A Critical Discourse Analysis: The Representation of 'Homosexual', 'Lesbian', and 'Gay' Collocates Words in Cola and BNC Corpus

Restu Anggi Gustara

Applied Linguistics, Yogyakarta State University, Indonesia.

e-mail: ranggig18@gmail.com

Abstract

This is a Critical Discourse Analysis of the collocation of 'homosexual', 'lesbian', and 'gay' terms in the corpus data of Corpus of Contemporary American English (COCA) and British National Corpus (BNC). By conducting Halliday's theory, this study aims to find out the representation of three terms, 'homosexual', lesbian', and 'gay', also the ideology, from the collocation words. As a combined study between Critical Discourse Analysis and Corpus Linguistics, qualitative and quantitative data were used. By using corpus analysis as the method, the researcher analyzes the ideology based on the collected collocates words. The result of the analysis shows that 'homosexual', 'lesbian', and 'gay' has a linear relationship. Those three terms are used in a different area of public text, which is 'homosexual' is more acceptable in the academic term and 'lesbian' and 'gay' are mostly used in the non-academic term. Even though COCA and BNC show the different amount of their existence, they share the same collocation: *rights, relationship, lifestyle, identity, activist,* and *couple*.

Keywords

Critical Discourse Analysis, corpus, homosexual, lesbian and gay

1 Introduction

The social practice has become one of interesting thing in the area of discourse analysis. This area has a big contribution as a useful framework, especially in the study of language and ideology relationship. Even though, as fundamental notions, discourse is not that easy to be defined (Van Dijk, 2009). It is because the discourse itself has become a multidimensional social phenomenon. One of that multidimensional phenomenon is the study between discourse analysis and linguistics. This linguistics part can be micro or macro analysis. In linguistics, there is also exist the method in collecting data, corpus linguistics.

In Partington et al (2004), the combined study between corpus linguistics and discourse analysis has known as Corpus-Assisted Discourse Studies (CADS). It is about quantitative and qualitative data in one study. By combining these two data, it is possible for the study to get the advantages of both quantitative and qualitative data. As the corpus-based, in its work, CADS is ready to reveal the hidden meaning

Citation: Gustara, R. A.: A Critical Discourse Analysis: The Representation of 'Homosexual', 'Lesbian' and 'Gay' Collocates Words in Cola and BNC Corpus. In. D. Ekawati, et al (eds.): Proceeding of The American Studies International Conference 2018, Vol. 2. pp. 123–132. UGM Digital Press Social Sciences and Humanities, Yogyakarta (2019).

Published: May, 2019

https://doi.org/10.29037/digitalpress.42269

in a word. in its study in ideology, it is already known that discourse analysis is a qualitative and a description of phenomena. The use of corpus linguistics can be said as a method to help discourse in explaining the ideology of the text. The marriage between discourse and Halliday's linguistics theory makes this combination become acceptable. CADS has a big contribution in analyzing discourse but it is not as critical as Corpus Linguistics Critical Discourse Analysis (CLCDA), even though both are known as an approach in discourse analysis and work in the computerized corpora in their analysis. As in Baker & McEnery (2015), it is said that in analyzing data, CADS has less critical than CLCDA.

In this study, the researcher adopts Halliday's view of collocation to the analysis of lexical. The focus is on terms which specifically consider the collocation words in two corpora, COCA and BNC. It examines the relationship between three terms, that are 'homosexual', 'gay', and 'lesbian'. Those three terms actually share the same area of phenomena, LGBT (Lesbian, Gay, Bisexual, Transgender). The use of these three terms sometimes unclear, consider the meaning of them are quite the same. In Cambridge: Advanced Learner's Dictionary, 'homosexual' is defined as a "*person, especially a man, who is sexually attracted to people with the same sex and not to people of the opposite sex*" (Cambridge: Advanced Learner's Dictionary, 3rd edition). Whether lesbian is a "*woman who is sexually attracted to other women*" (Cambridge: Advanced Learner's Dictionary, 3rd edition). The interesting part of this study is the definition of 'gay' in its dictionary. Gay is defined as a "*homosexual person, especially a man*". This means that the term 'gay' and 'homosexual' has the relational meaning. Actually, those three terms work as the same area, but in defining them, 'lesbian' has a different form of it. 'lesbian' is not defined as a homosexual, but only defined as a *woman who…*, not a homosexual. In its definition, both in 'homosexual' and 'gay' terms, always emphasized the idea of '*especially a man*'. This makes the phenomena become the main problem in this study.

