Multiparametric profiling of a linguistic construction: linguoquantitative and machine-learning aspects

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

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

Abstract

This paper discusses the results of the linguoquantitative multiparametric profiling of linguistic constructions, focusing on English 'detached nonfinite/ nonverbal with explicit subject'-constructions and adopting cognitive-quantitative construction grammar as a theoretical and methodological foundation. Despite extensive research into the linguistic diversity of the syntactic patterns under analysis, a comprehensive parametrization of their linguistic profiles to identify the determining properties that influence their linguistic behavior in present-day English has yet to be carried out. Thus, the statistical platform R was used to achieve two goals: 1) to perform a linguoquantitative parametrization of the formal properties of English 'detached nonfinite/ nonverbal with explicit subject'-constructions based on the corpus data; 2) to verify the results of the linguoquantitative parametrization in a machine learning model and establish the properties with the most significant capacities to differentiate between the DNF/NVES-constructions beyond the corpus. The results of the study prove that the operationalized parameters (factors/factor values) demonstrate different determinative capacity; thus, the linguistic profiles of the DNF/NVES-constructions are distinguished by high, medium, and low linguistic homogeneity, which allows their classification according to the degree of proximity/remoteness in the constructional network. The operationalized linguistic parameters are quite reliable in differentiating types of the DNF/NVES-constructions beyond the corpus.

Keywords

cognitive-quantitative construction grammar, clause-level construction, multiparametric profiling, statistical analysis, machine-learning experiment

1. Introduction

The development of modern linguistics, particularly construction grammar, is accompanied by a discussion about increasing the objectivity of research data and finding ways to improve research precision [1, p. 149; 2]. As a result, the methodology for analyzing linguistic phenomena is being refined and statistically reliable tools are being actively

CLW-2024: Computational Linguistics Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2024), April 12–13, 2024, Lviv, Ukraine

* Corresponding author.
† These authors contributed equally.

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

employed to verify scientific theories and hypotheses. Traditional methods of language analysis are being complemented by innovative quantitative corpus-linguistic methods [3, 4]. The use of objective linguistic quantitative analysis tools specifically designed for processing large amounts of language data increases the degree of evidential support for the obtained results, revealing new data that would be difficult to identify through conventional empirical and interpretive approaches. Cognitive-quantitative construction grammar studies apply advanced quantitative-corpus methods to parameterize a construction's linguistic profile [5-8].

This study ***aims*** to discuss the results of the linguoquantitative multiparametric profiling of linguistic constructions, focusing on English '*detached nonfinite/ nonverbal with explicit subject'-constructions* (*DNF/NVES-constructions*) and adopting *cognitive-quantitative construction grammar* as a theoretical and methodological foundation. With this in mind, the following ***objectives*** are attained: 1) to perform a linguoquantitative parametrization of the formal properties of English *'detached nonfinite/ nonverbal with explicit subject'-constructions* based on the corpus data; 2) to verify the results of the linguoquantitative parametrization in a machine learning model and establish the properties with the greatest capacity to differentiate between the *DNF/NVES-constructions* beyond the corpus. By integrating quantitative corpus linguistics and machine learning approaches, this study advances understanding of the determining properties that define the linguistic behavior of the analyzed constructions in present-day English.

2. Related Works

The parameterization method, widely used in engineering and exact sciences, is also applied in various fields of linguistics, such as linguistic modeling (Yatsenko (2011)), lexicology and lexicography (Boychuk (2011), Ivakhnenko (2016), Kupriianov (2019)), stylistics and genre studies (Romanchenko, Stryi (2022)), analysis of the individual writer's style (Buk (2021), Davydenko (2014), Tkachenko (2018)), cognitive linguistics (Harmash (2015)), corpus studies (Luchyk, Ostapova (2017)), and forensic linguistics (Azhniuk (2017)). Quantitative corpus-based studies employ parameterization to identify, define, and quantify the essential properties ("diagnostic features" [9, p. 35]) of a linguistic unit.

English *'detached nonfinite/nonverbal with explicit subject'-constructions* ([[$_{\text{AUG}}$***with***] [$_{\text{NP}}$***the bats***] [$_{\text{XP}}$***taking turns*** to be the starved victim]]; [[$_{\text{NP}}$***heart***] [$_{\text{XP}}$***thumping***]]; [[$_{\text{AUG}}$***despite***] $_{\text{NP}}$[***oil***] [$_{\text{XP}}$***being the lifeblood of industrial (modern) society***]]; [[$_{\text{AUG}}$***what with***] [$_{\text{NP}}$***my three sons***] [$_{\text{XP}}$***being away in the Army***]]) as complex clause-level *constructions* possess several idiosyncratic properties that set them apart from other complex syntactic units. Different aspects of these syntactic patterns in both diachrony and synchrony have been studied from the standpoint of various linguistic approaches and frameworks such as *traditional grammar* (Stump (1985), Quirk, Greenbaum, Leech, Svartvik (1985), Kortmann (1991)), *generative grammar* (Riemsdijk (1981), Beukema, Hoekstra (1984), Felser, Britain (2007), Nakagawa (2011)), *corpus linguistics* (van de Pol (2012, 2014), van de Pol & Petré (2015)), *systemic functional grammar* (He, Wu (2015), He, Yang (2015)), and *construction grammar* (Riehemann, Bender (1999), Bouzada-Jabois, Guerra (2016)). In addition, the analyzed syntactic units have been considered in the dimensions of *linguotypology* (Haff (2012), Hasselgård (2012)), *translation studies* (Davydiuk (2010)) and *discourse structure analysis* (Asher, Lascarides (2003)). Although several studies have been undertaken, the linguistic versatility of nonfinite/nonverbal syntactic patterns with an explicit subject in English raises a number of questions that have not yet been finally resolved. Primarily, most research has concentrated on the qualitative rather than quantitative aspects of these units, resulting in a gap in understanding their functional and contextual characteristics. Moreover, a comprehensive parametrization of their linguistic profiles to identify the determining properties that influence their linguistic behavior in present-day English and may reflect speakers' preferences in categorizing their linguistic experience has yet to be performed.

