Register Distribution of English Detached Nonfinite/ Nonverbal with Explicit Subject Constructions: a Corpus-Based and Machine-Learning Approach

Viktoriia Zhukovska[1], Oleksandr Mosiiuk[1] and Solomija Buk[2]

[1] *Zhytomyr Ivan Franko State University, Velyka Berdychivska str. 40, Zhytomyr, 10008, Ukraine*
[2] *Ivan Franko National University of Lviv, Universytetska str. 1, Lviv, 79000, Ukraine*

Abstract

This article presents the findings of a quantitative corpus-based analysis of the register distribution of detached nonfinite/ nonverbal with explicit subject constructions in present-day English. Despite substantial research on the linguistic diversity of the syntactic patterns under study, no quantitative corpus and machine-learning analysis of the distribution of all their types in modern English registers has been presented. Thus, the statistical platform R was employed to accomplish two goals: 1) to undertake a quantitative corpus-based analysis of register distribution of the analyzed clauses in the BNC corpus and 2) to assess the possibility of occurrence of the clauses under research in the registers of present-day English on the basis of a machine-learning model. The findings of this study provide compelling evidence for the applicability of an integrated quantitative corpus linguistic and machine learning analysis for investigating the linguistic behavior of complex clause-level constructions, such as English detached nonfinite/ nonverbal with explicit subject constructions. The obtained results refute the prevalent view in contemporary English grammars that detached nonfinite/ nonverbal with explicit subject constructions have a limited scope of use and demonstrate that the analyzed syntactic patterns expand their register distribution, penetrating both the written and spoken registers of contemporary formal and informal English discourse.

Keywords

Cognitive-quantitative construction grammar, clause-level construction, statistical platform, integrated approach

1. Introduction

Since the 1990s, the field of linguistics has undergone a significant methodological shift. It gradually reopened the empirical methods of corpus and experimental linguistics, transforming itself from a primarily rationalist discipline. This shift in methodology has changed the way linguists approach and analyze language, allowing for more accuracy and precision in their analysis. Consequently, the application of quantitative methods appears to have altered “the ecology of methodology in linguistics research” [1, p. 4], and text corpora have become “the alpha and omega of linguistics” [2, p. 8]. The use of corpus data has become indispensable in many areas of language study [3, p. 114], including those traditionally favoring a rationalist approach, such as syntax.

This research focuses on detached nonfinite/ nonverbal clauses with an explicit subject in English. The following contexts from the British National Corpus (BNC) [4] illustrate the clauses under consideration (1–5):

(1) *Katherine sat silently for a long moment*, [$_{\text{ØAUG}}$[$_{\text{NP}}$***her eyes***] [$_{\text{XP}}$***growing perceptibly wider***]], [[$_{\text{NP}}$***the color***] [$_{\text{XP}}$***draining from her cheeks***]] (BNC; FNT);

(2) Nathan was standing defiantly, [$_{\text{ØAUG}}$[$_{\text{NP}}$***hands***] [$_{\text{XP}}$***in pockets***]], near the window (BNC; AD9);

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine

EMAIL: victoriazhukovska@gmail.com (V. Zhukovska); mosxandrwork@gmail.com (O. Mosiiuk); solomija@gmail.com (S. Buk)
ORCID: 0000-0002-4622-4435 (V. Zhukovska); 0000-0003-3530-1359 (O. Mosiiuk); 0000-0001-8026-3289 (S. Buk)

(3) *The Coroner was in his early forties, a gaunt, greying man,* [[$_{\text{AUG}}$***with***] [$_{\text{NP}}$***thick spectacles***] [$_{\text{XP}}$***perched at the very end of his nose***]] (BNC; CES);

(4) *Its recommendation for a global oil strategy has so far received no recognition,* [[$_{\text{AUG}}$***despite***] [$_{\text{NP}}$***oil being***] [$_{\text{XP}}$***the lifeblood of industrial (modern) society***]] (BNC; B1W);

(5) She had a head start, of course, [[$_{\text{AUG}}$***what with***] [$_{\text{NP}}$***her mother***] [$_{\text{XP}}$***being immaculate too***]] (BNC; HGJ).

The ***purpose*** of this study is to investigate the register distribution of detached nonaugmented ($_{\text{ØAUG}}$) and augmented ($_{\text{AUG}}$) (*with, without, despite, what with*) nonfinite and nonverbal clauses with explicit subject in present-day English. With this in mind, two ***objectives*** are attained: 1) to undertake a quantitative corpus-based analysis of register distribution of the analyzed clauses in the BNC and 2) to assess the possibility of occurrence of the clauses under research in the registers of present-day English on the basis of a machine-learning model. Integrating quantitative corpus linguistics and machine learning approaches, this study contributes to a better understanding of the distributional properties of detached nonfinite/ nonverbal clauses with explicit subject in contemporary English.

2. Related Works

Researchers from various linguistic trends and schools have repeatedly drawn attention to English nonfinite/ nonverbal clauses with explicit subject: *descriptive grammar* (Quirk et al. (1985), Timofeeva (2011), Kortmann (2013)); *generative grammar* (Riemsdijk (1978), McCawley (1983), Beukema, Felser, Britain (2007)); *corpus linguistics* (Duffley, Dion-Girardeau (2015), Fonteyn, van de Pol (2015)); *functional systemic grammar* (He, Yang (2015), Khamesian (2016)); *construction grammar* (Riehemann, Bender (1999), Bouzada-Jabois, Pérez-Guerra (2016)). Despite the number of studies conducted, the linguistic diversity of the investigated nonfinite/ nonverbal clauses raises a number of questions that have not yet been conclusively answered by previous research. Specifically, the distribution of the clauses in the registers of present-day English needs to be further examined using the most recent developments of the cognitive-quantitative linguistic framework.

English detached nonfinite/ nonverbal clauses with explicit subject are conventionally regarded as rare, archaic Latinisms mainly used in official discourse [5, p. 250; 6, p. 95]. Therefore, the 'formal vs. informal' nature of a text or communication situation serves as the main criterion for differentiating the spheres of their use. According to R. Quirk and his co-authors, nonfinite/ nonverbal clauses with an overt subject are rather formal and uncommon in contemporary English [7, p. 1120]. This assertion is supported by the diachronic studies, which show that in Old English such clauses were predominantly Latin borrowings and were most likely used by educated people in official texts [8]. During the Middle English and New English periods (at least until 1660), these syntactic patterns were characteristic of classical, bookish, and scientific style and were widely used in religious and legal texts [9]. Modern usage recognizes such clauses as stylistically marked syntactic structures used more frequently in written, especially in formal and narrative, texts [10, p. 122], than in spoken ones, except for a few cliché expressions such as *present company excepted, all told, weather/ time permitting, God willing* [7, p. 1120]. Otherwise, subordinate clauses are almost always used in speech where nonfinite/ nonverbal clauses with explicit subject may appear in writing.

