Analysis of Gender-Marked Units: Statistical Approach

Anna Hadzalo [0000-0001-6812-1093]
Lviv Polytechnic National University, 12 Bandera str., Lviv, Ukraine, 79013
anna.gadzalo@gmail.com

Abstract. Gender-marked units always are in the interest of many linguists, as this type of figurative language units has a significant value when it comes to the verbalization of concepts. Gender differences in lexical units and their research makes a huge impact in the description of the male and female linguistic picture of the world. This article is an attempt to analyze the difference in the usage of these units by men and women. The modern statistical approach in linguistic research was used. The effectiveness of statistical methods has been investigated.

Keywords: gender-marked units, statistical research, corpus-based analysis.

1 Introduction

In modern multi-vector society, the concept of gender has received many interpretations. In the 1970s, the term ‘gender’ was defined as ‘sociocultural gender’, while ‘sex’ was understood as ‘biological sex’. American scientist Robert Jesse Stoller was one of the first psychiatrists at the congress to put forward the concept of distinguishing between these definitions [1]. It was he who interpreted ‘gender’ as a category denoting psychological, social, and cultural differences, while ‘sex’ represent a physiological (biological) difference that did not necessarily correlate with anatomical features. In addition, in *Sociology*, Anthony Giddens states that the difference between gender and sex is fundamental because many differences between men and women are of non-biological origin [2].

The interest in the study of gender aspects of language and speech has grown due to two phenomena. The first phenomenon is the tendency to analyze the social environment, which influences communication that is namely, the human factor in language. The second is the active development of the feminist movement and its direct influence on the actualization of gender-specific research. The concept of ‘gender’ is used to describe the grammatical category and the usage of grammatical forms. The main two questions that linguists are exploring are the following. The first one is how a man and a woman are reflected in a language, how people of different sex manifests in the language, what are the features attributed to men and women and in which semantic areas they are most prevalent, which linguistic mechanisms underpin these processes. The second question is the differences in speech and communicative behavior of different sexes in general.

2 Gender and Language

Analysis of the link between language and gender is a field of research in a large number of sciences, including applied linguistics, linguistic anthropology, cultural studies, feminist studies, psycholinguistics, sociolinguistics, and others. Linguistic genderology (or gender linguistics) is one of the relatively new multidisciplinary sciences that studies the linguistic manifestation of gender as a sociocultural construct.

Gender affiliation has a strong imprint on the language at all levels. The article *Language and the Place of Woman* by Robin Lakoff, published in 1972, states that the use of form and the way we express ourselves are governed by our perception of the real world [3]. Deborah Tannen, in the book *You Just Don't Understand Me*, describes how gender differences affect the style of communication and differ by the reasons, goals, rules, and ways of interpreting the conversation [4]. Daniel Maltz, Ruth Broker [5] and Pamela Fishman [6] devoted their work to the same research object. During communication, speakers independently create their gender through language, and for this reason, a wide variety of styles and ways of expression can often be observed.

The fundamental concept that served for the development of gender studies is the statement of Wilhelm von Humboldt that language categorizes the world only by its inherent system of thoughts and feelings [7]. The logical continuation of it lies in the linguistic relativity hypothesis. Edward Sapir and Benjamin Whorf assume that people speaking by different languages perceive the world differently and think differently [8].

In spite of the active development of gender-oriented research, it is important to note that the tactics of speech behavior are difficult to match to their specific implementation by men and women. Therefore, gender value is often hyperbolized and may lead to a misunderstanding of its role in communicative behavior [9].

3 Statistical Analysis in the Gender-Related Linguistic Research

The linguistic methodology for gender studies has many disadvantages. The main of which is the study of linguistic phenomena of only one level and a lack of consistency in the research such as collecting of factual basis, its heterogeneity, or the interpretation of the obtained results based on the small amount of test material. Elena Horoshko notes that the limited empirical material and the small number of conducted linguistic experiments can probably explain the high level of contradictory data in the field of gender linguistics [10: 266]. The adequacy of such conclusions could be in question. In light of this information, it is clear that gender linguistic still needs to improve already existing or new research methods. Olena Kharchenko mentions that depending on the purpose and subject of the research combination of methods is possible and even required [11]. It has to be done despite the established scientific tradition because only then the heuristic potential of the methods can be expanded. Olena Levchenko also notes that the results of gender studies in language are often unreasonable due to the lack of use of quantitative methods [12].

