rEFLections
Vol 29, No 3, September - December 2022

People or Persons?: A Corpus-based Study

PORNTHIP SUPANFAI
Mahasarakham University, Thailand
Author email: pornthip.su@msu.ac.th

Article information

Article history:
Received: 11 Jul 2022
Accepted: 18 Nov 2022
Available online: 22 Nov 2022

Keywords:
British national corpus
Collocation
People
Persons
Semantic preference
Synonym

Abstract

The study aims to investigate the similarities and differences between nominal synonyms people and persons focusing on collocations and semantic preferences. The data are drawn from the Longman Dictionary of Contemporary English (online version) and the original British National Corpus. The results of the study demonstrate that the two nouns share five statistically significant collocates and five semantic preferences including health, age, employment status, socioeconomic status, and thoughts and feelings. However, they also display distinctive semantic preferences. While people shows semantic preferences for negative actions, numbers, and ethnicity words, persons frequently collocates with words from the semantic set of legislation. The analysis of collocations and semantic preferences also confirms a high degree of formality of persons as indicated in the dictionary. Even so, the corpus data show that despite a high degree of formality, persons is not restricted to formal contexts but can also be used in informal contexts such as fiction.

INTRODUCTION

Recently, lexical semantics or the study of word meanings has been extensively investigated using a corpus linguistic approach as it is argued that the semantic and functional properties of a word can be revealed by the word's distributional characteristics (Gries, 2010). This concept has been applied particularly in the area of synonymy (Gries, 2010). Unlike traditional studies of synonyms which centre on description of semantic similarities between synonymous words, corpus-based studies of synonyms seek to identify the actual use of these words in discourse and focus on differences rather than similarities (Storjohann, 2010). These studies mainly rely on co-occurrence information at both lexical level i.e. collocation and syntactic level i.e. colligation as criteria for distinguishing synonyms (Gries, 2010).

The present study originated from my curiosity. As an English language learner and teacher, I have always been curious about when it is more appropriate to use *people* rather than *persons* and vice versa. This is because both words are the plural forms of *person* and have a similar meaning of 'human beings'. When I asked my colleagues about the differences between the two words, the only answer I often received from them was that *persons* has a higher degree of formality than *people*. For instance, it is common to see the use of *persons* instead of *people* in a notice indicating the occupancy limit in a lift. Even so, I still wonder if there are differences other than the difference in degrees of formality. This has therefore motivated me to conduct the present study. In particular, the present research aims to investigate the similarities and differences between the synonymous nouns *people* and *persons* in terms of collocations and semantic preferences through data drawn from a large collection of authentic British English i.e. the British National Corpus. However, in order to obtain as much information as possible about the words under study, I started by comparing and contrasting the accounts of the two words provided in a learner dictionary before investigating them in the corpus in more detail. It is hoped that the analysis will shed more light on the similar and different semantic properties between the words *people* and *persons* and be a valuable addition to the study in the field of lexical semantics.

This paper is structured as follows. First, a brief overview of synonymy, collocation, semantic preference, and previous studies on English synonyms is given. Then, the methodology and the data utilised in the study are presented. This is followed by the report and discussion of the results and the limitations of the study. Finally, the concluding remarks are given.

SYNONYMY, COLLOCATION, AND SEMANTIC PREFERENCE: A BRIEF OVERVIEW

1. Synonymy

Synonyms are “different phonological words which have the same or very similar meaning” (Saeed, 2003). Some examples of synonymous pairs are *lift/elevator, gain/obtain*, and *strong/ powerful*. Even though there are many synonyms in English, true or exact synonyms, that is, words that mean exactly the same and can always be substituted for each other, are very few in number (Denham & Lobeck, 2010). In fact, most English synonyms can be different in several aspects (Palmer, 1981, as cited in Saeed, 2003, p. 65).

Synonyms can vary according to dialects and registers (Denham & Lobeck, 2010; Saeed, 2003). Dialectal variation refers to the place where synonymous words are used. For instance, Irish people use the word *press* to refer to a piece of furniture commonly known by British people as a *cupboard* (Saeed, 2003). This synonymous pair *press/cupboard* thus exemplifies the dialectal variation between Irish English and British English. Register variation refers to the styles of context where the words in synonymous pairs occur e.g. informal, formal, or literary. For example, *wife* is more formal than *missus*, or cop and *copper* are slang terms for the more formal term *police officer* (Saeed, 2003). Regarding this point, Denham and Lobeck (2010) further argue that as English-speaking people consider words of Latin or Greek origin more valuable and prestigious than words with Anglo-Saxon roots, they prefer to use the former over the latter in formal, academic contexts. Examples of synonymous pairs containing words of Latin or Greek origin and their Anglo-Saxon equivalents are *attempt/try, difficult/hard, sanitary/clean*, and *cathedral/church*. In each pair, the first word is derived from Latin or Greek.

Synonyms can also be differentiated based on connotation or the speaker’s attitude (Phoocharoensil, 2010; Saeed, 2003). For instance, *naive* and *gullible* are more negative than their counterpart *ingenious*, or while *fuzz* and *flatfoot* reveal negative attitude towards the police, *cop* is neutral (Saeed, 2003). In addition to the aforementioned criteria, synonyms can also be distinguished according to collocation and semantic preference (Partington, 1998; Xiao & McEnery, 2006). These two concepts are discussed in the following sections.

2. Collocation

Collocation has been studied for a long time (McEnery & Hardie, 2012; Xiao & McEnery, 2006). The term was first used by Firth (1957, as cited in Sinclair et al., 2004, p. 3) who proposed the concept of meaning by collocation, arguing that one level of meaning of a word is, as McEnery and Hardie (2012) put it, "not contained within the word itself, considered in isolation, but rather subsist in the character associations that the word participates in, alongside other words or structures with which it frequently co-occurs" (p. 123). In other words, it is a meaning a word obtains from another word it frequently appears with.

The concept has been extensively studied and operationalised by neo-Firthian corpus linguists, especially by John Sinclair (McEnery & Hardie, 2012). Sinclair et al. (2004) define collocation as "the co-occurrence of two items in a text within a specified environment," (p. 10) and introduce the technical terms including *node, collocate*, and *span*. The node refers to the word under study. The collocate refers to any word that occurs with the node within the specified span which is defined as "the amount of text within which collocation between items is said to occur" (Sinclair et al., 2004, p. 10). Hence, according to this account, collocation is viewed as a loose pattern of co-occurrence in which a node and a collocate can co-occur in any order, and they do not have to appear adjacently (McEnery & Hardie, 2012).

