UDC 004.383.8[001.8:80]

Viktoriia V. Zhukovska
PhD in Linguistics, Associate Professor
Department of Cross-Cultural Communication and Applied Linguistics
Zhytomyr State Ivan Franko University, Zhytomyr, Ukraine
ORCID ID 0000-0002-4622-4435
victoriazhukovska@gmail.com

Oleksandr O. Mosiiuk
PhD in Pedagogical Sciences, Associate Professor,
Department of Computer Sciences and Information Technologies
Zhytomyr State Ivan Franko University, Zhytomyr, Ukraine
ORCID ID 0000-0003-3530-1359
mosxandrwork@gmail.com

STATISTICAL SOFTWARE R IN CORPUS-DRIVEN RESEARCH AND MACHINE LEARNING

Abstract. The rapid development of computer software and network technologies has facilitated the intensive application of specialized statistical software not only in the traditional information technology spheres (i.e., statistics, engineering, artificial intelligence) but also in linguistics. The statistical software R is one of the most popular analytical tools for statistical processing a huge array of digitalized language data, especially in quantitative corpus linguistic studies of Western Europe and North America. This article discusses the functionality of the software package R, focusing on its advantages in performing complex statistical analyses of linguistic data in corpus-driven studies and creating linguistic classifiers in machine learning. With this in mind, a three-stage strategy of computer-statistical analysis of linguistic corpus data is elaborated: 1) data processing and preparing to be subjected to a statistical procedure, 2) utilizing statistical hypothesis testing methods (MANOVA, ANOVA) and the Tukey post-hoc test, and 3) developing a model of a linguistic classifier and analyzing its effectiveness. The strategy is implemented on 11 000 tokens of English detached nonfinite constructions with an explicit subject extracted from the BNC-BYU corpus. The statistical analysis indicates significant differences in the realization of the factors of the parameter "Part of speech of the subject". The analyzed linguistic data are employed to build a machine model for the classification of the given constructions. Particular attention is devoted to the methodological perspectives of interdisciplinary research in the fields of linguistics and computer studies. The potential application of the elaborated case study in training undergraduate, master, and postgraduate students of Applied Linguistics is indicated. The article provides all the statistical data and codes written in the R script with comprehensive descriptions and explanations. The concluding part of the article summarizes the obtained results and highlights the issues for further research connected with the popularization of the statistical software complex R and raising the awareness of specialists in this statistical analysis system.

Keywords: corpus linguistics; machine learning model; linguistic classifier; statistical software R; RStudio; grammatical construction; linguistic parameter; univariate analysis of variance (ANOVA); multivariate analysis of variance (MANOVA); the Tukey test; linear discriminant analysis; methodological aspects of interdisciplinary studies.

1. INTRODUCTION

Theoretical assumptions. State-of-the-art science is marked by the transition to an alternative model of knowledge production. If the traditional model is characterized by disciplinarity, homogeneity, hierarchy, and domineering of academic communities, the distinctive features of the new model are interdisciplinarity, heterogeneity and heteroarchy. Generated in the applied empirical dimension, this model has its own clearly defined methods and practices, involves various forms of knowledge transfer, and departs from the standard

system of knowledge organization [1, p. 16-17]. In this context, the development of linguistics has been marked by an exponential growth of empirical research, motivated by the increasing need to use natural language mechanisms in information and computer systems. As a result, the methodology of linguistic analysis is being refined, and sophisticated computer technologies alongside statistically reliable tools are being actively utilized to verify scientific hypotheses and findings [2, p. 2].

This "quantitative turn" has already become endemic in linguistics, facilitated by the two interrelated factors: 1) the rapid progress of digitalized linguistic corpora and crowdsourcing sites that give linguists access to 'big data' and 2) significant progress in the development of free open-source computer programs for statistical data analysis (especially the statistical software environment R). Computerized processing of large amounts of empirical data significantly increases the objectivity of linguistic results and reveals new data that are difficult to obtain by adopting traditional introspective and interpretative methods [3, p. 127].

Corpus linguistics is the newest rapidly developing branch of applied linguistic studies, with text corpora being the most popular linguistic and information resources, "the alpha and omega of linguistics" [4, p. 8]. The spheres of corpus linguistic interest include 1) analysis of (usually) large collections of natural language data stored in electronic format and equipped with specialized computer software and 2) investigation of applied linguistic issues in communicative processes, concentrating on language as the process of meaningful communication in language. Corpus linguistics has direct ties with computational and cognitive linguistics. Computational linguistics provides effective tools for processing corpus data – sophisticated computer software to quickly and efficiently process large amounts of language data, search for language units, sort retrieved results, and annotate texts. Corpus linguistics is shaped by a fundamental cognitive usage-based commitment: "language, represented in an infinite number of texts, is the only reality that must be studied without reducing it to a limited set of structural schemes, invariants and ideal paradigms" [5, p. 28]. Consequently, corpora are conceived as powerful language information systems and are actively employed to address a wide range of research issues in almost all fields of linguistics, such as lexicography, grammar, lexicology and semasiology, stylistics, translation studies, pragmatics, sociolinguistics, psycholinguistics, language variation, language acquisition, and literary studies.

The field of corpus research has significantly advanced with the integration of sophisticated statistical software packages to analyze corpus data and build machine learning models on their basis.

Review of previous research. One of the analytical tools used for quantitative processing of empirical data in linguistics is the statistical data analysis system "R" ((R Development Core Team) [6], CRAN (Comprehensive R Archive Network)) [7]. The statistical software R is a free open-source program that provides numerous libraries to solve problems of varying complexity. In Ukrainian linguistics, R is practically not employed, with the rare exception of some publications on using R in corpus linguistic research [8]. Utilizing R in corpus-driven and usage-based studies is becoming increasingly popular in Western linguistics, as evidenced by a growing body of publications in the field. In recent years, comprehensive overviews of R application in the field of linguistics, and more specifically in corpus research, have been offered in V. Brezina [9], S. Th. Gries [10] [11], G. Desagulier [12], R. H. Baayen [13], N. Levshina [14], to name but a few. Numerous studies have addressed the issues of integrating statistical models into language processing using computer systems of statistical analysis, including R: J. Klavan, M.-L. Pilvik, K. Uiboaed [15], D. Divjak, A. Arppe [16].

A brief account of the current literature highlights the increasing interest in employing the statistical software complex R in language studies, especially in corpus and usage-based approaches. This **paper aims** to demonstrate the functionality and advantages of statistical software R for data analysis in linguistic studies and the development of machine learning models based on corpus statistical information.

The **key objectives** of the study are:

1) to substantiate the general strategy of applying the statistical complex R in linguistic research;

2. to demonstrate the effectiveness of utilizing machine learning methods and data analysis on the example of processing numerical information from a linguistic corpus;

3. to describe the methodological and educational significance of the elaborated interdisciplinary research.

2. RESEARCH METHODS

General and special research methods are employed to achieve the objectives of the present study. General research methods include reviewing the targeted literature on the application of statistical methods and analytical tools of computer data processing in language research; comparing the development of the specified issues in Ukraine and abroad; and making generalizations on the underresearched questions.

Special statistical methods of linguistic data analysis include data transformation – logarithmation; testing of statistical hypotheses – univariate analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA); identification of the grammatical constructions with the statistically significant differences in the one-way analysis of variance – the Tukey test.

Linear discriminant analysis is utilized to build a machine learning model for the classification of the analyzed constructions, and the effectiveness of the constructed model is tested on the basis of a confusion matrix.

