Journal Homepage: -www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI:10.21474/IJAR01/3158
DOI URL: http://dx.doi.org/10.21474/IJAR01/3158
ISSN NO. 2320-5407

RESEARCH ARTICLE

USING ANTCONC: A CORPUS-BASED TOOL, TO INVESTIGATE AND ANALYSE THE KEYWORDS IN DICKENS' NOVEL 'A TALE OF TWO CITIES'.

Mustafa Khalid Saleh Al-Rawi[1,2].

1. MA English Language and Applied Linguistics (ELAL).
2. Assistant Lecturer at the department of English Language, Cihan University – Sulaimanyah.

Manuscript Info

Manuscript History

Received: 14 December 2016
Final Accepted: 18 January 2017
Published: February 2017

Key words:-
Corpus Linguistics, AntConc tools, keywords, keynessand aboutness, and text analysis

Abstract

Keywords analysis is one of the important methods in corpus linguistics. It has the privilege to analyse texts/corpora in terms of statistical significance, by comparing two or more texts/corpora. This study regarded the keywords that are notable statistically in Dickens' 'A tale of two cities' using Laurence Anthony's AntConc tools. It aimed to find out a number of keywords of the mentioned Victorian novel (the node) compared with 35 Victorian novels (as a reference corpus). This study briefly shows and discusses a number of keywords of this novel as well as showing the limitations of this method of analysis. Quantitative and qualitative approaches are used; the quantitative analysis is by Laurence Anthony's AntConc. concordancer whereas the qualitative analysis would be individual. Log-likelihood method, significant keywords as well as a number of negative keywords are selected and analysed.

Corresponding Author:-Mustafa Khalid Saleh Al-Rawi.
Address:-MA English Language and Applied Linguistics (ELAL).

Introduction: -

The term "Keyword" is one of the research methods used in corpus linguistics which means: "words that are claimed to have a special status, either because they express important evaluative social meanings or because they play a special role in text or text - type" (Stubbs, 2010: 21). From this definition, keywords are the units that carry or participate in discovering the meaning of a text, this is from the perspective of Linguistics; from the social perspective, they are the units that have partial reference the culture and society, and also would represent the text type. There is an issue regarding the relationship between these two perspectives. Although in keywords analysis analysts correlate semantics with social aspects, the meaning of a text's keywords (semantics) tends not to be strongly related to the social and cultural world (Ibid, 2010). Words are the smallest units of meanings, but their meaning could be affected by a number of reasons: the combination with other words, the intention and purpose of saying/writing these words, as well as the participants who say/write them; and in social world, all these reasons are taken into consideration, therefore the social analysis is not strongly related to the individual semantic meaning of these keywords.

Stubbs (2010) argued that corpus linguistics studies are predominantly weak in social theory in the contrary of speech act theory – which means: regarding an utterance as a type of action which performs a function in language (*The Oxford English Dictionary*) - which provides 'powerful' social theory. However, it might be more sufficient to use empirical research of real data in studying the social theory. According to Stubbs (2010), there are three concepts

of keywords: the first concept is originated from 'cultural studies', the second is originated from 'lexico-grammar work', and finally, the concept that derived from the 'comparative quantitative corpus analysis'. This essay focuses on the third concept of keywords which means searching for words that are notable statistically in a text or a number of texts by using Laurence Anthony's AntConc tools. More specifically, this essay focuses on the keywords of Charles Dickens' novel 'A tale of two cities' (taken as a node corpus), compared with 35 Victorian novels (as a reference corpus). It is interesting to find out the keywords of this novel to know what is different in its language, and what could they reveal in terms of understanding it. This essay briefly shows and discusses a number of keywords of this novel as well as showing the limitations of this method of analysis.