Statistical methods, as specifically targeted, are considered as one of the most effective and valid methods of researching gender-marked units in applied linguistics. Statistical methods are a set of techniques and principles according to which the collection, systematization, processing and interpretation of statistical data is carried out in order to obtain scientific and practical conclusions.

Statistical methods are widely used in modern linguistics. Among Ukrainian researchers, these methods are also popular. Valentyna Perebyinis [13], Maria Zayats [14], Svitlana Romanyuk [15], Hanna Sytar [16] analyzed the basics of statistics that may be useful for linguists and gave examples of their application. Viktor Levytskyi adds that statistical methods can be used for natural language processing, machine translation systems creation and teaching optimization [17]. The usage of these methods is relevant today due to the rapid development of information technology and the increasing computerization of society.

American linguist James Pennebaker in 2001 introduced text analysis program LIWC (Linguistic inquiry and word count), which later, Matthew Newman improved by creating a large corpus (over 500, 000 textfiles of different fields and styles) [18-19]. The researchers computed gender as an independent variable for the entire dataset. Argamon Shlomo conducted a similar experiment using the BNC corpus as a dataset [20]. However, the statistical approach is still poorly used by linguists in gender studies.

4 Data Collection and Analysis

In the modern sense, the corpus is always a computer database and in the process of its creation and research, there is doubtless the need for the use of special programs. In addition, linguistic software allows conducting research on a larger amount of material and with a greater degree of confidence in the objectivity of the received data. The effectiveness of linguistic software usage is unquestionable.

Corpus is a reliable material for statistical analysis. We consider using the corpus as the most appropriate for a selection of the factual research base since only its order and system can ensure the accuracy of quantitative data and the persuasive conclusions [16]. For our research, we gathered gender-marked units, ones that mention the gender of the referent [21: 44]. The tokens were proofread and encoding was checked, with the help of KWIC concordance for conjunctions '*as a boy', 'as a girl'* was created. The sample did not include the soi-disant surreal similes, as well as similes based on realities, such as (like she did, as my father, etc.).

The gender aspect of any discourse is primarily concerned with the grammatical category. This category is a distinctive feature of the grammatical structure in the Indo-European language family. However, word inflection in the English language consequently led to the loss of ancestral oppositions in the grammatical system of the English noun. The generic differences of the noun are revealed mostly by the personal pronouns (he, she) or by lexical units. Man, boy, masculine, husband, son, father, mister represent the male gender. Female gender can be represented by such nouns as woman, girl, feminine, wife, daughter, mother, lady, miss, etc.

5 Analysis Using Corpora Capabilities

For our research, we use the most reliable and widely used corpora British National Corpus and Contemporary Corpus of American English. BNC is considered as a model corpus because it is composed of 100 million words that have been gathered from the texts of 1980-1990. COCA is the most widely used corpus of English, it contains one billion words of text dated 1990-2019 and is perfectly balanced. BNC represents the texts of British English, while COCA represents American English. Both corpora represent a wide range of genres (even transcribed spoken, academic, fiction, etc.).

With the help of both corpora, we are researching the peculiarities of collocations '*as a girl', 'as a boy'.* One of the most interesting criteria in statistics is a relative frequency (empirical probability) of an event. Mathematically, it is an absolute frequency normalized by the total number of events. Used corpora are counting relative frequency automatically. The collected data demonstrates a quantitative preponderance of colocations in the COCA corpus. To depict the frequency distribution table was used (See Table 1).

