ISSN: 2278-4012, Volume:9, Issue:2, APRIL 2020

# A CORPUS-BASED STUDY ON KEYWORDS AND COLLOCATIONS OF KEYWORDS IN BUSINESS ENGLISH NEWSPAPERS

**Nguyen Thu Hang, M. A.**

Danang University of Foreign Language Studies,

The University of Da Nang, Viet Nam

## Abstract

The study aims at investigating the most frequently used keywords in business articles and identify the patterns of collocations of keywords used in business articles. One hundred articles from business columns in "The econimist" newspapers published between the year 2019 and 2020 were chosen and the raw data was then analyzed thanks to the application of the Antconc 3.5.8. software developed by Laurence Anthony in 2014.

Within the adpotion of corpus-based research method, the findings of the study presented a list of 100 keywords that were most commonly used in business newspapers with detailed investigation into their semantic domains and grammatical functions. In the next stage, collocations of the top 10 keywords were identified and analyzed in terms of semantics and grammatical structures. The results of the study can be very beneficial to a wide range of subjects, including researchers, teachers, leanrers, language educators and many other stakeholders in the field of Business English.

Keywords: corpus linguistic; collocations; keywords; business newspapers; AntCont software

## 1.Introduction

As scientific sub-discipline of applied linguistics, corpus linguistics identifies the general principles of building corpora (text blocks) with the use of modern computer technology, developing solutions to the problem of a collection of real language phenomena. In other words, corpus linguistics "helps to examine linguistic phenomena from the point of view of frequency data and the relation of models" (Mahlberg, 2013, p.7).

Within the 21st century, along with the birth and development of linguistics, computer science and digital technology, corpus linguistics have experienced a vigourous development in the current information technology era.There have been various beneficial applications of corpus linguistics in many fields of science, especially in the field of applied linguistics. Thanks to corpus linguistics, linguists, teachers and leaners can get access to different corpora for a wide variety of purposes including teaching, learning, modelling, and contrastive analysis, etc.

ISSN: 2278-4012, Volume:9, Issue:2, APRIL 2020

One of the commonly used methods of corpus linguistics is the extraction of keywords in a corpus. The identifications of keywords of a corpus is very necessary since it can act as a means in order to determine the linguistic features of a particular kind of genre. The study aims at identifying the keywords that are most frequently used in Business English articles and the collocations of the keywords. Further analysis is implemented so as to identify semantic features and grammatical functions of the keywords and collocations of keywords.

## 2. Literature review

### 2.1 Keywords

There has been a considerable amount of literature on keywords over time. Keywords are defined as words that appear more or less frequently in one corpus than in a reference corpus (Scott, 1997; Scott & Trible, 2006). Keywords can demonstrate an artistic tendency about choosing a theme, theme, or author style, genre style, or functional variation. In addition, Keywords are automatically generated by computers by comparing the word list of the target corpus and that of a comparative corpus.

To sum up, keyword analysis has been widely employed in linguistics and other relevant fields for a wide variety of purposes including language education, stylistics and discourse analysis (Scott 2008, O'Keeffe et al. 2007, Goh & Lee 2008, McEnery 2009).

### 2.2 Collocations and collocates

There is a common assumption in teaching vocabulary that the more words a learner knows, the larger the learner's vocabulary knowlege is. However ,words are rarely used alone but they go with each other and keep their companies to make their own combinations. In considering the ability of words to combine, Wilkins (1972, p.126) states that "in every language, there are items which co-occur with high frequency, others which co-occur as the need aries and other whose co-occurrence seems imposible". In order to achieve naturalness in speech and writing, we tend to use common and regular ones .

The term "collocations" has been defined by many linguists in different ways. McCathy (1991,p.158) defines collocation as the likelihood of co-occurrence between lexical items. Collocation, in his view, is "the binding force between the words of a language". In line with McCathy, Nation (1990, p.32) believes that collocations are "words that often occur together " and the collocations of a word are the company it keeps. In addition , the importance of collocations in language learning has been supported by many scholars. As stated by McCathy (1990), collocations has a significant position in lexical acquisition among language leaners. The mastery of collocations is beneficial in enhancing learners' fluency and word use in contexts (Shin & Nation, 2008).

