



# Corpus Stylistics: Style and Corpora Applications

Khalil A. A. Abdulqader

Ph.D. Scholar, Department of linguistics, AMU, Aligarh, 202002, India

Mohammed Yassin Mohd Aba Sha'ar

Ph.D. Scholar, Department of English, AMU, Aligarh, 202002, India

Ahmed Hassan Ali Murshed

Ph.D. Scholar, Department of English, AMU, Aligarh, 202002, India

Abdulqawi Hamood Alyasery

Ph.D. Scholar, Department of English, AMU, Aligarh, 202002, India

## Abstract

To begin with, the term corpus stylistics is one branch of stylistics that combines methods and principles from both corpus linguistics and literary stylistics. It brings the methods of corpus linguistics to the practice of stylistics and the study of literary texts. More particularly, it refers to the statistical study of style, i.e. study of the relative frequency of elements in the text. This paper will define and differentiate between corpus stylistics and corpus linguistics. In fact, stylistics claims that it provides linguistic substantiation for interpretation of literary texts. But as corpus analysis is par excellence a means of exposing textual characteristics in accurate detail, it seems logical to assume that it must be interrelated to the stylistic project. Significantly, corpus stylistics is applicable in capturing the essence of a particular author, studying the frequency of lexical elements, author attribution, interpretative stylistics and language learning. This paper explicates the way in which corpus stylistics is applied in research. It extensively will examine its value and contribution in the area of stylistics. The main contention of this paper is to delineate corpus stylistics methodology and its applications in language learning, corpus linguistics and literary studies.

**Keywords:** Corpus Linguistics, Corpus Stylistics, Corpora, Stylistics, Language Learning

## Introduction

To begin with, the term corpus refers to a collection of texts that form a body and are stored in an electronic database. Essentially, corpus stylistics is a branch of computational linguistics which has begun developing in the late 1960s. It significantly helps in looking for particular characteristics (words, sentences and the frequency of items) in a particular text. Meanwhile, corpus linguistics is sometimes utilized interchangeably with corpus stylistics. However the difference between them is that corpus stylistics utilizes qualitative applications and tools to analyze texts with the help of computational methods. They both contribute in understanding literary texts or any other select data. However, this paper shows how corpus stylistics takes place using different methods and tools. It expounds the way in which we can identify the linguistic items and constructions.

In fact, corpus stylistic analysis depends mainly on qualitative processes and result in objective results which linguists and stylisticians look for. Therefore, corpus stylistic analysis assists in dealing with the very long texts, identifying and disclosing many features that may not be revealed in the manual analysis.

Corpus stylistic analyses employ different computational methodologies to investigate keywords, collocations, concordances, semantic domains, multidimensional analysis, n-grams (lexical bundles) etc, by using software toolkits such as Wmatrix3, Wordsmith, AntConc, etc. These software tools work as analysis systems to analyze literary texts. The dominant area of corpus stylistics concerns the lexis of literary works. As corpus stylistics has matured, some critics have used it along with other approaches and theories. Corpus stylistics concerned with the syntax of literary texts remains quite rare. In fact, corpus stylistics is a methodology that is capable of different applications and involves a number of different methodologies, linked through the use of computers and quantitative analysis but quite different in their aims, practice and results.

### **What is a Corpus?**

The term “Corpus” is mostly used to refer to a relatively large collection of naturally-occurring texts, which have been stored in machine-readable form (McEnery & Hardie, 2012). In this form, the texts are then studied using various computer programs, in the branch of linguistics known as “corpus linguistics”. A corpus can also be defined as a systematic collection of naturally-occurring texts (of both written and spoken language). “Systematic” means that the structure and contents of the corpus follows certain extralinguistic principles (‘sampling principles’ i.e. principles on the basis of which the texts included were chosen). For example, a corpus is often restricted to certain text types, to one or several varieties of English, and to a certain time span. If several subcategories (e.g. several text types, varieties etc.) are represented in a corpus, these are often represented by the same amount of text. “Systematic” also means that information on the exact composition of the corpus is available to the researcher (including the number of words in each category and in the whole corpus, how the texts included in the corpus were sampled etc.).

Although “corpus” can refer to any systematic text collection, it is commonly used in a narrower sense today, and is often only used to refer to systematic text Collections that have been computerized. The word “Corpus” has always been used by linguists to indicate a collection of naturally-occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape-recordings, which have been collected for linguistic study (Hunston, 2002). We can say that corpus is a body or collection of linguistic data for use in research. Criteria for defining a corpus are listed below:

- A corpus consists of authentic, naturally-occurring data;
- A corpus is collected according to explicit design criteria;
- A corpus is representative of a particular language or genre;
- A corpus is prepared for a specific linguistic or socio-pragmatic purpose.