Keyness: -

Keyness is "a quality possessed by words, words clusters, phrases etc., a quality which is not language -dependent but text -dependent" (Scott, 2010: 43). Therefore, words or keywords are not stand individually as a quality, but they are considered to be a quality in regard to the whole text or maybe in a number of texts, with also regard to social and cultural aspects. Because keywords are statistically significant in a text/texts (in comparison with usually a larger corpus), they tend to indicate the text's 'aboutness'- simply, what the text is about- as well as the text's style (Scott, 2010; Groom, 2010). On the one hand, keywords appear to provide an understanding about the whole or maybe just part of a text/ texts. On the other hand, these keywords would to a great extent participate in revealing aspects of styles of specific text/ texts which are different in comparison with the styles of others. As Scott (2010) pointed out, the use of the word 'key' is metaphorical. From this sense, key is what allow an individual to access a position which has not accessed before, or as he put it "enabling device" (2010: 44) i.e. as mentioned previously, they (keys/keywords) tend to provide an understanding of a text/texts.

Finding out the keyness of words in a text/ texts individually is not an easy process; the use of an automatic analysis (keywords tools) solves this issue. However, it has been claimed that the results of individual versus automatic analysis are not the same, for instance: Scott (2010) claimed that because keywords tools process analysis differently, their results are different (from that is of individual) quantitatively and qualitatively. It is not usually the case, sometimes the results could be the same even if the process is different and it is worth mentioning that because these two processes are different, in a sense, they are not comparable; as both of them has its advantages and disadvantages. The main differences in these both are: individuals seem not to agree consistently on the keywords of specific text/ texts picked by automatic tools, and automatic tools would provide keywords that are difficult for individual to notice (Ibid, 2010). The typical way of analysis is taking both processes together, using automatic tools to provide quantitative analysis and individual analysis using qualitative approach. However, there are a number of limitations regarding both approaches for example: the statistical process in comparing two or more texts is not always accurate, it tends to have a number of problematic issues considering the choice of keywords; taking, for example, a word from a text which is considered as a key in comparison to a reference corpus, but this word might tell nothing about the aboutness or the style of a text i.e. it might not be a key in this specific text.

This inaccuracy problem could be avoided by choosing the right reference corpus (Scott, and Tribble, 2006). In terms of qualitative analysis, the problem is not accuracy but it could, to an extent, be regarded as subjectivity. Individuals do have differences in perceptions, so, their analysis seems to be bias towards their perspectives. Avoiding this issue could happen only by perceiving the surroundings in objective ways.

Choosing the right reference corpus: -

Choosing a reference corpus is an important issue. It affects all the statistical processes and results. The most important issues taken into consideration when choosing a reference are: the size of the corpus and its content. The larger the reference corpus is taken the more and accurate keywords detection (Scott, and Tribble, 2006). Moderate reference corpus may be sufficient as Scott (2010) argues, Sardinha (2004 cited in Scott, and Tribble, 2006) argued that the reference corpus should be five times larger than the node one and if it is more than that, it would be more accurate. Logically speaking, if the reference corpus is slightly larger than the node one, then there would be no significance in keywords and it is not a rule regarding the size but the more significant results wanted, the larger corpus should be used. In terms of content, there must be a relation between the node and the reference corpora. It would not be appropriate, for example, using a medical corpus as a reference to analyse the keywords of a literary work. It can be done, but the results would be inaccurate. Therefore, using 'genre specific' is a key aspect in finding the accurate results. In addition, using period specific which means the node and the reference corpora are originated in the same period of time, because social aspects change through time and it tends to have an influence on language; of course unless if the study is about the change of language through time or the effects of social aspects on

language, both of these are exceptions from the period specific principle. In this study the node text is a novel in the Victorian era 'A tale of two cities' and the reference corpus is 35 Victorian novels, so, the two issues (size and content) and time are taken into consideration.

