The concept of language in American media discourse on the basis of the chi-square test

Myroslava Kovaliuk
Department of English
Yuriy Fedkovych Chernivtsi National University
Chernivtsi, Ukraine
m.kovaliuk@chnu.edu.ua

Nadiia Yesypenko
Department of English
Yuriy Fedkovych Chernivtsi National University
Chernivtsi, Ukraine
n.yesypenko@chnu.edu.ua

Nina Pechko
Foreign Languages and Translation Department
Lesya Ukrainka Volyn National University
Lutsk, Ukraine
pechko.nina@vnu.edu.ua

Larysa Taranenko
Department of Theory, Practice and Translation of English
National Technical University of Ukraine "Igor Sikorsky Kyiv Politechnic Institute"
Kyiv, Ukraine
larysataranenko@gmail.com

Abstract*−Language is a kind of reproduction of reality and the world of images, thanks to which a person forms and accumulates knowledge about the world picture. It (the picture of the world) can be replenished with new knowledge, corrected or regulated by human behavior. The concept of language, actualized in contemporary English-language media discourse, is an integral mental unit, the structure of which reflects the configuration of its culturally significant cognitive features that reflect the experience of language use by speakers through the prism of universal and culturally specific knowledge about language. The reflection of such a socio-cultural phenomenon as language in media discourse will help to identify the actual cognitive features of the language concept interpreted by a journalist as a representative of the American linguistic community. Linguistic and statistical methods are relevant in solving such issues of cognitive linguistics as the actualisation of nominal units by linguistic means. The use of linguistic and statistical methods in our article allows us to bring the study of the verbalization of the language image from subjective perception to the objective space of interdependence and subordination of the language system. The aim of the article is to express the concept of language using the chi-square test in American media discourse.

***Keywords*−language, chi-square test, quantitative analysis, lexical-semantic classes, accompanying lexeme, concept, discourse, discourse analysis.**

I. Introduction

Modern linguistics shares a common basis for studying language with computational linguistics. It is concerned with the theoretical and structural aspects of language, with the aim of describing and explaining its properties. Computational linguistics is applied, providing data-driven insights into language phenomena, developing practical tools and algorithms for processing and analyzing natural language [1].

The cognitive approach to linguistics considers language as a system of signs that help to encode and transform information. Language is studied as a universal cognitive mechanism and tool [2]. Language is a means of organizing, processing and transmitting information. It is not studied autonomously, but from the point of view of the reflection of the surrounding world in human consciousness and ways of conceptualizing the world, general principles of categorization and mechanisms of information processing, as well as from the point of view of how language reflects the cognitive experience of a person and the influence of the environment. Being an integral part of cognition, language reflects the interaction of communicative, functional and cultural factors.

Language is a systemic phenomenon, so the use of linguistic and statistical analysis in the study of linguistic phenomena is legitimate. Since language is a sign system, we consider it appropriate to use statistical methods to study it. However, even as a sign system, it differs from those systems that have clear mathematical laws. Mathematical formulas do not fully cover all the regularities characteristic of a particular linguistic phenomenon, but they allow us to practically support certain conclusions and laws of language functioning [3].

For a successful quantitative analysis of the material, we use the chi-square criterion to calculate the arithmetic mean of the distance between lexical items. The analysis of the concept of language in American media discourse identifies the most relevant areas of its expression in American linguistic culture.

II. Related works

Linguists study the importance of knowing and speaking a language. Despite sharing the same goal, they have different theoretical preferences as a function of their beliefs about what the true nature of language is. The study of language as a body of linguistic concepts did not begin until the rise of cognitive and cultural linguistics. Language plays an important role in the categorisation and conceptualisation of experience [4]. We see language as a semiotic (“meaning-making”) tool, which is the basis of learning [5]. As learners encounter multiple discursive communities in their everyday lives, it is also necessary to consider how these discourses/language uses are legitimised and disseminated across communities [6].

The concept is embodied in discourse as it exists in both the mental and material spheres. Engaging media discourse in a broad linguistic and cultural context and using linguistic and statistical methods to analyze the concept verbalized in it opens up a new perspective for studying concept actualization in a culturally labelled context [7].