3. RESEARCH RESULTS

3.1. Grammatical constructions and the strategy for computer and statistical analysis of corpus data

Following a new cognitive-quantitative vector in grammar research, this study incorporates theoretical and methodological tenets of construction grammar, specifically its usage-based version, and quantitative corpus linguistics. As a cognitive grammar theory, construction grammar rests on the premise that language should be described as a structured inventory of form-meaning pairings, collectively referred to as *constructions* [17] - [18]. All language units – from morphemes to abstract clausal patterns – are viewed as language signs, whose formal properties are largely determined by their semantics or functions. From the usage-based construction grammar perspective, the most appropriate way of establishing the linguistic properties of a particular construction is to analyze its occurrence in a corpus. The frequency of individual language units and their sequences is considered an important property of a human language [19], determining the degree of their entrenchment in a given speech community [20]. Consequently, input data crucially influence the formation of the mental grammar of speakers.

In this paper, we investigate English detached nonfinite constructions with an explicit subject, for example [[$_{NP}$***hands***] [$_{XP}$***in pockets***]]; [[$_{AUG}$***with***] [$_{NP}$***thick spectacles***] [$_{XP}$***perched*** at

every end of his nose]]; [[$_{AUG}$***despite***] $_{NP}$[***oil***] [$_{XP}$***being*** the lifeblood of industrial (modern) society]]; [[$_{AUG}$***without***][$_{NP}$***insects***][$_{XP}$***crawling*** in my hair]]; [[$_{AUG}$***what with***] [$_{NP}$***my three sons***] [$_{XP}$***being*** away in the Army]]. These syntactic patterns represent a nonfinite and nonverbal secondary predication of a syntactically independent configuration. They are part of a minimally two-clause syntactic structure consisting of a matrix clause and a punctuationally separated nonfinite or nonverbal clause with its own overt subject. The clauses are of a fixed binary structure [NP XP], where NP represents a secondary subject (Subj), different from the subject of the matrix clause SBJ_M, and (XP) is a secondary predicate (Pred), expressed by a nonfinite verb form (NF) (participle I (PI), participle II (PII), infinitive (to-Inf)) or non-verbal part of speech (VL) (noun phrase (NP), adjective phrase (AdjP), adverbial phrase (AdvP) or prepositional phrase (PP)), and connected with a matrix clause through augmentors (aug) (*with, without, despite, what with*) or asyndetically (øaug).

In the light of construction grammar, the clauses are identified as abstract clausal constructions specified in the given scheme FORM: [[aug/øaug] [SBJ_{NP}] [$PRED_{NF/VL}$]] ↔ MEANING: […]$_{FUNCTION}$, where their meaning is considered not as coded semantics but as their general syntactic function in a sentence.

The [[aug/øaug][$Subj_{NP}$][$Pred_{NF/VL}$]]-*constructions* constitute a taxonomic constructional network organized around the most abstract constructional scheme – macro-construction *(dtcht-Subj* $Pred_{nf/vl}$*-cnx),* whose properties are inherited by more specific meso-constructions (*dtcht-øaug- Subj* $Pred_{nf/vl}$*-cnx, dtcht-aug-Subj* $Pred_{nf/vl}$*-cnx {AUG: with, what with, without, despite}*) and further acquired by individual micro-constructions (*dtcht-**øaug**-Subj* $Pred_{nf/vl}$*-cxn, dtcht-**with**-Subj* $Pred_{nf/vl}$*-cxn, dtcht-**despite**-Subj* $Pred_{nf/vl}$*-cxn, dtcht-**without**–Subj* $Pred_{nf/vl}$*-cxn, dtcht- **what_with**-Subj* $Pred_{nf/vl}$*-cxn {NF: PI, PII, to-Inf; VL: NP, AdjP, AdvP, PP})* and instantiated in concrete realized constructions – *constructs* ([*his cheeks burning suddenly*], [*with thick spectacles perched at the end of his nose*], [*hands in pockets*]…).

As grammatical constructions, the analyzed syntactic patterns are characterized by a set of parameters (morphosyntactic, positional, relational, referential, functional, distributional, and lexico-semantic) realized in particular factors at a specified language level. To statistically assess the degree of proximity/ remoteness of (micro-)constructions in the network and quantitatively verify the determining parameters (factors) that condition the internal functional dynamics and variability of the constructional network, the strategy for computer and statistical analysis of the research corpus data was designed (Fig. 1). Statistical analysis was performed by applying the statistical software environment R. The results obtained through statistical analysis were further adopted to develop a machine learning model.

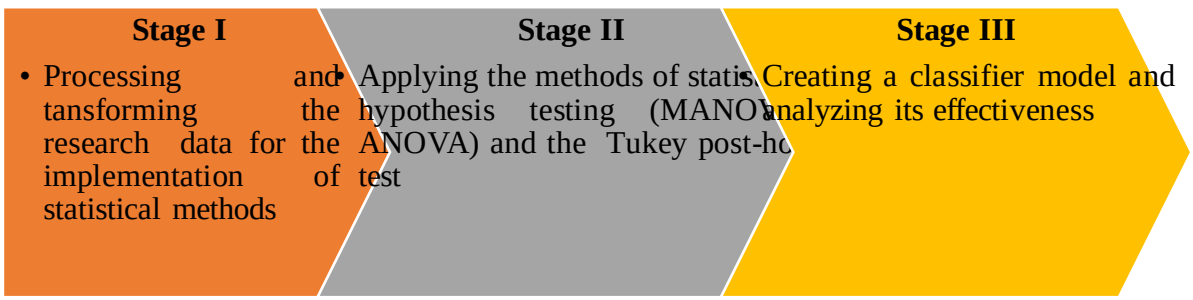

Fig. 1 The main stages of the strategy for computer and statistical analysis of the research data

The strategy for computer and statistical analysis of the linguistic data drawn from the BNC-BYU corpus includes three stages. Stage 1 is aimed at preparing the corpus data to be subjected to statistical processing. At Stage 2, the statistical methods for hypothesis testing (MANOVA, ANOVA) and the post-hoc Tukey test are employed on the research data. Stage 3 is devoted to creating a classifier model and testing its effectiveness.

Owing to space limitations, in this paper, we present the implementation of the suggested computer and statistical strategy to analyze only one parameter, namely "Part of speech representation of the subject" ("SubjPOS"). The parameter "SubjPOS" is manifested in the factors: "nominal subject" ("SubjN"), expressed by a noun or noun phrase, and "pronominal subject" ("SubjPrn"), expressed by a pronoun. The factor "nominal subject" ("SubjN") acquires the meanings: "common nouns" ("SubjNCmn") and "proper names" ("SubjNProp"). The factor "pronominal subject" is realized by "personal pronouns" ("PrnPers"), "indefinite pronouns" ("PrnIndf"), "reflexive pronouns" ("PrnRefl"), "demonstrative pronouns" ("PrnDem"), and "negative pronouns" ("PrnNeg").

3.2. Utilizing R in statistical corpus-driven research

The analysis of the [[aug/øaug][$Subj_{NP}$][$Pred_{NF/VL}$]]-*constructions* is carried out on authentic English usage-data collected from the well-balanced British National Corpus in December 2020 [21]. The data are retrieved automatically using the BNC-BYU's search engine. The total sample includes 11 000 tokens (constructs) of the constructions *dtcht-**øaug**-Subj $Pred_{nf/vl}$-cxn, dtcht-**with**-Subj $Pred_{nf/vl}$-cxn, dtcht-**despite**-Subj $Pred_{nf/vl}$-cxn, dtcht-**without**–Subj $Pred_{nf/vl}$-cxn, dtcht-**what_with**-Subj $Pred_{nf/vl}$-cxn*.

The quantitative representations of the linguistic factors within the "SubjPOS" parameter for each of the constructions under scrutiny are presented in Table 1. As can be seen, some issues may complicate a subsequent statistical procedure and therefore need to be settled beforehand. These issues are specified as follows.