ISSN: 2278-4012, Volume:9, Issue:2, APRIL 2020

It is obvious that there is a variety of views on the definition of collocations. However , there is still a common core of agreement to be found, that is the consensus on the co-occurrence of words. All these linguists share the view that collococations are the habitual associations of individual words . Based on this point , it is advisable to come up with a working definition for this study, which is that collocations refer to words that keep company with one another or collocations refer to the way some words occur together. In the same light, collocate refers to a word that occurs with another with a higher frequency than chance.

### 2.3.Review of corpus linguistics and corpus-based research on keywords and collocations of keywords

Corpus linguistics has undergone developments over recent decades with considerable impacts on linguistic studies. In 1963, Brown corpus which was built for the first time in Brown University (USA) includes 1 million English - American word units and phrases from print texts selected in 1961. Since the birth of Brown corpus, a series of corpora have been introduced. One typical corpus was Lancaster / OsloBergen (LOB) containing about 1 million word units and phrases used with morphological diagram.

In addition, a large number of English corpora were constructed for specialized purposes. To be more specific, Bank of English corpus was constructed in 1997 with 320 million word units and phrases. In addition, ICLE corpus included 200 million word units as written form for foreigners. In a word, series of English corpora have been construted and used over time for both research in English and teaching English as a foreign language.

Corpus linguistics, up to the present time, has been playing an increasingly important role in the global economy when different fields of science and technology flourish. Various corpora have been widely used by linguists, applied linguistic experts, lecturers and experts from many fields field of natural science and social science.

It can be said that corpus linguistics helps pave the ways for the identification of keywords or concepts in a corpus and the analysis of collocational patterns of keywords or word clusters. We can determine the context to the left and to the right of keywords, which enable us to identify regular patterns that occur in the corpus (Granger, 2012; Barnbrook, Mason & Krishnamurthy , 2013).

Over many decades, attempts have been made with the the puposes of identifying keywords or relatedness between keywords in various corpora. For instance, 11 most frequent keywords from letters corpus were found out by Lukac Hacker(2015) and keywords in insurance research articles were under investigation by Khamphairoh and Tangpijaikul (2012). In another perspective with a different research method, Pojanapunya (2016) determined the top 30 keywords of the target corpus of research articles in applied linguistics. In this study, the relatedness of these keywords was measured with a kind of mutual information statistic measurement that shows the strength of asociation between each pair of the keywords.

ISSN: 2278-4012, Volume:9, Issue:2, APRIL 2020

## 3. Research purpose

As a matter of fact, Business English newspapers belong to a very common genre; therefore, it is a necessity to understand the basic concepts or the keywords with their collocates in the newspapers that are read by many worldwide everyday. Moreover, the study of keywords and colocations in the field of Business English is still very limited. This study hence aims at building a keyword list of the most commonly used keywords in the corpus of articles in business English newspapers and making an insight into the collocations of the keywords. An investigation into the keywords and collocations of keywords can contribute to demonstrating parts of the linguistic features of the corpus of Business English newspapers under study.

## 4.Methodology

In the first step, in order to identify the keywords and collocations of keywords in Business newspapers, a lot of thought was given to the choice of the newspapers. “The economist” was chosen as it is one of the most prestigous and reliable business newspapers worldwide. A total number of 100 articles to form the research corpus were obtained from business columns in “The econimist” newspapers published between 2019 and 2020 so that the data can be as updated as possible for the highest validity of the study. The the research corpus was named Business English newspapers (BEN) corpus by the author. There were 89,403 word tokens and 11,399 word types in the analysis by AntConc. The BEN corpus provide data of keywords and collocations of keywords for quantitative analysis and qualitative analysis is based upon the contexts of the keywords under investigation.