Previous research on the distributional properties of English nonfinite and nonverbal clauses with explicit subject has primarily presented synchronic or diachronic accounts on some types of nonfinite clauses in texts of specific registers [5; 6; 10], however, no comprehensive quantitative corpus and machine-learning analysis of the distribution of all their types in modern English registers has been presented.

3. Theoretical and methodological background

The presented study is based on the theoretical and methodological foundations of the new framework of contemporary grammar studies – cognitive-quantitative construction grammar (CQCxGr). The framework is built on the synergy of the *theoretical tenets* advocated by cognitive linguistics (Langacker (1987; 1991); Janda, (2013)) and the constructionist approach (specifically, the updated version of the Berkley construction grammar (Fillmore (1988); Östman, Fried, (2004)), cognitive construction grammar (Goldberg (1995; 2006; 2019)), and usage-based construction grammar (Hoffmann (2016)*;* Hilpert (2019)) and the *methodological principles* of quantitative linguistics (Levytskyi (2007)), quantitative corpus linguistics (Gries, Stefanowitsch (2004, 2013); Brezina (2018)), automatic speech processing (Darchuk (2013)), and experimental linguistics (Gillioz, Zufferey (2020)), thereby providing a competent qualitative-quantitative approach for examining general and idiosyncratic features of language units.

The epistemological guidelines of cognitive-quantitative construction grammar entail providing an explanation for the semiotic phenomena of language and speech on their mental basis and developing a psychologically plausible description of language as one of the many cognitive and social systems available to humans. From the framework's perspective, language constitutes a repertoire of generalized 'form-meaning' pairings – *constructions* – of various degrees of schematicity and complexity. As non-compositional, (fully) productive, cognitively entrenched (automated), and complex units, *constructions* are holistic semiotic models for language representation – syntax, morphology, and vocabulary – stored in a *constructicon*, a structured inventory of taxonomic *networks of constructions* [11; 12].

A comprehensive account of the linguistic properties of a particular *construction* is the result of the analysis of interrelated parameters of its form and meaning (prosodic, morphological, syntactic, semantic, distributional, functional, pragmatic, etc.). The research toolkit of the cognitive-quantitative construction grammar is determined by a usage-based orientation to language study, extensive corpus data reference, active use of quantitative methods, and the application of specialized computer programs for processing massive arrays of linguistic data. From the usage-based perspective, the mental *constructicon* of speakers emerges as a result of recurrent interaction with language expressions (*constructions*), with frequency of occurrence playing a key role in the cognitive entrenchment of *constructions*. Consequently, corpus data are employed to explore *constructions* that conceptualize fundamental human experience and/or are frequently used in a language community. Large arrays of linguistic data cannot be efficiently analyzed without specialized software, which encourages researchers in the field to utilize high-tech, sophisticated methods of quantitative analysis with computer support. These methods open up new avenues for language research and have the potential to solve numerous theoretical and practical aspects of language research.

From the perspective of the cognitive-quantitative construction grammar, detached nonfinite/ nonverbal clauses with explicit subject acquire the status of grammatical *constructions*, which we nominate as *'D(etached) N(on)F(inite)/ N(on)V(erbal) (with) E(xplicit) S(ubject) constructions'* (hereinafter referred to as *DNF/NVES-constructions*). *DNF/NVES-constructions* as syntactic structures that include a nonfinite/nonverbal clause belong to the class of syntagmatically complex *clause-level constructions*.

The *DNF/NVES-constructions* constitute a taxonomic constructional network, with every node representing an individual type of *construction*. The given taxonomic network is organized around a constructional schema (*macro-construction – (dt-SubjPred*$_{NF/NV}$*–cxn)*), the characteristics of which are inherited by less abstract *meso-constructions* and further acquired by individual *micro-constructions* (*dt-**øaug**-Subj Pred*$_{NF/NV}$*–cxn, dt-**with**-Subj Pred*$_{NF/NV}$*–cxn, dt-**despite**-Subj Pred*$_{NF/NV}$*–cxn, dt-**without**–Subj Pred*$_{NF/NV}$*–cxn, dt-**what_with**-Subj Pred*$_{NF/NV}$*–cxn* {*N(on)F(inite):* PI, PII, to-Inf; *N(on)V(erbal):* NP, AdjP, AdvP, PP}). As *micro-constructions* are the most linguistically rich patterns, they serve as the basis for linguistic and quantitative analysis.

4. Experiment: corpus, data, statistical software R and computer-quantitative procedure

Today, corpora are used to solve a variety of issues, ranging from linguistic analysis to data mining and machine learning. Corpora have paved the way for the development of programs for automatic text translation and natural language processing, as they provide a large-scale data set to facilitate the training and testing of various algorithms. The use of specialized statistical software to examine data sets from certain corpora and then construct machine learning models from them is one of the most significant developments in the field.

In this study, linguo-quantitative analysis is carried out for English *DNF/NVES-constructions*, collected from the British National Corpus (https://www.english-corpora.org/) [4]. The data were retrieved automatically between 2018 and 2020 using the in-built BNC search engine. In total, the queries yielded 650 724 tokens that were then manually inspected to avoid spurious hits and formally similar but functionally different constructions. After removing the false hits, the database includes 11 000 tokens for analysis.

The analysis of the distributional characteristics of the *DNF/NVES-constructions* involves a quantitative linguistic corpus analysis of their distribution in the registers of contemporary English, reflected in the parameter "Register distribution" (RegDSTN).

The basis for distinguishing the registers of the contemporary English language is the typology of registers established by the developers of the British National Corpus, where registers are "language varieties associated with a particular combination of situational characteristics and communicative purpose" [13, p. 436]. The BNC includes texts of such registers as spoken *(Spoken)*, newspaper *(Newspaper)*, magazine *(Magazine)*, fiction *(Fiction)*, academic *(Academic)*, nonacademic (popular academic) *(Non-academic)* and unclassified *(Miscellaneous)* texts. Each of the registers is represented by a number of genres. For instance, *Fiction* is represented by drama (*W_fict_drama*), poetry (*W_fict_poetry*), and prose (*W_fict_prose*). Unclassified texts include advertisements, biographies, e-mails, school and university essays, etc. (For more information on the codification of genres in the British National Corpus, see [14]) (6–9):

(6) [*With the grass being so long*]$_{\text{DNF/NVES-construction}}$, you know? (SP:PS066) (unclear) grass (unclear) wasn't it (BNC; KBP) – *S_conv*;

(7) [*His breath ragged*]$_{\text{DNF/NVES-construction}}$, [*his eyes near wild*]$_{\text{DNF/NVES-construction}}$, he stared at her, and it came to him then that he wanted it all: the house, the money, and Theda, too (BNC; HE) – *W_fict_prose*;

(8) [*Despite these views being diametrically opposed*]$_{\text{DNF/NVES-construction}}$, both exist simultaneously in attitudes to retired people (BNC; CE1) – *W_ac_soc_science;*

(9) Both parties feeling that they have achieved an agreement they can live with, [*without it being constantly undermined*]$_{\text{DNF/NVES-construction}}$ (BNC; CFV) – *W_advert*.

