci. 2013. "The use of Corpus for Close Synonyms." *Journal of Language and Linguistic Studies* 9.1: 128-44.

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.

Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge University Press.

Palmer, Frank Robert. 1976. "Semantics: A New Outline.". Cambridge University Press.

Partington, Alan. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. John Benjamins Publishing.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Sinclair, John. 1996. 'The search for units of meaning'. *Textus* IX: 75-106.

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

Wallgrün, Jan Oliver, Alexander Klippel, and Timothy Baldwin. 2014. "Building a Corpus of Spatial Relational Expressions Extracted from Web Documents". *Proceedings of the 8th workshop on geographic information retrieval*. Accessed from https://dl.acm.org/doi/pdf/10.1145/2675354.2675702

Xiao, Richard, and Tony McEnery. 2006. "Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective." *Applied linguistics* 27.1: 103-29.

Xu, Sen, and Alexander Klippel. 2012. "Developing Nearness Models from Geocoding Spatial Entities in a News Corpus." *GIScience 2012, extended abstracts*.

APPENDIX A: First 50 collocates from the Compare search near_j / close_j + _nn*

WORD 1 (W1): NEAR (0.15)						WORD 2 (W2): CLOSE (6.87)					
	WORD	W1	W2	W1/W2	SCORE		WORD	W2	W1	W2/W1	SCORE
1	FUTURE	4628	2	2,314.0	15,907.7	1	FRIEND	3264	0	6,528.0	949.6
2	TERM	903	0	1,806.0	12,415.4	2	ATTENTION	1805	0	3,610.0	525.1
3	COLLAPSE	112	0	224.0	1,539.9	3	RELATIONSHIP	1333	0	2,666.0	387.8
4	CERTAINTY	107	0	214.0	1,471.2	4	LOOK	1040	0	2,080.0	302.6
5	FALL	82	0	164.0	1,127.4	5	RANGE	1009	0	2,018.0	293.5
6	TEARS	72	0	144.0	989.9	6	CALL	779	0	1,558.0	226.6
7	MONOPOLY	65	0	130.0	893.7	7	RACE	644	0	1,288.0	187.4
8	EXTINCTION	59	0	118.0	811.2	8	SECOND	617	0	1,234.0	179.5
9	DEATH	413	4	103.3	709.8	9	TIES	1176	1	1,176.0	171.1
10	BANKRUPTCY	50	0	100.0	687.5	10	QUARTERS	544	0	1,088.0	158.3
11	PANIC	49	0	98.0	673.7	11	CALLS	442	0	884.0	128.6
12	IMPOSSIBILITY	46	0	92.0	632.5	12	RELATIONSHIPS	434	0	868.0	126.3