### **Corpus Linguistics**

Corpus linguistics has been defined as “the empirical study of language relying on computer-assisted techniques to analyze large, principled databases of naturally-occurring language (Conrad, 2000:548)”. More recently McEnery and Hardie (2012) have Provided a definition of corpus linguistics with a more instrumental focus; for them, “Corpus linguistics is a field which deals with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions.”

Corpus linguistics is a method of carrying out linguistic analyses. As it can be used for the investigation of many kinds of linguistic questions and as it has been shown to have the potential to yield highly interesting, fundamental, and often surprising new insights about language, it has become one of the most widespread methods of linguistic investigation in recent years. In short, corpus linguistics refers to the analysis of naturally occurring language on the basis of computerized corpora. The analysis is usually performed with the help of the computer, i.e. with specialized software, and takes into account the frequency of the phenomena investigated. Corpus linguistics offers to stylistics a set of methods and theories for exploring large quantities of texts while stylistics offers to corpus linguistics a set of theories, models and analytical frameworks that may be of value in the interpretation of corpus data and corpus-derived results. Text types in corpus linguistics as follows: written texts, transcripts of spoken texts, spoken texts with sound files, signed texts, images and gestures.

As a modern approach to language studies, corpus linguistics has witnessed that the number and depth of many corpus approaches to the study and investigation of the English language is continuously increasing. Three different stages can be observed in the history of corpus linguistics. The first stage in corpus linguistics starting in the 1960s stressed on developing computerized general corpora consisting of different types of spoken and written texts. For example, the Brown corpus consists of 2000 word samples of 500 texts which are spread across 15 categories. The second stage of general corpora, which began in the 1980s, was developed in the 1990s, taking advantage of the technological advances of computing. It has produced mega-corpora such as the 450 million-word Bank of English Corpus, or the 100 million-word British National Corpus. Unlike many early corpora, they contain complete texts rather than sections of texts. The third wave starting in the 2000s focused on both developing Giga-corpora (i.e., corpora of texts over a billion word using Websites and Newswire text as data sources), and small specialized corpora designed for studies of Academic and professional English.

### **Corpus Stylistics**

Wales (1989) points out that corpus stylistics is a branch of computational linguistics. It is developed in the late 1960s and helps to examine certain characteristics of the data like words length and sentences based on statistical and computer aided instruments to study and investigate a number of issues related to style. It has become popular over the past decade, with Leech and Short (2007:286) noting “a corpus turn” in stylistics. It brings the methods of corpus linguistics to the practice of stylistics and the study of literary texts. In more specific, corpus stylistics is best used to describe a particular focus of corpus linguistics, one in which the research questions being investigated and/or the analytical frameworks used are primarily concerned with issues of style.

McIntyre (2015) states “Corpus stylistics is simply corpus linguistics with a different object of study (Literature as opposed to non-literary languages”. Moreover, he demonstrates that the difference between them is that corpus stylistics is not only borrowing tools from corpus linguistics, but it makes it unique by using qualitative tools and techniques of stylistics to analyze texts with the help of computational methods (Jaafar, 2017). Corpus stylistics helps to investigate certain features that cannot be understood without using the tools and techniques of computers. It also helps to analyze long texts more efficiently with the employment of quantitative and qualitative methodologies.

Corpus stylistics refers to the statistical study of style, i.e. study of the relative frequency of elements in a text. It helps in investigating the measurement of the length of the words and sentences used in texts and analyzing several authors' frequency distributions of word-length. In more specific, statements about the

relative frequency of various linguistic items in a corpus have become very accurate. Corpus stylistics is also used widely for authorship attribution and the dominant area of corpus stylistics concerns the lexis of literary texts. Leech and Short (2007) express some appreciation of what they call the “Corpus turn” in stylistics and reject worries that computer analysis is taking over stylistics. They demonstrate the ability to search rapidly and accurately for such elements as vocabulary, collocations and some simple syntactic categories but explain that more abstract levels of analysis require manual annotation of texts. They also observe that corpus stylistics is only statistically valid where quite large amounts of text are used.

Some important uses of statistical analysis of style through corpora are: Education, e.g. EFL textbook writing; Establishment of authorship, e.g. of unascribed manuscripts and Interpretive stylistics, e.g. study of the writer’s ideology, point of view, and keywords.