Open-class versus closed-class keywords: -

Open-class words are the content words which allow an addition of new words, whereas closed-class ones are the functional words which do not allow any additions and it contains the grammatical and functional words. In keyword analysis, it is generally considered by discourse analysts that open-class keywords represent the aboutness of a specific test/texts or corpus, while closed-class words represent the style of a specific test/texts or corpus (Groom,2010; Scott, and Tribble, 2006). However, analysing the meaning of a corpus depends on the sequences of words not on the form of these words (Groom, 2010). In addition, Sinclair (1991 cited in Groom, 2010: 62) mentioned that "[m]ost everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text". Therefore, analysing text/texts, using keywords, does not depend on the meaning of the words themselves, but on their meaning within the context. There is a slight distinction between closed-class and open-class keywords (Scott, and Tribble, 2006). Therefore, Groom (2010) argued that this distinction is regarded as a reason in skipping the analysis of closed-class keywords. It is possible to say that this issue is not a rule of thumb; it significantly depends on the text/texts or corpus. In a type of text, for instance, one can find the use of functional words (closed-class) is a representative of the style as well as the aboutness of this text but in other texts s/he may find these words represent just the style and not the aboutness. However, closed-class keywords would not be treated as the same as open-class keywords in analysing the aboutness, open-class keywords are always the prime indicators of aboutness (Groom, 2010; Baker, 2006).

There are two different methods of analysis: the quantitative and the qualitative, along with the distinction between aboutness and style. For closed-class keywords, quantitatively, they tend to have a statistical significance as they are grammatical and functional words; but, qualitatively, "these words can tell us almost nothing about the meanings and values expressed in a specialized corpus" (Groom, 2010: 63). Although these words are considered meaningless in terms of analysis, they appear to have the ability to reveal an indication of the meaning of a specific text/texts or corpus as well as the indication of the style; not when they are analysed individually, but when they are analysed contextually (Ibid, 2010). For open-class keywords, quantitatively, they, statistically, seem to be slightly lesser than the closed-class ones except for the proper nouns which always come first. Qualitatively, those keywords are the representative or the principal of meanings. Proper nouns tend to be also meaningless while analysing, except for a number of studies which require the analysis of them. Groom (2010: 70) points out that, considering closed-class keywords "irrelevant" or less important would be "wrong". It is possible to say that only a number of closed-class keywords could be 'tractable' to semantic (meaning) analysis, for example: the words 'the' and 'a', generally, provide almost nothing to indicate the aboutness. However, the word 'of' in Groom's (2010) study provided plenty of indication to the aboutness. In this study, both classes of keywords would be concentrated on to see whether the closed-class keywords in this novel provide indications of aboutness or not.

Bottom-up versus Top-down approaches of analysis of discourse: -

Bottom-up and Top-down approaches are contrasted. Simply, top-down approach is used when there is already an existed hypothesis or 'framework' of analysis and the reason of analysis is to support or prove the effectiveness of this framework. Whereas bottom-up approach is used when an analyst starts his/her analysis with the data (corpus/text(s) in order to reach or create a suitable framework of analysis (Biber, et al. 2007). Typically, the case of keywords analysis tends to be a bottom-up approach. However, it could be possible and effective to work with these two approaches in keywords analysis. In literary work analysis, which is the concern of this paper, 'corpus stylistics' appears to be the comprehensible approach of analysis. Corpus stylistics is the cooperation of methods of corpus linguistics and literary stylistics – where corpus linguistics is the bottom-up and literary stylistics is the top-down (Mahlberg, 2010; 2012). It is possible to claim that the relation between these approaches is similar to the relation between quantitative results and qualitative analysis, in a way that both of these relations could address difficulties and limitations of keywords analysis.