The software used for this corpus-based analysis of keywords and collocations of keywords is AntConc software developed by Anthony (2014), as presented in Figure 1. The software tools used for statistical research involve Keywords and Concordance.

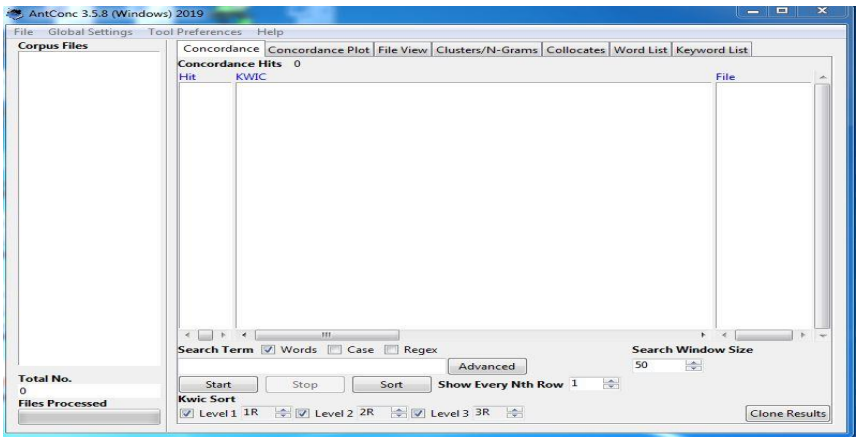


*Figure 1. The AntConc software developed by Anthony (2014)*

ISSN: 2278-4012, Volume:9, Issue:2, APRIL 2020

In the next stage, the tool of Keyword in AntConc software was utilized to obtain data on the frequency of the top keywords in BEN corpus. In terms of this function, Antconc can compare the words that appear in the target corpus with those in the reference corpus in order to create a list of keywords that are frequent in the target corpus (Anthony , 2014). Different kinds of corpora have been utilized as both target corpus and reference one by various researchers, such as Brown corpus, British National Corpus (BNC), Lancaster / OsloBergen (LOB), etc. In this study, Brown corpus was selected as the reference corpus since it has been commonly used as reference corpus and it is a typical representative of general English. Moreover, "rating scale approach" (Chung and Nation, 2004) was also applied by the reasearcher in combination with Keyword tool so as to maintain that those keywords have meanings related to Business Engish.

Within the next stage, with the list of top 100 keywords in BEN corpus in hand, thanks to Concordance tool of the software, the reasearcher could investigate the context and concordance lines of the keywords in BEN corpus. This kind of concordance data can reflect the use of collocations of keywords in contexts.

## 5. Findings and discussions

### 5.1. Keywords

The results from the software demonstrated a list of 100 most frequently used keywords whose meaning related to business in "The economist" newspapers. It can be seen from Table 1 that these keywords in general denote the basic concepts and notions in the field of business.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | firm | 21 | bank | 41 | consumers | 61 | investment | 81 | managers |
| 2 | market | 22 | price | 42 | sanctions | 62 | partners | 82 | privacy |
| 3 | workers | 23 | businesses | 43 | stake | 63 | venture | 83 | regulation |
| 4 | investors | 24 | customers | 44 | assets | 64 | laws | 84 | airlines |
| 5 | industry | 25 | trade | 45 | consultancy | 65 | vision | 85 | industries |
| 6 | sales | 26 | risk | 46 | contracts | 66 | content | 86 | margins |
| 7 | shares | 27 | investment | 47 | jobs | 67 | rival | 87 | profit |
| 8 | corporate | 28 | value | 48 | alliance | 68 | centre | 88 | regulators |
| 9 | pay | 29 | media | 49 | competition | 69 | softbank | 89 | staff |
| 10 | executives | 30 | global | 50 | value | 70 | charges | 90 | strategy |
| 11 | costs | 31 | economy | 51 | authorities | 71 | profit | 91 | trading |
| 12 | fund | 32 | prices | 52 | crisis | 72 | returns | 92 | directors |
| 13 | money | 33 | markets | 53 | manage | 73 | deals | 93 | mergers |
| 14 | capital | 34 | users | 54 | networks | 74 | earnings | 94 | recession |
| 15 | boss | 35 | brand | 55 | debt | 75 | founder | 95 | stockmarket |
| 16 | legal | 36 | demand | 56 | retail | 76 | prosecutors | 96 | buyers |
| 17 | services | 37 | bosses | 57 | risks | 77 | software | 97 | competitors |
| 18 | shareholder | 38 | startups | 58 | scale | 78 | suppliers | 98 | conglomerate |
| 19 | employees | 39 | revenues | 59 | commerce | 79 | analysts | 99 | downturn |
| 20 | cash | 40 | brands | 60 | credit | 80 | capitalism | 100 | stakes |