1. The data are presented in the form of interval discrete values, which indicate the frequency of observations of a particular construction (*dtcht-**øaug**-Subj $Pred_{nf/vl}$-cxn, dtcht-**with**-Subj $Pred_{nf/vl}$-cxn, dtcht-**despite**-Subj $Pred_{nf/vl}$-cxn, dtcht-**without**–Subj $Pred_{nf/vl}$-cxn, dtcht-**what_with**-Subj $Pred_{nf/vl}$-cxn*) in the BNC-BYU corpus.
2. The presence of missing values negatively affects the application of statistical methods.
3. The difference between the minimum and maximum values is sufficiently large, which also affects further processing.

To simplify further calculations and avoid erroneous conclusions, several transformations are performed to make the collected data more convenient to operate. First, all missing values are replaced by zero values, because zero is an ordinary number and does not change the content of the analyzed data. Next, logarithmation of the data is carried out [22], which allows the operation of continuous interval data, based on a logarithmic scale. Data standardization is performed using the formula $\ln(x_{ij} + const)$, where x_{ij} is the value in the table, and 2 represents the value of *const*. Generally speaking, any other positive number can be used instead of 2, except for 1. Since $\ln(0 + 1) = 0$, and this again results in zeros in the values that we try to exclude from the calculations. Further calculations are carried out on the standardized data in Table 2.

To begin with, we check whether there are statistically significant differences between the grammatical constructions within the "SubjPOS" parameter and identify the determining factors for each syntactic pattern. The calculations are performed using the R statistical package and its freely distributed libraries. The factors of the parameter (independent variables) are presented in the columns, and their values are in the rows. Multivariate analysis

of variance (MANOVA), a generalization of one-way analysis of variance (ANOVA) [14], is carried out to statistically substantiate the differences between the constructions in terms of realization of the "SubjPOS" parameter. The following statistical hypotheses are formulated:

*H0: The differences between the constructions (dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn, dtcht-**with**-Subj Pred$_{nf/vl}$-cxn, dtcht-**despite**-Subj Pred$_{nf/vl}$-cxn, dtcht-**without**–Subj Pred$_{nf/vl}$-cxn, dtcht-**whata_with**-Subj Pred$_{nf/vl}$-cxn) within the "SubjPOS" parameter are insignificant, and the identified dependencies are random.*

Table 1

Quantity of the constructions within the "SubjPOS" parameter

Construction	Factors of the "SubjPOS" parameter	Pred$_{NF}$			Pred$_{VL}$			
		PI	PII	to-Inf	NP	AdjP	AdvP	PP
*dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn*	Common nouns (*NCmn*)	1694	499	9	54	375	57	304
	Proper nouns (*NProp*)	575	–	–	–	1	–	1
	Personal pronouns *(PrnPers)*	265	–	1	6	–	–	14
	Indefinite pronouns (*PrnIndf*)	223	10	–	8	–	–	32
	Reflexive pronouns (*PrnRefl*)	64	25	–	26	–	–	–
	Demonstrative pronouns (*PrnDem*)	185	–	–	–	–	–	1
	Negative pronouns (*PrnNeg*)	53	–	1	–	–	–	3
*dtcht-**with**-Subj Pred$_{nf/vl}$-cxn*	Common nouns (*NCmn*)	3238	992	251	12	234	187	351
	Proper nouns (*NProp*)	401	29	6	–	34	9	26
	Personal pronouns *(PrnPers)*	71	7	–	1	1	13	10
	Indefinite pronouns (*PrnIndf*)	36	5	7	1	8	4	2
	Reflexive pronouns (*PrnRefl*)	6	2	3	–	3	–	1
	Demonstrative pronouns (*PrnDem*)	–	–	–	–	–	–	–
	Negative pronouns (*PrnNeg*)	7	5	20	–	20	–	1
*dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn*	Common nouns (*NCmn*)	21	1	1	–	1	1	2
	Proper nouns (*NProp*)	8	1	1	–	–	1	–
	Personal pronouns *(PrnPers)*	14	–	–	–	–	–	–
	Indefinite pronouns (*PrnIndf*)	1	–	–	–	–	1	–
	Reflexive pronouns (*PrnRefl*)	–	–	–	–	–	–	–
	Demonstrative pronouns (*PrnDem*)	–	–	–	–	–	–	–
	Negative pronouns (*PrnNeg*)	–	–	–	–	–	–	–
*dtcht-**without**–Subj Pred$_{nf/vl}$-cxn*	Common nouns (*NCmn*)	51	6	–	1	1	5	–
	Proper nouns (*NProp*)	6	–	1	–	–	–	1
	Personal pronouns *(PrnPers)*	14	–	–	–	–	2	4
	Indefinite pronouns (*PrnIndf*)	9	–	3	–	–	–	1
	Reflexive pronouns (*PrnRefl*)	2	–	–	–	–	–	–
	Demonstrative pronouns (*PrnDem*)	–	–	–	–	–	–	–
	Negative pronouns (*PrnNeg*)	–	–	–	–	–	–	–
*dtcht-**despite**-Subj Pred$_{nf/vl}$-cxn*	Common nouns (*NCmn*)	108	72	130	1	8	10	1
	Proper nouns (*NProp*)	8	1	–	–	–	–	–
	Personal pronouns *(PrnPers)*	10	–	–	–	–	–	–
	Indefinite pronouns (*PrnIndf*)	–	–	–	–	–	–	–
	Reflexive pronouns (*PrnRefl*)	–	–	–	–	–	–	–
	Demonstrative pronouns (*PrnDem*)	–	–	–	–	–	–	–
	Negative pronouns (*PrnNeg*)	–	–	–	–	–	–	–

*H1: The differences between the constructions (dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn, dtcht-**with**-Subj Pred$_{nf/vl}$-cxn, dtcht-**despite**-Subj Pred$_{nf/vl}$-cxn, dtcht-**without**–Subj Pred$_{nf/vl}$-cxn, dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn) within the "SubjPOS" parameter are significant, and the identified dependencies are important and regular.*

The calculations are performed using the following program, written in RStudio environment.

```
library('openxlsx')
file = file.choose()
tab <- read.xlsx(file, sheet = 1, startRow = 1, colNames = TRUE, rowNames = FALSE)
manova_test <- manova(cbind(NCmn, NProp, PrnIndf, PrnRefl, PrnDem, PrnNeg) ~
as.factor(Factor), data=tab)
summary(manova_test)
```