A word can collocate with lexical words. For example, in the BNC *cheese* is found to collocate with lexical items such as *feta, ricotta*, and *parmesan* (McEnery & Hardie, 2012). A word can also collocate with grammatical categories. This type of co-occurrence is often referred to as colligation (McEnery & Hardie, 2012). Many English nouns, for example, typically colligate with the definite article *the*. In addition to describing co-occurrence of a word with some grammatical categories, colligation also refers to "patterns of consistent co-occurrence of a word with different syntactic contexts (McEnery & Hardie, 2012, p. 130). Hoey (2005), for instance, examines the colligational properties of the noun *consequence* in the clause in the corpus of texts mainly drawn from the Guardian, finding that the noun prefers the Complement function and the Adjunct function but avoids the Object function.

There are two approaches to collocation and colligation extraction: non-statistical approach and statistical approaches (McEnery & Hardie, 2012). For the non-statistical approach, collocations and colligations are identified manually by the linguist via concordance lines. The statistical approach, on the other hand, involves the utility of statistics for collocation and colligation extraction. Both methods are accepted and widely used in collocational and colligational studies.

3. Semantic preference

Semantic preference is a concept closely related to collocation (Phoocharoensil, 2021). Stubbs (2001) defines semantic preference as "a relation, not between individual words, but between a lemma or word-form and a set of semantically related words" (p. 65). Put simply, while collocation is a relation between a node and a single collocate, semantic preference is a relation between a node and a set of collocates that are related in meaning. A well-cited study on semantic preference is a study conducted by Stubbs (2001). In his work, Stubbs (2001) investigates the adjective *large* in a 200-million-word corpus, finding that approximately 25% of the collocates of *large* are words denoting sizes and quantities such as *number(s), scale, part*, and *amounts.* The concept is also extensively examined by Partington (1998, 2004). For example, Partington (1998) investigates the adjective *sheer*, finding that the word co-occurs with words from the semantic sets of magnitude, force, persistence, strong emotion, and physical quality. He goes on to compare the collocational behaviour of *sheer* to that of the words considered to be synonyms of *sheer* including *pure, complete,* and *absolute*, pointing out that none of them share the same semantic preferences as *sheer*. To illustrate, *pure* is used to modify material substances and proper nouns. *Complete* collocates with words expressing absence, change, and destruction such as *ban, contrast,* and *collapse*, whereas *absolute* co-occurs with what Partington (1998) calls hyperbolic nouns such as *shame, chaos,* and *disgrace*.

Frequently discussed together with semantic preference is semantic prosody. The term semantic prosody was first introduced to the public by Louw (1993), who defines it as "a consistent aura of meaning with which a form is imbued by its collocates" (p. 157). He explains this concept using the adverb *utterly*. Examining the concordance lines of *utterly*, Louw (1993) discovers that many collocates that appear on the right of the word are negative in meaning such as *arid, confused, demolished*, and *destroying*, thereby giving *utterly* a negative semantic prosody. Louw (1993) classifies semantic prosody into positive and negative, arguing that there appear to be more negative prosodies than positive prosodies.

Two important points can be made based on Louw's account of semantic prosody. First, the semantic prosody of a word can be regarded as a word's tendency to co-occur with words with positive meaning or negative meaning. Second, similarly to semantic preference, semantic prosody is also a collocational phenomenon for the reason that the semantic prosody of a word is determined by the collocational patterns of that word. That said, one possible distinction between the two concepts can be made, that is, while "a semantic preference may be in favour of any definable semantic field, a semantic prosody is always either for positive evaluation or for negative evaluation" (McEnery & Hardie, 2012, p. 137). It must be noted at this point that semantic prosody is a controversial concept (Zhang, 2009). The account of semantic prosody presented in this study is by no means comprehensive. Detailed discussion of the concept can be found in Partington (2004, 2014), Sinclair (2004), Stubbs (2001), and Stewart (2010).

PREVIOUS STUDIES ON ENGLISH SYNONYMS

The discussion above shows the close relationship between collocation, semantic preference, and semantic prosody. We see that the semantic preference and the semantic prosody of a word are identified based on the word's collocates. For this reason, many corpus-based studies of synonyms rely on the words' collocational patterns, semantic preferences, and semantic prosodies as criteria for identifying possible similarities and differences between the words under study. Some studies extract collocations using collocation statistics. In other studies, collocations are identified manually by the researcher. In addition to collocation, semantic preference, and semantic prosody, other criteria for distinguishing synonyms are adopted such as distribution across genres and colligation. From my observation, most studies utilise the data derived either from the Contemporary Corpus of American English (COCA) or from the British National Corpus (BNC). Generally, these studies discover that synonyms cannot *always* be used interchangeably. Below is a brief overview of some recent corpus-based studies of English synonyms.

Ajmal et al. (2022) study the adjectives *beautiful, handsome*, and *pretty* focusing on their meanings and collocations. The data are obtained from the Longman Dictionary of Contemporary English (6th Edition) and the BNC. The collocates extracted from the corpus are manually identified from 100 randomly-selected concordance lines. The data reveal that *beautiful* has the most noun collocates and *handsome* has the most adverb collocates. The results of the study also show that the three adjectives are very much different in their collocations as they share only two noun collocates i.e. *woman* and *face* and only one adverb collocate i.e. *very*, which can modify almost any adjective.

Phoocharoensil (2021) investigates the verbal synonyms *persist* and *persevere* with an emphasis on their distribution across genres, collocations, semantic preferences and semantic prosodies in COCA, consulting the Longman Dictionary of Contemporary English (6th Edition) as the supplementary data. The study reveals that *persist* and *persevere* occur most frequently in academic texts and webpages respectively. The analysis of the statistically significant collocates shows that whereas *persist* collocates with nouns denoting sociocultural problems, false or unconfirmed belief, health problems, and disaster, *persevere* co-occurs with nouns associated with Christianity and powerful countries. In terms of semantic prosody, the concordance analysis of the two verbs suggests that *persist* commonly occurs in negative environments but *persevere* does not.

Phoocharoensil and Kanokpermpoon (2021) examine the similarities and differences between the verbs *increase* and *rise* in terms of distribution across genres and collocations in COCA. The study reveals that *increase* occurs with the highest frequency in academic texts and *rise* in popular magazines. The collocates are extracted statistically. The two verbs are found to share many adverb collocates but relatively fewer noun collocates. Despite the shared collocates and semantic preferences, the verb *rise* is characterised by its co-occurrence with the subject nouns associated with nature in a noun + verb combination to describe an upward movement of a natural entity such as the sun, the moon, and the sea.