Media discourse is a reflection of the politics of media institutions and is part of the cultivation of concepts in ways that are open to study. It shapes public opinion, influences social norms and frames people's perceptions of the world. For this reason, the development of a number of indicators of the prevailing winds of the general symbolic environment is necessary for informed political decision-making and reliable interpretation of the formation of and reactions to social ideas [7].

Quantitative research is probably the approach most commonly found in cognitive linguistics and much of social research in general. Discourse analysis is usually qualitative

979-8-3503-6046-2/23/$31.00 ©2023 IEEE

in nature, although recent developments towards using methods derived from computational linguistics to support discourse analysis have also given it quantitative dimensions [8].

III. Methods

The methodological basis of the article is the scientific conceptions within the following scientific areas: cognitive semantics [9], discourse theory and discourse analysis [10]; quantitative analysis [11]. The combination of linguistic and statistical methods makes it possible to consider this concept in a holistic statistically verified picture of its functioning in American media discourse with the reflection of systemic connections between the elements of its conceptual structure.

Within quantitative linguistics, quantitative and linguistic statistical methods are distinguished. Quantitative methods determine the frequency of use of units in different speech genres and their compatibility with other units of the language. Linguistic and statistical methods allow us to trace the relationship between language units, determine the probability and selective nature of their joint use. Quantitative and linguistic statistical methods are widely used at all levels of language, as they can be used to express the quantitative composition of phonemes, sounds, letters, syllables, morphemes, words, phrases and syntactic structures.

The frequency of use of a linguistic unit depends to a certain extent on each of its systemic characteristics. In most cases, it is a combination of interrelated systemic properties that affects the frequency of use of a unit in a text. As a result, the presence or absence of interdependence between the frequencies of certain linguistic characteristics indicates the presence or absence of a relationship between them. The features of different styles and genres, communicative orientation of the text, certain author's writing features have an impact on the frequency of use of language units [11].

The means of language are a way of expressing a concept. The semantic features of a word represent the content of the concept, and the word itself reflects its name. However, one word does not convey all the features of the concept, except for those that are directly relevant to a particular reproduction. To systematize the spheres of expression of the concept nominated in the language, the accompanying words that clarify or explain the concept are analyzed. Actualization of the concept by defining the word-name of the concept and its accompanying lexemes makes it possible to create a linguistic and mental profile of the object or phenomenon behind the concept. Each concept's name is structured to indicate what it means, and each word that accompanies it is a fragment of that concept's profile.

In our article, we use chi-square test to identify those lexical and semantic classes (LSCs) of lexemes accompanying the concept name that are dominant in media discourse and thus establish the most relevant areas of concept expression in American media discourse texts. The spheres of expression of the concept of language in media discourse are built on the basis of classifying the nouns, verbs and adjectives accompanying the nominal lexeme into lexical-semantic classes and identifying statistically relevant indicators of the frequency of use of these LSCs in discourse. The accompanying words are selected within the syntactic framework of a single sentence in which the nominal lexeme of the concept is used. The most common formula for calculating chi-square (χ^2) is:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \qquad (1)$$

where O − actually existing (empirical) values; E − theoretically expected values; Σ sign means the sum [11].

An excess of χ^2 indicates the superiority of the empirical use of a particular lexical-semantic class over the theoretically expected one. Therefore, we are talking about the selectivity of the use of a particular lexical-semantic class of accompanying nouns, i.e. the importance of LSC for the verbalization of the concept of language under study.

IV. Experiment

Lexemes play a nominative role in the verbalization of a concept. Language units are in a mutual hierarchical relationship. Verbs, nouns and adjectives belong to the higher language units in the hierarchy. In order to determine the lexical-semantic classes of the accompanying words to the concept name of language, we have identified the accompanying nouns, verbs, adjectives to the concept name lexeme and classified them into lexical-semantic classes of nouns, verbs, adjectives. A lexical-semantic class is a set of linguistic units accompanying the concept name that reveal one common topic. The material of the study is 1 500 nominations of the concept language and 18 769 accompanying lexemes (verbs − 5 785.5, nouns − 9 891, adjectives − 3 092.5), selected from 10 sources of American media in 2021−2022). In our study, the lexical composition of the analyzed fragments of American media discourse was systemized and divided into lexical-semantic classes of nouns, adjectives and verbs. Quantitative counts of the usage of these lexemes are presented in Table 1.