The analysis of the *DNF/NVES-constructions* in terms of their distribution by registers is carried out in the factors "spoken texts" (RegSpkn), "fiction texts" (RegFict), "magazine texts" (RegMag), "newspaper texts" (RegNews), "non-academic texts" (RegNonAc), "academic texts" (RegAc), and "unclassified texts" (RegMisc).

The frequency of constructions in the corpus indicates the degree of their entrenchment in the language community and correlates with the number of tokens associated with the corresponding parameter/ factor. The verification of the data retrieved from the BNC and the establishment of statistically significant indicators are carried out using a three-stage linguistic and quantitative procedure that involves the consistent application of the following statistical metrics: 1) multivariate analysis of variance (MANOVA), 2) one-factor analysis of variance (ANOVA) and 3) Tukey's multiple comparison method. The obtained results are used to build a machine learning model (linear discriminant analysis) to predict the register distribution of the *DNF/NVES-constructions* outside the corpus. All quantifications are performed using the statistical data analysis platform *R*.

The statistical platform *R* is one of the most frequently used software applications in philological research. This software is distributed as an open source program with a large number of free libraries designed to solve problems of varying levels of complexity.

5. Results/ Discussion

5.1. Register distribution of the *DNF/NVES-constructions*: a quantitative corpus-based analysis

A sample of 11 000 tokens of the micro-constructions of the network of English *DNF/NVES-constructions (dt-**øaug**-Subj Pred$_{NF/NV}$–cxn, dt-**with**-Subj Pred$_{NF/NV}$–cxn, dt-**despite**-Subj Pred$_{NF/NV}$–cxn, dt-**without**–Subj Pred$_{NF/NV}$–cxn, dt-**what_with**-Subj Pred$_{NF/NV}$–cxn)* has been quantitatively processed.

Table 1 displays the raw frequencies of the analyzed micro-constructions depending on the type of the nonfinite and nonverbal predicate in the registers of the BNC.

Table 1
Raw frequencies of the micro-constructions within the “RegDSTN” parameter

Micro-construction	**Factors of the “RegDSTN” parameter**	**Frequency of a micro-construction**						
		Pred$_{NF}$			**Pred$_{NV}$**			
		PI	PII	to-Inf	NP	AdjP	AdvP	PP
*dt-**øaug**-* *Subj Pred$_{NF/NV}$–cxn*	spoken texts (RegSpkn)	63	–	–	1	–	–	8
	fiction texts (RegFict)	1681	434	5	33	337	48	262
	magazine texts (RegMag)	116	17	–	7	1	–	4
	newspaper texts (RegNews)	188	9	1	3	2	2	26
	non-academic texts (RegNonAc)	295	18	2	15	4	2	32
	academic texts (RegAc)	258	15	–	22	4	2	4
	unclassified texts (RegMisc)	458	41	3	13	28	3	20
*dt-**with**-* *Subj Pred$_{NF/NV}$–cxn*	spoken texts (RegSpkn)	56	7	2	1	4	13	5
	fiction texts (RegFict)	559	193	49	2	68	73	70
	magazine texts (RegMag)	319	77	33	1	21	26	32
	newspaper texts (RegNews)	723	138	35	5	34	25	51
	non-academic texts (RegNonAc)	639	185	64	–	66	29	61
	academic texts (RegAc)	456	156	34	–	32	3	60
	unclassified texts (RegMisc)	996	285	67	5	75	44	112
*dt-**what_with**-* *Subj Pred$_{NF/NV}$–cxn*	spoken texts (RegSpkn)	6	–	–	–	–	–	–
	fiction texts (RegFict)	18	2	2	–	1	3	1
	magazine texts (RegMag)	6	–	–	–	–	–	–
	newspaper texts (RegNews)	5	–	–	–	–	–	2
	non-academic texts (RegNonAc)	–	–	–	–	–	–	–
	academic texts (RegAc)	4	–	–	–	–	–	–
	unclassified texts (RegMisc)	5	–	–	–	–	–	–
*dt-**without**–* *Subj Pred$_{NF/NV}$–cxn*	spoken texts (RegSpkn)	5	–	–	–	–	–	1
	fiction texts (RegFict)	18	–	3	–	–	6	3
	magazine texts (RegMag)	11	1	–	–	–	–	–
	newspaper texts (RegNews)	2	–	–	–	–	–	–
	non-academic texts (RegNonAc)	8	1	–	–	–	1	–
	academic texts (RegAc)	19	2	1	–	1	–	–
	unclassified texts (RegMisc)	19	2	–	1	–	–	1
*dt-**despite**-* *Subj Pred$_{NF/NV}$–*	spoken texts (RegSpkn)	2	–	1	–	–	–	–
	fiction texts (RegFict)	26	3	10	1	–	3	–
	magazine texts (RegMag)	12	7	14	–	3	–	–
	newspaper texts (RegNews)	32	6	16	–	–	2	–
	non-academic texts (RegNonAc)	13	20	33	–	1	1	–

academic texts (RegAc)	22	23	32	–	1	1	–
unclassified texts (RegMisc)	19	17	24	–	3	3	1

According to the data in Table 1, there is a clear connection between the frequency of micro-constructions with a particular type of predicate and a register. The unaugmented *dt-**øaug**-Subj Pred*$_{NF/NV}$*–cxn* micro-construction tends to be most strongly associated with the fiction register, demonstrating a significantly lower frequency of use in academic/non-academic texts, newspaper and magazine texts, and it appears to be least frequently used in spoken texts. The augmented *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn* micro-construction performs nearly identically in fictional, non-academic, and newspaper texts. However, if newspaper and magazine articles are taken to represent mass media discourse generally, the indicators change slightly. The *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn* micro-construction is used most frequently in mass media texts and is almost as prevalent in popular academic and fictional texts. The highest usage rates of the augmented *dt-**despite**-Subj Pred*$_{NF/NV}$*–cxn* are found in academic texts, followed by popular academic and newspaper texts. The least frequently micro-construction occurs in informal speech. The high frequency of the augmented *dt-**without**–Subj Pred*$_{NF/NV}$*–cxn* and *dt-**without**–Subj Pred*$_{NF/NV}$*–cxn* micro-constructions in literary texts is correlated with the lowest frequency in colloquial texts. However, in academic writing, *dt-**without**–Subj Pred*$_{NF/NV}$*–cxn* is slightly more common.