### **Corpus Stylistics Methodology**

Corpus stylistics is not a theory, but rather a methodology which is capable of different applications. It involves a number of different methodologies that are linked through the use of computers and quantitative analysis, but quite diverse both in their practice, aims and results. Methodology in corpus stylistics demonstrates that simple things may characterize different styles such as average of word length, average of sentence length, type: token ratio (vocabulary richness) in which number of types equals number of different words and number of tokens equals total number of words; vocabulary growth (homogeneity of text). The corpus methodology can contribute to the stylistic approach to literary study of texts and the corpus stylistics can offer new insights or new forms of research. The corpus stylistic research applies corpus methods to the analysis of literary texts, giving particular emphasis to the relationship between linguistic description and literary appreciation (Mahlberg, 2013).

More complex analysis can give a more interesting result such as specific syntactic structures, degree of modification in NPs, types of verbs (e.g. verbs of persuasion, speech verbs, action verbs, descriptive verbs), distribution of pronouns (1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup> person), etc. Quite developed mathematical techniques can give an overall picture, for example, factor analysis identifies from a (big) range of variables which ones best identify or characterize differences. The corpus stylistic methods that contribute in analyzing texts are:

### **Multidimensional Analysis**

Multidimensional analysis focuses on collecting a huge range of measures of a wide variety such as some simple word counts, syntactic features and classes and subclasses of N, V, Adj, Adv. It also helps in factor analysis like choosing a range of features to measure, see which ones are correlated. Biber et al. (1998) used statistical measures to identify stylistic factors that co-occurred, and could therefore be definitional of text types and genres, e.g. conjunctions like therefore, nevertheless and use of passive together indicate more formal style. Biber (1998) developed his method of multidimensional analysis. It is used to identify the linguistic characteristics of different spoken and written language registers.

### **Concordance**

Corpora is useful not only for counting frequencies of features, but also for Concordancing and collocation. Concordancing functions the following: it lists occurrences of word in context, it identifies syntactic use of word, identifies range of meanings and relative frequency of different uses or meanings. Concordance is also known as (KWIC) list (Key word in context). It allows us to see the immediate

environment in which a word appears. Listings can be customized to show what you want more clearly e.g. sorted according to next or previous word, showing more or less context, etc. This screenshot given below shows concordances of the word ( Russia ) in the novel ‘Animal Farm’ by George Orwell.

Screenshot (1) Seven examples of successive concordances of the word ‘Russia’

Hit	KWIC
1	was a kind of blasphemy. Events in Russia and events elsewhere were to be judged
2	the Soviet régime. I have never visited Russia and my knowledge of it consists only
3	hodoxy is an uncritical admiration of Soviet Russia. Everyone knows this, nearly everyone acts on
4	means much more than that admiration for Russia happens to be fashionable at this moment.
5	idea of Socialism as the belief that Russia is a Socialist country and that every
6	them. In Spain as well as in Russia the nature of the accusations (namely, conspiracy
7	dictators, that it can apply only to Russia, to the Animal Farm: A Fairy Story -5-

From the above screenshot, it is obvious from the right context that the successive concordances of the word ‘Russia’ has an important significance. The word ‘Russia’ occurs (7) times indicating the importance of the theme of Russian revolution in the novel. The concordances or the occurrences of this word in the text lead the reader to this interpretation.

### Collocation

Collocation shows what words occur together and compares distribution of close synonyms vocabulary in context. It is a term coined by Firth (1957) to characterize (part of) his theory of meaning “You shall judge a word by the company it keeps”. Sinclair (1991) defines collocation as “the occurrence of two or more words within a short space of each other in a text. The relationship a lexical item has with items that appear with greater than random probability in its (textual) context (Hoey, 1991). Collocation tries to distinguish between general and more usual collocations vs. technical and more personal ones, e.g. in a general corpus time collocates with save, waste, spend, fritter away, etc., but in a sports corpus reports time collocates with half, full, injury, extra, first, second, third, etc.

### Wmatrix3

Wmatrix3 is a software tool that works as an analysis system to analyze texts. It also works as a concordance tool that can calculate keywords, a key part of speech tags and key semantic categories in a text. The tool has both simple and advanced interfaces. The simple one consists of four functions. These include finding a list of words with their frequencies, seeking for a particular word, word clouds that allow seeing the significance of the words which are illustrated by the font size in the tool. The word clouds are represented in different sizes of large fonts and small fonts. The larger items are more significant and have more importance. While the advanced interface has other different tags such as the semantic and part of speech tags. In short, Wmatrix3 is a tool that is used to conduct the analysis of texts by examining keywords and key semantic domains. It is necessary for new users or researchers who are interested in using Wmatrix3 to conduct a corpus stylistic study to know its major basics.