Table 1 Quantitative distribution of accompanying lexemes according to the part-of-speech criterion

Part of speech	**American media discourse**	
	amount	***%***
noun	9 891	53
verb	5 785.5	31
adjective	3 092.5	16
Total	18 769	100

Using linguistic statistical techniques (χ^2-test), we determine the degree of intensity of the combination of two words in a text at the level of semantic compatibility ("word + word subclass") [12]. We define accompanying words as words that often occur in a text fragment in the immediate vicinity of the nominal lexeme of a concept and are related to each other in terms of meaning. The cognitive-semantic or associative connection of the accompanying words exists when the accompanying lexemes are repeated.

Mathematical formulas do not fully cover all the regularities characteristic of a particular linguistic phenomenon. Nevertheless, linguistic statistics provides reliable results that represent the most typical characteristics of linguistic phenomena and describe the exceptions that can be traced in the language data. Deviations from the average are not acceptable in linguistic studies; they are a priori predictable [11].

The use of the chi-square test makes it possible to distinguish the spheres of expression of the concept of language in the statistical processing of the frequencies of use of accompanying words with the nominal lexeme of the concept by lexical-semantic classes. For a reliable statistical analysis of the quantitative data on the use of each lexical-semantic class, it is worth using the chi-square test, since we are not sure whether the frequency of lexical-semantic classes usage in American journalism exceeds the theoretically expected value, which would emphasize the selectivity of one or another class. In case of a sample difference in frequencies, which gives a χ^2 value that does not exceed 3.84 at df = 1, we consider the frequency differences to be random. If the sum of χ^2 at df = 1 is greater than 3.84, the frequency difference indicates certain regular phenomena and is considered significant. Thus, the frequency of use of a certain lexical-semantic class in one of the ten journalistic publications differs from the frequency of use of the specified lexical-semantic class in other publications. When analyzing the statistical data, we take into account only those lexical-semantic classes for which the χ^2 value is significant, which means that their use in discourse is selective. This shows that the thematic sphere represented by lexical-semantic class is relevant for the embodiment of the language in the minds of the Americans.

Among the lexical-semantic classes of the accompanying nouns to the name of the concept of language, according to the χ^2 index, we can observe five cases of exceeding the theoretically expected values for the LSC “education” (χ^2 = 10.02 (The Chicago Tribune), χ^2 = 6.48 (The Wall Street Journal), χ^2 = 85.75 (The New York Times), χ^2 = 61.49 (The Newsday), χ^2 = 70.25 (The Washington Post). A number of lexical-semantic classes demonstrate three times the empirical frequency of use such as: “language name” (χ^2 = 36.56 (The Chicago Tribune), χ^2 = 15.28 (The Newsday), χ^2 = 11.46 (The Washington Post)); “country/nationality” (χ^2 = 21.88 (The Houston Chronicle), χ^2 = 4.24 (USA Today), χ^2 = 5.75 (The New York Post)); “media/art” (χ^2 = 21.17 (The Houston Chronicle), χ^2 = 3.48 (USA Today), χ^2 = 52.30 (The New York Post)); “technology” (χ^2 = 56.01 (The Chicago Tribune), χ^2 = 18.37 (The Washington Post), χ^2 = 24.84 (USA Today)).