This paper brings corpus as a methodology to manipulate the data. The corpora that are contributed in this research are Corpus of Contemporary American English (COCA) and British National Corpus (BNC). The corpus COCA was latest updated in December 2017. There are 20 million words added from 2016 and 2017. It is the largest English corpus that is accessible and free to use. The data sources in this corpus are divided into five sections, there are spoken, fiction, popular magazines, newspapers, and academic texts. It consists of 560 million words by 20 million words each year until now. The second corpora are BNC. The British National Corpus is a corpus that comes from Oxford University Press. It is formulated in 1980 and contains 100 million words which consist of spoken, fiction, newspapers, magazines and academic sections. The data from COCA and BNC contribute as the participants of this research. The procedures are outlined in some steps, there are:

Step 1: designing the questions of the research
Step 2: Compiling corpus (in this research, this part is not really necessary)
Step 3: Investigating the subject of this research, by looking at its history or culture of this topic.
Step 4: finding the appropriate corpus to this research
Step 5: Listing the keyword, making the frequency, and examining the frequencies to the topic
Step 6: Generating the collocation list
Step 7: making an analysis
Step 8: Reading and studying the chosen sites that are related to the term
Step 9: checking the relation between the findings and the theoretical framework
Step 10: refining the question of the research
Step 11: making related further corpus analysis (if any)
(Baker et al., 2008; Hardt-Mautner, 2009; Partington, 2008 as cited in Haider, 2017)

While the collocation of words has already observed in many studies, this present study offers the different subject of the investigation. It is still a lexical but the different area will complete the newest of this study. the researcher tries to find out the ideology inside those three terms (homosexual, lesbian, and gay). Besides, the researcher also tries to find out the relationship between 'homosexual', 'lesbian', and 'gay' by the collocation words around them, also the area of their existence.

2 Theoretical Framework

Discourse is a multidimensional phenomenon standing individually among the other branch of language study. As stated by van Dijk (2009) that defining a discourse cannot be done yet. The multidimensional itself which can make discourse has so many thought and perspectives. One says that discourse is produced as a result of an act of communication and it is general in the language (Richards et al., 1992, p. 111). The result of an act of communication can be words, phrase, and sentences, but discourse comes to a larger unit such as paragraphs, conversations, even interviews. Conversations and interviews happen every single day.

Those are must be used in person to communicate with others. That is why discourse also can be said as 'language in use' (Brown & Yule, 1983, p. 1). This idea reflects the thought of one of the popular paradigms, that is a functionalist paradigm. This paradigm never leaves the point of purpose and function of language in human life in analyzing language. While the purpose and function of language in human life always connected with culture and society. Those are well organized in the way of speaking. For discourse in this paradigm, language is used to defining something and to doing something (Richardson, 2007, p. 24; emphasize in original). That is why in interpreting texts, the researcher needs to understand what the speakers do and how they are connected with larger socio-cultural, interpersonal, and material contexts.

The next paradigm, structuralist paradigm, interpret discourse as a language over the clause (Stubbs, 1983, p. 1). The focus of this paradigm must be on the structure of language, that is organization and cohesion. But still, even though this paradigm seeing discourse in structure, the structuralist do not leave a little seeing of social ideas in producing of usage and interpretation of language.

Those two ideas of discourse more or less give some sights about the position of discourse in language study. So it can be said that discourse analysis is something that works in the surface of discourse. When we go deeper, it can be more critical than analyzing the use of language as a paragraph or conversation in cohesion area. The more critical analysis of discourse is known as Critical Discourse Analysis. It comes with ideology and power stored in text. For Richardson (2007, p. 1; emphasis in original), critical discourse analysis is both theory and method in analyzing the language use that used by individuals and institutions. As power and ideology mentioned above, critical discourse analysis or CDA usually works on politics or social phenomena that has sensitive issues. Different from the other branches of language study, CDA walks in a different way to analyzing and theorizing the concept in the whole field.

There are some principals come from Fairclough and Wodak (1997, p. 271-280). The eight principals are:

1. Critical discourse analysis examining the social and cultural processes and structures in linguistics as addressed as social problems.
2. The power relations in discourse are not proprietary. It can be said that power relations are negotiable. It used language as a control in negotiating.
3. Society and culture stand by the discourse does not only stand for social relations but also a part of that relations which produces them.
4. In discourse, ideologies are the most produced term. It includes how society is represented and constructed.
5. Discourse cannot be separated from intertextual relations and sociocultural knowledge.
6. There is a complex connection between social and cultural structures and process and properties of texts produced by critical discourse analysis.
7. Critical discourse analysis can be both interpretative and explanatory and the growth are influenced by the new contextual sources.
8. The social action in critical discourse analysis can be said as an influencer in the movement of communicative and socio-political practices.