3. Theoretical and methodological background

Cognitive-quantitative construction grammar (CQCxG) is a novel research framework in cognitive-quantitative grammar studies. This framework triangulates the theoretical and methodological underpinnings of cognitive-semiotic grammar approaches with analytical and research tools of quantitative corpus linguistics to investigate general and idiosyncratic properties of linguistic *constructions*. Cognitive-quantitative construction grammar revitalizes the traditional concept of a *construction*, promoting it to the status of the basic unit for language representation and analysis. *Constructions* are conceptualized as holistic semiotic models, emergent cognitively entrenched symbolic units conventionally used in a language community, and exhibit pairings of generalized form and meaning/function (plane of expression and plane of content) [10]. *Constructions* embrace all language levels, from morphemes and abstract clausal patterns to text types and genres, ultimately forming an organized inventory of constructional networks (*constructicon*), constantly updated and adjusted by language usage [11-13]. An in-depth examination of the linguistic properties of a particular linguistic *construction* can be performed by analyzing its essential form/meaning properties (prosodic, morphological, syntactic, semantic, distributional, functional, pragmatic, etc.).

The most effective analytical and research tool for examining the essential properties of a *construction* is a comprehensive methodology for multiparametric constructional profiling. In quantitative corpus studies, 'profiling' refers to the process of establishing specific linguistic properties at a particular language level based on quantitative indicators of this property (parameter) realization in the corpus, whereas the 'profile' of a linguistic unit is a set of established quantitative indicators [14, 15]. The set of these properties determines the linguistic behavior of a *construction*.

Multiparametric profiling is based on the procedure of a linguoquantitative parameterization. The procedure entails identifying and statistically verifying a set of essential linguistic properties (parameters/ factors/ factor values) of the plane of expression (form) and the plane of content (meaning/function) of a linguistic *construction* that constitute its linguistic profile. Thus, a *construction's* linguistic profile is an inventory of its formal and semantic properties (parameters) (morphosyntactic, positional, relational, referential, distributional, syntactic-functional, collocational-collexeme and cognitive-semantic), along with corresponding quantitative indicators obtained through their linguoquantitative verification in corpus data. Linguistic *parameters* of a *construction* are realized in linguistic features at a particular language level – *factors*, which are then manifested in specific language categories – *factor values*.

Nonfinite/nonverbal syntactic patterns with an explicit subject are considered syntagmatically and semantically complex clause-level *constructions* in CQCxG and are referred to as *"D(etached) N(on)F(inite)/N(on)V(erbal) (with) E(xplicit) S(ubject)"-constructions* (*DNF/NVES-constructions*). The argument-predicate structure of the *DNF/NVES-constructions* minimally consists of a predicate expressed by a nonfinite (NF)/nonverbal (NV) phrase (XP) and a subject (the external argument of the nonfinite/nonverbal predicate) expressed by a (pro)nominal phrase (NP). These clause-level *constructions* are partially schematic, represented by obligatory lexically unspecified slots [Subj$_{\text{NP}}$] and [Pred$_{\text{NF/NV}}$], with an open slot for an augmentor [Aug/ØAug] that in present-day English is expressed by a limited number of units {AUG: *with, without, despite, what with*}. The *constructions* represent a syntactically independent configuration, detached from a matrix clause by intonation or a punctuation mark. The morphosyntactic arrangement of the components is displayed as [[Aug/ØAug][Subj$_{\text{NP}}$][Pred$_{\text{NF/NV}}$]] (e.g., [$_{\text{Aug}}$with][$_{\text{Subj}}$***her eyes***$_{\text{NP}}$][$_{\text{PredNV}}$***open***$_{\text{Adj}}$] (BNC, GOS); [$_{\text{Aug}}$despite] [$_{\text{Subj}}$***desparate attempts***$_{\text{NP}}$][$_{\text{PredNF}}$***to revive***$_{\text{to-Inf}}$ her] (BNC, JYB); [$_{\text{Aug}}$what_with][$_{\text{Subj}}$***delays***][$_{\text{PredNF}}$***getting started***$_{\text{PII}}$] (BNC, HPP); [ØAug][$_{\text{Subj}}$***heart***$_{\text{NP}}$][$_{\text{PredNF}}$***thumping***$_{\text{PI}}$ widely] (BNC, EWH). The *DNF/NVES-constructions* constitute a taxonomic constructional network in which individual *constructions* are projected onto the network as nodes with different degrees of schematicity, lexical specification, and productivity [16].

The quantitative multiparametric constructional profiling methodology is employed to establish the essential properties of the *DNF/NVES-constructions* that determine their linguistic behavior in contemporary English. Multiparametric profiling of the *DNF/NVES-constructions* entails linguoquantitative parameterization of their linguistic (formal and semantic/functional) properties that comprise their constructional multiparametric linguistic profiles, followed by verification of the obtained data through a machine learning experiment.

4. Experiment: corpus sample, statistical software R, and research algorithm

The procedure for linguoquantitative parameterization of a constructional profile is applied to a research sample of the DNF/NVES-constructions obtained from the British National Corpus [17]. The sample includes 11,000 corpus contexts that instantiate five DNF/NVES-constructions (dt-øaug-SubjPredNF/NV–cxn, dt-with-SubjPredNF/NV–cxn, dt-despite-SubjPredNF/NV–cxn, dt-without-SubjPredNF/NV–cxn, dt-what_with-SubjPredNF/NV–cxn), manifested in 35 predicate specifications {NF: VPPI, VPPII, VPto-Inf; NV: NP, AdjP, AdvP, PP}. The sample size is adequate to be regarded as reliable for linguistic quantitative profiling since the derived indicators are characterized by a 1,9% relative error. A 5% error is considered acceptable in linguistic and statistical research, although an error of 20-30% is also permitted [18, p. 28].