Table 2

Standardized data of the constructions within the "SubjPOS" parameter

Construction		Factors of the "SubjPOS" parameter						
		Subj$_{NCmn}$	*Subj$_{NProp}$*	*Subj$_{PrnPers}$*	*Subj$_{PrnIndf}$*	*Subj$_{PrnRefl}$*	*Subj$_{PrnDem}$*	*Subj$_{PrnNeg}$*
*dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn*	Pred $_{PI}$	7,436028	6,357842	5,587249	5,4161	4,189655	5,231109	4,007333
	Pred $_{PII}$	6,216606	0,693147	0,693147	2,484907	3,295837	0,693147	0,693147
	Pred $_{to\text{-}Inf}$	2,397895	0,693147	1,098612	0,693147	0,693147	0,693147	1,098612
	Pred $_{NP}$	4,025352	0,693147	2,079442	2,302585	3,332205	0,693147	0,693147
	Pred $_{AdjP}$	5,932245	1,098612	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{AdvP}$	4,077537	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{PP}$	5,723585	1,098612	2,772589	3,526361	0,693147	1,098612	1,609438
*dtcht-**with**-Subj Pred$_{nf/vl}$-cxn*	Pred $_{PI}$	8,083329	5,998937	4,290459	3,637586	2,079442	0,693147	2,197225
	Pred $_{PII}$	6,901737	3,433987	2,197225	1,94591	1,386294	0,693147	1,94591
	Pred $_{to\text{-}Inf}$	5,533389	2,079442	0,693147	2,197225	1,609438	0,693147	3,091042
	Pred $_{NP}$	2,639057	0,693147	1,098612	1,098612	0,693147	0,693147	0,693147
	Pred $_{AdjP}$	5,463832	3,583519	1,098612	2,302585	1,609438	0,693147	3,091042
	Pred $_{AdvP}$	5,241747	2,397895	2,70805	1,791759	0,693147	0,693147	0,693147
	Pred $_{PP}$	5,866468	3,332205	2,484907	1,386294	1,098612	0,693147	1,098612
*dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn*	Pred $_{PI}$	3,135494	2,302585	2,772589	1,098612	0,693147	0,693147	0,693147
	Pred $_{PII}$	1,098612	1,098612	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{to\text{-}Inf}$	1,098612	1,098612	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{NP}$	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{AdjP}$	1,098612	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{AdvP}$	1,098612	1,098612	0,693147	1,098612	0,693147	0,693147	0,693147
	Pred $_{PP}$	1,386294	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147
*dtcht-**without**–Subj Pred$_{nf/vl}$-cxn*	Pred $_{PI}$	3,970292	2,079442	2,772589	2,397895	1,386294	0,693147	0,693147
	Pred $_{PII}$	2,079442	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{to\text{-}Inf}$	0,693147	1,098612	0,693147	1,609438	0,693147	0,693147	0,693147
	Pred $_{NP}$	1,098612	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{AdjP}$	1,098612	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{AdvP}$	1,94591	0,693147	1,386294	0,693147	0,693147	0,693147	0,693147
	Pred $_{PP}$	0,693147	1,098612	1,791759	1,098612	0,693147	0,693147	0,693147
***despite**-Subj Pred$_{nf/vl}$*	Pred $_{PI}$	4,70048	2,302585	2,484907	0,693147	0,693147	0,693147	0,693147
	Pred $_{PII}$	4,304065	1,098612	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{to\text{-}Inf}$	4,882802	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{NP}$	1,098612	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147

Table 2

Standardized data of the constructions within the "SubjPOS" parameter

Construction		Factors of the "SubjPOS" parameter						
		$Subj_{NCmn}$	$Subj_{NProp}$	$Subj_{PrnPers}$	$Subj_{PrnIndf}$	$Subj_{PrnRefl}$	$Subj_{PrnDem}$	$Subj_{PrnNeg}$
	Pred $_{AdjP}$	2,302585	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{AdvP}$	2,484907	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147
	Pred $_{PP}$	1,098612	0,693147	0,693147	0,693147	0,693147	0,693147	0,693147

The results of the calculation in the RStudio console are as follows.

```
         Df Pillai approx F num Df den Df  Pr(>F)
as.factor(Factor) 4 1.5559  2.4555   28  108 0.0005129 ***
Residuals     30
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The obtained results show that *Pr(F > F*)* is 0,0005129 and significantly less than 0,01; therefore, the null hypothesis is rejected, and the alternative hypothesis is accepted: *The differences between the constructions (dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn, dtcht-**with**-Subj Pred$_{nf/vl}$-cxn, dtcht-**despite**-Subj Pred$_{nf/vl}$-cxn, dtcht-**without**–Subj Pred$_{nf/vl}$-cxn, dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn) within the "SubjPOS" parameter are significant, and the identified dependencies are important and regular.*

To examine the influence of each of the specified factors on a construction in more detail, one-way analysis of variance (ANOVA) [14], p. 171] is performed. The two statistical hypotheses are formulated:

*H0: The differences between the constructions (dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn, dtcht-**with**-Subj Pred$_{nf/vl}$-cxn, dtcht-**despite**-Subj Pred$_{nf/vl}$-cxn, dtcht-**without**–Subj Pred$_{nf/vl}$-cxn, dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn) within the factor "NCmn" ("NProp"/ "PrnPers"/ "PrnIndf"/ "PrnRefl"/ "PrnDem"/ "PrnNeg") of the "SubjPOS" parameter are insignificant, and the identified dependencies are random.*

*H1: The differences between the constructions (dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn, dtcht-**with**-Subj Pred$_{nf/vl}$-cxn, dtcht-**despite**-Subj Pred$_{nf/vl}$-cxn, dtcht-**without**–Subj Pred$_{nf/vl}$-cxn, dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn) within the factor "NCmn" ("NProp"/ "PrnPers"/ "PrnIndf"/ "PrnRefl"/ "PrnDem"/ "PrnNeg") of the "SubjPOS" parameter are significant, and the identified dependencies are important and regular.*

The results for the linguistic factor "*NCmn*" are as follows:

```
      Df Sum Sq Mean Sq F value  Pr(>F)
Factor    4 108.13 27.031  13.04 2.87e-06 ***
Residuals  30 62.18  2.073
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The obtained results show that the analyzed constructions differ in terms of subjects expressed by common nouns ($p<0,01$). The one-way analysis of variance indicates the existence of differences, but does not specify where these differences are best manifested. One way to solve this issue is to run the Tukey post-hoc test [11].

The script performing all the calculations, including the ANOVA test, is provided below.

```
anova_item <- aov(NCmn ~ Factor, data = tab)
summary(anova_item)
TukeyHSD(anova_item, ordered = FALSE, conf.level = 0.95)
```

The result of the command performing the Tukey test is presented below.

```
Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = NCmn ~ Factor, data = tab)

$Factor
                    diff       lwr       upr     p adj
what_with-despite  -1.6089541 -3.84104924 0.6231410 0.2502620
with-despite        2.6939280  0.46183284 4.9260231 0.0118880
øaug-despite        2.1338836 -0.09821157 4.3659787 0.0663458
without-despite    -1.3275573 -3.55965244 0.9045378 0.4345041
with-what_with      4.3028821  2.07078696 6.5349772 0.0000409
øaug-what_with      3.7428377  1.51074254 5.9749328 0.0003101
without-what_with   0.2813968 -1.95069832 2.5134919 0.9959717
øaug-with          -0.5600444 -2.79213954 1.6720507 0.9483242
without-with       -4.0214853 -6.25358041 -1.7893902 0.0001133
without-øaug       -3.4614409 -5.69353600 -1.2293457 0.0008497
```

As can be seen from the results, the differences in the subjects expressed by common nouns are statistically significant for the pairs of compared constructions 1) *dtcht-**with**-SubjPred$_{nf/vl}$–cxn* and *dtcht-**despite**-SubjPred$_{nf/vl}$–cxn*, 2) *dtcht-**with**-SubjPred$_{nf/vl}$–cxn* and *dtcht-**what_with**-SubjPred$_{nf/vl}$–cxn*, 3) *dtcht-**øaug**-SubjPred$_{nf/vl}$–cxn* and *dtcht-**what_with**-SubjPred$_{nf/vl}$–cxn*, 4) *dtcht-**without**-SubjPred$_{nf/vl}$–cxn* and *dtcht-**with**-SubjPred$_{nf/vl}$–cxn*, 5) *dtcht-**without**-SubjPred$_{nf/vl}$–cxn* and *dtcht-**øaug**-SubjPred$_{nf/vl}$–cxn* (level of significance $p < 0,01$). The use of common nouns in the [Subj] slot is most noticeable in the lingual profile of a ***with***-augmented construction (*dtcht-**with**-SubjPred$_{nf/vl}$–cxn*) and distinguishes this structure from ***despite***-, ***without***- and ***what_with***-augmented constructions.