*Table 1. The top 100 keywords in BEN corpus.*

ISSN: 2278-4012, Volume:9, Issue:2, APRIL 2020

Concerning grammatical functions of the keywords, these keywords fall into three categories of parts of speech with the biggest proportions of the keywords being nouns (97%), followed by adjectives and verbs with 2% and 1% respectively. It is obvious that nouns take up dominant percentages in the list of keywords. This finding is in line with the results of keywords in insurance research articles by Khamphairoh and Tangpijaikul (2012) and those in applied lingusitics by Pojanapunya (2016) with the dominance of nouns over other parts of speech.

In terms of semantic analysis, the top 100 keywords can be classified into different semantic domains including:

1.Keywords denoting people and organizations in business

e.g. *employees, shareholder, bosses, founder, prosecutor, buyers, analysts....*

2. Keywords denoting activities and events in business

e.g. *downturn, crisis, recession...*

3. Keywords denoting money and finance

e.g. *cash, money, pay, shares, prices, debt...*

4. Keywords denoting concepts or subjects in business

e.g. *brand, contracts, startups, markets, capital...*

The classification above is done by the author based on semantic meanings of the keywords; nevertheless, there can be many other ways of classification in different perspectives. What's more, there may be overlap in this kind of classification as some keywords can belong to more than one semantic domains.

**5.2. Collocations of keywords in BEN corpus**

With the support of Concordance tool in the software, the the reasearcher could investigate the concordance lines of the keywords and construct the most frequent collocations of the 10 most frequent keywords in the keyword list. For instance, the following figure demonstrates the collocations of the keyword "*market*" from the concordance lines performed by AntConc 3.5.8. software.