From a logical perspective, on the one hand, the bottom-up approach of analysis seems to be more difficult and subjective while working with alone and there is a possibility of wrong or odd findings. Therefore, taking the framework (top-down approach) – in this case, the literary stylistic methods – would be a guidance mostly towards significant findings. On the other hand, taking the top-down approach alone would be a time consuming; as Short (1996 cited in Mahlberg, 2010: 295) put it "analysing a long novel … could take a lifetime". Moreover, there are plenty of other limitations which would be mentioned later. To avoid such limitations, this essay would take these both approaches into consideration. In terms of top-down approach, this novel is on the time of French revolution, and the story is taking place in two cities, London and Paris. It has a sub story and the concentration is mainly on the characters of this story. It has a psychological effect on the readers because of the use of different techniques such as: 'metaphors', 'comic relief' (humour), 'repetition' …etc. (Newlin, 1998). The language of this novel is highly stylistic and philosophical; moreover, it has a subjective element, by which it differs according to the characters' perspectives (the view points of the characters are different because of the class distinction) (Ibid, 1998). The understanding of this brief summary would provide guidance and an idea about the analysis of keywords

Methodology: -

The aim of this study is to find out a number of keywords of the Victorian novel 'A tale of two cities' and what these keywords could tell about this novel. Quantitative and qualitative approaches are used; the quantitative analysis is by Laurence Anthony's AntConcconcordancer whereas the qualitative analysis would be individual. Top-down and bottom-up approaches of analysis are used in this analysis to provide clearer insights and provide possible hypotheses, but the focus would be more on bottom-up approach. The node text (A tale of two cities) was downloaded from the Project Gutenberg website whereas the reference corpus was taken from one of the corpus linguistics lectures, which is also downloaded from the Project Gutenberg website. The reference corpus consists of 35 Victorian novels. By using AntConc, Log-likelihood method, significant keywords as well as a number of negative keywords are selected and analysed.

KeywordsAnalysis: -

Table 1:- the significant keywords in this study

Rank	Freq	Keyness	Keyword
1	223	649.786	doctor
2	138	592.878	prisoner
3	63	362.907	citizen
4	60	297.528	spy
5	2011	268.943	his
6	76	212.373	streets
7	622	163.047	mr
8	134	137.103	business
9	20	131.850	patriots
10	22	115.954	tribunal
11	13	99.646	citizeness
12	54	87.528	faces
13	233	86.668	miss
14	98	84.531	until
15	20	75.729	wot
16	221	63.054	its
17	10	62.737	wos
18	4999	59.761	and

Table 2:- the negative keywords in this study

Rank	Freq	Keyness	N. Keyword
1	18	240.434	mrs
2	1987	158.423	i
3	1045	137.901	her
4	27	65.414	oh

The word 'doctor', other than its normal role, emphasizes the role of social class, the well educated people and also there is a logical tendency with the term doctor as whenever there is an emphasis on doctors there must be patients, it might be a representative of the ill situation at that time i.e. sick people as a sign of not only physical sickness but the poverty and depression of maybe most of the lower or working class at that timeThe word 'prisoner' in this novel is not surprisingly being a keyword as the story deals with prison and court. However, it also has a metaphoric use as in 'must positively find the prisoner Guilty' and in 'innocent prisoner' these two examples are not only metaphoric, but also a representation of injustice. Moreover, the word 'spy' used in this novel as also a representative of the world of injustice.

As citizen means "someone who has the right to live permanently in a particular country and has the right to the legal and social benefits of that country …" (Macmillan dictionary). The use of this word in this context has two senses: the first one is the sense of formality, this word has been mentioned once in the first book of the novel and the rest 62 times were mentioned at the third book because the word occurred in the context of court which tends to be more formal than any other situation. It also carries positive aspects (most collocates are positive semantically). The second sense, this word is used metaphorically as good quality content must be existed in people (it is the 'container' conceptually; whoever supports the French revolution is called citizen/ess).

The use of the possessive pronoun 'his' is an indication of the major theme of this novel; it seems to be a male dominance. Although it does not deal with only one hero and it is not only a one story novel, it seems that the focus is on one hero of the sub story. For more support, it is obvious that the use of female possessive pronoun 'her' (which is negative keyword in this novel) is only half times than 'his' which also indicates the male dominance theme.