METHODOLOGY

The words *people* and *persons* were first examined for their definitions and usage in the Longman Dictionary of Contemporary English (LDOCE) (online version). The LDOCE was selected on the grounds that it is compiled using the Longman Corpus Network (Pearson, n.d.). This indicates that the definitions and usage of a word provided in the dictionary represents the authentic use of that word. One may argue at this point that it seems pointless then to conduct a corpus-based analysis of the words that have already been studied via a corpus linguistic approach, as the results obtained from the analysis could be similar to the accounts of the words provided in the dictionary. I would argue that my analysis could *also* return additional information not present in the LDOCE and more importantly make the similarities and differences between the words under study become more distinct, as will be discussed later.

The target nouns were subsequently investigated for their collocations and semantic preferences in the original British National Corpus (BNC). The original BNC or the BNC1994 is a 100-million-word general corpus of contemporary British English (Aston & Burnard, 1998). Ninety percent of the corpus is made of written texts derived from a wide range of sources. The remaining ten percent is made of spoken texts drawn from formal and informal encounters. Each text in the BNC was assigned a genre label (Lee, 2001). The written texts were classified into 46 sub-genres, and the spoken texts were categorised into 24 sub-genres. The study utilised all the written and spoken data in the corpus. The BNC1994 data can be accessed via different interfaces. In this study, the corpus data were explored via the CQPweb (Hardie, 2012), a web-based corpus analysis system developed at Lancaster University, due to the system's flexibility, user-friendly interface, and various analysis options.

It should also be noted at this point that I chose the original BNC over the new BNC or the BNC2014 because when I started to conduct the study, the new BNC was only available via LancsBox and could only be searched via Key Word in Context (KWIC) (CASS, n.d.). This limitation in functionalities could thus make it difficult for me to extract statistically significant collocates and complete my analysis. Using the original BNC, I could fulfil my objective of identifying the similarities and differences between the target nouns. Moreover, the availability of various tools in the CQPweb could assist me to analyse the data conveniently. For this reason, I chose the original BNC despite the fact that the data contained in the corpus may not be as up-to-date as those available in the new BNC.

I chose to identify the collocates using the statistical approach due to the large dataset. The collocates were generated based on lemmas rather than wordforms. They were extracted within a 4-4 window span as proposed by Sinclair et al. (2004). The statistical significance test adopted in the study was Mutual Information (MI). MI is a measure of collocation strength; that is, it indicates the strength of the relationship between two words (Hunston, 2002). A collocate with an MI score of three or greater than three is considered significant (Hunston, 2002). Thus, in the study, for a collocate to pass the significance test, it must have an MI score of at least three. However, as collocates with very high MI scores tend to be less frequent words, the reliance on the MI score alone may result in the inclusion of infrequent occurrences (Hunston, 2002). To avoid this shortcoming, two more criteria for the inclusion of collocates for analysis were set. First, a collocate must occur with the node at least 20 times. This practice is in line with Xiao and McEnery (2006). In their analysis of collocates of groups of near synonyms in the original BNC, Xiao and McEnery (2006) set the minimum frequency of co-occurrence at 20, arguing that this adjustment has allowed them "to use the MI score safely" (p. 105). Second, a collocate must occur with the node in at least ten different texts. This practice helps to reduce the possibility of including a collocate with the high frequency of co-occurrence but is restricted to a particular text. The collocates that passed the significance test and met the specified criteria were subsequently grouped on the basis of semantic similarity. In addition to analysing the statistically significant collocates, I investigated the concordance of the target words where relevant.

RESULTS AND DISCUSSION

This section presents the results of the study. It begins with the results obtained from the LDOCE, followed by the findings from the BNC.

1. Dictionary data

The meanings and usage of the word *people* as presented in the LDOCE are as follows.

1. persons [plural] used as a plural of 'person' to refer to men, women, and children
   e.g. How many people were at the meeting?
2. people in general [plural] people in general or people other than yourself
   e.g. I don't care what people think.
3. country/race [countable] the people who belong to a particular country, race, or area
   e.g. He pledged that he would never lie to the American people.

The word *persons* does not have its own dictionary entry, but it is presented in the entry for the word *person*. The meanings and usage of *person as* explained in the LDOCE are as follows.

1. (*plural* people) a human being, especially considered as someone with their own particular character
   e.g. The police are appealing for any person who was in the area at this time to contact them.
2. *(plural* persons) *formal* or law someone who is not known or not named
   e.g. murder by person or persons unknown

It is also noted in the dictionary that the plural form of *person* is usually *people*. The plural form *persons* is used in official English such as official notices, documents, or statements.

We can see from the dictionary data that *people* has a wider range of meanings and usage than *persons*. To illustrate, only the word *people* can refer to humans in general other than yourself and members of a nation. However, when *people* and *persons* are used as the plural forms of *person*, they express a similar meaning of human beings. Despite the similarity in meaning, the words are used in different contexts. While *people* is commonly used in ordinary contexts, *persons* is preferred in formal and legal contexts, especially to refer to someone unknown. The meanings and usage of *people* and *persons* based on the dictionary are summarised in Table 1.

Table 1

Meanings and usage of *people* and *persons*

Meanings	*People*	*Persons*
men, women, and children	✓	✓ *(*formal/legal*)*
people in general or people other than yourself	✓	
the people who belong to a particular country, race, or area	✓	
someone who is not known or not named		✓ *(*formal/legal*)*

2. Corpus data

There are 121,591 instances of *people* in 3,599 different texts in the BNC, whereas *persons* occur only 3,969 times in 996 different texts. The greater number of *people* is likely due to the word's broader range of meanings. Based on the criteria for collocation extraction, 72 collocates were obtained for *people*. However, two collocates were excluded from the analysis. One is an acronym; the other is *pe*, the unit tagged in the BNC as an unclassified item which is not a word of the English lexicon. The observation of the concordance lines shows that *pe* mostly occurs as part of the stuttering pronunciation of *people* in the spoken part of the BNC. Seventy collocates thus remained for the analysis of *people*. For *persons*, 23 collocates met the criteria and were all included in the analysis. The smaller number of the collocates of *persons* is likely due to the relatively lower frequency of the word in the corpus.