13	COMPLETION	45	0	90.0	618.7	13	WATCH	333	0	666.0	96.9
14	WATER	42	0	84.0	577.5	14	EXAMINATION	330	0	660.0	96.0
15	POVERTY	41	0	82.0	563.7	15	GAMES	322	0	644.0	93.7
16	WALL	40	0	80.0	550.0	16	SCRUTINY	294	0	588.0	85.5
17	REPEAT	38	0	76.0	522.5	17	ASSOCIATION	258	0	516.0	75.1
18	ABSENCE	38	0	76.0	522.5	18	FRIENDSHIP	222	0	444.0	64.6
19	DISASTER	36	0	72.0	495.0	19	FRIENDS	3394	8	424.3	61.7
20	STANDSTILL	34	0	68.0	467.5	20	INSPECTION	207	0	414.0	60.2
21	SURFACE	34	0	68.0	467.5	21	COOPERATION	200	0	400.0	58.2
22	SIDE	134	2	67.0	460.6	22	ELECTION	396	1	396.0	57.6
23	RIOT	32	0	64.0	440.0	23	ENCOUNTERS	383	1	383.0	55.7
24	CAPACITY	27	0	54.0	371.2	24	ASSOCIATES	170	0	340.0	49.5
25	DARKNESS	52	1	52.0	357.5	25	CONTACT	678	2	339.0	49.3
26	SILENCE	50	1	50.0	343.7	26	COLLABORATION	162	0	324.0	47.1
27	RETIREMENT	47	1	47.0	323.1	27	RACES	162	0	324.0	47.1
28	VACUUM	22	0	44.0	302.5	28	ASSOCIATE	157	0	314.0	45.7
29	UNANIMITY	21	0	42.0	288.7	29	ALLIES	138	0	276.0	40.1
30	MIRACLE	21	0	42.0	288.7	30	FRIENDSHIPS	136	0	272.0	39.6
31	PERFECTION	20	0	40.0	275.0	31	TABS	135	0	270.0	39.3
32	DOUBLING	20	0	40.0	275.0	32	GAME	265	1	265.0	38.5
33	OBSESSION	19	0	38.0	261.2	33	CONNECTION	261	1	261.0	38.0
34	HYSTERIA	19	0	38.0	261.2	34	OBSERVATION	128	0	256.0	37.2
35	MELTDOWN	18	0	36.0	247.5	35	SUPERVISION	127	0	254.0	36.9
36	STARVATION	18	0	36.0	247.5	36	ALLY	241	1	241.0	35.1
37	TRAGEDY	18	0	36.0	247.5	37	CONNECTIONS	117	0	234.0	34.0
38	DEPRESSION	17	0	34.0	233.7	38	PROXIMITY	1381	6	230.2	33.5
39	DESTRUCTION	17	0	34.0	233.7	39	LINKS	113	0	226.0	32.9
40	CORNER	17	0	34.0	233.7	40	LID	104	0	208.0	30.3
41	MISS	162	5	32.4	222.7	41	ADVISER	99	0	198.0	28.8
42	SPACECRAFT	30	1	30.0	206.2	42	SHOT	94	0	188.0	27.3
43	CAMPUS	15	0	30.0	206.2	43	STUDY	90	0	180.0	26.2
44	DROWNING	15	0	30.0	206.2	44	ANALYSIS	90	0	180.0	26.2
45	END	56	2	28.0	192.5	45	CONTACTS	88	0	176.0	25.6
46	HALT	14	0	28.0	192.5	46	CIRCLE	86	0	172.0	25.0
47	MISSES	137	5	27.4	188.4	47	AIR	168	1	168.0	24.4
48	ZERO	26	1	26.0	178.7	48	MONITORING	83	0	166.0	24.1
49	TOTAL	12	0	24.0	165.0	49	ELECTIONS	81	0	162.0	23.6
50	REPEATS	12	0	24.0	165.0	50	CONTEST	76	0	152.0	22.1