The same procedure is performed to check other factors of the "SubjPOS" parameter. The obtained data reveal that the use of proper nouns ("NProp"), indefinite pronouns ("PrnIndf"), reflexive pronouns ("PrnRefl"), and negative pronouns ("PrnNeg") is statistically significant in the unaugmented construction (*dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn*) and in ***with***-, ***without***-, ***despite***-, ***what_with***-augmented constructions. The Tukey post-hoc tests prove that most of the differences in the use of nominal ("SubjN") and pronominal ("SubjPrn") subjects are registered between the constructions *dtcht-**øaug**-SubjPred$_{nf/vl}$–cxn*, *dtcht-**with**-SubjPred$_{nf/vl}$–cxn* and the rest of the augmented constructions.

From the statistical analysis that has been carried out, it is possible to conclude: the grammatical constructions *(dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn, dtcht-**with**-Subj Pred$_{nf/vl}$-cxn, dtcht-**despite**-Subj Pred$_{nf/vl}$-cxn, dtcht-**without**–Subj Pred$_{nf/vl}$-cxn, dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn)* reveal significant differences in the realization of the POS parameter of their subjects; significant differences between the analyzed constructions are also proved by the one-factor analysis of variance for such factors of the "SubjPOS" parameter as common ("NCmn") and proper ("NProp") nouns, indefinite ("PrnIndf"), reflexive ("PrnRefl") and negative ("PrnNeg") pronouns, with the Tukey post-hoc tests pointing out that these differences are mostly manifested in two out of five constructions: the unaugmented construction (*dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn*) and ***with***-augmented construction (*dtcht-**with**-Subj Pred$_{nf/vl}$-cxn*).

However, for the factors “PrnPers” and “PrnDem” such differences are not registered. Thus, it can be assumed that the factors with statistically significant differences are distinguishing features that determine the variation of the *dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn, dtcht-**with**-Subj Pred$_{nf/vl}$-cxn, dtcht-**despite**-Subj Pred$_{nf/vl}$-cxn, dtcht-**without**–Subj Pred$_{nf/vl}$-cxn, dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn* in terms of part of speech representation of the subject constituent.

3.3. Machine learning tools in linguistic research

Problems of classification constitute the most significant area in modern machine learning [23]. Diverse approaches and methods are used to resolve them (naïve Bayesian classifier, linear classifier, neural network-based classifiers, etc.). The application of these methods requires rigorous analysis of the data structure and preliminary preparation. The task of classification is to predict the category to which the object belongs, which is described in advance by the predefined and predetermined features. In our study, we assess the possibility of using the specified linguistic factors as features that allow predicting the type of the construction (*dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn, dtcht-**with**-Subj Pred$_{nf/vl}$-cxn, dtcht-**despite**-Subj Pred$_{nf/vl}$-cxn, dtcht-**without**–Subj Pred$_{nf/vl}$-cxn, dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn*) within the “SubjPOS” on other data.

The essence of linear discriminant analysis is to find an additional axis (axes) passing through the entire set of points. Each point is a grammatical construction represented in the coordinate system of the factors so that their projections on it will provide maximum division into classes [24]. In general, the location of the axis is determined by a linear discriminant function, verifying the influence of each feature (in our case, a factor of the “SubjPOS" parameter) based on calculated weight coefficients.

To perform a linear discriminant analysis in R, we implement a specialized MASS package [25] - [26] to the data in Table 2. The code to run the calculations is represented below.

```
#Part01
library('openxlsx')
library('caret')
library('MASS')

#Part02
file = file.choose()
tab <- read.xlsx(file,sheet = 1, startRow = 1, colNames = TRUE,rowNames =
FALSE)

#Part03
set.seed(101)
training.pattern <- createDataPartition(y = tab$Factor, p = 0.75, list =
FALSE)
train.data <- tab[training.pattern, ]
test.data <- tab

#Part04
lda_data <- lda(Factor ~ ., data = train.data)
lda_data

#Part05
predictions <- predict(lda_data, test.data)
p1 <- predictions$class

conf_tab <- table(Predicted = p1, Actual = test.data$Factor)
```

```
conf_tab
```

The given code includes 5 essential parts, that are executed sequentially. The first two parts of the code load the modules necessary for the proper operation of the program ('openxlsx' allows to process data stored in Excel spreadsheets; 'caret' provides the instruments to manipulate data; 'MASS' contains tools for building classification models, including linear discriminant analysis) and upload the necessary data.

In the third part, training and test samples are formed. The training sample is organized using a random 75% selection from the main sample. The test sample is selected in such a way as to completely coincide with the original data, as the table contains data of only 35 records, which can affect the accuracy of the model.

The fourth and fifth parts of the code are responsible for the construction of the model, predicting and forming a confusion matrix, according to which the effectiveness of the model is evaluated.

The implementation of the algorithms yields the following results:

```
Call:
lda(Factor ~ ., data = train.data)

Prior probabilities of groups:
 despite what_with   with    øaug   without
   0.2    0.2    0.2    0.2    0.2

Group means:
     NCmn   Nprop   PrnPers  PrnIndf  PrnRefl  PrnDem  PrnNeg
despite  3.295575 1.0289644 0.9917738 0.6931472 0.6931472 0.6931472 0.6931472
what_with 1.078982 0.8958797 0.6931472 0.7607247 0.6931472 0.6931472 0.6931472
with   5.643849 3.0311544 2.0143510 2.1622796 1.3451510 0.6931472 1.9519190
øaug   5.297316 1.7724181 1.9229819 2.2511349 1.7096801 1.5170516 1.4658042
without  1.814336 0.9917738 1.3383473 1.0448494 0.8086717 0.6931472 0.6931472

Coefficients of linear discriminants:
         LD1    LD2    LD3     LD4
NCmn   -0.96209007 0.3074754 -0.4417541 -0.09665136
Nprop   1.25729791 -1.3793323 -1.2415410 1.34337166
PrnPers -0.84265305 0.5735341 1.6491720 -2.79813014
PrnIndf  0.76273962 -0.4403514 0.2120885 2.66082681
PrnRefl -0.20590083 0.5419865 1.3014925 -1.33678748
PrnDem  0.07938331 1.6698469 -1.5249577 1.39087349
PrnNeg  -1.35746749 -0.6037894 1.2383849 -1.72655019

Proportion of trace:
  LD1  LD2  LD3  LD4
0.5806 0.3106 0.0604 0.0483

      Actual
Predicted  despite what_with  with  øaug without
 despite     4     0  2     1    1
 what_with    2     6  0     0    3
 with      0     0  5     0    0
 øaug      1     0  0     5    0
 without     0     1  0     1    3
```

The data obtained require some explanation. The record Prior probabilities of groups shows that each construction produces the same effect, although these are somewhat idealized conditions. The record Group means presents average values for each of the factors for a particular construction. The record Coefficients of linear discriminants specifies a linear combination of a new axis, which will delimit the analyzed constructions as much as possible, and because there are 5 syntactic patterns under scrutiny, there will be one axis less, namely 4. After the record Proportion of trace, the delimitation of objects on each of the new axes is presented in descending order of conditional "force". The data indicate that the LD1 and LD2 axes are the

most strongly separated. However, a confusion matrix is more important for assessing the effectiveness of the model [27][28]. The confusion matrix is built using the commands:

```
conf_tab <- table(Predicted = p1, Actual = test.data$Factor)
conf_tab
```

The confusion matrix is the table 5×5 (Table 3), where the current values of the constructions are presented in the columns, and the predicted ones are given in the rows. The number of predicted structures will be placed at the intersection of the row and the column. The main diagonal of the matrix will display the number of correctly executed classifications by the newly constructed model. According to the results obtained, in our case there will be 23 of 35 records presented in the test sample. This allows assessing the effectiveness of the created classifier, namely Accuracy, i.e. the ratio of correctly executed predictions to the total number of constructions in the test sample. The calculations are presented in formula (1).

$$\text{Accuracy} = \frac{23}{35} = 0{,}657, \tag{1}$$

The obtained assessment is somewhat general and does not reflect in which cases the model works better, and where the classifier submits incorrect data. To resolve this issue, we will use such performance characteristics of the classifier model as precision and recall. Precision is defined as the ratio of the number of correct predictions to the number of all predictions made by the classifier for a particular construction (class). Recall is calculated as the ratio of the number of correctly classified objects to the number of current objects in a particular class (construction). The data for the analysis are presented in Table 3.