The statistical significance of the observed quantitative differences is examined using a three-stage computer and statistical strategy. At first, multivariate analysis of variance (MANOVA) is employed to statistically validate the differences between the constructions in terms of factors within "RegDSTN" parameter realization. Second, using one-factor analysis of variance (ANOVA), the statistically significant differences in the use of micro-constructions for each of the selected factors are determined. When such differences exist, Tukey's multiple comparison is used to confirm the results and determine which pairs of micro-constructions a particular factor is significant for. The calculations are performed with the statistical software *R* and its freely available libraries.

As seen in Table 1, the frequency of constructions is represented by discrete interval values, some data are missing, and the difference between the minimum and maximum values is substantial. Therefore, for the designed computer and statistical strategy to be implemented, the collected data must be standardized. Consequently, several data transformations are carried out. Initially, missing data are replaced with zero values. The data are then transformed logarithmically to produce continuous interval data using the formula: $\ln(x_{ij} + const)$, where x_{ij} is the table value and $const$ is set to 2. Because $\ln(0 + 1) = 0$, it is possible to set any positive number other than 1 [15, p. 5]. Subsequent calculations are performed on the standardized data provided in Table 2.

Table 2
Standardized data of the micro-constructions within the "RegDSTN" parameter

Micro-construction	**Factors of the "RegDSTN" parameter**	**Factors of the "RegDSTN" parameter**						
		RegSpkn	RegFict	RegMag	RegNews	RegNonAc	RegAc	RegMisc
*dt-**øaug**-Subj Pred*$_{NF/NV}$*–cxn*	Pred$_{\text{PI}}$	4,1743873	7,4283	4,7707	4,7875	5,6937	5,5607	6,1312
	Pred$_{\text{PII}}$	0,6931472	6,1862	2,9444	2,3979	2,9957	2,8332	3,7612
	Pred$_{\text{to-Inf}}$	0,6931472	1,9459	0,6931	1,0986	1,3863	0,6931	1,6094
	Pred$_{\text{NP}}$	1,0986123	3,5553	2,1972	1,6094	2,8332	3,1781	2,7081
	Pred$_{\text{AdjP}}$	0,6931472	5,826	1,0986	1,3863	1,7918	1,7918	3,4012
	Pred$_{\text{AdvP}}$	0,6931472	3,912	0,6931	1,3863	1,3863	1,3863	1,6094
	Pred$_{\text{PP}}$	2,3025851	5,5759	1,7918	3,3322	3,5264	1,7918	3,091
dt - ***with-***	Pred$_{\text{PI}}$	4,060443	6,3297	5,7714	6,5862	6,1269	6,9058	8,2324
	Pred$_{\text{PII}}$	2,1972246	5,273	4,3694	4,9416	5,0626	5,6595	6,9489

Micro-construction	Factors of the "RegDSTN" parameter	Factors of the "RegDSTN" parameter						
		RegSpkn	RegFict	RegMag	RegNews	RegNonAc	RegAc	RegMisc
	Pred$_{\text{to-Inf}}$	1,3862944	3,9318	3,5553	3,6109	3,5835	4,2341	5,6664
	Pred$_{\text{NP}}$	1,0986123	1,3863	1,0986	1,9459	0,6931	1,9459	2,7726
	Pred$_{\text{AdjP}}$	1,7917595	4,2485	3,1355	3,5835	3,5264	4,3438	5,7104
	Pred$_{\text{AdvP}}$	2,7080502	4,3175	3,3322	3,2958	1,6094	3,8286	5,3706
	Pred$_{\text{PP}}$	1,9459101	4,2767	3,5264	3,9703	4,1271	4,7362	5,9738
*dt-**what_with**-Subj Pred$_{NF/NV}$–cxn*	Pred$_{\text{PI}}$	2,0794415	2,9957	2,0794	1,9459	0,6931	1,7918	1,9459
	Pred$_{\text{PII}}$	0,6931472	1,3863	0,6931	0,6931	0,6931	0,6931	0,6931
	Pred$_{\text{to-Inf}}$	0,6931472	1,3863	0,6931	0,6931	0,6931	0,6931	0,6931
	Pred$_{\text{NP}}$	0,6931472	0,6931	0,6931	0,6931	0,6931	0,6931	0,6931
	Pred$_{\text{AdjP}}$	0,6931472	1,0986	0,6931	0,6931	0,6931	0,6931	0,6931
	Pred$_{\text{AdvP}}$	0,6931472	1,6094	0,6931	0,6931	0,6931	0,6931	0,6931
	Pred$_{\text{PP}}$	0,6931472	1,0986	0,6931	0,6931	0,6931	0,6931	0,6931
*dt-**without**–Subj Pred$_{NF/NV}$–cxn*	Pred$_{\text{PI}}$	1,9459101	2,9957	2,5649	1,3863	2,3026	3,0445	3,0445
	Pred$_{\text{PII}}$	0,6931472	0,6931	1,0986	0,6931	1,0986	1,3863	1,3863
	Pred$_{\text{to-Inf}}$	0,6931472	1,6094	0,6931	0,6931	0,6931	1,0986	0,6931
	Pred$_{\text{NP}}$	0,6931472	0,6931	0,6931	0,6931	0,6931	0,6931	1,0986
	Pred$_{\text{AdjP}}$	0,6931472	0,6931	0,6931	0,6931	0,6931	1,0986	0,6931
	Pred$_{\text{AdvP}}$	0,6931472	2,0794	0,6931	0,6931	1,0986	0,6931	0,6931
	Pred$_{\text{PP}}$	1,0986123	1,6094	0,6931	0,6931	0,6931	0,6931	1,0986
*dt-**despite**-Subj Pred$_{NF/NV}$–cxn*	Pred$_{\text{PI}}$	1,3862944	3,3322	2,6391	3,5264	2,7081	3,1781	3,0445
	Pred$_{\text{PII}}$	0,6931472	1,6094	2,1972	2,0794	3,091	3,2189	2,9444
	Pred$_{\text{to-Inf}}$	1,0986123	2,4849	2,7726	2,8904	3,5553	3,5264	3,2581
	Pred$_{\text{NP}}$	0,6931472	1,0986	0,6931	0,6931	0,6931	0,6931	0,6931
	Pred$_{\text{AdjP}}$	0,6931472	0,6931	1,6094	0,6931	1,0986	1,0986	1,6094
	Pred$_{\text{AdvP}}$	0,6931472	1,6094	0,6931	1,3863	1,0986	1,0986	1,6094
	Pred$_{\text{PP}}$	0,6931472	0,6931	0,6931	0,6931	0,6931	0,6931	1,0986