## AntConc

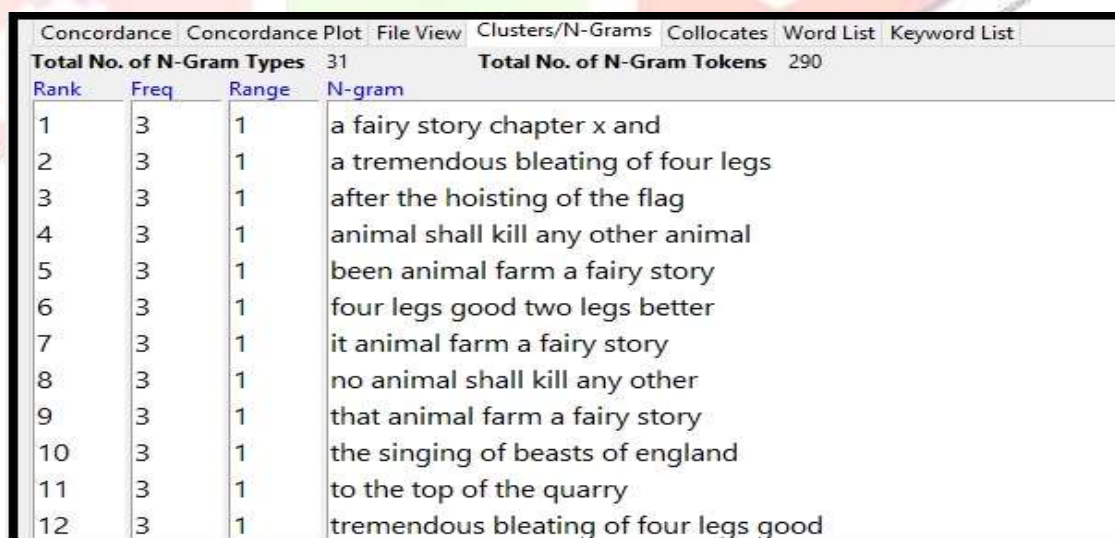
AntConc is a free software tool for extracting keywords, concordances and n-grams designed by Anthony (2011). It helps to identify and count long clusters that facilitate the qualitative analysis which then enables to a better understanding of the concerned text. In short, AntConc is a tool that investigates n-grams in texts and indicates their significance to the overall interpretation. AntConc as a tool for achieving the study has the following functions:

- 1) It examines keywords of the literary text. These keywords can help in guiding readers to the meaning of the text. They are regarded as the basis of the world text or the fictional world.
- 2) It provides the connection between the resulted keyword and key semantic domain analysis.
- 3) It looks for lengthy n-grams, and attempts to identify their significance, value and their connections to the whole meaning.

### Clusters or n-grams

Clusters (n-grams or lexical bundles) are regarded as repeated sequences of words. There are two kinds of clusters: short clusters and long clusters. Short clusters such as 2-grams, 3-grams and 4-grams while long clusters such as 8-grams, 7-grams, 6-grams and 5-grams. When the number of n-grams is lengthy, their frequency usually goes down (Mahlberg, 2013). The tool ‘AntConc’ reveals the number of grams in texts. This screenshot given below shows n-grams (6-grams) in George Orwell’s Animal Farm.

Screenshot (2) the Retrieval of 6-grams



Concordance		Concordance Plot		File View		Clusters/N-Grams		Collocates		Word List		Keyword List											
Total No. of N-Gram Types						31						Total No. of N-Gram Tokens						290					
Rank	Freq	Range	N-gram																				
1	3	1	a fairy story chapter x and																				
2	3	1	a tremendous bleating of four legs																				
3	3	1	after the hoisting of the flag																				
4	3	1	animal shall kill any other animal																				
5	3	1	been animal farm a fairy story																				
6	3	1	four legs good two legs better																				
7	3	1	it animal farm a fairy story																				
8	3	1	no animal shall kill any other																				
9	3	1	that animal farm a fairy story																				
10	3	1	the singing of beasts of england																				
11	3	1	to the top of the quarry																				
12	3	1	tremendous bleating of four legs good																				

As seen in the above screenshot, we observe six-word clusters with the frequency of occurrences of three times throughout the whole novel. These repeated sequences of words are vital to foster the interpretation of the novel.