APPENDIX B: First 50 collocates from the Compare search nearer_j / closer_j + _nn*

WORD 1 (W1): NEARER (0.03)						WORD 2 (W2): CLOSER (31.36)					
	WORD	W1	W2	W1/W2	SCORE		WORD	W2	W1	W2/W1	SCORE
1	HOME	19	0	38.0	1,191.6	1	LOOK	4011	1	4,011.0	127.9
2	TERM	9	0	18.0	564.4	2	INSPECTION	562	0	1,124.0	35.8
3	JUSHI	6	0	12.0	376.3	3	EXAMINATION	417	0	834.0	26.6
4	FUTURE	6	0	12.0	376.3	4	ATTENTION	362	0	724.0	23.1
5	TREES	4	0	8.0	250.9	5	TIES	233	0	466.0	14.9
6	HAND	3	0	6.0	188.1	6	SCRUTINY	191	0	382.0	12.2
7	DEATH	3	0	6.0	188.1	7	RELATIONSHIP	172	1	172.0	5.5
8	PILES	3	0	6.0	188.1	8	CONTACT	80	0	160.0	5.1
9	STARS	9	3	3.0	94.1	9	PROXIMITY	72	0	144.0	4.6
10	POINT	3	1	3.0	94.1	10	RELATIONSHIPS	68	0	136.0	4.3
11	END	3	1	3.0	94.1	11	RELATIONS	67	0	134.0	4.3
12	OBJECT	4	2	2.0	62.7	12	ANALYSIS	52	0	104.0	3.3
13	SUBURBS	4	2	2.0	62.7	13	CONNECTION	49	0	98.0	3.1
14	SIDE	4	3	1.3	41.8	14	COOPERATION	90	1	90.0	2.9
15	UNDERSTANDING	3	16	0.2	5.9	15	ROLE	40	0	80.0	2.6
16	VIEW	3	69	0.0	1.4	16	EYE	73	1	73.0	2.3
						17	STUDY	35	0	70.0	2.2
						18	RANGE	33	0	66.0	2.1
						19	READING	32	0	64.0	2.0
						20	SHOT	31	0	62.0	2.0
						21	RACE	31	0	62.0	2.0
						22	INTEGRATION	30	0	60.0	1.9
						23	COLLABORATION	29	0	58.0	1.8
						24	LINKS	28	0	56.0	1.8
						25	UNION	28	0	56.0	1.8
						26	WATCH	28	0	56.0	1.8
						27	INVESTIGATION	27	0	54.0	1.7
						28	RESEMBLANCE	24	0	48.0	1.5
						29	CALL	22	0	44.0	1.4
						30	ALIGNMENT	21	0	42.0	1.3
						31	MONITORING	21	0	42.0	1.3
						32	SUPERVISION	20	0	40.0	1.3
						33	WALK	36	1	36.0	1.1
						34	OBSERVATION	18	0	36.0	1.1
						35	TABS	17	0	34.0	1.1
						36	MATCH	16	0	32.0	1.0
						37	ASSOCIATION	15	0	30.0	1.0
						38	BOND	15	0	30.0	1.0
						39	TOUCH	15	0	30.0	1.0
						40	COORDINATION	14	0	28.0	0.9
						41	MAGAZINE	13	0	26.0	0.8
						42	SHAVE	13	0	26.0	0.8
						43	ACCESS	12	0	24.0	0.8
						44	VIEW	69	3	23.0	0.7
						45	ALLIANCE	11	0	22.0	0.7
						46	COMPARISON	11	0	22.0	0.7
						47	QUARTERS	11	0	22.0	0.7
						48	PARTNERSHIP	11	0	22.0	0.7
						49	PARALLEL	10	0	20.0	0.6
						50	REVIEW	10	0	20.0	0.6

APPENDIX C: First 50 collocates from the Compare search nearest_j / closest_j + _nn*