The word 'streets' is significant in this novel, but it tells a little about the aboutness, simply as the actions are taken place in the streets. However, "but, when the streets grew hot" and "once-peaceful streets" in these examples 'streets' has another meaning: metonymic way of representing weather and personifying the impersonal.

The word "Mr." is mentioned 622 times and it tells nothing but when comparing with the negative keyword "Mrs" which is mentioned only 16 times, one can assume that this novel is male dominance, though the word "miss" is significant (mentioned 233 times).

The word "business" used in different situations providing different meanings for instance: 'business eye' metaphoric use, 'man of business' used in two senses: men who have jobs and men who are in charge, 'business mind' means reasonable, 'it's not my business' means not my responsibility. In addition, this word provides positive sense, however there are some instances of the negative use of this word as in 'dreadful/murderous business".

Although the word 'patriots' is mentioned only twenty times, it is significant in this novel. The instances are seen from the second half of book two and book three, when the revolution started, because of the fact that it appears to

be semantically and logically linked with the revolution.

The word 'until' is used as a preposition as well as conjunction. The conjunction 'until' is used more that the preposition one, and it might be more interesting in the sense that a certain action is depending on another action, so, this would tell a little about the language of this novel.

The word 'wot' is the non-standard form of the word 'what' and it is significant in this novel. Although it is not frequent enough, it shows the colloquial language used by people at that time. And also shows the subjective element mentioned above, which is not all characters use this word instead of 'what'; 'what' was used 384 times while 'wot' was used 20 times.

It is not unusual for the word 'and' to be a keyword; it is the additive element, so, it might be possible to say that this novel has a significant number of extensions. As for the possessive 'its', it is mostly preceded by the prepositions 'in, of' and only in small number of instances it is preceded by 'on, at, for'.

The word 'wos' is used ten times and it is significant in this novel for the same reason of the word 'wot'. It is the colloquial term of the word 'was', but in seven instances it was used as: 'if it wos so' and two times without 'so', 'if it wos' and only one time it was used alone.

The word 'faces' is significant in this novel; it represents the normal use of the word and: the metonymic concept of 'part represents the whole', so, face represents the whole body as in:

'and faces hardened in the furnaces of suffering'. It represents the status of a person as in:
'the faces changed, from faces of pride to faces of anger and pain'.

The only aspect of the significance of the word 'tribunal' in this novel may lie in the element of formality i.e. the court tends to be more formal than the tribunal and the fact that at the times of French revolution, the revolutionary court was oppressive and informal. Therefore, this could be the reason of using this word instead of the word court.

The negative keywords 'I' and 'oh' can tell that the use of the first speaker pronoun is not significant and it could be because the focus is on the indirect speech. As for 'oh' a representative of sigh and grief, it is unusual to be negative in this novel (as it is sorrowful and melancholic).

Finally, the keywords used in this analysis provided a number of elements regarding the aboutness and the style of this novel and provided an understanding on the surface level. In order to widening this research to gain an understanding on a deeper level, there are a number of limitations will be mentioned later. If a researcher takes them all into consideration, the results would be deeper and more generalizable. In Addition, the closed-class keywords in this novel appear to have nothing about the aboutness.

* The examples are taken from AntConc directly and the whole text is referenced in the reference list.

Limitations and further study: -

There are a number of limitations in the study of keywords analysis other than the ones mentioned above. It is difficult to make 'generalizations' in a wide range of texts because it depends on the linguistic features of different registers and on the writer's style (in this case Charles Dickens whose style can be generalized in almost all his texts and that could provide clues to the analysis of a specific one(text) i.e. the style of a text tends to be subjective and the effect on readers/listeners may also be subjective – in a sense that text could affect readers/listeners differently (Mahlberg, 2012). A word could be a key that frequently occurs in one part of a text/corpus as in the case of the word 'citizens' in this novel. The keywords do not focus on grammatical, semantic or functional differences or even lexical similarities; they concentrate only on lexical differences (Baker, 2004a cited in Rayson, 2012). It is also possible that the selection of keywords could be subjective and that would to an extent frame the analysis in certain perspective. Mahlberg (2005: 17) pointed out that, generally, "corpuslinguistics is still a long way away from the creation of a unifying theory that accommodates individual findings within a broad framework".