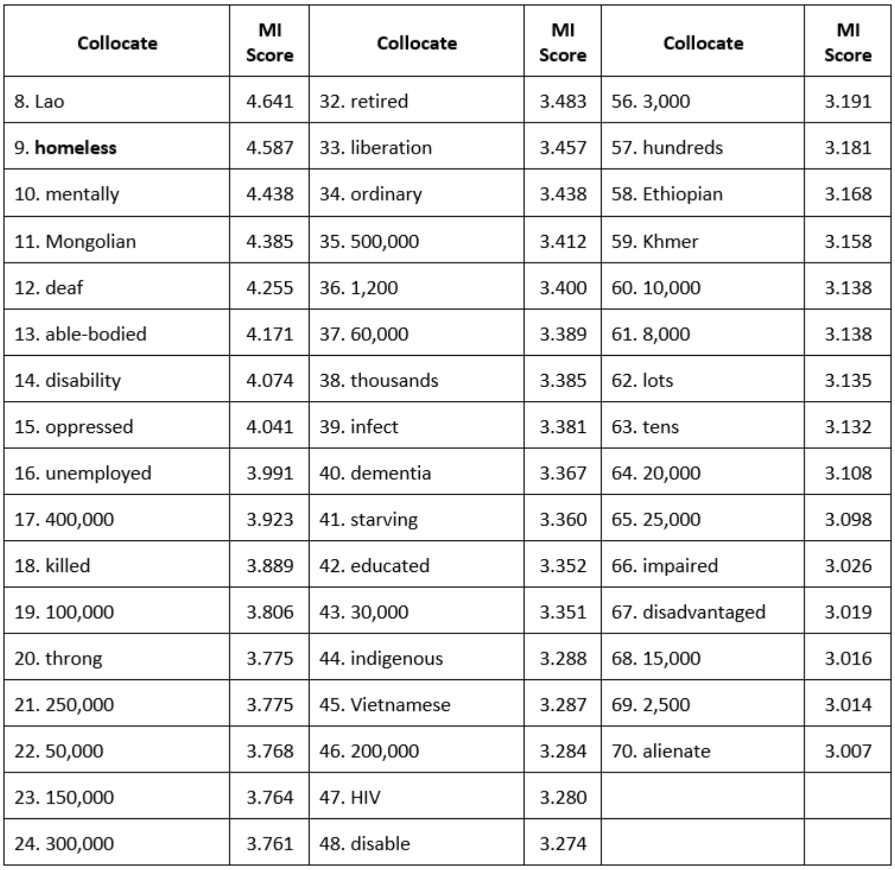
2.1 Collocational patterns of *people* and *persons*

Table 2 and Table 3 present the collocates of *people* and *persons* ranked by the MI score respectively. The collocates that are shared by *people* and *persons* are presented in bold.

Table 2

Statistically significant collocates of *people*

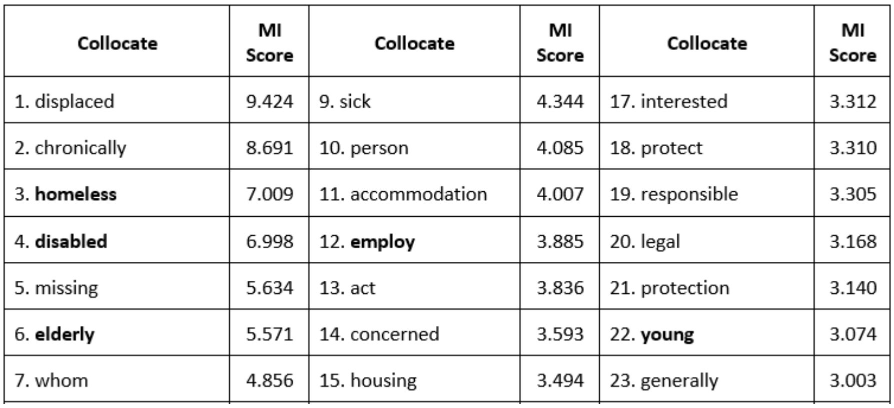
Collocate	MI Score	Collocate	MI Score	Collocate	MI Score
1. Azanian	6.664	25. 7,000	3.730	49. overweight	3.260
2. Hural	5.527	26. frail	3.719	50. kill	3.237
3. **disabled**	5.226	27. **young**	3.638	51. 5,000	3.230
4. **elderly**	4.945	28. 80,000	3.573	52. emigrate	3.223
5. handicapped	4.941	29. millions	3.547	53. 40,000	3.212
6. Eritrean	4.880	30. injure	3.542	54. **employ**	3.208
7. like-minded	4.747	31. 270	3.527	55. fewer	3.193



Collocate	MI Score	Collocate	MI Score	Collocate	MI Score
8. Lao	4.641	32. retired	3.483	56. 3,000	3.191
9. **homeless**	4.587	33. liberation	3.457	57. hundreds	3.181
10. mentally	4.438	34. ordinary	3.438	58. Ethiopian	3.168
11. Mongolian	4.385	35. 500,000	3.412	59. Khmer	3.158
12. deaf	4.255	36. 1,200	3.400	60. 10,000	3.138
13. able-bodied	4.171	37. 60,000	3.389	61. 8,000	3.138
14. disability	4.074	38. thousands	3.385	62. lots	3.135
15. oppressed	4.041	39. infect	3.381	63. tens	3.132
16. unemployed	3.991	40. dementia	3.367	64. 20,000	3.108
17. 400,000	3.923	41. starving	3.360	65. 25,000	3.098
18. killed	3.889	42. educated	3.352	66. impaired	3.026
19. 100,000	3.806	43. 30,000	3.351	67. disadvantaged	3.019
20. throng	3.775	44. indigenous	3.288	68. 15,000	3.016
21. 250,000	3.775	45. Vietnamese	3.287	69. 2,500	3.014
22. 50,000	3.768	46. 200,000	3.284	70. alienate	3.007
23. 150,000	3.764	47. HIV	3.280		
24. 300,000	3.761	48. disable	3.274		

Table 3

Statistically significant collocates of *persons*



Collocate	MI Score	Collocate	MI Score	Collocate	MI Score
1. displaced	9.424	9. sick	4.344	17. interested	3.312
2. chronically	8.691	10. person	4.085	18. protect	3.310
3. **homeless**	7.009	11. accommodation	4.007	19. responsible	3.305
4. **disabled**	6.998	12. **employ**	3.885	20. legal	3.168
5. missing	5.634	13. act	3.836	21. protection	3.140
6. **elderly**	5.571	14. concerned	3.593	22. **young**	3.074
7. whom	4.856	15. housing	3.494	23. generally	3.003