In the current study, the primary focus is on the plane of expression (form) of the investigated *constructions*, while the properties of the content plane need a different methodology. The inventory of formal parameters (factors/factor values) is determined by the linguistic and constructional nature of the *constructions* under study. The *DNF/NVES-constructions*, structurally complex clause-level *constructions*, are distinguished by seven parameters of the plane of expression, which define their morphosyntactic, relational, referential, syntactic-functional, positional, and distributional properties. The inventory of the specified parameters is not exhaustive, but it is sufficient for an objective examination of the linguistic behavior of the *DNF/NVES-constructions* in contemporary English.

A considerable number of parameters (factors/factor values) specified to describe linguistic profiles of the *DNF/NVES-constructions* in combination with a large amount of quantitative data cannot be objectively analyzed without using complex statistical procedures and appropriate computer programs for statistical processing of linguistic data. As a result, each linguistic parameter (factor / factor value) is submitted to computerized quantitative verification and subsequent qualitative interpretation.

One of the most widely used analytical tools for quantitative processing of empirical data in Western corpus-oriented linguistics and usage-based construction grammar is the statistical data analysis system R (R Development Core Team) [19]. It is a robust and freely distributed statistical software environment for data analysis, providing researchers with a comprehensive toolset for qualitative linguistic and statistical analysis and result visualization [20]. The software environment enables users to manipulate extensive amounts of multidimensional data, employing various processing techniques such as visualization, primary data analysis, matrix graph construction, scatter plots, etc. Additionally, it offers classification methods for organizing data, performing statistical verification, and mathematical modeling.

Parametrization of linguistic profiles of the *DNF/NVES-constructions* is carried out according to the following algorithm.

Step 1. Operationalization of the parameter by identifying the factors of its linguistic manifestation and defining the values that a particular factor acquires at the appropriate level of the linguistic structure.

Step 2. Quantitative analysis of the realization (frequency) of a particular parameter (factor/factor value) in the corpus.

Step 3. Statistical analysis of the data obtained using multivariate analysis of variance (MANOVA), one-factor analysis of variance (ANOVA), and Tukey's multiple comparison method, quantified with the computer statistical data analysis system R.

Step 4. Interpretation of quantitative indicators and identification of essential parameters (factors/factor values) that determine the degree of proximity/remoteness between *constructions* in a constructional network.

Step 5. Verification of the linguoquantitative data in a machine experiment to establish factors/factor values with the highest capacity to distinguish between the *constructions* beyond the corpus.

The application of the algorithm to the factors *"Part of speech representation of the subject"* (SubjPOS) of the morphosyntactic parameter and the factor *"Register Distribution"* (RegDSTN) of the distributional parameter of the *DNF/NVES-constructions* has already been extensively discussed in our prior works [21, 22]. The findings of our previous research provide the foundation for using statistical methods and machine-learning approach in the current study.

5. Results/ Discussion

5.1. Linguistic profiles of the *DNF/NVES-constructions*: a computerized linguoquantitative parametrization

The linguistic profiles of the *DNF/NVES-constructions* are parametrized through the quantitative verification of 7 parameters (morphosyntactic, relational, referential, syntactic-functional, positional, distributional, and punctuational), manifested in 12 factors and 34 factor values as shown in Table 1. The quantitative data of the operationalized parameters and their respective factors/ factor values retrieved from the BNC are standardized by logarithmization. Subsequently, the null (H0) and alternative (H1) statistical hypotheses are formulated for each of the identified factors:

H0: *The quantitative differences between the analyzed DNF/NVES-constructions (dt-**øaug**-SubjPred*$_{NF/NV}$*–cxn, dt-**with**-SubjPred*$_{NF/NV}$*–cxn, dt-**despite**-SubjPred*$_{NF/NV}$*–cxn, dt-**without**-SubjPred*$_{NF/NV}$*–cxn, dt-**what_with**-SubjPred*$_{NF/NV}$*–cxn) within the "FACTOR" are insignificant, and any detected quantitative differences are random.*

H1: *The quantitative differences between the analyzed DNF/NVES-constructions (dt-**øaug**-SubjPred*$_{NF/NV}$*–cxn, dt-**with**-SubjPred*$_{NF/NV}$*–cxn, dt-**despite**-SubjPred*$_{NF/NV}$*–cxn, dt-**without**-SubjPred*$_{NF/NV}$*–cxn, dt-**what_with**-SubjPred*$_{NF/NV}$*–cxn) within the "FACTOR" are significant, and the differences found are essential.*

The formulated hypotheses are tested using multivariate analysis of variance (MANOVA), and the quantified findings are displayed in Table 1.

Table 1
The results of statistical hypothesis testing for the factors of the operationalized parameters using MANOVA

Parameter	Factor	Null hypothesis	Alternative hypothesis
Morphosyntactic	Pronominal Subject Case (SubjPrnCASE)		accepted
	Subject Determiner (SubjDET)		accepted
	Predicate Part of Speech (PredPOS)	accepted	
	Nonfinite Predicate Aspect (PredASP)	numerical values are not calculated by the program	
	Nonfinite Predicate Voice (PredVoice)	accepted	

Punctuation	Punctuation marking (PUNC)	accepted
Positional	Position to Matrix Clause (SentPSN)	accepted
Referential	Coreference with Matrix Clause (CoREFR)	accepted
Distributional	Discourse Mode Distribution (DiscMdDSTN)	accepted
	Text Type Distribution (TxtTpDSTN)	accepted
Syntactic-functional	Syntactic Function to Matrix Clause (FSYN)	accepted
Relational	Syntactic Relation with Matrix Clause (SynREL)	accepted

Table 1 shows that the quantitative differences between the DNF/NVES-constructions are not present in 3 out of 12 factors. Next, the differences in the quantitative realizations of the factor values of the specified factors are established. The one-factor analysis of variance (ANOVA) is used to determine statistically significant differences in the realization of a particular factor value within the factors. Table 2 provides the results on statistically significant differences in the quantitative realization of a specific factor value.