The following lexical-semantic classes of the accompanying nouns are distinguished twice: “family relationships” (χ^2 = 48.86 (The Chicago Tribune), χ^2 = 13.47 (The New York Post)); “economics” (χ^2 = 99.21 (The Politico), χ^2 = 3.25 (The Star Tribune)); “sphere of language use” (χ^2 = 5.79 (The New York Times), χ^2 = 15.34 (The New York Post)); “institution/place of language use” (χ^2 = 81.16 (The Chicago Tribune), χ^2 = 8.09 (The Newsday)); “event / occasion” (χ^2 = 4.36 (USA Today), χ^2 = 4.42 (The New York Post)); “mental sphere” (χ^2 = 5.65 (The Wall Street Journal), χ^2 = 7.63 (The New York Times)); “person/group of people” (χ^2 = 7.87 (USA Today), χ^2 = 5.23 (The New York Times)); “religious sphere” (χ^2 = 17.72 (The Wall Street Journal), χ^2 = 4.26 (USA Today)); “transport/movement” (χ^2 = 10.25 (The Houston Chronicle), χ^2 = 11.60 (USA Today)); “food/drink” (χ^2 = 7.53 (The Chicago Tribune), χ^2 = 15.68 (The Politico)); “medicine” (χ^2 = 22.36 (The Wall Street Journal), χ^2 = 11.15 (The Houston Chronicle)); “law” (χ^2 = 4.53 (The Houston Chronicle), χ^2 = 62.08 (The Politico)); “language policy” (χ^2 = 6.82 (The Wall Street Journal), χ^2 = 6.34 (The Houston Chronicle)).

Fifteen cases of single deviation of χ^2 values are identified for the lexical-semantic classes of nouns and shown in Table 2.

TABLE 2 SINGLE DEVIATION OF THE CHI-SQUARE VALUE FOR LEXICAL-SEMANTIC CLASSES OF NOUNS

The lexical-semantic class of nouns	**χ^2 value**
“time/age”	χ^2 =4.21 (The Newsday)
“public sphere”	χ^2 = 65.48 (The Politico)
“science”	χ^2 = 9.46 (The New York Times)
“relationship/attitude”	χ^2 = 7.09 (The Politico)
“documents”	χ^2 = 10.28 (The Politico)
“artefacts/products of activity”	χ^2= 5.68 (The Newsday)
“measure/size”	χ^2 = 7.54 (The Houston Chronicle)
“animal/plant life”	χ^2 = 12.39 (The New York Post)
“body language/body parts”	χ^2 = 24.76 (The Houston Chronicle))
“homeland/nation”	χ^2 = 6.45 (TheStar Tribune)
“abstract concepts”	χ^2 = 5.22 (The Chicago Tribune)
“sports/game”	χ^2 = 58.08 (USA Today)
“purism”	χ^2 = 48.08 (The New York Post)
“productive/destructive activity”	χ^2 = 8.21 (The New York Times)
“sign/sound language”	χ^2 = 9.24 (The Houston Chronicle)

The excess of the χ^2 value for the lexical-semantic class “education” indicates the relevance of language use in the educational and training sphere, which is reflected in American media. The least important areas of language functioning include language purity, public and sporting activities, flora and fauna, sign language.

The analysis of the lexical-semantic classes of the accompanying verbs to the name of the concept of language revealed three cases of preferences for the lexical-semantic class of verbs “implementation of educational policy” (χ^2 = 6.21 (The Chicago Tribune), χ^2 = 6.75 (The Newsday), χ^2 = 28.77 (The Washington Post)). Two cases of exceeding the theoretically expected values of usage are identified for: the lexical-semantic class represented by the verbs “communication” (χ^2 = 8.93 (The Politico), χ^2 = 5.52 (USA Today)); LSC “acquisition/enrichment” (χ^2 = 20.34 (The Newsday), χ^2 = 10.35 (The Chicago Sun Times)); LSC “development /creation” (χ^2 = 12.63 (The Washington Post), χ^2 = 10.04 (The New York Post)); LSC “legal activity” (χ^2 = 15.36 (The Politico), χ^2 = 8.83 (USA Today)); LSC “medical activity” (χ^2 = 4.55 (The Star Tribune), χ^2 = 5.34 (USA Today)); LSC “sound” (χ^2 = 11.18 (The Houston Chronicle), χ^2 = 4.14 (The Politico)); LSC “structure” (χ^2 = 12.37 (The Politico), χ^2 = 4.24 (The Washington Post).