From Table 2 and Table 3, *people* and *persons* have in common five collocates which include *disabled, elderly, homeless, young*, and *employ*. However, despite these similarities, further examination of the concordance lines of *disabled* and *homeless* reveals distinctions between *people* and *persons* in terms of contexts of use. There exist many instances of the combinations *disabled people*, *disabled persons, homeless people*, and *homeless persons* in the data. These combinations are in fact the most frequent patterns of co-occurrence of *disabled* and *homeless* with the two target nouns. However, of the 107 instances of *disabled persons*, 51 occur as part of the name of an act or a bill. For example, ‘the Disabled Persons (Employment) Acts’ occurs three times, whereas ‘the Civil Rights (Disabled Persons) Bill’ was found nine times in the corpus. In a similar vein, of the 34 instances of *homeless persons*, 11 occur as part of ‘the Housing (Homeless Persons) Act’ and 3 in ‘the Homeless Persons Act’. No instances of *disabled people* and *homeless people*, on the other hand, were found to form part of the name of an act or a bill.

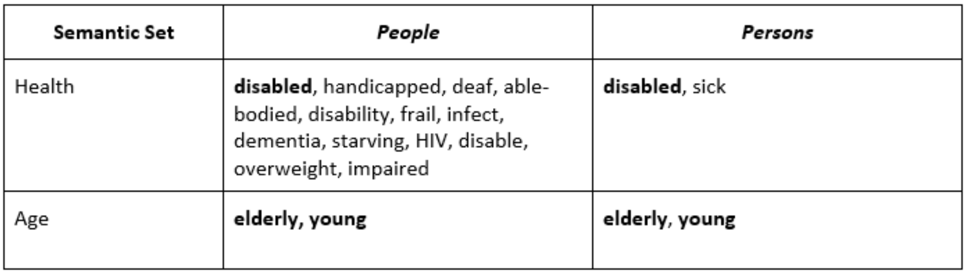
The most frequent patterns of co-occurrence of *people* with *young* and *persons* with *young* are *young people* and *young persons* respectively. Thirty-nine instances of *young persons* occur as part of ‘the Children and Young Persons Act’. However, unlike the case of *disabled* and *homeless*, there is one instance of *young people* occurring as part of the name of an act as in ‘the New German Children and Young People Act’. No discernible distinctions in terms of contexts of use were found between *people* and *persons* with regard to the collocates *elderly* and *employ*.

We can see from the analysis that even though *people* and *persons* do share five collocates, the contexts where the two words co-occur with two of these shared collocates appear to be different. Specifically, the co-occurrence of *persons* with the collocates *disabled* and *homeless* was frequently found in legal contexts, that is, as part of the name of an act or a bill. On the other hand, there are not any instances where *people* combines with each of these two collocates to form part of the name of an act or a bill. These findings thus lend support to the findings obtained from the dictionary that the noun *persons* has a higher degree of formality due to its frequent use in legal contexts.

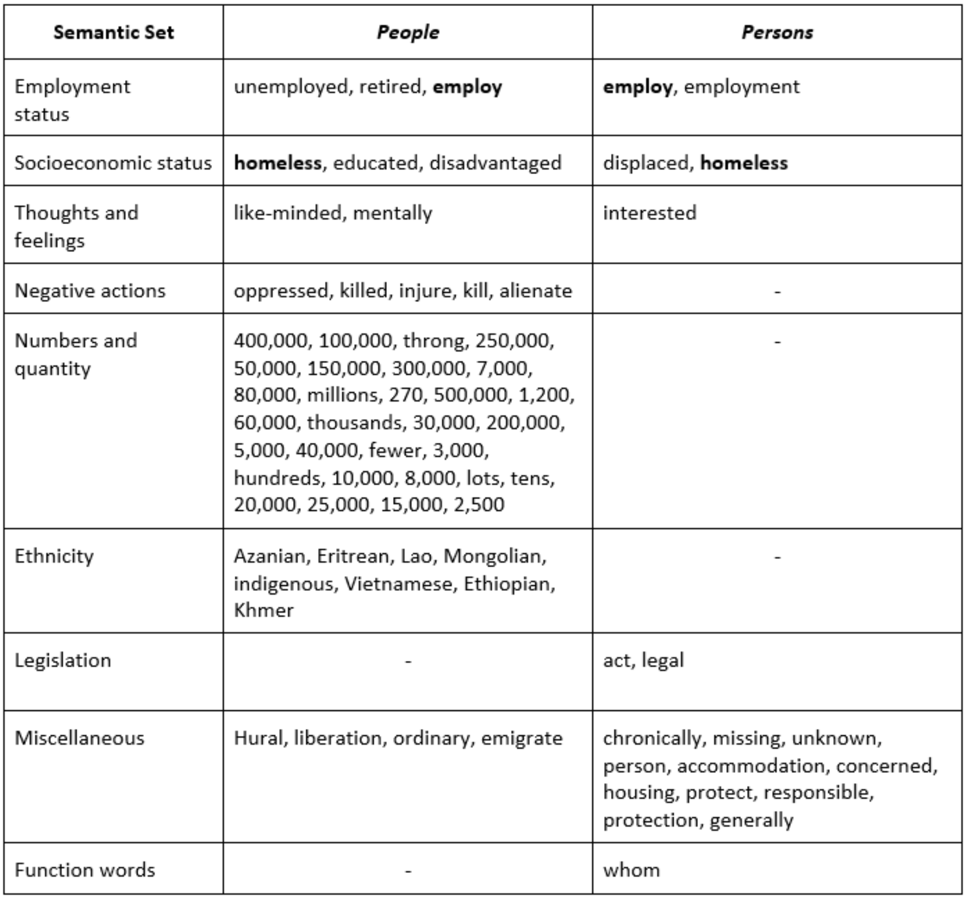
2.2 Semantic preferences of *people* and *persons*

Many collocates of *people* and *persons* are semantically related and are therefore grouped according to their semantic similarity. Table 4 below shows the collocates in different semantic sets. The collocates in each group are ranked according to their MI scores.