According to the table data, the constructed model is the most effective for the constructions *dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn* and *dtcht-**with**-Subj Pred$_{nf/vl}$-cxn* and is less effective for the *dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn* construction. These findings are confirmed by the value of the F-measure (harmonic mean for Precision and Recall) for each of the examined constructions (2).

$$\begin{aligned} &F_{\text{despite}} = 0{,}53; \\ &F_{\text{what with}} = 0{,}67; \\ &F_{\text{with}} = 0{,}83; \\ &F_{\text{aug}} = 0{,}77; \\ &F_{\text{without}} = 0{,}5. \end{aligned} \tag{2}$$

Table 3

Confusion matrix and Precision and Recall results

Constructions		Current values						
		despite	**what with**	**with**	**øaug**	**without**		
Predicted values	**despite**	4	0	2	1	1	*0,5*	*Precision*
	what with	2	6	0	0	3	*0,55*	
	with	0	0	5	0	0	*1*	
	øaug	1	0	0	5	0	*0,83*	
	without	0	1	0	1	3	*0,6*	
		0,57	*0,86*	*0,71*	*0,71*	*0,43*		
		Recall						

The obtained results are supported by their graphic representation. The graph based on LD1 and LD2 clearly indicates the differences between the analyzed constructions (Fig. 2). It is evident, that green and blue dots (representing *dtcht-**with**-Subj Pred$_{nf/vl}$-cxn* and *dtcht-**øaug**-*

Subj Pred$_{nf/vl}$-cxn) lie far from other dots (indicating ***despite***-, ***what_with***- and ***without***-augmented constructions) and are also separated from each other.

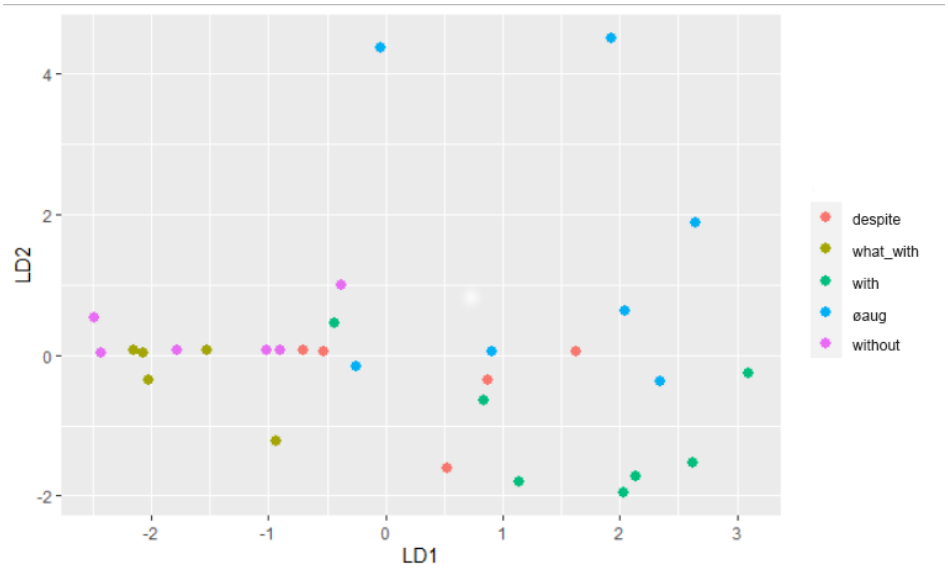

Fig. 2. Graphic representation of the linguistic classifier model

Summing up, the findings of this part of the study suggest that

1. The overall effectiveness of the machine learning model to solve the problem of classification of grammatical constructions (*dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn, dtcht-**with**-Subj Pred$_{nf/vl}$-cxn, dtcht-**despite**-Subj Pred$_{nf/vl}$-cxn, dtcht-**without**–Subj Pred$_{nf/vl}$-cxn, dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn*) within the "SubjPOS" parameter is insufficient (accuracy = 0,657). To increase the accuracy of the model, it is necessary to either enlarge the training sample or change the classification method.

2. Despite the insufficient overall accuracy, the model effectively classifies the *dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn* and *dtcht-**with**-Subj Pred$_{nf/vl}$-cxn* constructions, while other constructions are more difficult to classify.

3. The specified factors of the “SubjPOS” parameter prove to be distinguishing characteristics for two out of five analyzed constructions (namely, *dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn* and *dtcht-**with**-Subj Pred$_{nf/vl}$-cxn*). Provided an effective model is constructed, these factors can be used to determine types of grammatical constructions based on their numerical indicators.

3.4. Educational and methodological significance of the presented interdisciplinary research

The research methodology incorporating heterogeneous fields of science is quite complex and characterized by high requirements for the object, subject, and methods of research. Moreover, the results obtained at the intersection of scientific fields are undoubtedly more significant and have bigger perspectives for practical implementation. For instance, the integration of such fields of knowledge as philology, computer science, statistics, and data analysis has facilitated the emergence of recent frameworks of corpus and applied linguistics.

Systems of automatic text translation and human language recognition, developed on methods and tools of artificial intelligence, are also progressing quite dynamically. Thus, if to consider the presented case study (which serves as an illustration of the application of methods of statistical analysis, machine learning, and software R in the field of linguistics) in

the context of interdisciplinary research, the number of significant methodological aspects are singled out:

✓ The selection, analysis, and preparation of data obtained as a result of processing the linguistic corpus crucially influence the quality of research.

✓ Elaborating the strategy for application of statistical methods allows to substantiate and validate the obtained results.

✓ Developing a classification model and evaluating its effectiveness using machine learning methods provides an opportunity to apply the findings to similar situations in the analysis of statistical information from other language corpora.

The outlined aspects determine how effectively the data will be generalized, organized, and systematized. The use of statistical methods provides evidence and reliability of the results, which is also a very important component of modern research.

It should be noted that the presented case study on utilizing the statistical complex R can be adapted for research of data collected from corpora of other languages.

Application of R statistical software and machine learning methods in modern linguistic research is not merely of scientific importance but can also facilitate the training process at the educational-professional and educational-scientific levels in universities. According to the program of the specialty 035 Philology (specialization – 035.10 Applied Linguistics) at Zhytomyr State Ivan Franko University, students master a number of academic disciplines (Corpus linguistics, Intelligent Web Data Analysis, Programming, and Probability Theory), which lay the foundations for the integration of methods and tools of computer science into contemporary linguistic research [29]. In addition, while working on course and diploma projects, future specialists can use the methodological developments of the elaborated case study to conduct their own research implementing both the suggested statistical software or other modern software for data analysis. The knowledge gained from the conducted case study will increase their awareness of innovative technologies for processing vast amounts of linguistic data as well as develop their research skills for involving methods and tools from other fields of knowledge.