Words are automatically included as a discourse. The word itself has its own works in discourse analysis. In the structuralist paradigm, words play in some ways, one of them is collocation. The term collocation is a part of cohesion system that in line with repetition, synonymy, hyponymy, and meronymy. Collocation is the form of the enhancing lexical in the system of cohesion that proposed by Halliday (2004). It highlights the relation of the semantic. As related to the semantic, the term collocation is related to meaning, as it mentioned in Firth (1957, p. 194) as a "meaning by collocation". Collocations words sometimes used together in a text (McCarthy & O'dell, --). In connection with discourse analysis, a word can be the point of power and also an ideology that comes out from the discourse. It can be the sensitive issues spread in social phenomena.

https://doi.org/10.29037/digitalpress.42269

3 Results

3.1 Corpus of Contemporary American English (COCA)

In the Corpus of Contemporary American English (COCA), the term 'gay' becomes the most frequent word with the total 28.129 words in the corpus. The graphics below show the result of the finding of those three terms 'homosexual', 'gay', and 'lesbian':

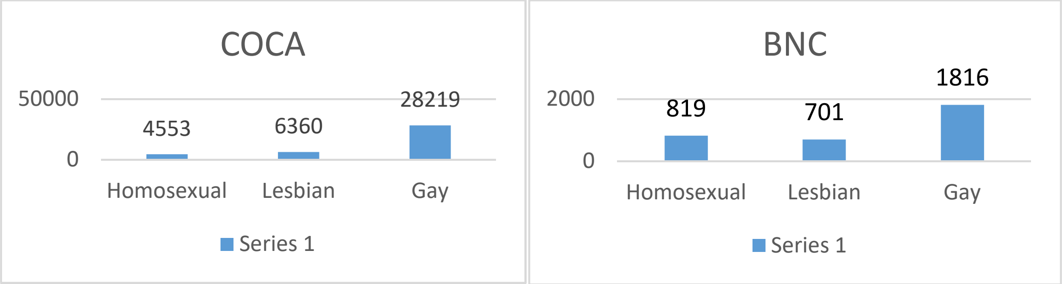

Fig. 1 Homosexual, Lesbian, and Gay in COCA and BNC

Based on the graphic 3.1 on the left, gay is the most frequent words with 28.129 words, then followed by the term 'lesbian' with 6.360 words and homosexual with 4.553 words. Different from COCA, BNC (graphic on the right) shows the result with the data less than 2000 words. There are significant differences between 'homosexual', 'lesbian' and 'gay'.

A very significant difference between 'homosexual', 'lesbian', and 'gay' whether in COCA and BNC are described more specifically in the graphic below:

3.1.1 'Homosexual' in COCA

There are spoken (1016 words), fiction (240 words), magazine (743 words), newspaper (756 words) and academic (1798 words) source in COCA. The most dominant appearance of 'homosexual' is in the academic source. The total of 'homosexual' term in the academic source is 1798 from 4.553 total words of 'homosexual'. This appearance is followed by many collocations. There are five main collocations (without considering the punctuation and preposition) that follow the term 'homosexual', there are: heterosexual, acts, bisexual, men, and behavior (table academic collocation). There is a sentence with one of main collocations in the data result:

> ...to refuse medical treatment, 246 and the right to engage in private, consensual homosexual activity.247 # Unless the Court intends to overrule all of these decisions, it is…
> ... The scale items originated from open-ended interviews regarding sexual attitudes with heterosexual and homosexual couples and have been shown to have excellent reliability and validity (Catania, McDermott…
> ...human rights abuses, including a law signed in October that imposes life imprisonment for homosexual acts. Njie attended the University of Texas, according to his LinkedIn account, …

3.1.2 'lesbian' in COCA

The data result shows the most appearance of the term 'lesbian' in five sources and the academic source become the most appearance source of 'lesbian' with 1881 words from total 6360 'lesbian' words. There are 1311 words in Spoken, 686 in Fiction, 1022 in Magazine, and 1460 in Newspaper. The appearance of the term 'lesbian' has five main collocations, there are gay, bisexual, women, transgender, and youth. Some of them are in the sentence below:

> ...that she has a monopoly on compassion for women's rights, gay rights, lesbian rights, but she took all this money from countries like Saudi Arabia where they…
> ...suicide attempts among high school students - and an even greater reduction among gay, lesbian and bisexual... # As one of the first records of human music, infant-directed…

> … I had as much diversity as possible. If there are five white women with lesbian mothers, I'll maybe just include two of them. # What do you…

3.1.3 'gay' in COCA

The data result shows that the spoken data become the most appearance source of 'gay' word with 8001 words, while the others such as Fiction gets 2100 words, Magazine gets 5112 words, Newspaper gets 7821 words and Academic gets 5095 words. This is different from the last two terms, 'homosexual' and 'lesbian' which academic source become the most appearance source. In the data sources, the word 'gay' collocates with five main collocation words, there are marriage, rights, lesbian, men, and community. Some of them are:

> …think the AIDS epidemic helped people, including gay people, realize the importance of gay marriage? CLEVE-JONES# Absolutely. When I was young, marriage was just, you…
> …But - so how did going to Quaker meetings lead to your discovery of the gay rights movement? CLEVE-JONES# The Quakers were very welcoming. And even back then, …
> … the things that he thought were really negative and wrong like the prohibitions against being gay or lesbian because he came out and he wasn't going to deal with that…

3.2 Corpus of British National Corpus (BNC)

The next details show the result of 'homosexual', 'lesbian', and 'gay' finding in another corpus, that is BNC.

3.2.1 'homosexual' in BNC

The term 'homosexual' in BNC mostly appear in academic source with the amount of 345 from the total 816 words found. In academic source, there are five main collocations that follow this term; men, behavior, among, HIV and infection. The sentences below are the example:

> …whose image has been marred by scandals. Lambeth Council is trying to encourage more homosexual men and women to adopt children or become foster parents. A plan approved by…
> …. The data examined in this study give a consistent picture of behavior changes among homosexual men which have led to a continuance of HIV transmission in this group in England…
> …. This group differs from other groups at risk of infection with HIV, like homosexual men and men with hemophilia, as their lifestyle often causes a high background incidence…

3.2.2 'lesbian' in BNC

The term 'lesbian' has the most appearance in the non-academic source. Five main words that collocate with 'lesbian' are gay, rights, community, black, and workers. The sentence below is an example:

> …Even in the so-called' permissive' 1960s, there were no gay switchboards or lesbian lines, and Chad Varah made it part of his mission to encourage self-acceptance in…
> …served to reinforce my own feelings that we in the UK must fight to preserve lesbian and gay rights which were so hard won and which Mrs. Thatcher seems determined to…
> …focusses upon incest and child abuse in a black family, and portrays a black lesbian relationship as tender and liberating. Through anger, self-confidence, love and eventual forgiveness…

3.2.3 'gay' in BNC

The most appeared 'gay' word is in the non-academic source with 752 words. The five main collocations are Lesbian, men, lesbians, rights, and community.

> …on education was being spent elsewhere; for example, on supporting lesbian or' gay' groups. The old notion of partnership between central and local government had long…
> …-- notably poor, black men on welfare in the US since Reagan, and gay men in the era of AIDS. She does not idealize those groups, nor…
> …and gay film-makers are as capable of producing' negative' images of lesbians and gay men as their heterosexual counterparts. Some would argue that they be actively encouraged to…

COCA also shows the graphic of the use of those three terms 'homosexual', 'lesbian', and 'gay' in 1990 until 2017.

https://doi.org/10.29037/digitalpress.42269

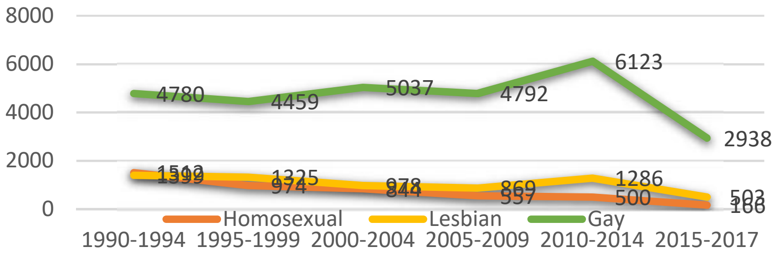

Fig. 2

In 1990 until 1995, the word 'gay' and 'homosexual' experience a decreasing moment. The increasing and the decreasing of the use of both terms are slightly the same even though the amount of them is not the same. Different graphic shows in the 'lesbian' term. There is no significant increase or decrease in the use of this term, but there is a little increasing usage in 2010 until 2014.