On the first stage, statistically significant differences in the frequency distribution of the *DNF/NVES-constructions* in the BNC registers are measured. Multivariate analysis of variance (MANOVA) [25, p. 198] is employed to statistically substantiate the differences between the micro-constructions (*dt-**øaug**-Subj Pred$_{NF/NV}$–cxn, dt-**with**-Subj Pred$_{NF/NV}$–cxn, dt-**despite**-Subj Pred$_{NF/NV}$–cxn, dt-**without**–Subj Pred$_{NF/NV}$–cxn, dt-**what_with**-Subj Pred$_{NF/NV}$–cxn*) in terms of realization of the "RegDSTN" parameter. In Table 2, the factors of the "RegDSTN" parameter (independent variables) are displayed in columns, and their values are represented in rows. The following statistical hypotheses are developed:

H0: The differences between the micro-constructions within the "RegDSTN" parameter are insignificant, and the measured dependencies are random.

H1: The differences between the micro-constructions within the "RegDSTN" parameter are significant, and the measured dependencies are important and regular.

The program is run in the RStudio console to perform MANOVA:

```
   library('openxlsx')

   file = file.choose()
   tab <- read.xlsx(file, sheet = 1, startRow = 1, colNames = TRUE, rowNames = FALSE)
   manova_test <- manova(cbind(RegSpkn, RegFict, RegMag, RegNews, RegNonAc, RegAc,
RegMisc) ~ as.factor(Construction), data=tab)
   summary(manova_test)
```

The results of MANOVA test are as follows.

```
                        Df Pillai approx F num Df den Df    Pr(>F)    
as.factor(Construction)4 2.0144   3.9132     28    108 1.651e-07 ***
Residuals               30                                          
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The results show that *Pr(F > F*)* is 1.651^{e-07} and significantly less than 0,01; therefore, the null hypothesis is rejected, and the alternative hypothesis is accepted: The differences between the micro-constructions (*dt-**øaug**-Subj Pred$_{NF/NV}$–cxn, dt-**with**-Subj Pred$_{NF/NV}$–cxn, dt-**despite**-Subj Pred$_{NF/NV}$–cxn, dt-**without**–Subj Pred$_{NF/NV}$–cxn, dt-**what_with**-Subj Pred$_{NF/NV}$–cxn*) within the “RegDSTN” parameter are significant, and the measured dependencies are important for distinguishing their prevalent spheres of usage.

The second stage is aimed at examining the impact of each of the specified factors within the “RegDSTN” parameter on the analyzed constructions. For this purpose, one-way analysis of variance (ANOVA) [16, p. 171] is carried out. For each of the factors specified, the two statistical hypotheses are reformulated:

H0: The differences between the micro-constructions within the factor “RegSpkn” (“RegFict”/ “RegMag”/ “RegNews”/ “RegNonAc”/ “RegAc”/ “RegMisc”) of the the “RegDSTN” parameter are insignificant, and the identified dependencies are random.

H1: The differences between the micro-constructions within the factor “RegSpkn” (“RegFict”/ “RegMag”/ “RegNews”/ “RegNonAc”/ “RegAc” “RegMisc”) of the the “RegDSTN” parameter are important and regular.

The following are the results obtained for the factor *“RegSpkn”*:

```
            Df Sum Sq Mean Sq F value Pr(>F)  
Construction  4  9.017  2.2543   3.409 0.0206 *
Residuals     30 19.838  0.6613                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The results indicate statistically significant differences at the 95% significance level (Pr(>F)=0,0206<0,05) between the micro-constructions under study within the factor “RegSpkn”, i.e. the frequency of certain micro-constructions in spoken texts can be a factor differentiating them from the rest of the analyzed constructions. In general, the results of the ANOVA test revealed statistically significant differences among the micro-constructions based on the factors “literary texts” (Pr(>F)=4,22^{e-06}<0,001), “magazine texts” (Pr(>F)=0,000454<0,001), “newspaper texts” (Pr(>F)=1,54^{e-05}<0,001), and “non-academic texts” (Pr(>F)=3,97^{e-05}<0,001).

As ANOVA indicates the presence of differences but does not specify where these differences are best manifested, the third stage requires the application of the Tukey post-hoc test. All the calculations, including the ANOVA test, are performed by the script provided:

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

Below is the output of the command that calculates the Tukey test:

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

Fit: aov(formula = RegSpkn ~ Construction, data = tab)

$Construction
                          diff         lwr        upr     p adj
what_with-despite    0.04109744 -1.21970747 1.30190235 0.9999808
with-despite         1.31966450  0.05885959 2.58046941 0.0366813
øaug -despite        0.62821869 -0.63258622 1.88902360 0.6043795
```