Table 2
Results of the single-factor analysis of variance of the factor values

Factor	**Pronominal Subject Case (SubjPrnCASE)**						
Factor values	*Nominative case (Nom)*			*Accusative case (Acc)*			
	–			–			
Factor	**Subject Determiner (SubjDET)**						
Factor values	*Definite article (ArtDef)*	*Possessive pronoun (PrnPoss)*	*Demonstrative pronoun (PrnDem)*	*Indefinite article (ArtIndef)*	*Indefinite pronoun (PrnIndef)*	*Singular noun (NSing)*	*Plural noun (NPl)*
	+	+	+	+	+	+	+
Factor	**Punctuation marking (PUNC)**						
Factor values	*Coma(PUNCcm)*			*Other punctuation marks (PUNCothr)*			
	+			+			
Factor	**Position to Matrix Clause (SentPSN)**						
Factor	*Sentence initial*	*Sentence medial*	*Sentence final*	*Sentence split*			

values	*(SentInit)*	*(SentMid)*	*(SentFin)*	*(SentSpl)*	
	+	+	+	+	
Factor	**Coreference with Matrix Clause (CoREFR)**				
Factor values	*Full coreference (CorrefFull)*	*Partial coreference (CorrefPart)*	*No coreference (NonCorref)*		
	–	+	+		
Factor	**Discourse Mode Distribution (DiscMdDSTN)**				
Factor values	*Spolken (Spkn)*	*Written (Wrtn)*			
	+	+			
Factor	**Text Distribution (TxtTpDSTN)**				
Factor values	*Narrative texts (TxtNar)*	*Non-narrative texts (TxtNonNar)*	*Fiction texts (TxtLit)*	*Non-fiction texts (TxtNonLit)*	
	+	+	+	+	
Factor	**Syntactic Function to Matrix Clause (FSYN)**				
Factor values	*Extension (Extn)*	*Elaboration (Elbn)*	*Enhancement (Enhnt)*		
	+	+	+		
Factor	**Syntactic Relation with Matrix Clause (SynREL)**				
Factor values	*Augmentor with (AugWith)*	*Augmentor without (AugWithout)*	*Augmentor despite (AugDespite)*	*Augmentor what with (AugWhatwith)*	*Non augmented (øAug)*
	+	+	+	+	+

The statistical analysis of the linguistic profiles of the *DNF/NVES-constructions* did not reveal statistically significant differences in the realization of 3 factor values (*Nominative case* (Nom), *Accusative case* (Acc), *Full coreference* (CorrefFull)) out of 32. The quantitative correlations between these linguistic properties do not differentiate the linguistic profiles of the *DNF/NVES-constructions* and suggest general regularities of the subject's linguistic embodiment and the reference relations between the *DNF/NVES-constructions* and the corresponding matrix clauses.

The one-way ANOVA indicates the existence of differences but does not explain where these differences are most prominent. To solve this issue, the post hoc Tukey test is employed to prevent erroneous rejection of the null hypothesis.

The Tukey's multiple comparison method detects the *DNF/NVES-constructions* that exhibit statistically significant differences in the realization of factor values. The Tukey's test is quantified by comparing the indicators for a specific factor value in pairs of *constructions*. It enables the establishment, with a 95% confidence level, which linguistic features are determining for specific *constructions*. The multiple comparison method is used to analyze ten pairs of the *DNF/NVES-constructions*: 1) *dt-**what_with**-SubjPred$_{NF/NV}$–cxn* and *dt-**despite**-SubjPred$_{NF/NV}$–cxn*; 2) *dt-**with**-SubjPred$_{NF/NV}$–cxn* and *dt-**despite**-SubjPred$_{NF/NV}$–cxn*; 3) *dt-**øaug**-SubjPred$_{NF/NV}$–cxn* and *dt-**despite**-SubjPred$_{NF/NV}$–cxn*; 4) *dt-**without**-*

SubjPred$_{NF/NV}$*–cxn* and *dt-**despite**-SubjPred*$_{NF/NV}$*–cxn*; 5) *dt-**with**-SubjPred*$_{NF/NV}$*–cxn* and *dt-**what_with**-SubjPred*$_{NF/NV}$*–cxn*; 6) *dt-**øaug**-SubjPred*$_{NF/NV}$*–cxn* and *dt-**what_with**-SubjPred*$_{NF/NV}$*–cxn*; 7) *dt-**without**-SubjPred*$_{NF/NV}$*–cxn* and *dt-**what_with**-SubjPred*$_{NF/NV}$*–cxn*; 8) *dt-**øaug**-SubjPred*$_{NF/NV}$*–cxn* and *dt-**with**-SubjPred*$_{NF/NV}$*–cxn*; 9) *dt-**without**-SubjPred*$_{NF/NV}$*–cxn* and *dt-with-SubjPred*$_{NF/NV}$*–cxn*; 10) *dt-**without**-SubjPred*$_{NF/NV}$*–cxn* and *dt-**øaug**-SubjPred*$_{NF/NV}$*–cxn*.