WORD 1 (W1): NEAREST (0.68)						WORD 2 (W2): CLOSEST (1.46)					
	WORD	W1	W2	W1/W2	SCORE		WORD	W2	W1	W2/W1	SCORE
1	PHONE	36	0	72.0	105.4	1	FRIENDS	1465	3	488.3	333.5
2	HEALTH	15	0	30.0	43.9	2	ADVISERS	168	0	336.0	229.4
3	RIVER	14	0	28.0	41.0	3	ALLY	166	0	332.0	226.7
4	INTEGER	13	0	26.0	38.1	4	AIDES	114	0	228.0	155.7
5	AIRLOCK	12	0	24.0	35.1	5	FRIEND	592	3	197.3	134.7
6	CROSS	12	0	24.0	35.1	6	ALLIES	276	2	138.0	94.2
7	DOLLAR	11	0	22.0	32.2	7	CONFIDANTS	52	0	104.0	71.0
8	FUTURE	11	0	22.0	32.2	8	ADVISER	48	0	96.0	65.6
9	FOOT	10	0	20.0	29.3	9	ASSOCIATES	92	1	92.0	62.8
10	PAY	10	0	20.0	29.3	10	CONFIDANT	41	0	82.0	56.0
11	SECOND	10	0	20.0	29.3	11	COLLEAGUES	31	0	62.0	42.3
12	ROCK	10	0	20.0	29.3	12	AIDE	29	0	58.0	39.6
13	PUB	10	0	20.0	29.3	13	RELATIONSHIPS	28	0	56.0	38.2
14	STREAM	10	0	20.0	29.3	14	PROXIMITY	27	0	54.0	36.9
15	TRAILHEAD	10	0	20.0	29.3	15	CONFIDANTE	27	0	54.0	36.9
16	ROAD	73	4	18.3	26.7	16	ADVISORS	50	1	50.0	34.1
17	SHELF	9	0	18.0	26.4	17	ATTENTION	24	0	48.0	32.8
18	TRAUMA	9	0	18.0	26.4	18	RACE	22	0	44.0	30.0
19	FREEWAY	9	0	18.0	26.4	19	ADVISOR	21	0	42.0	28.7
20	GARBAGE	9	0	18.0	26.4	20	ELECTION	21	0	42.0	28.7
21	ENTRANCE	9	0	18.0	26.4	21	ANALOGUE	20	0	40.0	27.3
22	CHAIR	51	3	17.0	24.9	22	ASSOCIATE	20	0	40.0	27.3
23	OCEAN	16	1	16.0	23.4	23	RELATIONSHIP	19	0	38.0	25.9
24	TRASH	16	1	16.0	23.4	24	RACES	18	0	36.0	24.6
25	CAFE	8	0	16.0	23.4	25	SUPPORTERS	17	0	34.0	23.2
26	BUSH	8	0	16.0	23.4	26	FINISH	16	0	32.0	21.9
27	EVACUATION	8	0	16.0	23.4	27	CONFIDANTES	15	0	30.0	20.5
28	DRUGSTORE	8	0	16.0	23.4	28	PALS	15	0	30.0	20.5
29	POLICE	56	4	14.0	20.5	29	ANALOGY	29	1	29.0	19.8
30	BOOKSTORE	14	1	14.0	20.5	30	FOLLOWERS	13	0	26.0	17.8
31	ELEVATOR	7	0	14.0	20.5	31	DISCIPLES	12	0	24.0	16.4
32	MILLION	7	0	14.0	20.5	32	CIRCLE	12	0	24.0	16.4
33	MOSQUE	7	0	14.0	20.5	33	CHILDHOOD	12	0	24.0	16.4
34	GARAGE	7	0	14.0	20.5	34	STATES	12	0	24.0	16.4
35	INTERNET	7	0	14.0	20.5	35	RESEMBLANCE	12	0	24.0	16.4
36	TUBE	7	0	14.0	20.5	36	COLLABORATORS	11	0	22.0	15.0
37	TELEVISION	7	0	14.0	20.5	37	CONNECTION	11	0	22.0	15.0
38	TOILET	7	0	14.0	20.5	38	PARTNERS	10	0	20.0	13.7
39	RETAILER	7	0	14.0	20.5	39	TIES	19	1	19.0	13.0
40	MALL	13	1	13.0	19.0	40	TOBACCO	9	0	18.0	12.3
41	CLIFF	13	1	13.0	19.0	41	SCRUTINY	9	0	18.0	12.3
42	WALL	64	5	12.8	18.7	42	COUSIN	9	0	18.0	12.3
43	VET	6	0	12.0	17.6	43	PARALLEL	34	2	17.0	11.6
44	UNIVERSITY	6	0	12.0	17.6	44	DISTANCE	17	1	17.0	11.6
45	RAILWAY	6	0	12.0	17.6	45	ELECTIONS	8	0	16.0	10.9
46	STAIRWELL	6	0	12.0	17.6	46	THINGS	29	2	14.5	9.9
47	CLUSTER	6	0	12.0	17.6	47	VOTE	7	0	14.0	9.6
48	COMPUTER	6	0	12.0	17.6	48	ENCOUNTER	7	0	14.0	9.6
49	CONTINENT	6	0	12.0	17.6	49	GIRLFRIENDS	7	0	14.0	9.6
50	DISPENSING	6	0	12.0	17.6	50	FRIENDSHIPS	7	0	14.0	9.6