```
without-despite      0.07994511 -1.18085980 1.34075002 0.9997284
with-what_with       1.27856706  0.01776215 2.53937197 0.0455838
øaug-what_with       0.58712125 -0.67368366 1.84792616 0.6625967
without-what_with    0.03884767 -1.22195724 1.29965258 0.9999846
øaug-with           -0.69144581 -1.95225072 0.56935910 0.5145378
without-with        -1.23971939 -2.50052430 0.02108552 0.0557355
without- øaug       -0.54827358 -1.80907849 0.71253134 0.7160811
```

The results indicate that the following pairs of compared micro-constructions have the greatest differences in use in spoken texts (level of significance p < 0,05): 1) *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn* and *dt-**despite**-Subj Pred*$_{NF/NV}$*–cxn*; 2) *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn* and *dt-**what_with**-Subj Pred*$_{NF/NV}$*–cxn*.

The Tukey's multiple comparison method applied to other factors revealed that the greatest number of statistically significant differences are found in the factor "literary texts" (RegFict) between 6 pairs of micro-constructions, in the factor "newspaper texts" (RegNews) and "academic texts" (RegAc) between 4 pairs, in the factor "magazine texts" (RegMag) and "non-academic texts" (RegNonAc) between 3 pairs, and in the factor "spoken texts" between 2 pairs of micro-constructions.

Among the constructions, the greatest statistically significant differences are recorded between *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn* construction and *without*-, *despite*-, *what_with*-augmented micro-constructions. The indicators of *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn* and *dt-**what_with**-Subj Pred*$_{NF/NV}$*–cxn* differ in all of the identified factors. The *dt-**with**-Subj Pred*$_{NF/NV}$*– cxn* and *dt-**despite**-Subj Pred*$_{NF/NV}$*–cxn* constructions do not differ only in the factor "academic texts", the *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn* and *dt-**without**–Subj Pred*$_{NF/NV}$*–cxn* do not differ in the factor "spoken texts", but the differences between these micro-constructions in the other factors are significant.

The unaugmented micro-construction *dt-**øaug**-Subj Pred*$_{NF/NV}$*–cxn* demonstrates statistically significant differences with *despite*-, *what_with*-, *without*-augmented constructions in terms of occurrence in fiction texts, but does not show differences in this factor with *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn*.

Multiple Tukey's comparisons revealed no statistically significant differences with respect to the analyzed factors between the *dt-**despite**-Subj Pred*$_{NF/NV}$*–cxn, dt-**without**–Subj Pred*$_{NF/NV}$*–cxn* and *dt-**what_with**-Subj Pred*$_{NF/NV}$*–cxn* micro-constructions. The findings demonstrate that the distribution of these micro-constructions across registers in the BNC is homogeneous, with a general tendency toward low usage in spoken text types and prevalence in written texts.

Based on the results of a three-stage linguistic and quantitative procedure on the BNC, it is evident and statistically proven that certain registers (factors) have a greater influence on the occurrence of the *DNF/NVES-constructions* in them, i.e., certain micro-constructions tend to occur more frequently in texts of certain registers than others. At this point of our research, the question arises: Can the established register distribution of the *DNF/NVES-constructions* be extrapolated beyond the BNC? To answer this question, the data obtained in the quantitative corpus-based procedure are subjected to machine-learning modeling. The machine-learning model will probabilistically predict the distribution of the analyzed constructions in present-day English usage.

5.2. Register distribution of the *DNF/NVES-constructions*: a machine-learning approach

To predict the register distribution of the *DNF/NVES-constructions* beyond the analyzed corpus, it is essential to assess the viability of building a machine learning model to classify the constructions based on statistical test results. Consequently, linear discriminant analysis [17, p. 667] is employed: 1) to build a model for classifying *dt-**øaug**-Subj Pred*$_{NF/NV}$*–cxn, dt-**with**-Subj Pred*$_{NF/NV}$*–cxn, dt-**despite**-Subj Pred*$_{NF/NV}$*–cxn, dt-**without**–Subj Pred*$_{NF/NV}$*–cxn, dt-**what_with**-Subj Pred*$_{NF/NV}$*–cxn* constructions, given that the statistical indicators will determine their register distribution in the corpus; 2) to identify the variables (i.e., factors "spoken texts" (RegSpkn), "fiction texts" (RegFict), "magazine texts" (RegMag), "newspaper texts" (RegNews), "non-academic texts" (RegNonAc), "academic texts" (RegAc), and "unclassified texts" (RegMisc)) that contribute most to the separation of constructions.

The objective of linear discriminant analysis is to find an additional axis (axes) that will pass through the entire set of points (each point is a construction represented in the coordinate system of categories) so that their projections on it will provide the greatest possible separation between classes [18]. The location of such an axis is determined by a linear discriminant function, which defines the impact of each feature (specifically, the corpus register) using the calculated coefficients.

The data from Table 2 and the specialized package MASS [19; 20] are used to build the model for linear discriminant analysis in *R*. Presented is the code performing the calculations:

```
library('openxlsx')
library('caret')
library('MASS')
file = file.choose()
tab <- read.xlsx(file,sheet = 1, startRow = 1, colNames = TRUE,rowNames = FALSE)
set.seed(101)
training.pattern <- createDataPartition(y = tab$Category, p = 0.75, list = FALSE)
train.data <- tab[training.pattern, ]
test.data <- tab
lda_data <- lda(Category ~ ., data = train.data)
lda_data
predictions <- predict(lda_data, test.data)
p1 <- predictions$class
conf_tab <- table(Predicted = p1, Actual = test.data$Category)
conf_tab
```

Since the authors explain each command in detail in the article [15], only the analysis of the results is provided here.

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

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

Group means:
            RegSpkn  RegFict    RegMag   RegNews  RegNonAc     RegAc   RegMisc
despite   0.8762492 1.737047 1.7674338 1.8781271 2.0408021 2.1356103 2.2607577
what_with 0.6931472 1.212066 0.6931472 0.6931472 0.6931472 0.6931472 0.6931472
with      2.2070640 4.247804 3.5437580 3.9939997 3.4336548 4.4862832 5.7835696
øaug      1.5415935 5.145737 1.9986316 2.3981321 2.7966955 2.3428092 3.2672570
without   0.9695185 1.460676 1.0726917 0.8086717 1.0965419 1.2681451 1.3357226

Coefficients of linear discriminants:
                LD1        LD2         LD3        LD4
RegSpkn  -1.0291488  0.1787200 -1.71021035 -0.1257059
RegFict  -1.2299897  1.1801267 -0.03610643  0.2861375
RegMag   -1.2218508 -1.1909062  0.94767175  1.0348112
RegNews   2.3230869 -0.5053956  1.67740983  1.8661933
RegNonAc -2.1161569  0.7294720  0.46712888 -1.6034093
RegAc    -0.5048724 -1.0978915 -0.92176444 -0.8206950
RegMisc   3.7993088  0.8826389 -0.68935723 -0.5295016

Proportion of trace:
   LD1    LD2    LD3    LD4
0.8144 0.1663 0.0148 0.0044


           Actual
Predicted   despite what_with  with  øaug without
despite         6         0     0     0       1
what_with       1         7     0     0       2
with            0         0     7     0       0
øaug            0         0     0     6       0
without         0         0     0     1       4
```

As demonstrated by the results, the greatest separation exists along the LD1 and LD2 axes. The weight coefficients enable to determine the contribution of each variable (register) to distinguishing between the objects (more precisely, *dt-**øaug**-Subj Pred*$_{NF/NV}$*–cxn*, *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn*, *dt-**despite**-Subj Pred*$_{NF/NV}$*–cxn*, *dt-**without**–Subj Pred*$_{NF/NV}$*–cxn*, *dt-**what_with**-Subj Pred*$_{NF/NV}$*–cxn* micro-constructions). According to the obtained results for the first linear discriminant function, the most significant for the separation of the *DNF/NVES-constructions* are “unclassified texts” (RegMisc), “newspaper texts” (RegNews) and “non-academic texts” (RegNonAc). In the LD2 function, the most significant are “fiction texts” (RegFict), “magazine texts” (RegMag) and “academic texts” (RegAc).