ISSN: 2278-4012, Volume:9, Issue:2, APRIL 2020

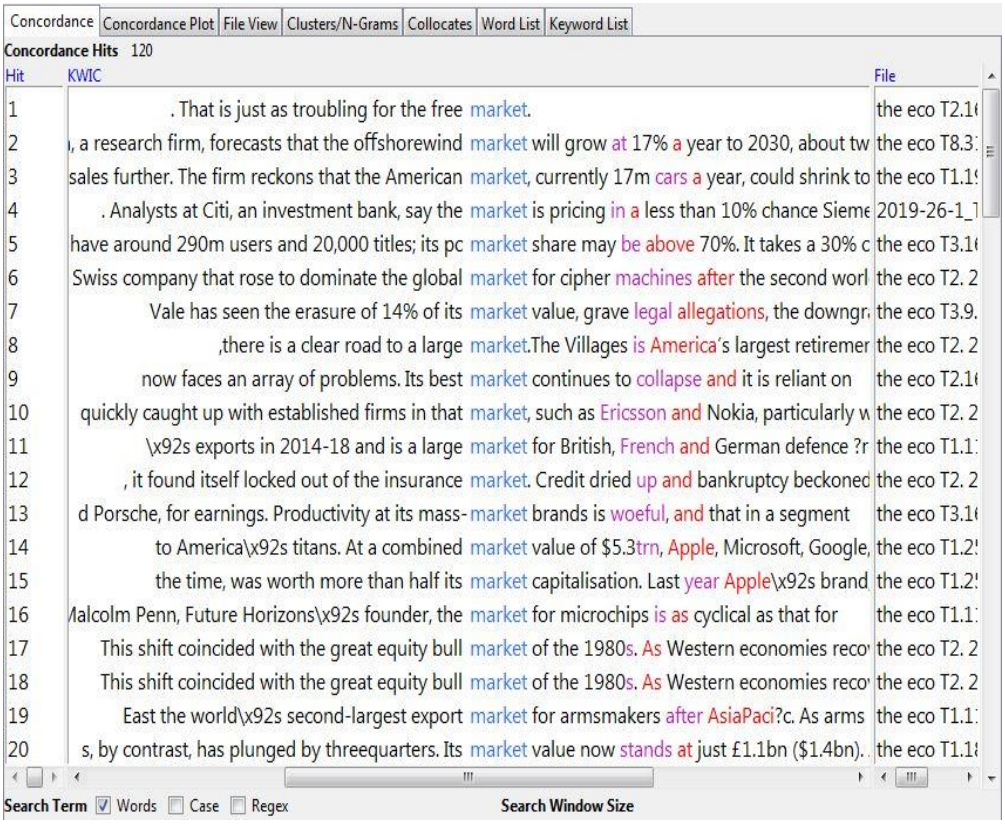


*Figure 2. The concordance lines of the keyword "market"*

Based on the concordance lines of the top 10 keywords, a list of the most frequent 2-word and 3-word collocations of the keywords was generated (as seen in table 2).

| No. | Keyword | 2-word collocations | 3-word collocations |
|---|---|---|---|
| 1 | firm | research firm<br>holding firm<br>energy firm<br>techinvesting firm<br>local firm<br>law firm<br>security firm | venture capital firm<br>upstart telecom firms |
| 2 | market | offshorewind market<br>export market<br>mass market<br>insurance market<br>bull market<br>combined market<br>global market | combined market value<br>combined market share<br>powerful market force<br>big bull market<br>global retail market |

| | | | |
|---|---|---|---|
| 3 | investors | potential investors<br>active investors<br>private-sector investors<br>foreign investors<br>domestic investors<br>intitutitional investors | climate concerned investors<br>wary bond investors<br>attitude of investors<br>convinced shrewd investors |
| 4 | industry | fragmented industry<br>tourism industry<br>cottage industry<br>offshore-wind industry<br>telecom industry<br>car industry<br>energy industry | chip industry analysis<br>ache inducing industry<br>high return industry<br>tourism industry council<br>private equity industry |
| 5 | sales | online sales<br>arms sales<br>declining sales<br>global sales<br>recurring sales<br>retail sales<br>direct sales | annual sales revenues<br>total retail sales<br>sales of services<br>retail unit sales |
| 6 | share | market share<br>revenue share<br>executive share<br>firm share<br>fair share<br>median share | firm's share prices<br>executive share prices<br>executive share award<br>win market share |
| 7 | corporate | corporate culture<br>corporate disater<br>corporate governance<br>corporate raiders<br>corporate tax<br>corporate crises | corporate climate resilience<br>corporate governance code<br>corporate governance reforms<br>corporate pension funds<br>corporate tax cut |
| 8 | pay | executive pay<br>ceo pay<br>pay structure<br>pay compensation<br>pay scheme<br>pay packet | pay of executive<br>final pay packet<br>senior executive pay<br>pay attention to |

| | | | |
|---|---|---|---|
| **9** | executive | senior executive<br>media executive<br>chief executive<br>industry executive<br>top executive<br>gold executive | global chief executive<br>senior executive pay<br>executive share awards<br>executive share options |
| **10** | costs | marginal cost<br>production cost<br>additional cost<br>low cost<br>financial cost<br>big cost | cost of electricity<br>low labour cost<br>high labour cost<br>push up cost |