1. **Similar collocation**

The term 'homosexual', 'lesbian' and 'gay' become the keyword of this analysis because they are experiencing a semantic overlap. It is represented by some the same collocative words from those three terms. In COCA, the common collocates they shared are *rights, relationship, lifestyle, identity, activist,* and *couple*.

2. **Pinned collocation**

The unique collocates refer to the term/collocation which has different word choice from the others. The table below shows the collection of collocation words in each term:

Table 1 Pinned Collocation

Term	COCA	BNC
Homosexual	Abortion, rape, victim, priest, panic, disordered, morally, pornography	Infection, infected, tendencies, rape, diagnosed, sinful, injecting, passive, disease, discrimination, prayer, aids
Lesbian	Catholics, discrimination, Christian, defenders	London, film, politics, feminism, Jewish, policies, supporting, anti-discrimination, disabilities
Gay	Abortion, liberation	Breeze, culture, equality, beggar, Catholics

3.3 Area of collocation

The collocate words indicate some area of the word such as life, people or place that associated with the word itself. There are some domains exist in this finding both COCA and BNC:

Table 2 Area of collocation

Domain	Collocates word
Job	Soldier, lecturers, singers, writer, workers, film-maker, businessmen
Religion	Sinful, Catholics, Christians, prayer, Jewish
Medical	AIDS, disordered, genital, virgin, HIV, infection, HIV-1, drug, patients, infected, diagnosed, injecting, disease, psychologist, disabilities

Criminal	Sodomy, abortion, rape, priest, panic, scandal, crimes, policies, beggar
Human Rights	Marriage, rights, activist, victim, copyright, feminist, identity, discrimination, advocates, feminism, liberation, equality, campaign, supporting, anti-discrimination
Social-self	Exclusively, immoral, morally, stigma, morality, affair, radical
	Force, panic, desire, feelings, protection, defenders, cowards

4 Discussion

The finding revealed the 'more than expected' data from both corpora, COCA and BNC. The three terms 'homosexual', 'lesbian', and 'gay' are decreasing in year by year. Some of the public texts are no longer mention those three terms, even though in 2010 until 2014 there are in both 'lesbian' and 'gay' had a little increasing.

In COCA, both 'homosexual' and 'lesbian' have the most frequent appearing source in the academic source. It indicates that there are many scholars that do some research about homosexual and lesbian. Those two terms become the most investigated than the term 'gay'. The term 'gay' itself has the most frequent appearing source in the spoken data. It already becomes a big issue since the decretal of the latest ex-president of America, Barrack Obama in legalizing LGBT in their country. This term may be the talks or even debate by almost all the TV shows in the world. The term 'gay' also already exist in soap opera and movies. By the most appearance in the spoken data, it means that the term 'gay' is more familiar in society rather than 'homosexual' and 'lesbian'. Otherwise, in an academic source, the term 'gay' does not in the investigated theme. There are some relationships between the amount of 'gay' term in spoken data and newspaper and academic source. The term that exists in the newspaper usually indicates that that term is popular, but when there is a little amount in academic, this term has become a negative issue that cannot be investigated at that time. It can be said that investigating this term is harmful or risky.

In BNC, the data findings show the contrasting differences between those three terms. 'homosexual' term experiences many discussions in the academic source, whether 'lesbian' and 'gay' receive a big amount of appearance in the non-academic source. It indicates that the term 'homosexual' is more academic rather than 'lesbian' and 'gay' words.

In the collocation area, there are some parts that have a big contribution to interpreting the term 'homosexual', 'lesbian', and 'gay' by looking at the collocation. Those parts are first the collected collocation data, second is the similar collocation data, and then pinned collocation and the last is an area of the collocation. Those parts are set manually by looking at the semantic profiles proposed by Orpin (2005).

In the collection of collocation data, most of them are collocating each other. It can be said that those three terms explaining each other and have relational meaning. It means that the word 'homosexual', 'lesbian' and 'gay' do not have a collocational relationship with each other. By table 3.1 whether in COCA or BNC, each of them included in their own box, 'homosexual' collocates with 'lesbian' and 'gay' and vice versa. in Merriam Webster, homosexual is defined as 'tendency to direct sexual desire to persons in the same sex', it can be lesbian or gay. But in reality, 'homosexual', 'lesbian', and 'gay' are in the linear relationship. It is also proved in the similar collocation they share. 'homosexual', 'lesbian', and 'gay' shared some similar words that indicate events or common ground of those three terms. The shared collocation word such as *rights, identity,* and *activist* indicate the fight of them for the legality of their identity. The want of legality means that 'homosexual', 'lesbian', and 'gay' are different from others.