The confusion matrix [21, p. 217-218; 22], however, is more important for assessing the model’s effectiveness. It is constructed using the commands:

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

The confusion matrix is a five-by-five table (Table 3), with the columns representing the actual values of the constructions and the rows displaying the predicted values. The number of predicted constructions is located at the intersection of the row and column. The main diagonal of the matrix represents the number of correct classifications performed by the newly constructed model. The obtained results show that these are 30 out of 35 records in the test sample. This allows for determining one of the estimates of the created classifier’s effectiveness, namely $_{\text{Accuracy}}$, i.e., the proportion of accurate predictions to the total number of test sample constructions. Formula (1) describes the calculation:

$$Accuracy = \frac{30}{35} = 0{,}857, \tag{1}$$

The conducted classification of the constructions is reasonably accurate, indicating that the created model is significantly more effective than the model created in the study of the *DNF/NVES-constructions* based on part-of-speech data [15]. To estimate the precision and recall values for this model the F-measure (harmonic mean of Precision and Recall) for each construction is calculated. Each calculation is displayed in the table (Table 3).

The data in Table 3 show that the created model is effective for classifying all the analyzed constructions, except for ***without***-augmented one. The value of F-measure (harmonic mean for Precision and Recall) for each construction confirms this (2).

$$\begin{aligned} &\mathrm{F}_{\text{despite}} = 0{,}86; \\ &\mathrm{F}_{\text{what\_with}} = 0{,}82; \\ &\mathrm{F}_{\text{with}} = 1; \\ &\mathrm{F}_{\text{øaug}} = 0{,}92; \\ &\mathrm{F}_{\text{without}} = 0{,}67; \end{aligned} \tag{2}$$

Table 3

Confusion matrix and Precision and Recall results

		Actual values					
		despite	what_with	with	øaug	without	*Precision*
Predicted values	**despite**	6	0	0	0	1	*0,857*
	what_with	1	7	0	0	2	*0,7*
	with	0	0	7	0	0	*1*
	øaug	0	0	0	6	0	*1*
	without	0	0	0	1	4	*0,8*
	Recall	*0,857*	*1*	*1*	*0,857*	*0,57*	

Importantly, a number of notable outcomes emerge from the analysis of the collected data:

1) The overall efficiency (*Accuracy = 0.857*) of the machine learning model for classifying *dt-**øaug**-Subj Pred*$_{NF/NV}$*–cxn*, *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn*, *dt-**despite**-Subj Pred*$_{NF/NV}$*–cxn*, *dt-**without**–Subj Pred*$_{NF/NV}$*–cxn*, *dt-**what_with**-Subj Pred*$_{NF/NV}$*–cxn* micro-constructions in the BNC registers is quite

high, therefore it can be used to predict the occurrence of the *DNF/NVES-constructions* in the specified registers outside the analyzed corpus.

2) The model is the most effective in separating between *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn* and *dt-**øaug**-Subj Pred*$_{NF/NV}$*–cxn* micro-constructions. Less effectively it classifies *dt-**despite**–Subj Pred*$_{NF/NV}$*–cxn* and *dt-**what_with**-Subj Pred*$_{NF/NV}$*–cxn* micro-constructions. The least effective the model separates *dt-**without**–Subj Pred*$_{NF/NV}$*–cxn* micro-construction.

3) “Unclassified texts” (RegMisc), “newspaper texts” (RegNews), “non-academic texts” (RegNonAc), “fiction texts” (RegFic), “magazine texts” (RegMag), and “academic texts” (RegAc) have the most weight in separating the micro-constructions.

The dot diagram displays the results of the linear discriminant analysis for the register distribution of the *DNF/NVES-constructions* in the BNC (Fig. 1):

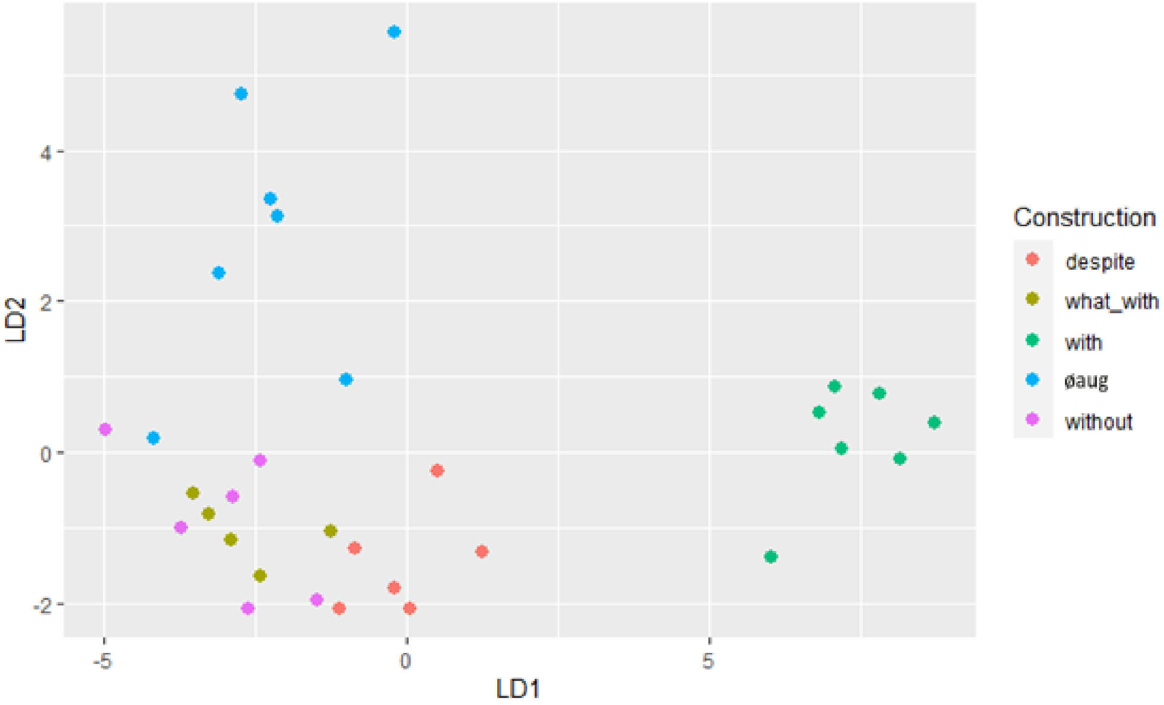

Figure 1: Graphic representation of the linear discriminant analysis of register distribution for the *DNF/NVES-constructions* in the BNC