Table 4
Semantic preferences of *people* and *persons*



Semantic Set	*People*	*Persons*
Health	**disabled**, handicapped, deaf, able-bodied, disability, frail, infect, dementia, starving, HIV, disable, overweight, impaired	**disabled**, sick
Age	**elderly, young**	**elderly**, **young**



Semantic Set	***People***	***Persons***
Employment status	unemployed, retired, **employ**	**employ**, employment
Socioeconomic status	**homeless**, educated, disadvantaged	displaced, **homeless**
Thoughts and feelings	like-minded, mentally	interested
Negative actions	oppressed, killed, injure, kill, alienate	-
Numbers and quantity	400,000, 100,000, throng, 250,000, 50,000, 150,000, 300,000, 7,000, 80,000, millions, 270, 500,000, 1,200, 60,000, thousands, 30,000, 200,000, 5,000, 40,000, fewer, 3,000, hundreds, 10,000, 8,000, lots, tens, 20,000, 25,000, 15,000, 2,500	-
Ethnicity	Azanian, Eritrean, Lao, Mongolian, indigenous, Vietnamese, Ethiopian, Khmer	-
Legislation	-	act, legal
Miscellaneous	Hural, liberation, ordinary, emigrate	chronically, missing, unknown, person, accommodation, concerned, housing, protect, responsible, protection, generally
Function words	-	whom

Table 4 shows that based on the statistically significant collocates *people* and *persons* have in common five semantic preferences including health, age, employment status, socioeconomic status, and thoughts and feelings. It is possible to assume that these shared semantic preferences involve the overlapping meanings between *people* and *persons* of human beings. Interestingly, almost all of the collocates of *people* with the exceptions of *abled-bodied* and *educated* and all of the collocates of *persons* from the semantic groups of health and socioeconomic status are negative in meaning. These findings seem to suggest that *people* and *persons* have a tendency to occur in negative environments; in other words, the two words appear to display a negative semantic prosody. That is, based on the corpus data when language users use the words *people* and *persons* in discourse, it seems common for them to talk about the *poor* rather than the good health and the *low* rather than the high socioeconomic status of the people/persons mentioned. For example, they tend to talk about disabled, handicapped, or homeless people/persons.

In addition to the collocates associated with poor health and low socioeconomic status, *people* was also found to frequently co-occur with words related to negative actions, especially actions causing harm or danger to someone's life, which are expressed by the collocates *killed* (adj.), *injure*, and *kill* (v.). None of these collocates were found on the collocational list of *persons.* The frequent co-occurrence of *people* with these negative collocates not only strengthens the word's negative semantic prosody but also distinguishes it from *persons*. That is, although the two words display a negative semantic prosody, *people* can occur in a broader range of negative contexts. Observation of the concordance of *people* further reveals a colligational property of the word; when used with words expressing negative actions, *people* tends to be the receivers rather than the actors of these *unpleasant* actions. In other words, it seems common for language users to talk about unpleasant rather than pleasant actions that *happen* to the people mentioned as shown in the following concordance lines.

1 had returned from a sojourn in New York to aid his **oppressed** *people* in the Spanish civil war, in which he lost his life
2 disaster is when a plane falls out of the sky and 296 *people* get **killed**. It damages our pride and reputation but there
3 A man in his 20s also died and a further 90 *people* were **injured**. Last night a third boy, a man

It is worth noting at this point that even though the words *people* and *persons* frequently occur in negative environments, they can *also* appear in neutral or positive environments as reflected by the presence of the neutral or positive collocates such as *retired, elderly, young, employ, able-bodied,* and *educated*. Therefore, the findings should be interpreted as the *typical* use (Hunston, 2002) of *people* and *persons* rather than the only use of the words.

People and *persons* also have different semantic preferences. In addition to its distinctive co-occurrence with words associated with negative actions, *people* is also characterised by its tendency to collocate with numbers, indefinite determiners including *fewer* and *lots (of)*, and nouns indicating a large quantity without being exact e.g. *millions, thousands,* and *hundreds*. In fact, the semantic group of numbers and quantity is the largest group for *people*. The observation of concordance shows that in almost all cases these numbers and quantity collocates quantify the word *people*. Below are some examples taken from the concordance lines.

4 towns and villages. He thinks that in Chiang Mai alone **30,000** *people* are infected with HIV. A series of test was carried out
5 now. I didn't like what Joyce said about **lots** of *people* being interested in the house. It's got me worried
6 You can't fool hundreds of thousands, or **millions** of *people* at the same time with absolute garbage for record after record

No numbers or quantity words were found on the collocational list of *persons*. We may consider these findings as an important distinction between *people* and *persons*. However, as we know that numbers and quantity words can be used to modify any countable noun in English, to gain further insights into the use of *persons*, I investigated the use of the word in 200 randomly-selected concordance lines. I found many instances of *persons* being used with numbers in various contexts such as academic prose, non-academic prose, religion, and conversation. This suggests that numbers can *also* co-occur with *persons*. However, unlike the co-occurrence of numbers with *people*, the co-occurrence of numbers with *persons* is likely due simply to chance alone; otherwise, some numbers would have passed the significance test (MI ≥ 3) and appeared on the collocational list of *persons*. To put it another way, although numbers can occur with both *people* and *persons*, based on the significance test, it is much more common for language users to select *people* over *persons* despite the meaning overlap when they want to indicate the number of human beings they refer to.

People is also distinguished from *persons* by its semantic preference for ethnicity. This semantic preference clearly involves the distinctive meaning of *people* as members of a nation. The investigation of the concordance lines demonstrates that in many cases these ethnicity collocates co-occur with *people* in political and military contexts. To be more precise, they combine with *people* to form the name of a political organisation or party, or an army as in 'the Azanian People's Liberation Army', 'the Lao People's Revolutionary Party', and 'the Khmer's People National Liberation Front'. Nevertheless, there are also instances where *people* and the ethnicity collocate co-occur but do not form part of the name of a political party or an army as shown in the concordance lines below.

7 means we lose at least half of this annually. The **Vietnamese** *people* are trying to re-use the barren land and replant the affected areas
8 proposals were "subject to the will and decision of the **Ethiopian** *people*" and it was anticipated that the details would be put forward

The word *persons* is characterised by its tendency to collocate with the words associated with legislation including *act* and *legal*, all of which are absent from the collocational list of *people*. The examination of the concordance lines also shows that most co-occurrences of *persons* with each of these two collocates appear in the academic prose of politics, law, and education. In 104 out of 154 instances, *act* refers to a law passed by a country's government. Of the 33 instances of co-occurrence of *persons* with *legal*, 26 appear in legal contexts.