The second example of screenshot exhibits n-grams (5-grams) in George Orwell’s Animal Farm.

Screenshot (3) the Retrieval of 5-grams

Concordance		Concordance Plot	File View	Clusters/N-Grams	Collocates	Word List	Keyword List	
Total No. of N-Gram Types			56	Total No. of N-Gram Tokens				507
Rank	Freq	Range	N-gram					
1	3	1	a tremendous bleating of four					
2	3	1	after the hoisting of the					
3	3	1	animal shall kill any other					
4	3	1	as soon as they were					
5	3	1	been animal farm a fairy					
6	3	1	but at this moment the					
7	3	1	fairy story chapter x and					
8	3	1	from side to side and					
9	3	1	in the middle of the					
10	3	1	it animal farm a fairy					
11	3	1	it did not seem strange					
12	3	1	legs good two legs better					
13	3	1	no animal shall drink alcohol					
14	3	1	no animal shall kill any					
15	3	1	on the other side of					
16	3	1	shall kill any other animal					
17	3	1	skipping from side to side					
18	3	1	that animal farm a fairy					
19	3	1	the animals on the farm					

Apparently and in allusion to the above screenshot, we observe five-word clusters with the frequency of occurrences of three times throughout the whole novel. These clusters of five words contribute a significant role in the interpretation and understanding the text. They lead in a direct way to the basic ideas and themes in the text. Moreover, they reveal something about the style of George Orwell in the novel 'Animal Farm'.

## Style and Corpora Applications

### Stylometry

Stylometry or stylometrics attempts to capture the essence of a particular author's style by reference to a variety of quantitative criteria, usually lexical, called discriminators. It studies frequently occurring features such as word/sentence length; choice and frequency of words and vocabulary richness. It also deals with the ideal situation for authorship studies, e.g., when there are large amounts of undisputed text or few contenders for the authorship of the disputed text(s). Burrows (1997) defines 'Style' as whatever marks the distinct identity of an author or a school, a set of loosely consistent features. These features include phenomena that can be divided and counted and hence may be captured by stylometric techniques. These techniques typically focus on the most common words. Hoover (2007:176) states that "Stylometric techniques assume that word frequencies are largely outside the author's conscious control because they result from habits that are stable enough to create a verbal fingerprint".

### Author Attribution

Corpus stylistics has been widely to tackle problems of authorship. It contributes in establishing the author of an unascrbed manuscript. Rudman (2006) states "when used for authorship identification, corpus methodology is based on the hypothesis that every writer has a unique and verifiable style." Burrows (2007) observes that evidence of authorship pervades whatever anybody writes provided

appropriate procedures are employed in the analysis of an appropriate set of texts, it can almost always be elicited. Stylometry also finds application in forensic authorship analysis, which looks for legal evidence for authorship attribution (smith, 1989).

## Language Learning

A simple way to look at lexis is to compute the frequency with which certain words and items appear. Frequency had a role in language learning in the days before electronic corpora existed. In fact, the 'Corpus revolution' made frequency information available regarding to language use in a totally unprecedented way. Dictionaries of frequency and frequency-based grammatical information are becoming more and more available and new sources of frequency information from the web are being tapped. Various types of knowledge exist in recent language textbooks (collocational, semantic and grammatical) are getting to be frequency based. Generally speaking, corpora represents real usage of language. Besides, "more frequent" means and equals more important" in many aspects of language learning. Language learning concerns with the following subjects: word frequency, real usage, translation, grammar, teaching collocations and decoding.

## Interpretive Stylistics

In interpretive stylistics, programs like tools of Wordsmith and Wmatix3, and other windows-based applications allow researchers to derive a list of keywords (words which occur significantly more often than expected in texts when compared to a corpus of reference). Keywords are a powerful and quick technique of analysis. It has used to examine discourses relating to specific Social and cultural issues and the ideology behind authors or texts.

## Benefits of Corpus Analyses

- 1) It deals with relatively objective data in huge amounts.
- 2) It achieves descriptive research method.
- 3) It supports FL learning with high proficiency.
- 4) It grants insight into morphosyntactic patterns.
- 5) It provides insights into processes of language change.
- 6) It grants insight into underlying discourses.