Table 1. Relative frequency in collocations '*as a girl', 'as a boy'* in BNC and COCA

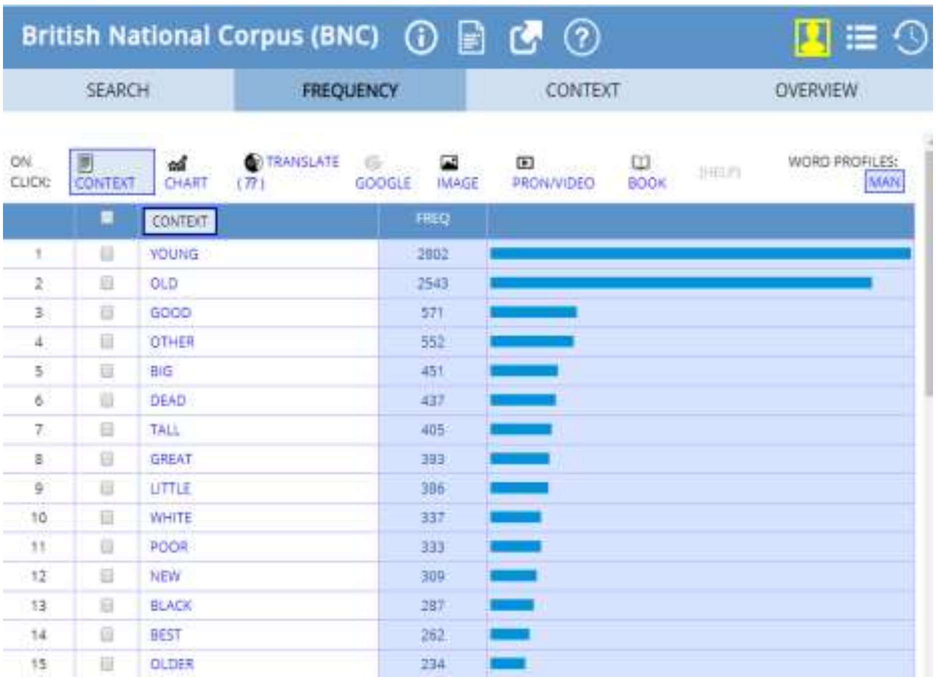
	Corpus Name	**Relative frequency (10^{-2})**
Collocation	**BNC**	**COCA**
as a girl	0,57	1,12
as a boy	2,25	2,62

The interesting results were extracted by using if the function *Collocates* which allows narrowing the search. The most frequently used collocations indicating girls' appearance and behavior (adj + as a girl) in both corpora are 'delicate', 'graceful', 'pretty', 'high', 'fine'. The most frequently used collocations indicating boys' are 'short', 'young/little/junior', 'strong'.

The *'man'* is more actively nominated and more clearly presented which testifies gender asymmetry. The unit is the largest and represents characteristics of a person in general (*a man of principle, man's sin, man's creation*) and features that are inherent in man *(to be a man, a man of muscles, man of courage).* The example of the instrument usage is depicted in Figure 1.

The transfer of stereotypical features forms the largest amount of similes found in both corpora. For example, by the way people of different genders dress, look or behave: *he cried like a girl, to utter as a girl, his voice is high as girl's, she laughs as a man, she shoots and fights as a boy, Alexander sobbed like a girl.*

Presented similes are a bright example of the gender stereotypes existing in language. Linguistic gender stereotypes include 'male' and 'female' features or expectations of certain characteristics and actions from representatives of a particular gender. It should be noted that, especially recently, gender stereotypes are based on a significant exaggeration of the differences between men and women.



	■	CONTEXT	FREQ
1	□	YOUNG	2802
2	□	OLD	2543
3	□	GOOD	571
4	□	OTHER	552
5	□	BIG	451
6	□	DEAD	437
7	□	TALL	405
8	□	GREAT	393
9	□	LITTLE	386
10	□	WHITE	337
11	□	POOR	333
12	□	NEW	309
13	□	BLACK	287
14	□	BEST	262
15	□	OLDER	234

Fig.1. The collocations *'boy'* + adj represented in BNC

To analyze the genre distribution of collocations '*as a girl'*, *'as a boy'* in corpora we can use the *Chart* function. Using this function was investigated that collocation '*as a boy'* is mostly used in fiction prose (647), magazines (356) and newspapers (334). The distribution by genres of collocation '*as a girl'* is the most productive in fiction (373), magazines and newspapers (189). The least productive genres are non-fiction and academic. In Figure 2.we can see the relative frequency distributed by genres.