According to the degree of proximity/remoteness by the number of statistically significant differences in the realization of the specified factor/factor values, we distinguish constructional linguistic profiles with high, medium, and low linguistic homogeneity.

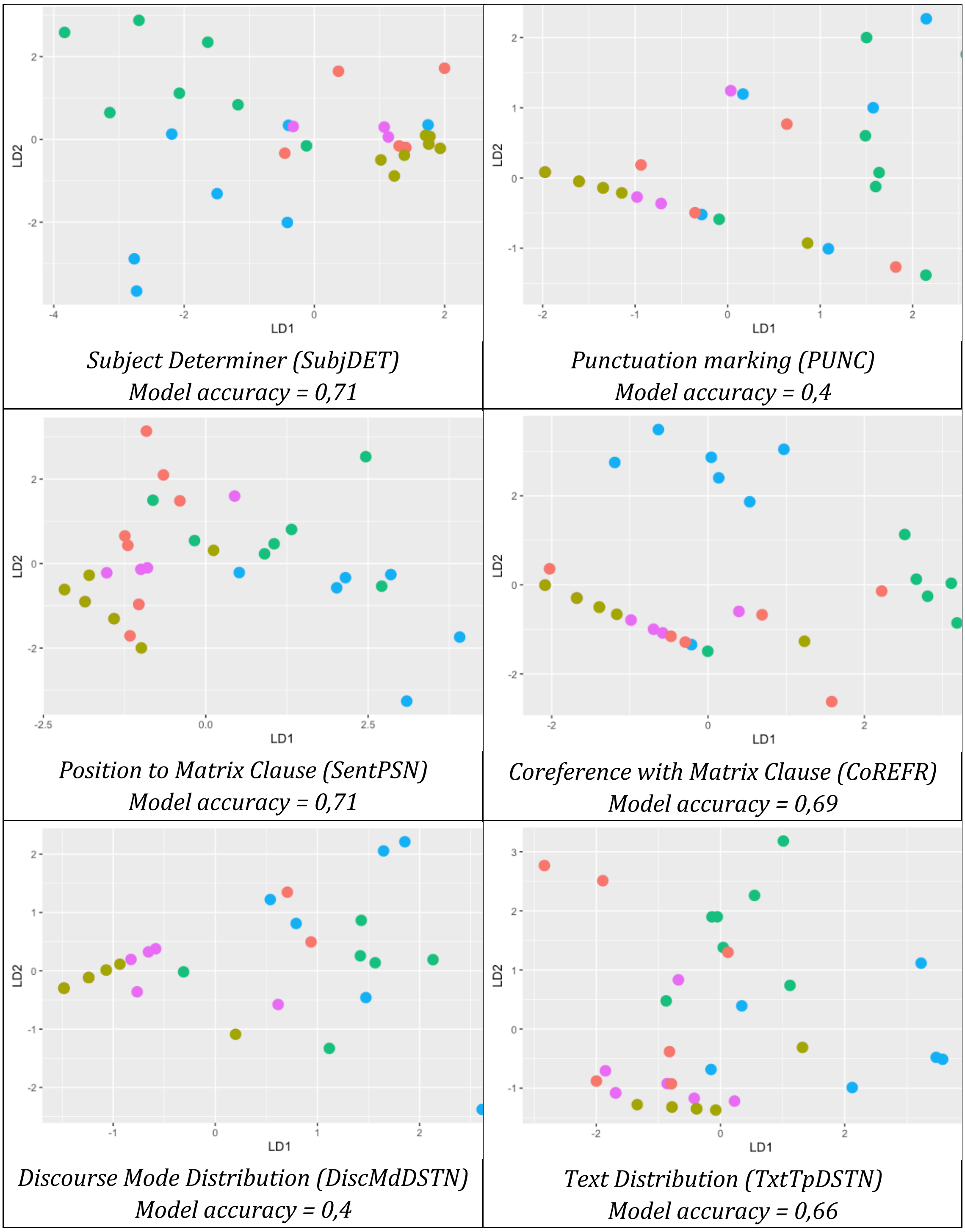
The *constructions dt-**despite**-SubjPred*$_{NF/NV}$*–cxn*, *dt-**without**-SubjPred*$_{NF/NV}$*–cxn*, *dt-**what_with**-SubjPred*$_{NF/NV}$*–cxn* form a group with high linguistic homogeneity, showing no statistically significant differences in the analyzed factors/ factor values realization. The medium linguistic homogeneity group consists of *constructions* such as *dt-**øaug**-SubjPred*$_{NF/NV}$*–cxn* and *dt-**with**-SubjPred*$_{NF/NV}$*–cxn*, which revealed statistically significant differences in 6 factor values. The group with low linguistic homogeneity includes two subgroups of the *constructions*: 1) subgroup *dt-**with**-SubjPred*$_{NF/NV}$*–cxn* and *dt-**despite**-SubjPred*$_{NF/NV}$*–cxn*, *dt-**without**-SubjPred*$_{NF/NV}$*–cxn*, *dt-**what_with**-SubjPred*$_{NF/NV}$*–cxn*, where significant differences were recorded between the ***with***- and ***what_with***-augmented *constructions* by 31 factor values, between the ***with-*** and ***despite-*** augmented *constructions* by 30, and between the ***with-*** and ***without***-augmented *constructions* by 27; 2) *dt-**øaug**-SubjPred*$_{NF/NV}$*–cxn* та *dt-**despite**-SubjPred*$_{NF/NV}$*–cxn*, *dt-**without**-SubjPred*$_{NF/NV}$*–cxn*, *dt-**what_with**-SubjPred*$_{NF/NV}$*–cxn*, between which differences in 13, 17, and 20 factor values were registered, respectively.