The only function word found on the collocational lists is *whom*. It is a statistically significant collocate of *persons*. The concordance analysis shows that among the 103 instances where *whom* and *persons* collocate, *whom* functions as the relative pronoun that introduces the relative clause that follows and modifies *persons* in as many as 102 instances. Moreover, among these instances of relative clause, there are 70 instances where the relative pronoun is preceded by the preposition as presented in the concordance lines below.

9 where property is subject to a trust, the *persons* <u>to</u> **whom** it belongs shall be regarded as including any person having
10 of the University they represent. To a great majority of those *persons* <u>with</u> **whom** they come into contact it is the only opportunity afforded

The use of preposition and the relative pronoun together at the beginning of the relative clause is common in formal English (Oshima & Hogue, 2006). Therefore, the tendency of *persons* to co-occur with *whom* and with the collocates from the semantic set of legislation further confirms a high degree of formality of the noun.

The collocates that do not seem to have a clear connection in meaning and do not fit in any of the semantic categories are classified into the miscellaneous group. It is interesting to see the singular *person* on the collocational list of *persons*. The two words co-occur 137 times. One observable pattern of co-occurrence is *person or persons*, which was found 71 times. The pattern is fixed in terms of word order as no instances of *persons or person* were found. However, the phrase can be preceded by different noun determiners such as *a, the,* or *any.* Further observation of the concordance lines reveals 11 instances of *person or persons unknown.* In addition to these instances, there are two other instances where the co-text around the phrase clearly indicates the absence of the identity of the person(s) mentioned as shown in the concordance lines below.

11 in this column. If not, would the *person (or persons)* responsible please make themselves known to us so that their contribution
12 can say very little about the individual identity of the *person or persons* in communication with us; their name, gender, age,

The surrounding co-text of other instances of *person or persons* also suggests, though not as clearly as it does for the two examples above, that the person(s) mentioned can be anyone with or without specific conditions who we do not know as shown in the following concordance lines.

13 an unsecured personal loan or a credit card debit lies with the *person or persons* who were party to the agreement. A point to watch here
14 suspense short story has as its object to put a *person or persons* into danger and give revelation to one or more people

These findings thus correspond to the data gained from the dictionary that *persons*, particularly the combination *person or persons*, is used to refer to someone not known. However, the corpus data also supply additional information to the dictionary. Examining the contexts where *person or persons* occur, I discovered that the phrase was found most frequently in fiction and academic prose of politics, law, and education. It occurs 15 times in each of these two genres. The corpus data therefore suggest that *person or persons* is not restricted to formal or legal contexts. To obtain further insights into the use of *person or persons* in the fiction genre, I examined all of the 15 instances of the phrase in extended co-text. I found that in 11 instances, the phrase occurs in informal contexts where the narrator or the character in the story talks about illegal actions such as murder, harassment, and ransom, as presented in the following concordance lines.

15 **murdered?**' 'Hit on a head by a *person or persons* unknown. They told me at the hospital that you've worked
16 any reason, Mr Quinn, to believe that a *person or persons* unknown might have attempted to interfere with the **ransom** exchange, to

The contexts where *person or persons* appear in the other four instances are unrelated to illegal actions or legal contexts. Two of them are presented below.

17 rule of the game that 'Smee' and the *person or persons* who have found 'Smee' have to keep quiet. This

18 likely to him was that Adam had allowed some undesirable *person or persons* access to the place and it was these vagrants or hippies -

The concordance investigation thus appears to confirm that *person or persons* can *also* be used in informal contexts. Even though the corpus data also show that in fiction the phrase is frequently used in the context where illegal actions are the object of discussion, it is part of the narration or conversation of the character rather than part of formal explanation of rules or legal systems as when the phrase occurs in academic prose of politics, law, and education.

LIMITATIONS AND POSSIBLE FUTURE RESEARCH

This study has some limitations. The study identified the similarities and differences between *people* and *persons* based on collocations and semantic preferences. However, as Wijitsopon (2021) argues, findings based on analysis of semantic preference alone should not be treated as an absolute distinction between synonymous words. Future research may therefore focus on concordance analysis and investigation of the discourse functions of the words in more detail. Furthermore, the study used the original BNC. This corpus may not be able to represent the present-day British English as clearly as the new BNC does. It is possible for future research to repeat this study using the new BNC. The results obtained from the BNC1994 and the BNC2014 can also be compared and contrasted in order to identify possible diachronic change in the language i.e. British English with regard to the use of *people* and *persons*. Most importantly, as the statistic used to calculate the significance of a collocation can in part determine the collocates extracted (McEnery & Hardie, 2012), different results may be yielded if other statistics such as log-likelihood, t-score, or z-score are applied in future studies.

CONCLUSION

The analysis of collocations and semantic preferences has allowed me to identify the semantic similarities and differences between *people* and *persons*. The findings are consistent with previous research which claims that synonyms can be distinguished based on their collocational patterns and semantic preferences and that they cannot always be substituted for each other (e.g. Jarunwaraphan & Mallikamas, 2020; Jiranathiporn, 2018; Partington, 1998; Phoocharoensil, 2021; Phoocharoensil & Kanokpermpoon, 2021). In the study, I started by conducting a contrastive analysis of the accounts of the two words provided in the Longman Dictionary of Contemporary English (online version). The original BNC was subsequently examined for more detailed information about the collocational patterns and semantic preferences of the words. The corpus data have yielded valuable insights into the authentic use of *people* and *persons* which can be summarised as follows.

With regard to collocations, the two words were found to occur with both different and similar collocates. However, the number of distinctive collocates is larger than that of the shared collocates. This seems to suggest a high degree of difference between the two words in terms of collocations. Moreover, of the five shared collocates, two were found to have a tendency to co-occur with *people* and *persons* in different contexts.

For semantic preferences, *people* and *persons* were found to have in common five semantic preferences. It is noticeable that most collocates from the semantic sets of health and socioeconomic status are negative in meaning. This can suggest that both words display a negative semantic prosody or a tendency to occur in negative environments. However, *people* was found to occur in a wider range of negative contexts than *persons*.

The analysis of collocations and semantic preferences has revealed a major distinction in terms of degrees of formality between the two words. The corpus data appear to indicate that *persons* has a higher degree of formality than *people* as it was found frequently in formal contexts, particularly in legal work. These findings correspond to the dictionary data which indicate that *people* is used in ordinary contexts while *persons* is preferred in formal contexts. Even so, the corpus analysis additionally reveals that despite the higher degree of formality, *persons* is also used in informal contexts. We have seen in the corpus data many instances of the use of *persons* in fiction. In addition to the information on degrees of formality and contexts of use, the corpus analysis has also yielded information about collocations, semantic preferences, and semantic prosodies, as mentioned earlier.