As can be seen from Fig. 1, the data demonstrate a clear distinction between two micro-constructions – *dt-**with**-Subj Pred*$_{NF/NV}$*–cxn* and *dt-**øaug**-Subj Pred*$_{NF/NV}$*–cxn*, lending credibility to the results of Tukey’s aposteriori tests. Therefore, it is reasonable to assume that these micro-constructions will be distinguished similarly in the investigated registers outside the BNC. However, it is also necessary to consider the difficulty of distinguishing between the *dt-**despite**–Subj Pred*$_{NF/NV}$*–cxn*, *dt-**without**–Subj Pred*$_{NF/NV}$*–cxn* and *dt-**what_with**-Subj Pred*$_{NF/NV}$*–cxn* micro-constructions, which reduces the overall accuracy of the model (<0.9) and prevents us from drawing conclusions about the model’s overall effectiveness for the entire language. To address the limitations of this model, additional research is required, such as the construction of a new lda model with a larger sample size or the implementation of an alternative method.

6. Conclusions

The results of this study conclusively demonstrate the applicability of an integrated quantitative corpus linguistic and machine learning analysis to the investigation of the linguistic behavior of complex clause-level *constructions*, such as English *detached nonfinite/nonverbal with explicit subject constructions*.

The analysis of the register distribution of the English *DNF/NVES-constructions* reveals that these syntactic patterns are more productive in present-day English usage than the data of diachronic studies indicate. At the current stage, the English *DNF/NVES-constructions* exhibit a steady tendency toward further improvement and development, as evidenced by the analysis of their distribution by registers, based on a representative sample from the British National Corpus. The observed results refute the prevalent view in modern English grammars that these constructions have a limited scope of use and prove that the *DNF/NVES-constructions* expand their distribution, penetrating both the written and spoken registers of contemporary English discourse.

The findings presented in this paper show the need for future investigations. Clearly, additional studies of the analyzed syntactic patterns from the cognitive-quantitative construction grammar standpoint will be of great interest. The proposed computerized linguo-quantitative strategy will be used to investigate other linguistic parameters (positional, referential, functional, etc.) of the analyzed *constructions* and statistically validate the determining parameters (factors) that affect the functional dynamics and variability of the network of *detached nonfinite/ nonverbal with explicit subject constructions* in present-day English.

7. References

[1] Sh. Liao, L. Lei, What we talk about when we talk about corpus: A bibliometric analysis of corpus-related research in linguistics (2000–2015), Glottometrics 38, 2017. 1–20.

[2] G. Desagulier, Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics, Springer International Publishing, Cham, 2017.

[3] В.В. Жуковська, Лінгвістичний корпус як новітній інформаційно-дослідницький інструментарій сучасного мовознавства, Вчені записки ТНУ імені В.І. Вернадського. Серія: Філологія. Соціальні комунікації. Том 31 (70), №3 (2020) 113–119. doi: 10.32838/2663-6069/2020.3-1/20

[4] British National Corpus (BNC), 2023. URL: https://www.english-corpora.org/bnc/

[5] Q. He, B. Yang, A Corpus-based approach to the genre and diachronic distributions of English absolute clauses, Journal of Quantitative Linguistics 22 (2015) 250–272. doi: 10.1080/09296174.2015.1037160

[6] N. Aljović, Non-finite Clauses in English: Formal Properties and Function. Sarajevo, 2017.

[7] R. Quirk, S. Greembaum, G. Leech, J. Svartvik, A Comprehensive Grammar of the English Language, Longman, New York, 1985.

[8] O. Timofeeva, Latin Absolute constructions and their Old English equivalents: Interfaces between form and information structure, in: A. Meurman-Solin, M. J. Lopez-Couso, B. Los (Eds.), Information Structure and Syntactic Change in the History of English. Oxford Academic, New York, 2012. pp. 228–242. doi: 10.1093/acprof:oso/9780199860210.001.0001

[9] N. van de Pol, Between copy and cognate: the origin of absolutes in Old and Middle English, in: L. Johanson, M. Robbeets (Eds.), Copies versus Cognates in Bound Morphology. Brill Academic Publishers, Leiden, Boston, 2012, pp. 297–322.

[10] C. B. Bouzda-Jabois, Nonfinite supplements in the recent history of English, Universida de Vigo, Tese de Doutoramento, 2020.

[11] A. E. Goldberg, Explain me this: Creativity, Competition, and the Partial Productivity of Constructions, Princeton University Press, Princeton, 2019. doi: 10.1515/9780691183954

[12] T. Hoffmann, Construction Grammar, Cambridge University Press, Cambridge, 2022.

[13] L. Goulart, B. Gray, Sh. Staples, A. Black, A. Shelton, D. Biber, J. Egbert, S. Wizner, Linguisitc perspectives on register, Annual Review of Linguistics 6:1 (2020) 435–455 doi: 10.1146/annurev-linguistics-011718-012644

[14] D. Lee, Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle, Language Learning & Technology 5 (3) (2001) 37–72.

[15] V. V. Zhukovska, O. O. Mosiyuk, Statistical software R in corpus-driven research and machine learning, Information Technologies and Learning Tools 86 (6) (2021) 1–18. doi: 10.33407/itlt.v86i6.4627

[16] N. Levshina, How to do linguistics with R. John Benjamins Publishing, Amsterdam, 2015.

[17] A. Basirat, M. Tang, Lexical and morpho-syntactic features in word embeddings –- A case study of nouns in Swedish, in: Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018), volume 2, Funchal, Madeira, Portugal, 2018, pp. 663–674. doi: https://doi.org/10.5220/0006729606630674.

[18] Sthda.com, Discriminant Analysis Essentials in R, Articles, STHDA, 2021 URL: http://www.sthda.com/english/articles/36-classification-methods-essentials/146-discriminant-analysis-essentials-in-r/#linear-discriminant-analysis---lda

[19] Cran.r-project.org. Package MASS. 2021. URL: https://cran.r-project.org/web/packages/MASS/MASS.pdf

[20] M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, et al. Package “caret”: Classification and Regression Training. 2023. URL: https://cran.r-project.org/web/packages/caret/caret.pdf

[21] A. Luque, A. Carrasco, A. Martín, A. de Las Heras, The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition 91 (2019) 216 – 231. doi: https://doi.org/10.1016/j.patcog.2019.02.023

[22] S. Narkhede, Understanding Confusion Matrix, 2021, URL: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62