Concerning master and postgraduate students' training, a comprehensive study of the stages of the R software application for the analysis of linguistic data is important for promoting the use of statistical methods and tools to test research assumptions and hypotheses. Moreover, considering the possibilities of the software package R, future scholars in the field of philology will master the standards of scientific research, globally practiced by scientists, and particularly in English-speaking countries. More than that, statistical and analytical quantifications in linguistic corpus-oriented studies prevailingly involve vast applications of the software R. The most apparent advantages of R, compared to such costly and resource-intensive software as Statistica and SPSS, include free distribution and access to the library of additional modules which enhance its capabilities and allow to quickly adapt to specific tasks. Despite the undeniable benefits of the software R, linguists in this country are still not active proponents of it.

Thus, the presented research does not only reveal the linguistic properties of the detached nonfinite constructions in present-day English but also demonstrates (and therefore popularizes) the effectiveness of the R complex and the tools of machine learning and data analysis in philological research.

4. CONCLUSIONS AND FUTURE RESERACH

From the research that has been undertaken, the following conclusions can be drawn.

1. The development of modern computer technologies and software in the field of statistical data analysis has expanded the possibilities of their application in various fields of

science, particularly in linguistics they have advanced the development of corpus and computational linguistics.

2. Statistical software R is one of the most accessible and efficient analytical tools for processing vast arrays of digitalized language data.

3. The results of a statistical analysis of the "SubjPOS" parameter of English detached nonfinite constructions with an explicit subject (*dtcht-**øaug**-Subj Pred$_{nf/vl}$-cxn, dtcht-**with**-Subj Pred$_{nf/vl}$-cxn, dtcht-**despite**-Subj Pred$_{nf/vl}$-cxn, dtcht-**without**–Subj Pred$_{nf/vl}$-cxn, dtcht-**what_with**-Subj Pred$_{nf/vl}$-cxn*) demonstrate the advantages of the functionality of R complex for testing statistical hypotheses, data analysis, and construction of machine learning models. The conducted research reveals how the study of language phenomena and processes can benefit from the application of statistical software and specialized open libraries.

4. Carrying out philological research that involves machine learning and data analysis practices, it is necessary to take into account methodological aspects, including selection, analysis, preparation of numerical data, elaboration of strategy for the application of statistical methods, etc.

5. The presented interdisciplinary case study is of significant educational value. It is worth considering with undergraduate students, masters and graduate students as an example of effective application of recent advances in information technology, machine learning, and data analysis in linguistic corpus-oriented research.

The findings presented in this paper open new vistas for future research. Obviously, further studies incorporating methods and tools of machine learning based on the statistical software complex R into corpus-driven linguistics will be of considerable interest. In our future research, we intend to validate the suggested strategy for statistical and computer analysis to investigate other linguistic parameters of the grammatical constrictions under study and statistically verify the determining parameters (factors) that condition functional dynamics and variability of the network of detached nonfinite constructions with an explicit subject in present-day English.

REFERENCES (TRANSLATED AND TRANSLITERATED)

[1] R. Fox, "The Contribution of Linguistics Towards Transdisciplinarity in Organizational Discourse." *International Journal of Transdisciplinary Research*, no. 1(4), pp.16 – 34, 2009. (in English)

[2] L. A. Janda, *Cognitive linguistics: the quantitative turn*. Berlin: De Gruyter Mouton, 2013. doi: https://doi.org/10.1515/9783110335255. (in English)

[3] L. A. Janda, "Linguistic profiles: A quantitative approach to theoretical questions." *Language and Method*, no. 3, pp.127-145. 2016. (in English)

[4] G. Desagulier, *Corpus linguistics and statistics with R. Introduction to quantitative methods in linguistics*. Cham: Springer International Publishing, 2017. doi: https://doi.org/10.1007/978-3-319-64572-8. (in English)

[5] M. V. Kopotev, *Principles of syntactic idiomaticity*. Helsinki: Helsinki University Press, 2008. (in Russian)

[6] *The R Project for Statistical Computing*. [Online]. Available: http://www.R-project.org/ (in English)

[7] *Comprehensive R archive network*. [Online]. Available: https://cran.r-project.org.

[8] V. V. Zhukovska, O. O. Mosiiuk, & V. V. Komarenko, (2018). "Using R in the research by future philologists." *Information Technologies and Learning Tools*, vol.66(4), pp.272-285, 2018. doi: https://doi.org/10.33407/itlt.v66i4.2196. (in Ukrainian)

[9] V. Brezina, *Statistics in corpus linguistics*. Cambridge: Cambridge University Press, 2018. doi: https://doi.org/10.1017/9781316410899. (in English)

[10] S. Gries, *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement (Open linguistics series)*. New York, London: Continuum International Publishing Group Ltd., 2003. (in English)

[11] S. Gries, *Statistics for Linguistics with R: A Practical Introduction (Mouton Textbook)*. Berlin/Boston: De Gruyter Mouton., 2013. (in English)

[12] G. Desagulier, *Corpus linguistics and statistics with R*. Cham: Springer., 2017. doi: https://doi.org/10.1007/978-3-319-64572-8. (in English)

[13] R. Baayen, *Analyzing linguistic data*. Cambridge: Cambridge University Press. 2008. doi: https://doi.org/10.1017/CBO9780511801686. (in English)

[14] N. Levshina, *How to do linguistics with R*. Amsterdam: John Benjamins Publishing., 2015. doi: https://doi.org/10.1075/z.195. (in English)
[15] J. Klavan, M. Pilvik, & K. Uiboaed, "The use of multivariate statistical classification models for predicting constructional choice in spoken, non-standard varieties of Estonian." *SKY Journal of Linguistics*. [Online], no. 28, pp.187-224. 2015. Available: http://www.linguistics.fi/julkaisut/SKY2015/SKYJoL28_Klavan.pdf (in English)
[16] D. Divjak, & A. Arppe, 2013. "Extracting prototypes from exemplars What can corpus data tell us about concept representation?" *Cognitive Linguistics*, no.24(2), pp.221-274, 2013. doi: https://doi.org/10.1515/cog-2013-0008. (in English)
[17] A. E. Goldberg, *Explain me this: Creativity, Competition, and the Partial Productivity of Constructions*. Princeton/ Oxford : Princeton University Press, 2019. doi: https://doi.org/10.1515/9780691183954. (in English)
[18] M. Hilpert, "Constructional Approaches," in *The Oxford Handbook of English Grammar*. B. Aarts, J. Bowie, G. Popova (eds). Oxford: Oxford University Press, pp.106-123. 2020. doi: https://doi.org/10.1093/oxfordhb/9780198755104.013.13. (in English)
[19] J. Bybee, "From usage to grammar: The mind's response to repetition." *Language*, no.82, pp.711 – 733, 2006. (in English)
[20] J. Bybee, "Usage-based Theory and Exemplar Representations of Constructions", in *The Oxford Handbook of Construction Grammar*, T. Hoffmann, G. Trousdale (eds.) Oxford: Oxford University Press, pp.49 - 69, 2013. (in English)
[21] *BNC-BYU*. (2020, Dec. 20). [Online]. Available: www.english-corpora.org/bnc/. (in English)
[22] A. B. Shipunov, E. M. Baldin, P. A. Volkova, A. I. Korobeinikov, S. A. Nazarova, S. V. Petrov, V. G. Sufiyanov, (2021, July 27). *Visual statistics. Use R!*, [Online]. Available: https://cran.r-project.org/doc/contrib/Shipunov-rbook.pdf (in Russian)
[23] Yu. V. Nikolskyi, V. V. Pasichnyk, Yu. M. Shcherbyna, *Artificial intelligence systems*, Lviv, 2015. (in Ukrainian)
[24] *Discriminant Analysis Essentials in R - Articles - STHDA*. (2021, July 27). [Online]. Available: http://www.sthda.com/english/articles/36-classification-methods-essentials/146-discriminant-analysis-essentials-in-r/#linear-discriminant-analysis---lda. (in English)
[25] *Package MASS*. (2021, July 27). [Online]. Available: https://cran.r-project.org/web/packages/MASS/MASS.pdf. (in English)
[26] M. Kuhn, Building predictive models in R using the caret package. *Journal of Statistical Software*, no.28(5). 2008. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v028i05/v28i05.pdf. (in English)
[27] L. Coelho, and W. Richert, *Building Machine Learning Systems with Python*. Packt Publishing, 2013. (in English)
[28] S. Narkhede, *Understanding Confusion Matrix*. (2021, July 27). [Online] Medium. Available: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62. (in English)
[29] Educational-professional program «Applied Linguisitcs» (2021, July 27). [Online]. Available: https://eportfolio.zu.edu.ua/media/ StudyProgram/99/6dx45d.pdf (in English)