The area of the collocation words of 'homosexual', 'lesbian', and 'gay' represent where those three terms usually discussed. The term 'homosexual' mostly appears in medical term. It collocates with the word AIDS, HIV, disease, infection, and drug. It indicates that 'homosexual' is the term that way more negative or more medical than the others. It also shows that 'homosexual' term is something that negative because it related to something threatened, panic, priest, and victim. It is related to table 3.5 and graphic 3.3, 3.6 that 'homosexual' mostly discussed in the academic area rather than in the fiction or magazine.

There is something unique in the collection of collocating words in 'lesbian'. The word 'feminism' appear in the search result. The appearance of the word feminism indicates that the 'lesbian' term is whether in the positive or negative vibes for feminism. Related to the other collocation, such as rights, activist and discrimination, this term is included as the positive vibes for lesbian rights. The term 'gay' is mostly collocates with something about the legality of their identity. The words such as liberation and activist show that 'gay' has a bigger contribution to their rights rather than 'lesbian'. For discourse analysis, it may be enough to find out the relationship and the surface ideology between 'homosexual', 'lesbian', and 'gay' term, but the limitedness of the application that the researcher has and the formula to formulate the data seem not to fulfill the entity of corpus analysis.

5 Conclusions

This study demonstrated the collaboration between Critical Discourse Analysis as the qualitative data with Corpus Analysis as the quantitative data as a tool. The study found that there was a linear relationship between the term ‘homosexual’, ‘lesbian’, and ‘gay’. The term ‘lesbian’, ‘homosexual’, and ‘gay’ are not collocating each other because their relationship is vice versa. From those three terms, the word ‘homosexual’ is the more applicable for every academic and medical issue, whether in daily life, ‘lesbian’ and ‘gay’ are more acceptable.

The suggestion was also made for the further research to go deeper and wider for this investigation that this may be better served by collocation formulas in collecting the quantitative data of ‘homosexual’, ‘lesbian’, and ‘gay’ term.

References

1. Baker, P., & McEnery, T. (2015). A Corpus-Based Approach to Discourses of Refuges and Asylum Seekers in UN and Newspaper Texts. *Journal of Language and Politics*, 4(2), 197-197
2. Brown, G. and Yule, G. 1983. Discourse Analysis. Cambridge/London/New York: Cambridge University Press. p. 1
3. Cambridge University Press. (2008). Cambridge Advanced Learner’s Dictionary, 3rd Edition. Retrieved on December 22, 2017
4. Fairclough, N., and Wodak, R. (1997). Critical Discourse Analysis. In T.A. van Dijk (ed.). Discourse as Social Interaction. London: Sage. (pp. 271-280)
5. Firth, J. (1957). *Paper in Linguistics, 1934-1951*. Oxford: Oxford University Press
6. Haider, A.S. (2017). Using corpus linguistic Techniques in critical discourse studies: Some comments on the combination. Department of Linguistics. The University of Canterbury.
7. Halliday, M. (2004). *An Introduction to Functional Grammar 3rd (Revised by Matthiessen, C.M.I.M) ed.).* London: Hodder Arnold.
8. http://corpus.byu.edu/ Retrieved on December 20, 2017
9. McCarthy, M, & O’dell, F. (--). *English Collocation in Use*.: Cambridge: Cambridge University Press
10. Orpin, D. (2005). Examining the Ideology of Sleaze. *International Journal of Corpus Linguistics*, vol 10:1, pp 37-61
11. Partington, A., Morley, J., & Haarman, L. (2004). *Corpora and Discourse*. Bern: Peter Lang.
12. Richards, J.C., Platt, J., and Platt, H. (1992). *Longman Dictionary of Language Teaching and Applied Linguistics* (2nd ed). Harlow, Essex: Longman.
13. Richardson, John E. (2007) *Analyzing Newspapers: An Approach from Critical Discourse Analysis*. London: Palgrave
14. Stubbs, M. (1983*). Discourse Analysis: the sociolinguistic analysis of natural language*. Oxford: Basil Blackwell.
15. Van Dijk, T. A. (2009). “Critical discourse studies: A sociocognitive approach.” In R. Wodak & M. Meyer (Eds.). Methods of critical discourse analysis (Vol. 2, pp. 62-86)
16. Van Dijk, T.A. (1993). “Principles of Critical Discourse Analysis”. *Discourse and Society* 4:249-83

https://doi.org/10.29037/digitalpress.42269