The statistical analysis of linguistic profiles of the DNF/NVES-constructions demonstrates that some factors/factor values significantly influence their linguistic behavior. Determining parameters (factors/factor values) of the plane of expression of the analyzed constructions define the degree of proximity and remoteness of the constructions. To validate the obtained results, a machine learning experiment is conducted to establish the operationalized factors with the most significant capacity to differentiate between the DNF/NVES-constructions beyond the corpus.

5.2. Linguistic profiles of the *DNF/NVES-constructions*: a machine-learning model

The use of machine learning technologies alongside traditional statistical approaches in language study is becoming more prevalent [23, 24]. Researchers have employed algorithms for machine learning to investigate strategies to increase classifier accuracy in evaluating readability [25], as well as assess the prediction powers of ML systems in cross-linguistic vowel categorization [26]. These approaches gave been also used to address problems in the field of Natural Language Processing (NLP) [27]. Drawing on our previous research [21, 22], we utilize linear discriminant analysis (LDA) to determine the factors with the greatest potential for separation between the *DNF/NVES-constructions* outside the corpus (the British National Corpus) in present-day English. The specialized package MASS [28, 29] is used to build the model for linear discriminant analysis in *R* and all graphs are produced with ggplot2 library for R programming language [30].

Figure 1 displays the results of the distribution of the *DNF/NVES-constructions* for each factor, where significant differences were found with a one-factor ANOVA. All data are best divided along the first two axes LD1 and LD2, simplifying the graphical representation of information and enabling result comparison.



Subject Determiner (SubjDET)
Model accuracy = 0,71

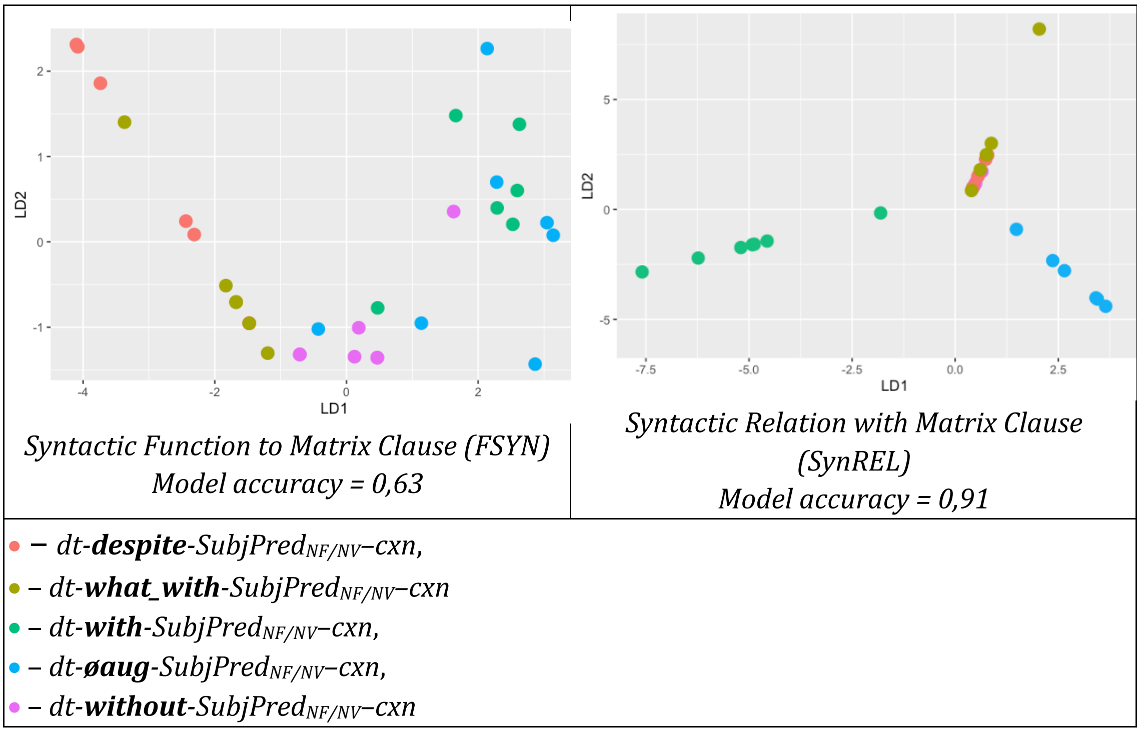
Punctuation marking (PUNC)
Model accuracy = 0,4

Position to Matrix Clause (SentPSN)
Model accuracy = 0,71

Coreference with Matrix Clause (CoREFR)
Model accuracy = 0,69

Discourse Mode Distribution (DiscMdDSTN)
Model accuracy = 0,4

Text Distribution (TxtTpDSTN)
Model accuracy = 0,66



Syntactic Function to Matrix Clause (FSYN)
Model accuracy = 0,63

Syntactic Relation with Matrix Clause (SynREL)
Model accuracy = 0,91

● – *dt-**despite**-SubjPred*$_{NF/NV}$*–cxn*,
● – *dt-**what_with**-SubjPred*$_{NF/NV}$*–cxn*
● – *dt-**with**-SubjPred*$_{NF/NV}$*–cxn*,
● – *dt-**øaug**-SubjPred*$_{NF/NV}$*–cxn*,
● – *dt-**without**-SubjPred*$_{NF/NV}$*–cxn*

Figure 1: Graphic representation of the linguistic classifier model.

Figure 1 shows a distinct separation of *dt-**with**-SubjPred*$_{NF/NV}$*–cxn* and *dt-**øaug**-SubjPred*$_{NF/NV}$*–cxn* from other *constructions* submitted to the linear analysis. However, in the factor *"Syntactic Function to Matrix Clause"* (FSYN), the *dt-**despite**-SubjPred*$_{NF/NV}$*–cxn* is also distinguished.