*Table 2. The top 10 keywords with their 2-word clusters and 3-word cluster*

In terms of semantic analysis, collocations of keywords can denote some kinds of meanings that more or less correspond to the grammatical functions of the components and collocations structures as folllows:

- **Collocates to the left of the keywords can refer to industries or lines of business:**

In the BEN corpus, as seen from the table above, the collocates to the left of many keywords (such as *firm, investors, market)* tend to denote industries or lines of business with mostly Noun + Noun structures.

e.g. *research firm, energy firm, law firm, insurance market, tourism industry*

- **Collocates to the left of the keywords may refer to characteristics/qualities of subjects**

With Adjective + Noun formation, the collocates to the left of the keyword seem to denote characteristics of subjects since these collocates function as adjectives that stand to the left of the keywords.

e.g. *local firm, global sales, holding firms, fair share*

- **Collocates of keywords can convey metaphors**

In some cases of collocations, with MIP (a method for identifying metaphors in discourse) developed by Steen et al.(2007), "*bull market*" and "*gold executive*" could be identified as metaphors. As a matter of fact, a bull symbolizes strength and power and gold can signify luxury and high quality. This finding can be supported by many studies by McCloskey (1983) Krugman (1985), Biccieri (1988), Marshall (1920) and Charteris-Black (2000) who highly recognized the significannce and dominance of metaphors in business discourse.

- **Collocates to the left of the keywords can denote possessions**

e.g. *executive pay , executives' pay*

Apostrophes serve two chief grammatical functions as an apostrophe replaces letters that have been omitted in contractions (i.e., can't) and apostrophes refer to possession of the preceding nouns (Hacker, 2003). In the examples above, apostrophes form possessive cases of subjects. In possessive nouns, apostrophes precede -s (i.e., executive's) or follow -s in the case of plural nouns ending in -s (i.e., executives').

**Collocates to the left of the keywords can demonstrate methods of operation or activity**

With (present participle + noun or past participle + noun) strutures, collocates of keywords may demonstrate methods of operation or activity.

e.g. *combined market, recurring sales, techinvesting firm, holding firm, rising cost.*

The grammatical functions of present participles and past participles are highly confirmed by the vast majority of literature. To be more specific, present participles and past participles, which are in an attributive position before nouns, play the roles of adjectives that express states and describe of features the subject of the sentences. P*resent participle* is often used in order to express an active action whereas *past participle* is employed so as to express an passive action (Quirk et al., 1985; Declerck ,1991; Huddleston & Pullum et al., 2002 & De Smet (2010).

Regarding grammatical structures of collocations of keywords found in BEN corpus, the following summary classification could be recommended as seen in table 3:

| | **Structures of collocations** | **Examples in BEN corpus** |
|---|---|---|
| Form 1 | Adjective + Noun | e.g. *median share* |
| Form 2 | Adjective + Noun + Noun | e.g. *annual sales revenues* |
| Form 3 | Noun + Noun | e.g. *corporate tax* |
| Form 4 | Noun + 's + Noun | e.g. *executive's pay* |
| Form 5 | Noun + Noun + Noun | e.g. *corporate governance reforms* |
| Form 6 | Gerund + Noun | e.g. *recurring sales* |
| Form 7 | Past participle + Noun | e.g. *combined market share* |
| Form 8 | N + Gerund + N | e.g. *ache-inducing industry* |
| Form 9 | Noun + Hyphen+ Noun + Noun | E.g. *private-sector investor* |

| | | |
|---|---|---|
| Form 10 | Noun + preposition + Noun | e.g. *cost of electricity* |
| Form 11 | Verb + Noun + preposition | e.g. *pay attention to* |
| Form 12 | Verb + preposition + Noun | e.g. *push up cost* |

*Table 3. Summary of collocational patterns of keywords in BEN corpus*

## 6. Conclusion

In conclusion, it can be observed from BEN corpus that the the dominant proportion of the 100 most frequent keywords belong to nouns, with very small percentages devoted to adjectives and verbs. The top 100 keywords may refer to some semantic domains in the field of business and, the collocations of the top 10 keywords can, in the same way, denote some groups of meanings. Moreover, it is notable that there is a wide range of different grammatical structures of the 2-word and 3-word collocations of keywords. Above all, the findings from the study can become a source of teaching and learning materials of Business English, resources for linguists and researchers and references of stylistics in this particular genre.