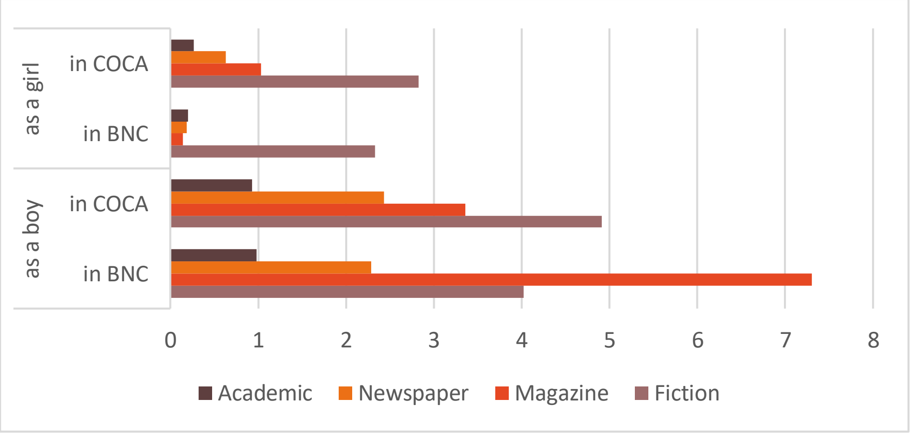
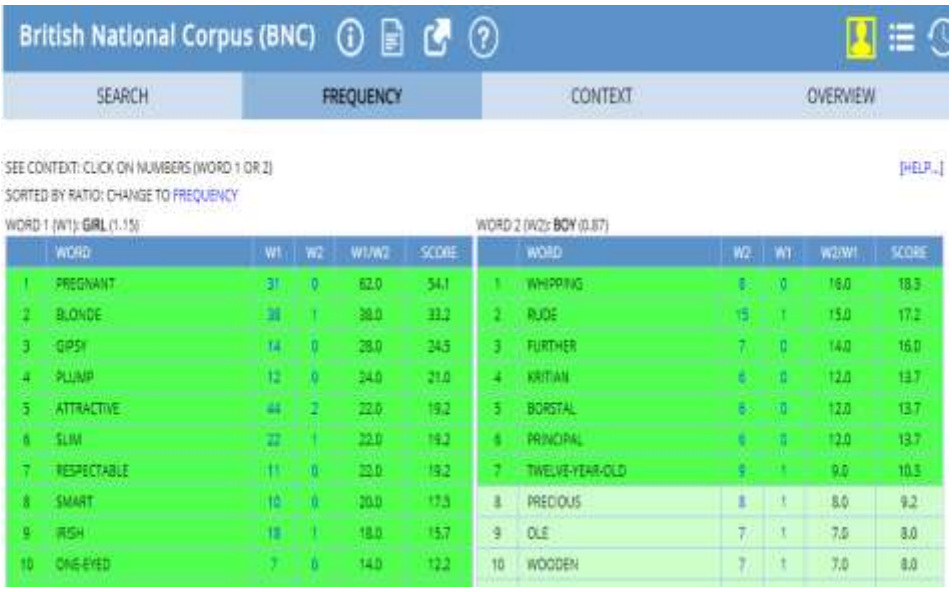


Fig. 2. The relative frequency distribution of collocations *'as a boy'* and *'as a girl'* by genres represented by BNC and COCA

The corpora also have a function that allows seeing the collocation in the usage and how they are described. For this reason, we are using function *Compare* and adding in the below bar left- and right-side context. The first column shows the collocates for ‘*girl’, ‘boy’*. The second and third columns show the frequency of the word compared to one another. The score is the ratio of W1 to W2 column compared to the overall ratio. It is important to note that if you want to get full-blown examples it is important to use the ‘*position*’ tag and add a part of speech you want to see before or after the analyzed word. The list could be sorted by ratio and frequency. In Figure 3 we can see the collocates of two words and how they differ in meaning and usage.