The confusion matrices are generated for all specified factors to validate the results reached from the analysis of graphic materials. Due to space limitations, only the confusion matrices with the highest and lowest model accuracy are presented in Table 3 and Table 4. The analysis of the confusion matrices reveals that for ***with-*** and ***øaug***-augmented *constructions* the Recall and Precision values are pretty high, particularly for models with an overall accuracy of 0,7 and higher. It suggests a high probability of their correct extraction by the created machine learning models. The specified factors may be ranked based on the model accuracy: the factor *'Syntactic Relation with Matrix Clause'* with the model's accuracy of 0,91 is characterized by the highest capacity to differentiate the analyzed *constructions*. The factors *'Subject Determiner'* (0,71), *'Position to Matrix Clause'* (0,71), *'Coreference with Matrix Clause'* (0,69), *'Text Distribution'* (0,66), *'Syntactic Function to Matrix Clause'* (0,63) are less reliable in differentiating between the *DNF/NVES-constructions*. The factors *'Punctuation marking'* (0,4) and *'Discourse Mode Distribution'* (0,4) show the lowest differentiating capacity.

The results of the machine learning experiment are very similar to those obtained by the linguoquantitative parameterization of the constructional profiles. However, the overall efficiency of the constructed machine learning model to solve the problem of distinguishing the types of the *DNF/NVES-constructions* beyond the corpus is not sufficient. The model effectively distinguishes the *dt-**øaug**-SubjPred*$_{NF/NV}$*–cxn* and *dt-**with**-SubjPred*$_{NF/NV}$*–cxn* *constructions*, despite insufficient overall accuracy. However, *dt-**despite**-SubjPred*$_{NF/NV}$*–cxn*, *dt-**what_with**-SubjPred*$_{NF/NV}$*–cxn,* and *dt-**without**-SubjPred*$_{NF/NV}$*–cxn* are more challenging to classify. Provided an effective model is constructed, the specified factors of the operationalized linguistic parameters can be used to distinguish types of the *DNF/NVES-constructions* in the present-day English usage.

Table 3
Syntactic Relation with Matrix Clause (Model accuracy = 32/35 = 0,91)

		Actual values					
		despite	***what_with***	***with***	***øaug***	***without***	*Precision*
Predicted values	**despite**	5	0	0	0	0	*1*
	what_with	0	6	0	0	0	*1*
	with	0	0	7	0	0	*1*
	øaug	0	0	0	7	0	*1*
	without	2	1	0	0	7	*1*
		0,71	*0,86*	*1*	*1*	*1*	
		Recall					

Table 4
Discourse Mode Distribution (Model accuracy = 14/35 = 0,4)

		Actual values					
		despite	***what_with***	***with***	***øaug***	***without***	*Precision*
Predicted values	**despite**	0	1	0	0	3	*0*
	what_with	5	5	0	1	2	*0,38*
	with	0	0	4	2	1	*0,57*
	øaug	2	0	2	4	0	*0,5*
	without	0	1	1	0	1	*0,33*
		0	*0,71*	*0,57*	*0,57*	*0,14*	
		Recall					

6. Conclusions

The formal parameters (factors/factor values) are determined by the linguistic and constructional nature of the *DNF/NVES-constructions* as complex clausal *constructions*. The analysis includes seven parameters of the expression plane of the *DNF/NVES-constructions*, which define their morphosyntactic, relational, referential, syntactic-functional, positional, punctuational, and distributional features. The operationalized parameters (factors/factor values) reveal different determinative capacities; thus, the linguistic profiles of the *DNF/NVES-constructions* are characterized by a certain degree of proximity/remoteness in the constructional network, which allows their categorization according to the degree of linguistic homogeneity: 1) high *(dt- **despite**-SubjPred*$_{NF/NV}$*–cxn, dt- **without**-SubjPred*$_{NF/NV}$*–*

*cxn, dt- **what_with**-SubjPred*$_{NF/NV}$*–cxn*); 2) medium (*dt- **øaug**-SubjPred*$_{NF/NV}$*–cxn* and *dt- **with**-SubjPred*$_{NF/NV}$*–cxn*); 3) low (subgroup *dt- **with**-SubjPred*$_{NF/NV}$*–cxn* and *dt- **despite**-SubjPred*$_{NF/NV}$*–cxn, dt- **without**-SubjPred*$_{NF/NV}$*–cxn, dt- **what_with**-SubjPred*$_{NF/NV}$*–cxn*; subgroup *dt- **øaug**-SubjPred*$_{NF/NV}$*–cxn* and *dt- **despite**-SubjPred*$_{NF/NV}$*–cxn, dt- **without**-SubjPred*$_{NF/NV}$*–cxn, dt- **what_with**-SubjPred*$_{NF/NV}$*–cxn*). The specified factors of the operationalized linguistic parameters can be utilized to differentiate types of the *DNF/NVES-constructions* beyond the corpus in current English usage. Differences in the quantitative realization of individual factors/factor values within one parameter of a specific *DNF/NVES-construction* are determined by intra-constructional variability. In contrast, quantitative differences in the realization of factors/factor values within one parameter between different *DNF/NVES-constructions* are determined by inter-constructional variability.

The results of this study indicate a need for further research on the discussed issues. In future studies, it will be interesting to generate other types of machine learning models and assess their effectiveness in distinguishing between the types of the *DNF/NVES-constructions* based on the data sets for the operationalized parameters/factors/factor values extracted from the corpus.

References

[1] L. A. Janda, Cognitive Linguistics in the Year 2015, Cognitive Semantics 1 (2015) 131–154. doi:10.1163/23526416-00101005.

[2] L. A. Janda, Quantitative Perspectives in Cognitive Linguistics, Review of Cognitive Linguistics 17(1) (2019) 7–28. doi: 10.1075/rcl.00024.jan.