Text of the article was accepted by Editorial Team 06.08.2021

СТАТИСТИЧНЕ ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ R У КОРПУСО-КЕРОВАНИХ ДОСЛІДЖЕННЯХ ТА МАШИННОМУ НАВЧАННІ

Жуковська Вікторія Вікторівна
кандидат філологічних наук, доцент,
доцентка кафедри міжкультурної комунікації та прикладної лінгвістики
Житомирський державний університет імені Івана Франка, м. Житомир, Україна
ORCID ID 0000-0002-4622-4435
victoriazhukovska@gmail.com

Мосіюк Олександр Олександрович
кандидат педагогічних наук, доцент кафедри комп'ютерних наук та інформаційних технологій
Житомирський державний університет імені Івана Франка, м. Житомир, Україна
ORCID ID 0000-0003-3530-1359
mosxandrwork@gmail.com

Анотація. Динамічний розвиток обчислювальної техніки, мережевих технологій та прикладного програмного забезпечення уможливлює широке використання спеціалізованих статистичних комплексів для вирішення різного типу і складності завдань не лише в межах класичних напрямів застосування інформаційних технологій (статистиці, інженерії, штучному інтелекті), а й у мовознавстві. Статистична система аналізу даних R є одним із найпопулярніших аналітичних інструментів оброблення великих масивів диджиталізованих мовних даних, особливо у квантитативно-корпусних розвідках Західної Європи та Північної Америки. Запропонована стаття розкриває переваги застосування функціоналу програмного комплексу R для виконання складних статистичних аналізів лінгвальних даних у корпусо-керованих дослідженнях та в машинному навчанні для створення лінгвістичних класифікаторів. З цією метою у роботі запропоновано стратегію комп'ютерно-статистичного аналізу лінгвальних корпусних даних, що складається з трьох послідовних етапів: 1) опрацювання й стандартизація даних для застосування статистичних методів, 2) застосування методів перевірки статистичних гіпотез (MANOVA, ANOVA) та апостеріорного тесту Тьюкі, 3) створення моделі лінгвістичного класифікатора та аналіз її ефективності. У результаті застосування запропонованої стратегії до 11 000 токенів англійських відокремлених нефінітних конструкцій з експліцитним суб'єктом, відібраних з корпусу BNC-BYU, встановлено статистично значущі відмінності в реалізації лінгвальних факторів параметру "Частиномовна приналежність суб'єкту" та побудовано машинну модель класифікації досліджуваних конструкцій у корпусному матеріалі. Окремим питанням розглянуто методологічні аспекти міжпредметних досліджень з лінгвістики та комп'ютерних наук та окреслено можливості практичного застосування презентованого кейсу в підготовці бакалаврів, магістрів та аспірантів у галузі прикладної лінгвістики. Стаття містить необхідні статистичні дані, представлені в таблицях, та код, написаний із застосуванням скрипту R. Усі матеріали супроводжуються детальним описом та поясненнями. У підсумку аналізуються отримані результати та окреслюються перспективи подальших досліджень, які пов'язуються з популяризацією статистичного програмного комплексу R та підвищенням обізнаності фахівців з цією статистичною системою аналізу.

Ключові слова: корпусна лінгвістика; модель машинного навчання; лінгвістичний класифікатор; статистична система аналізу даних R; RStudio; граматична конструкція; лінгвальний параметр; однофакторний дисперсійний аналіз (ANOVA); багатофакторний дисперсійний аналіз (MANOVA); апостеріорний тест Тьюкі; дискримінантний аналіз; методологічні аспекти міждисциплінарних досліджень.

СТАТИСТИЧЕСКОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ R В КОРПУСО-УПРАВЛЯЕМЫХ ИССЛЕДОВАНИЯХ И МАШИННОМ ОБУЧЕНИИ

Жуковская Виктория Викторовна
кандидат филологических наук, доцент
доцент кафедры межкультурной коммуникации и прикладной лингвистики
Житомирский государственный университет имени Ивана Франко, г. Житомир, Украина
ORCID ID 0000-0002-4622-4435
victoriazhukovska@gmail.com

Мосиюк Александр Александрович
кандидат педагогических наук,
доцент кафедры компьютерных наук и информационных технологий
Житомирский государственный университет имени Ивана Франко, г. Житомир, Украина
ORCID ID 0000-0003-3530-1359
mosxandrwork@gmail.com

Аннотация. Динамическое развитие вычислительной техники, сетевых технологий и прикладного программного обеспечения позволяет широко использовать специализированные статистические комплексы для решения различного типа и сложности задач не только в пределах классических направлений применения информационных технологий (статистике, инженерии, искусственном интеллекте), но и в языкознании. Как следствие, наблюдается экспоненциальное увеличение числа прикладных языковедческих исследований, в частности в таких технологически ориентированных отраслях, как корпусная и компьютерная лингвистика. Статистическая система анализа данных R

является одним из популярнейших аналитических инструментов обработки больших массивов диджитализированных языковых данных, особенно в квантитативно-корпусных исследованиях Западной Европы и Северной Америки. Предложенная статья раскрывает преимущества применения функционала программного комплекса R для выполнения сложных статистических анализов лингвальных данных в корпусоуправляемых исследованиях и в машинном обучении для создания лингвистических классификаторов. С этой целью в работе предложено стратегию компьютерно-статистического анализа лингвальных корпусных данных, которая включает три последовательных этапа: 1) разработка и стандартизация данных для применения статистических методов, 2) применение методов проверки статистических гипотез (MANOVA, ANOVA) и апостериорного теста Тьюки, 3) создание модели лингвистического классификатора и анализ ее эффективности. В результате применения предложенной стратегии к 11 000 токенов английских обособленных нефинитных конструкций с эксплицитным субъектом, отобранных из корпуса BNC-BYU, установлено статистически значимые различия в реализации лингвальных факторов параметра “Частеречная принадлежность субъекта” и построено машинную модель классификации исследуемых конструкций в корпусном материале. Отдельным вопросом рассмотрены методологические аспекты междисциплинарных исследований в лингвистике и компьютерных науках, а также указаны возможности практического использования представленного кейса в подготовке бакалавров, магистров и аспирантов в области прикладной лингвистики. Статья содержит необходимые статистические данные, представленные в таблицах, и код, написанный с применением скрипта R. Все материалы сопровождаются подробным описанием и объяснениями. В выводах анализируются полученные результаты и определяются перспективы дальнейших исследований, которые связываются с популяризацией статистического программного комплекса R и повышением осведомленности специалистов с этой статистической системой анализа.

Ключевые слова: корпусная лингвистика; модель машинного обучения; лингвистический классификатор; статистическая система анализа данных R; RStudio; грамматическая конструкция; лингвальный параметр; однофакторный дисперсионный анализ (ANOVA); многофакторный дисперсионный анализ (MANOVA); апостериорный тест Тьюки; дискриминантный анализ; методологические аспекты междисциплинарных исследований.