SEE CONTEXT: CLICK ON NUMBERS (WORD 1 OR 2)

SORTED BY RATIO: CHANGE TO FREQUENCY

WORD 1 (W1): GIRL (1.15)

	WORD	W1	W2	W1/W2	SCORE
1	PREGNANT	31	0	62.0	54.1
2	BLONDE	38	1	38.0	33.2
3	GIPSY	14	0	28.0	24.5
4	PLUMP	12	0	24.0	21.0
5	ATTRACTIVE	44	2	22.0	19.2
6	SLIM	22	1	22.0	19.2
7	RESPECTABLE	11	0	22.0	19.2
8	SMART	10	0	20.0	17.5
9	IRISH	18	1	18.0	15.7
10	ONE-EYED	7	0	14.0	12.2

WORD 2 (W2): BOY (0.87)

	WORD	W2	W1	W2/W1	SCORE
1	WHIPPING	8	0	16.0	18.3
2	RUDE	15	1	15.0	17.2
3	FURTHER	7	0	14.0	16.0
4	KRITIAN	6	0	12.0	13.7
5	BORSTAL	6	0	12.0	13.7
6	PRINCIPAL	6	0	12.0	13.7
7	TWELVE-YEAR-OLD	9	1	9.0	10.3
8	PRECIOUS	8	1	8.0	9.2
9	OLE	7	1	7.0	8.0
10	WOODEN	7	1	7.0	8.0

Fig. 3. The function *Compare* in BNC online

COCA corpus also allows tracing the trend of given collocations usage in the diachronic aspect. Judging by the frequency, we can make the conclusion that collocation ‘as a boy’ is more productive (2,62) than ‘as a girl’ (1,12) and was widely used in the text of 2005-2009. Also, it is important to mention that the period of the biggest usage of both collocations is 1995-1999 and the average usage of them for a million words is 1,87. The quantitative predominance of ‘as a boy’ is an example of gender asymmetry, the unevenness of reflection in the language itself (vocabulary, grammar, speech).

To get a deeper understanding of the overall picture of using gender-marked lexis in different periods and genres we have also investigated the usage of the word *‘gender’* itself. According to the obtained results *‘gender’* is more frequently used in COCA (58,42) than in BNC (19,47). However, compared to collocations *‘as a girl’* and *‘as a boy’* genre distribution is completely different. It is more used in academic texts, magazines and newspapers. Also, with the help of *Chart* we investigated that the period of these collocation usages is similar and the most productive is 2005-2009. These results may be related to the growing number of non-fiction literature and the greater interest of society in gender issues. The relative frequency of *'gender'* is depicted in the pie-chart diagram (See Figure 4).

Table 2. The relative frequency distribution of collocations *'as a boy'* and *'as a girl'* by time period represented in COCA

	Relative frequency (10^{-2})		
Time period	***'as a boy'***	***'as a girl'***	**Average**
1990-1994	2,57	0,92	1,75
1995-1999	2,72	1,11	1,92
2000-2004	2,48	1,11	1,79
2005-2009	2,76	1,07	1,91
2010-2014	2,45	1,26	1,86
2015-2019	1,74	0,85	1,30
Total	2,62	1,12	1,87