We may conclude from the study that the corpus analysis not only confirms but provides additional information regarding the definitions and usage of the target words as presented in the dictionary. However, this conclusion should not be interpreted as criticism of the dictionary. Due to limited space available in any dictionary, it can be very difficult for lexicographers to include all linguistic details of a word. In my view, the dictionary data and the corpus data have their merits. While the dictionary data provide a starting point for the study, the corpus analysis helps to highlight the similarities and differences between the words under study and supply additional information that could have been excluded from the dictionary due to insufficient space.

In addition to being a useful addition to the study in the area of lexical semantics, the findings are also useful for English language teaching. English language learners should be made aware of the similarities and especially the differences with respect to collocations, semantic preferences, semantic prosodies, degrees of formality, and contexts of use between the two words under study. They should also be taught and encouraged to seek additional information about vocabulary from corpora (Phoocharoensil, 2021). The linguistic information obtained from corpus analyses can assist English language learners to become more proficient English language users.

ACKNOWLEDGEMENT

The study reported in this paper is part of the research project *The Contrastive Corpus-based Study of English Near Synonyms: People and Persons*, funded by Faculty of Humanities and Social Sciences, Mahasarakham University, Contract No. 23/2562. The researcher is also grateful to the two anonymous reviewers for their valuable comments on the manuscript.

THE AUTHOR

Pornthip Supanfai is a lecturer at Department of Western Languages and Linguistics, Faculty of Humanities and Social Sciences, Mahasarakham University. She received her PhD in Linguistics from Lancaster University, UK. Her research interest is corpus linguistics.
pornthip.su@msu.ac.th

REFERENCES

Ajmal, M., Kumar, T., Ritonga, M., & Nukapangu, V. (2022). A corpus-based analysis of the adjectives and synonyms - beautiful, handsome, and pretty. *World Journal of English Language, 12*(2), 159-168. https://doi.org/10.5430/wjel.v12n2p159

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

CASS. (n.d.). *The British National Corpus 2014*. Retrieved May 6, 2022, from http://cass.lancs.ac.uk/bnc2014/

Denham, K., & Lobeck, A. (2010). *Linguistics for everyone*. Cengage Learning.

Firth, J. R. (1957). *Papers in linguistics*. Oxford University Press.

Gries, S. (2010). Behavioral profiles: A fined-grained and qualitative approach in corpus-based lexical semantics. *The Mental Lexicon, 5*(3), 323-346. http://doi.org/10.1075/ml.5.3.04gri

Hardie, A. (2012). CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics, 17*(3), 380-409. http://doi 10.1075/ijcl.17.3.04har

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.

Hunston, S. (2002). *Corpora in applied linguistics* (1st ed.). Cambridge University Press.

Jaruwaraphan, B., & Mallikamas, P. (2020). A corpus-based study of English synonyms: Chance and opportunity. *rEFLections, 27*(2), 218-245. https://so05.tci-thaijo.org/index.php/reflections/article/view/248710/168938

Jirananthiporn, S. (2018). Is this problem giving you trouble? A corpus-based examination of the differences between the nouns problem and trouble. *Thoughts, 2*, 1-25. https://so06.tci-thaijo.org/index.php/thoughts/article/view/163389/118109

Lee, D. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology, 5*(3), 37-72. https://dx.doi.org/10125/44565

Longman. (n.d.) People. In *Longman Dictionary of Contemporary English online*. Retrieved April 6, 2022, from https://www.ldoceonline.com/dictionary/people

Longman. (n.d.). Person. In *Longman Dictionary of Contemporary English online*. Retrieved April, 6, 2022, from https://www.ldoceonline.com/dictionary/person

Louw, W. E. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157-176). John Benjamins.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Methods, theory, and practice*. Cambridge University Press.

Oshima, A., & Hogue, A. (2006). *Writing academic English* (4th ed.). Pearson Longman.

Palmer, F. R. (1981). *Semantics* (2nd ed.). Cambridge University Press.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. John Benjamins.

Partington, A. (2004). Utterly content in each other's company: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics, 14*(2), 139-158. https://doi.org/10.1075/ijcl.9.1.07par

Partington, A. (2014). Evaluative prosody. In K. Aijmer & C. Rühlemann (Eds.), *Corpus pragmatics: A handbook* (pp. 279-303). Cambridge University Press.

Pearson. (n.d.). *Longman Corpus Network*. Retrieved September 12, 2022, from http://www.pearsonlongman.com/dictionaries/corpus/

Phoocharoensil, S. (2010). A corpus-based study of English synonyms. *International Journal of Arts and Sciences, 3*(10), 227-245.

Phoocharoensil, S. (2021). Semantic prosody and collocation: A corpus study of the near-synonyms persist and persevere. *Eurasian Journal of Applied Linguistics, 7*(1), 240-258. http://dx.doi.org/10.32601/ejal.911269

Phoocharoensil, S., & Kanokpermpoon, M. (2021). Distinguishing the near-synonyms 'increase' and 'rise': Genre and collocation investigation. *Kasetsart Journal of Social Sciences, 42*(4), 968-975. https://so04.tci-thaijo.org/index.php/kjss/article/view/255757

Saeed, J. I. (2003). *Semantics* (2nd ed.). Blackwell.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.

Sinclair, J., Jones, S., & Daley, R. (2004). *English collocation studies: The OSTI report*. Continuum.

Storjohann, P. (2010). Synonyms in corpus texts: Conceptualisation and construction. In P. Storjohann (Ed.), *Lexical semantic relations: Theoretical and practical perspectives* (pp. 69-94). John Benjamins.

Stewart, D. (2010). *Semantic prosody: A critical evaluation*. Routledge.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell.

Wijitsopon, R. (2021). A corpus-based study of Thai and English quantity word equivalents: Lǎay, several and many. *rEFLections, 28*(2), 188-207. https://so05.tci-thaijo.org/index.php/reflections/article/view/252502

Xiao, R. Z., & McEnery, T. (2006). Near synonym, collocation, and semantic prosody: A cross-linguistic perspective. *Applied Linguistics, 27*(1), 103-129. https://doi.org/10.1093/applin/ami045

Zhang, W. (2009). Semantic prosody and ESL/EFL vocabulary pedagogy. *TESL Canada Journal, 26*(2), 1-12. https://doi.org/10.18806/tesl.v26i2.411