## Conclusion

To sum up, corpus stylistics is one branch of stylistics and computational linguistics which uses and employs methods of corpus linguistics to the practice of stylistics in investigating and studying literary texts. Researchers depend highly on computers to access and gain information about certain data. Corpus stylistics employs computational methodologies and software toolkits to study the literary texts. It examines keywords, key semantic domains, frequency of items, collocations, concordances, multidimensional analyses, n-grams (lexical bundles), etc. It helps to find certain features that cannot be done or understood without using computer's techniques. Quantitative and qualitative methodologies are applied to analyze lengthy texts more efficiently. Corpus stylistics proves to be of great significance in conducting corpus stylistic analyses of literary texts. Methodologies of corpus stylistics work as perfect guides to investigate and understand elements of the text in a logical and systematic way. Both the corpus



and manual stylistic analyses work in glove, and they complement each other. Corpus stylistic analysis is conducted by using software tools such as Wmatrix3, Wordsmith, AntConc, etc. Wmatrix3 and Wordsmith are those two tools that are used to conduct the stylistic analysis by examining keywords, concordances, collocations, and key semantic domains. AntConc is a complementary tool that investigates and studies concordance and n-grams (lexical bundles) in literary texts and to point out their significance to the whole interpretation and to reveal something about the style of the literary text. Applications of corpora are stylometry, author's attribution, language learning and interpretive stylistics. Stylometry is the application that attempts to capture the essence of a particular author's style whereas in author attribution, corpus stylistics establishes the author of an unasccribed manuscript. In language learning, corpora are applied to look at and compute the frequency with which certain words and items appear. Interpretive stylistics applies corpora to allow researchers to derive a list of keywords, point of view and author's ideology in texts.

## References

1. Anthony, L. (2011) AntConc ( Version 3.2.4m)Computer Software. Tokyo, Japan: Waseda University. Web 5 August 2019, <http://antlab.sci.waseda.ac.jp/>
2. Biber, D., Conrad, S. and Reppen, R. (1998) Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press.
3. Burrows, J. (1997) Style. The Cambridge Companion to Jane Austen, Cambridge: Cambridge University Press.
4. Burrows, J. (2007) Textual Analysis. A companion to digital humanities. Oxford: Blackwell. 5. Firth, J., 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, Issue Special Volume of the Philological Society, pp. 1-32.
6. Hoey, M. (1991) Patterns of Lexis in Text. Oxford University Press.
7. Hoover, D. (2007) Corpus Stylistics, Stylometry and the styles of Henry James. Oxford: Blackwell.
8. Hunston, S. (2002) Corpora in Applied Linguistics. Cambridge: Cambridge University Press
9. Jaafar, E., A. (2017) Corpus Stylistic Analysis of Thomas Harris' The Silence of the Lambs, University of Baghdad, Khazar Journal of Humanities and Social Sciences, 20, 1, Baghdad, Iraq
- Leech, G. and Short, M. (2007) Style in Fiction. London and New York: Longman.
10. Leech, G. (2008) Language and Literature: Style and Foregrounding. London: Pearson Education.
11. Leech, G. & Short, M., 2007. Style in Fiction: A Linguistic Introduction to English Fictional Prose. 2nd ed. Harlow: Pearson Longman.
12. McIntyre, D. (2015) "Towards an Integrated Corpus Stylistics". University of Huddersfield, United Kingdom.
13. McEnery, T. & Hardie, A. (2012) Corpus Linguistics: Methods, Theory and Practice. Cambridge: CUP.
14. Mahlberg, M. (2013) Corpus Stylistics and Dickens' Fiction. London: Routledge. 15. Orwell, G. (1945) Animal Farm: A fairy Story text from [gutenberg.net.au](http://gutenberg.net.au) Prefaces from [orwell.ru](http://orwell.ru). Cover portrait from [k-1.com/Orwell](http://k-1.com/Orwell). (Eric Arthur Blair) Web 13 October 2018, [www.animal.farm/animal.farm](http://www.animal.farm/animal.farm) -pdf

- 16 .Rudman, D, L. (2006) Shaping the active, Autonomous and responsible modern retiree: An analysis of discursive technologies and their links with neo-liberal political rationality. United kingdom: Cambridge University Press.
- 17 .Sinclair, J. (1991) Reading Concordances. An Introduction. Harlow: Pearson.
- 18 .Smith, M. (1989) Forensic Stylometry: A theoretical basis for further developments of practical methods: Journal of the Forensic Science Society, 29(1), 5-33.
- 19 .Wales, K. (1989) A Dictionary of Stylistics: Routledge. London.