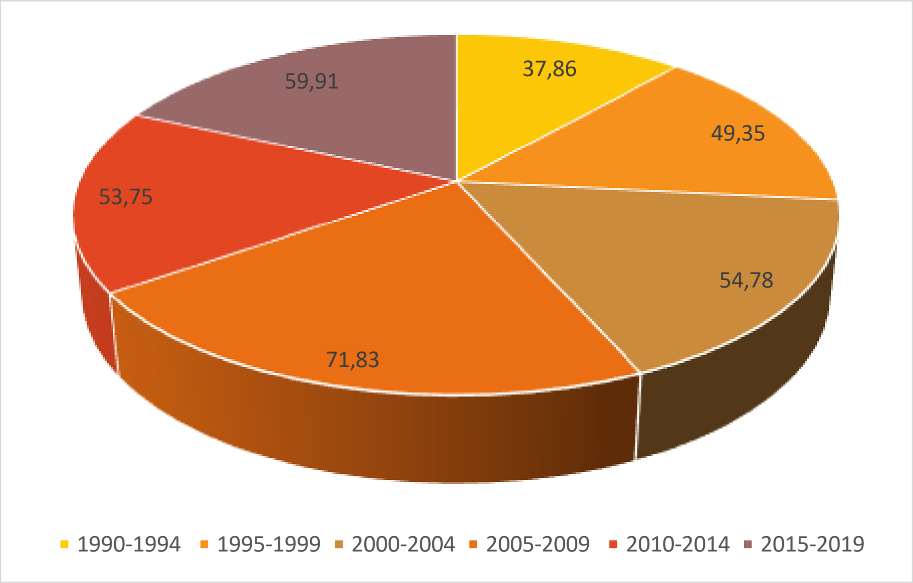


Fig. 4. The distribution of word *'gender'* represented in COCA

We can also use the additional features available in Sketch Engine for BNC corpus. For example. *Word Sketch* allows us to perform a quantitative analysis of the frequency and in the concept sphere *MAN* the most used units are boy, father and for *WOMAN* components mother and girl. *Thesaurus* allows you to mark words that often have similar collocation behavior. The word *'boy'* is most commonly found as husband (162) and father (209), and girl as woman (437), wife (260) and sister (48). Using *Wordlist* we track that the word girl has been used 230 times, other lemmas of the word girls - 144 and the adjective girlish - 5 times.

It is very important to mention the advantages of COCA in comparison to BNC. COCA is much bigger, it's volume is 20 million words each year 1990-2019. Among the advantages the latest updates released in March 2020 allows downloading the corpus for online use. Also, for the most frequent words users can hear the pronunciation, see videos with that word in the text, find related images from Google Images, an see a translation for their preferred target language.

For our research, we also have used the Google Ngram Viewer. This online service allows building frequency charts for language units based on a huge number of printed sources published since the 16th century and collected in the Google Books service. In addition, there is a possibility to make a search based on specialized text corpora. Google Ngram is a variation of syntax N-grams that are defined by paths in syntax dependency trees or component trees, rather than linear text structure. This service is useful for analyzing the frequency of syntax alternations in the corpus. The example of using Google Ngram is depicted in Figure 4. It shows that component *'gender'* is not as productive as *'boy'* or *'girl'* and it is interesting that they have similar usage tendencies.

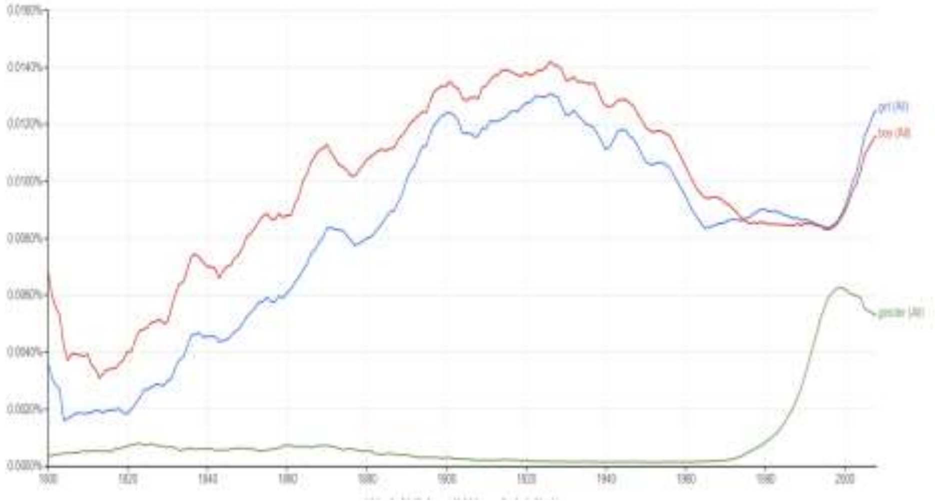


Fig. 1. Google Ngram viewer for words *'girl', 'boy', 'gender'*

6 Conclusions

Gender linguistic, as a young science, still needs advanced methods of research. One of the effective and valid methods are statistical since they can mathematically confirm the obtained results.

Statistical methods study the quantitative characteristics of linguistic phenomena in order to show a specific measure of phenomena, to find out how changes in numerical characteristics in certain conditions show the laws of linguistic development. However, the statistics obtained express only the qualitative content of the phenomena and it should be noted that a linguist should conduct the analysis or argumentation.

The achievement of corpus linguistics greatly facilitates the data collection process for analysis, as it provides linguistic corpora of texts that serve as a powerful information and linguistic support for learning different aspects of the language.