[3] B. Kortmann, Reflecting on the Quantitative Turn in Linguistics, Linguistics 59(5) (2021) 1207–1226. doi: 10.1515/ling-2019-0046.

[4] B. Winter, Statistics for Linguists. An Introduction Using R, Routledge, New York, London, 2021.

[5] K. Krawczak, The role of verb polysemy in constructional profiling: A cross-linguistic study of *give* in the dative alternation, in: M. Bouveret (Ed.), Constructional Approaches to Language, John Benjamins Publishing Company, Amsterdam, Philadelphia, 2021. pp. 75-96. doi: 10.1075/cal.29.

[6] J. E. Casal, Y. Shirai, X. Lu, English verb-argument construction profiles in a specialized academic corpus: Variation by genre and discipline, English for Specific Purposes 66 (2022) 94-107. doi: 10.1016/j.esp.2022.01.004.

[7] V. V. Zhukovska, Quantitative Corpus-Based Methods for Construction Grammar Research, Zhytomyr Ivan Franko State University Journal. Philological Sciences [Visnyk Zhytomyrskoho derzhavnoho universytetu imeni Ivana Franka. Filolohichni nauky] 1(99) (2023) 93–104. doi: 10.35433/philology.1(99).2023.93-104.

[8] P. Wyroślak, D. Glynn, Disentangling constructional networks: integrating taxonomic effects into the description of grammatical alternations, Linguistics Vanguard (2024). doi:10.1515/lingvan-2023-0035.

[9] D. Speelman, D. Geeraerts, Causes for causatives: The case of Dutch *doen* and *laten*, in: T. Sanders, E. Sweetser (Eds.), Causal Categories in Discourse and Cognition. Mouton de Gruyter, Berlin, New York, 2009, pp. 173–204.

[10] M. Hilpert, Constructional Approaches, in: B. Aarts, J. Bowi, G. Popova (Eds.), The Oxford Handbook of English Grammar, Oxford University Press, Oxford, 2020, pp. 106-123.

[11] T. Hoffmann, Construction Grammar: The Structure of English, Cambridge University Press, Cambridge. 2022.

[12] H. Diessel, The Constructicon: Taxonomies and Networks (Elements in Construction Grammar), Cambridge University Press, Cambridge, 2023.

[13] T Ungerer, S. Hartmann, Constructionist Approaches: Past, Present, Future (Elements in Construction Grammar), Cambridge University Press, Cambridge, 2023.

[14] D. Speelman, S. Grondelaers, D. Geeraerts, Profile-Based Linguistic Uniformity as a Generic Model for Comparing Language Varieties, Computers and the Humanities 37 (3) (2003) 317–337.

[15] D. Divjak, S. T. Gries, Ways of trying in Russian: Clustering Behavioral Profiles, Corpus Linguistics and Linguistic Theory 2(1) (2006) 23–60. doi: 10.1515/CLLT.2006.002.

[16] V. V. Zhukovska, Constructional Modeling in the Formalism of Cognitive-Quantitative Construction Grammar, Messenger of Kyiv National Linguistic University. Series "Philology" [Visnyk Kyivskoho Natsionalnoho Universytetu. Seriia "Filolohiia"] 26(2) (2023) 51–62. doi: 10.32589/2311-0821.2.2023.297670.

[17] M. Davis*, British National Corpus (BNC), 2004.* URL: https://www.english-corpora.org/bnc/.

[18] V. S. Perebyinis, Statistical methods for linguists [Statystychni metody dlia linhvistiv]. Nova knyha. Vinnytsia, 2002.

[19] R Core Team, R: A language and environment for statistical computing, 2024. https://www.r-project.org/.

[20] E. G. M. Hui, Learn R for Applied Statistics. With Data Visualizations, Regressions, and Statistics, Apress, Berkeley, 2019.

[21] V. Zhukovska, O. Mosiiuk, Statistical Software R in Corpus-Driven Research and Machine Learning, Information Technologies and Learning Tools 86(6) (2021) 1-18.

[22] V. Zhukovska, O. Mosiiuk, S. Buk, Register Distribution of English Detached Nonfinite/Nonverbal with Explicit Subject Constructions: a Corpus-Based and Machine-Learning Approach. In: CEUR Workshop Proceedings, 3396, 2023, pp. 63–76.

[23] R. Baayen, Analyzing linguistic data. Cambridge University Press, Cambridge, 2008. doi: 10.1017/CBO9780511801686.

[24] J. S. Th. Gries, Statistics for linguistics with R. Mouton de Gruyter, Berlin, New York, 2013.

[25] N. Sukhija, R. Priya, V. Arya, N. Kohli, A. Arya, Hybrid Ensemble Stacking Model for Gauging English Transcript Readability International Journal of Performability Engineering 19(11) 2023 719–727. doi: 10.23940/ijpe.23.11.p2.719727.

[26] G. P. Georgiou, Comparison of the Prediction Accuracy of Machine Learning Algorithms in Crosslinguistic Vowel Classification. Scientific Reports 13(15594) 2023 1–10 doi: 10.1038/s41598-023-42818-3.

[27] V. Addanki, S. Durgapu, K. Dorasanaiah, S. Abhishek, Safeguarding SMS: A Dynamic Duo Approach to Tackle Spam Using LDA and QDA. 2023 Innovations in Power and Advanced Computing Technologies (i-PACT). IEEE, 2023 1–6. doi: 10.1109/i-PACT58649.2023.10434334.

[28] Cran.r-project.org. Package MASS, 2024. URL: https://cran.r-project.org/web/packages/MASS/MASS.pdf

[29] 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

[30] H. Wickham, ggplot2. Elegant Graphics for Data Analysis, Springer, 2016. doi: 10.1007/978-3-319-24277-4.