For further study, it would be beneficial to include collocation analysis in researching keywords because perhaps it would provide a deeper analysis of keywords. Also using Scott's technique to find 'key keywords' would be

beneficial (Rayson, 2012) – key keywords are words that are keys in not only one text/corpus but in a large number of texts/corpora (Scott, and Tribble, 1996). This technique would support Groom's claim in focusing on the closed-class words (because key keywords frequently are from the closed-class group) as well as the phrasal/cluster analysis.

Conclusion: -

In conclusion, keywords analysis is one of the important methods in corpus linguistics. It has the privilege to analyse texts/corpora in terms of statistical significance, by comparing two or more texts/corpora i.e. analysing the texts/corpora using the most frequent words. Theoretically, this method would be the most significant method of analysis in corpus linguistics, because it is an inclusive approach as it focuses at the text/corpora as a whole and selects the most frequent words for the analysis. However, texts/corpora are not just language, there is always aspects other than language (Social, cultural and cognitive aspects) which influence the language. Inevitably, these aspects affect the context of texts/corpora. With the fact that the meanings of words in texts are different from their meanings in isolation, keywords might to an extent be influenced by those aspects. Therefore, using multi-approaches (Top-down/Bottom-up) would result more accuracy in terms of analysis. Both of these approaches support the analysis of this novel as well as the choice of the reference corpus. Although the findings are not generalizable because the number of keywords analysed is not enough, there are a number of indications about the aboutness and style of the language used and this brief analysis reveals a number of aspects that support the understanding of this text.

Reference list: -

1. Baker, P. (2006) *Using corpora in discourse analysis*, London: Continuum.
2. Biber, D., Connor, U. and Upton, T. A. (2007) *Discourse on the move: using corpus analysis to describe discourse structure*, Amsterdam: John Benjamins.
3. Groom, N. (2010) *Closed-class keywords and corpus-driven discourse analysis*, in M. Bondi, and M. Scott, (eds.) *Keyness in texts*, Amsterdam: John Benjamins.
4. *Macmillan Dictionary*http://www.macmillandictionary.com/.
5. Mahlberg, M. (2005) *English general nouns: a corpus theoretical approach*, Amsterdam: John Benjamins.
6. Mahlberg, M. (2010) *Corpus Linguistics and the Study of Nineteenth-Century Fiction*: *Journal of Victorian Culture*, 15(2), pp. 292-298.
7. Mahlberg, M. (2012) *Corpus Analysis of Literary texts: The encyclopaedia of AppliedLinguistics*.http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0249/full accessedon the 20th of March 2016.
8. Newlin, G. (1998) *Understanding A tale of two cities: a student casebook to issues, sources, and historical documents*, United States of America: Greenwood Press.
9. Rayson, P. (2012) *Corpus Analysis of Key Words: The encyclopaedia of AppliedLinguistics*.http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0247/full accessedon the 20th of March 2016.
10. Scott, M. and Tribble, C. (2006) *Textual patterns: key words and corpus analysis in language education*, Amsterdam: John Benjamins.
11. Scott, M. (2010) *Problems in investigating keyness, or cleaning the undergrowth andmarking out trails*, in M. Bondi, and M. Scott, (eds.) *Keyness in texts*, Amsterdam:John Benjamins.
12. Stubbs, M. (2010) *Three concepts of keyness*, in M. Bondi, and M. Scott, (eds.) *Keyness in texts*, Amsterdam: John Benjamins.
13. *The Oxford English Dictionary*. 2nd ed. (2005) App.
14. *The Project Gutenberg EBook of A Tale of Two Cities, by Charles Dickens*http://www.gutenberg.org/