The analyzed gender-marked units are a prime example of realization of gender stereotypes on lexical level that take their place in language. The research of lexical units extracted from corpora are created by applying stigmatic gender labels. However, in my opinion, the results should be verified on a wider fact sheet and on texts of other styles. Further study of gender stereotypes in speech interaction has necessitated their consideration in a broader context within the overall structure of communication.

The prospect of further research is to improve the methods of gender linguistics that can contribute to a higher degree of representativeness of research results.

7 References

1. Stoller, R.J:. Sex and Gender. In: The Development of Masculinity and Femininity, Vol. 1, Science house, New York, 400 (1968)
2. Giddens, A., Sutton W. P.: Sociology, 8th Edition, 1192 (2017)
3. Lakoff, R.: Language and woman’s place. In: Language in Society 2, 45–79 (1972)
4. Tannen, D.: You Just Don't Understand: Women and Men in Conversation. In: William Morrow Paperbacks, 342 (2007).
5. Maltz, D., Borker, R.: A Cultural approach to male-female miscommunication. In: Language and social identity, Cambridge University Press, New York, 196–216 (1982)
6. Fishman, P.: Interaction: The work women do. In: Social Problems, No 24, 397–406 (1978)
7. Losonsky, M.: Humboldt. On Language, On the Diversity of Human Language Construction and its Influence on the Mental Development of the Human Species. In: CUP, 25–64 (1999)
8. Carroll, J. B.: Language, thought, and reality: Selected writings of Benjamin Lee Whorf. In: Cambridge, MA: The Technology Press of MIT, New York, Wiley, 207–219 (1956)
9. Kirilina, A.: Gendernyye issledovaniya v zarubezhnoy i rossiyskoy lingvistike (Filosofskiy i metodologicheskiy aspekty). In: Obshchestvennyye nauki i sovremennost, No 4. 138–143 (2000)
10. Goroshko, E.: Informatsionno-kommunikativnoye obshchestvo v gendernom izmerenii. In: Kharkov: FLP Liburkina L.M, 816 (2009)
11. Kharchenko, O.: Metodolohichni pidkhody hendernykh doslidzhen v sotsiolohyi. In: Sotsiolohiya v sytuatsiyi sotsialnykh nevyznachenostey: Tezy dopovidey uchasnykiv I Konhresu Sotsiolohichnoyi asotsiatsiyi Ukrayiny, KH.: KHNU imeni V.N.Karazina,, 476 (2009)
12. Levchenko, O.: Linhvistychni doslidzhennya henderu v Ukrayini. In: Lyudyna. Kompyuter. Komunikatsiya: zbirnyk naukovykh prats, 74–83 (2017)
13. Perebyinis, V.: Statystychni metody dlia linhvistiv. In: Nova Knyha, Vinnystia, 176 (2013)
14. Zayats M., Zayats V., Metody zistavlennya statystychnykh kharakterystyk pry formuvanni vybirok u linhvistytsi. In: Visnyk Natsionalʹnoho universytetu "Lvivska politekhnika", No 67, 296–305 (2010)

15. Romaniuk, S.: Zastosuvannya statystychnykh metodiv u linhvistychnykh doslidzhennyakh.. In: Naukovi zapysky Natsionalnoho universytetu "Ostrozka akademiya", seriya «Filolohichna», Vyp.54, 134–137 (2015)
16. Sytar H.: Syntaksychni frazeolohizmy v rozrizi konstruktsiynoyi hramatyky. In: TOV «Nilan-LTD», Vinnytsa, 458 (2017)
17. Levitskiy, V.: Kvantitativnyye metody v lingvistike. In: Novaya kniga, Vinnitsa, 264 (2007)
18. Pennebaker, J.W., Francis, E.M., Roger, J.B.: Linguistic inquiry and word count: LIWC 2001. In: Mahway: Lawrence Erlbaum Associates71, 13 (2001)
19. Matthew, L.N.: Gender differences in language use: An analysis of14,000 text samples. In: Discourse Processes 45.3, 211–236 (2008)
20. Argamon, Sh.:Gender, genre, and writing style in formal writtentexts. In: Text23, 3 (2003)
21. Stavytska L.: Hender: mova, svidomist, komunikatsia. In: Instytut ukrayinskoyi movy NAN Ukrayiny, Kyyiv : KMM, 440 (2015)