The following lexical-semantic classes of verbs demonstrate one excess of the χ^2 value each: “mental activity” (χ^2 = 5.68 (The Wall Street Journal)); “possession/loss” (χ^2 = 8.88 (The Houston Chronicle)); “movement” (χ^2 = 38.31(The Politico)); “existence” (χ^2 = 43.62 (The Houston Chronicle)); “change of state” (χ^2 = 20.24 (The New York Times)); “scientific activity” (χ^2 = 5.53 (USA Today)); “animal/plant action” (χ^2 = 12.19 (The Star Tribune)); “implementing of language policy” (χ^2 = 8.58 (USA Today)); “functioning of technology” (χ^2 = 5.68 (USA Today)); “sporting activity” (χ^2 = 12.76 (The Newsday)); “control/power” (χ^2 = 5.69 (The Washington Post)); “destructive activity” (χ^2 = 5.52 (The New York Post)); “embodiment of body language” (χ^2 = 4.17 (USA Today)); “religious activity” (χ^2 = 11.36 (The Wall Street Journal)); “military activity” (χ^2 = 15.90 (The Politico)); “measuring” (χ^2 = 5.93 (The New York Times)); “purification of language” (χ^2 = 5.29 (USA Today)); “implementing of state policy” (χ^2 = 136.56 (The Politico)).

Thus, in the American press, the most important spheres of language functioning are educational, training, and speech activities, and the significant role of language in the productivity of actions is emphasized. Little attention is paid to the purity of the language, military, religious and scientific spheres, physical and mental activities.

The analysis of the lexical-semantic classes of adjectives accompanying the name of the concept of language shows two cases of excessive frequencies of use for the following LSCs: “feature of language” (χ^2 = 13.51 (The Chicago Tribune), χ^2 = 12.37 (USA Today)); “description of time characteristics” (χ^2 = 33.70 (The New York Times), χ^2 = 7.67 (The New York Post)); “description of age characteristics” (χ^2 = 45.93 (The Wall Street Journal), χ^2 = 8.33 (The Newsday)); “description of educational/scientific field” (χ^2 = 20.40 (The Star Tribune), χ^2 = 13.73 (The Newsday)); “name/type of language indication” (χ^2 = 15.42 (The Chicago Tribune), χ^2 = 29.21 (The Houston Chronicle)); “description of the physical/physiological state of objects” (χ^2 = 9.24 (The New York Post), χ^2 = 4.47 (The Houston Chronicle)); “description of the public sphere” (χ^2 = 4.89 (The Wall Street Journal), χ^2 = 92.70 (The Politico)); “description of appearance/parameters” (χ^2 = 4.27 (The Washington Post), χ^2 = 5.68 (The New York Post)); “body language/sound language description” (χ^2 = 12.47 (The Star Tribune), χ^2 = 12.79 (USA Today)); “colour description” (χ^2 = 4.52 (The Houston Chronicle), χ^2 = 8.09 (USA Today)).

One sample of usage was recorded for adjectives denoting “description of an emotional state” (χ^2 = 11.31 (The Star Tribune)); “positive evaluation” (χ^2 = 4.32 (The New York Post)); “negative evaluation” (χ^2 = 5.32 (The Politico)); “description of the action performed on the object” (χ^2 = 8.11 (The Houston Chronicle)); “country/nationality indication” (χ^2 = 14.33 (The Newsday)); “description of behaviour (character traits), attitude/relationships” (χ^2 = 6.13 (USA Today)); “description of quantity/size” (χ^2 = 45.50 (The Washington Post)); “description of purism” (χ^2 = 23.78 (The Politico)); “description of the nation/homeland” (χ^2 = 10.15 (The Wall Street Journal)); “description of proceduralism” (χ^2 = 5.15 (The Politico)); “technology description” (χ^2 = 4.51 (The Wall Street Journal)).

Thus, the analysis of the lexical-semantic classes of adjectives represents the importance of describing the type of language, educational, scientific, physical and public activities and the temporal and age characteristics of language functioning. The least typical associations with language are descriptions of character traits and relationships between people, using the established procedures, portraying countries and nationalities, indicating types of technology and language purism.