## REFERENCES

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer software]. Retrieved from http: //www.laurenceanthony.net/software/antconc/.

Barnbrook, G., Mason, O., & Krishnamurthy, R. (2013*). Collocation: Applications and implications*. London, England: Palgrave Macmillan.

Bicchieri, Cristina. (1988). *Should a Scientist Abstain from Metaphor?* Cambridge: Cambridge University Press.

Charteris-Black, J. (2000). Metaphor and vocabulary teaching in ESP economics. *ESP Journal*, 19.

Chung, T.M. and Nation, P. 2004. Identifying technical vocabulary. *System*, *32*, 251-263.

Declerck, R. (1991). *A comprehensive descriptive grammar of English*. Tokyo: Kaitakusha.

Goh, G-Y. and Lee, S-W. 2008. A Corpus-based analysis of the language of English news in Korea. *The Journal of Studies in Language* 23.4, 601-619.

De Smet, Hendrik. 2010. English -ing-clauses and their problems: The structure of grammatical categories. *Linguistics 48*, 1153–93.

Granger, S. (2012). How to use foreign and second language learner corpora. In A. Mackey & S.M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 7-29). London, England: Basil Blackwell.

Hacker, D. (2003). *A pocket style manual (3rd ed.)*. Boston: Bedford/St. Martin's.

Huddleston, R. & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Khamphairoh, T. & Tangpijaikul, M. (2012).Collocations of keywords found in insurance research articles: a corpus-based analysis. *Humanities Journal 19*, 166-188.

Krugman, P. (1995) *Development, Geography and Economic theory*. London: Cambridge MA.
Lukac, M. (2015). Linguistic prescriptivism in letters to the editor. *Journal of Multilingual and Multicultural Development, 37(3)*, 321-333.

Mahlberg, M. (2013). *Corpus Stylistics and Dickens's Fiction*. London and New York: Routledge.
Marshall, A. (1920). *The Principle of Economics*. London: Macmillan and Co.,Ltd
McCarthy, M. 1990. *Vocabulary*. Oxford: Oxford University Press.

McCarthy, M. (1991). *Discourse analysis for language teacher*. Cambridge: Cambridge University Press.

McCloskey, D.N.(1983) *The Rhetoric of Economics*. Economic Literature XXI.

McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.
Nation, I.S.P. (1990). *Teaching and learning vocabulary*. Newbury House.

O'Keeffe, A., M. McCarthy, and R. Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Pojanapunya, P., & Watson Todd, R. (2016). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*. Retrieved from https://www.degruyter.com/view/j/cllt.ahead-of-print/ cllt-2015-0030/cllt-20150030.xml?format=INT.

Quirk, R., Greenbaum, S. , Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Scott, M. (1997). PC analysis of key words – and key key words. *System, 25*(2), 233-245.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam, the Netherlands: John Benjamins.
Scott, M. (2008). *Oxford Wordsmith Tools 5.0*. Liverpool: Lexical Analysis Software.
Shin, D. and Nation, P. (2008). Beyond single words: the most frequency collocations in spoken English. *ELT Journal, 62(4),* 339-348.
Steen, G. J., Cameron, L. J, Cienki, A. J., Crisp, P., Deignan, A. , Gibbs, R., Grady,J., Kövecses, Z., Low, G. D. & Semino, E. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol, 22*, 1-39.
Wilkins, D. A. *Linguistics in Language Teaching*. Cambridge: Cambridge University Press.