V. Conclusions

The sphere of expression of the concept verbalized in the discourse is formed by using linguistic and statistical methods that help to establish the relationship and dependence between the accompanying words of the concept and its nominal lexeme used by lexical-semantic classes. The results of calculating the chi-square test determine the lexical-semantic classes that are most important for expressing a concept in American media discourse.

In order to identify lexical-semantic classes of the accompanying words for the concept name of language, we identified and counted accompanying nouns, verbs, adjectives to the concept name lexeme used in the analyzed fragments of American media discourse. The quantitative distribution of the accompanying lexemes according to the part-of-speech criterion showed that the noun prevails (9 891 cases of use). The verb ranks second in terms of frequency of use in media discourse (5 785.5 occurrences), and the adjective ranks third (3 092.5 occurrences). The semantic structure of the name of the concept characterizes its semantic content, and the accompanying words reflect the fragments of the subject profiling.

For a more reliable analysis of the frequency of use of lexical-semantic classes, the linguistic statistical method χ^2-test was used. In American media discourse, the most frequent lexical-semantic classes of nouns are “education”, “language name”, “country/nationality”, “media/art”, “technology”, lexical-semantic class of verbs “implementation of educational policy” and lexical-semantic classes of adjectives: “feature of language”, “name/type of language indication”, “description of educational/scientific field”, “description of the public sphere”, “description of age/time characteristics”, “body/sound language description”, “description of appearance/parameters”, “colour description”, “description of physical/physiological state of objects”.

Language is a tool for thinking and communication. It provides interaction between the sender of a verbal message and its recipient. It is a tool for modelling a picture of the world. It verbally embodies reality and the world of images, being a bridge between reality and a person. Language is formed in human consciousness and plays a significant role in the accumulation, assimilation, organization of knowledge, its processing and storage in memory. The function of language is to lexically express the components of the world picture, to exchange information, experience and express emotions. Language plays an important communicative role in society, as the exchange of information and experience is crucial in human communication.

References

[1] M. Zakaria Kurdi, Natural Language Processing and Computational Linguistics 2: Semantics, Discourse and Applications, vol. 2. John Wiley & Sons, 2017.

[2] H. Diessel, “Usage-based construction grammar” in Cognitive Linguistics – A Survey of Linguistic Subfields, E. Dąbrowska and D. Divjak, Eds. Berlin: De Gruyter, 2019, pp. 50−80.

[3] V. Brezina, Statistics in Corpus Linguistics: A Practical Guide. Cambridge: Cambridge University Press, 2018.

[4] X. Wen and Z. Fu, “Categorization” in The Routledge Handbook of Cognitive Linguistics, X. Wen, John R. Taylor, Eds. New York; London: Routledge, 2021, pp. 173−190.

[5] A. Gałkowski and M. Kopytowska, “Semiotic Perspectives on Language, Signification and Communication” in Current perspectives in semiotics: signs, signification, and communication, vol. 55, A. Gałkowski, M. Kopytowska, Eds. Internationaler Verlag der Wissenschaften, 2018, pp. 9−31.

[6] M. Guardado, Discourse, Ideology and Heritage Language Socialization, vol. 104. Berlin: De Gruyter Mouton, 2018.

[7] R. Facchinetti, “News discourse” in The Bloomsbury Handbook of Discourse Analysis, K. Hyland, B. Paltridge, L. Wong, Eds. London: Bloomsbury Academic, 2021, pp. 961−1048.

[8] J.−P. Metzger, Discourse: A Concept for Information and Communication Sciences, vol. 4. Wiley−ISTE, 2019.

[9] L. Talmy, Ten Lectures on Cognitive Semantics. Leiden: Brill, 2018.

[10] B. Paltridge, Discourse Analysis: An Introduction, 3rd Edition (Bloomsbury Discourse). Bloomsbury Publishing, 2021.

[11] Mario F. Triola Essentials of Statistics. New York: Pearson, 2019.

[12] O. Levy “Word Representation” in The Oxford Handbook of Computational Linguistics, R. Mitkov, Ed. Oxford: Oxford University Press, 2022, pp. 334−358.