
Hypothesis-Driven Corpus Querying at Scale: A Method and Its Application to *-nje* Nominalizations in BCMS

Predrag Kovačević¹[\[0000-0003-2545-9240\]](#), Marko Simonović²[\[0000-0002-9651-6399\]](#), Marija Đokić Petrović³[\[0000-0002-0568-5219\]](#)

¹ University of Novi Sad, Faculty of Philosophy, Department of English Studies, Serbia, pedjakovacevic90@gmail.com;

² University of Graz, Department of Slavic Studies, Austria, marko.simonovic@uni-graz.at;

³ Union University, School of Computing, Belgrade, Serbia, mdjokicpetrovic@raf.rs

Scientific paper

DOI: 10.18485/judig.2026.2.ch13

Abstract

This paper presents a reproducible method for hypothesis-driven corpus querying at scale and applies it to Bosnian/Croatian/Montenegrin/Serbian (BCMS) *-nje* nominalizations. Using pluralization and genitive complements as diagnostics of internal verbal structure, the study tests predictions concerning structural differences between perfective and imperfective deverbal nominals across four BCMS varieties. The paper introduces a fully automated workflow for large-scale corpus querying, enabling systematic testing of theoretically motivated grammatical hypotheses across 24,000 corpus queries. The results reveal stable asymmetries between perfective and imperfective nominalizations consistent with previous theoretical proposals, while also uncovering limited but systematic microvariation across BCMS varieties. Beyond its empirical contribution, the study provides a generalizable methodological framework for scalable and reproducible corpus-based grammatical research.

Keywords: nominalization, Slavic linguistics, BCMS, aspect, corpus linguistics, reproducibility

1. Introduction

A central question in the study of deverbal nominalizations concerns the extent to which verbal structure is preserved in related nominals. Deverbal nouns often combine nominal morphology with properties associated with verbs, such as event interpretation, argument structure, aspectual sensitivity, and modification by event-related expressions. The challenge is to determine which of these properties reflect genuine inherited verbal structure and which can arise independently in the nominal domain.

This paper addresses this question through a large-scale corpus study of Bosnian/Croatian/Montenegrin/Serbian (BCMS) *-nje* nominalizations. The empirical focus is the contrast between perfective- and imperfective-derived *-nje* nominalizations, using pluralization and postnominal genitive complements as quantitative diagnostics of internal verbal structure. The study has two goals. The theoretical goal is to test whether imperfective *-nje* nominalizations show stronger evidence of verbal structure than perfective *-nje* nominalizations. The methodological goal is to show that such hypotheses can be tested reproducibly at scale through automated corpus querying.

We introduce a workflow for hypothesis-driven corpus querying in which theoretically motivated CQL query templates are systematically generated and executed across large web corpora. The method is applied to 1000 aspectual pairs of BCMS *-nje* nominalizations, yielding 24,000 queries across srWaC, hrWaC, bsWaC, and meWaC.

The results show that imperfective nominalizations are less likely to pluralize and more likely to occur with genitive complements, while perfective nominalizations show the opposite tendency. This supports analyses according to which imperfective *-nje* nominalizations preserve more verbal structure. At the same time, we find limited microvariation across the four BCMS varieties.

2. Background

2.1. Deverbal nominalization and verbal structure

Following [Alexiadou \(2001\)](#), deverbal nouns can be divided into verbal gerunds, process nominals, and result nominals. The contrast can be illustrated with the English forms in [1](#) and [2](#). The

important point is that these forms differ not only in meaning, but also in the amount of verbal behavior they preserve.

- (1) a. the teacher careful*(ly) examining the students
b. the teacher's careful(*ly) examination (of the students)
c. the teacher's (*careful(ly)) exam (*of the students)

The verbal gerund in 1a exhibits the highest number of verbal properties of all deverbal derivations under discussion. It takes a direct object, *the students*, and it is modified by the adverb *carefully*, just like an ordinary verb phrase.

The process nominal in 1b shows considerably fewer verbal and more nominal characteristics. It occurs with possessive marking on the external argument, *the teacher's*, and its internal argument appears as an *of*-phrase, *of the students*, rather than as an accusative object. It also takes the adjective *careful*, not the adverb *carefully*. Nevertheless, it can still describe an event of examining, which is why it is not simply an ordinary object-denoting noun.

The result nominal in 1c is the most noun-like of the three. The noun *exam* does not normally preserve the verbal argument structure of *examine*: it cannot take *of the students* as an internal argument in the relevant eventive sense, and it does not combine naturally with event-related modification. It denotes an object or result rather than the event of examining.

The same contrast appears in pluralization, as shown in 2.

- (2) a. *the teacher frequently examinings the students
b. *the teacher's frequent examinations (of the students)
c. the teacher's exams

The verbal gerund in 2a cannot be pluralized: *examinings* is impossible in the intended verbal-gerund meaning. This is expected if verbal gerunds preserve too much verbal structure to behave like ordinary count nouns. Process nominals such as *examination* are also restricted in pluralization when they preserve an eventive reading with an internal argument, as in 2b. By contrast, result nominals such as *exam* pluralize freely, as shown in 2c, because they behave like ordinary count nouns.

In sum, nominalization involves embedding different amounts of verbal structure under nominal functional structure. Verbal gerunds preserve the largest verbal structure. They include verbal

projections such as *v*P and VoiceP, which makes it possible for them to license accusative internal arguments and combine with adverbial modifiers. Process nominals involve less verbal structure. They may contain event-related structure, but they lack the kind of active Voice/*v* configuration needed for accusative case assignment, so their internal arguments surface as genitive or *of*-complements. Result nominals are derived closest to the root. They lack intervening verbal structure and therefore behave like ordinary lexical nouns: they do not inherit verbal argument structure and they pluralize freely.

2.2. Deverbal nominalizations in Slavic: a puzzle

Slavic native deverbal nominalizations in *-nje*, *-će* and their cognates are arguably the most productive event-nominalizing formations across the language group. Examples from South Slavic are provided in 3.

- (3) a. *pisa-nje*
 ‘writing’ *BCMS/Slovenian*
- b. *pišuva-nje*
 ‘writing’ *Macedonian*
- c. *pisa-ne*
 ‘writing’ *Bulgarian*

Also West and East Slavic have similar formations, e.g. Czech *psaní* and Russian *pisanije*, both meaning roughly ‘writing’. Across the group, there are synchronic morphological reasons to connect these nominalizations to passive participles (Simonović and Arsenijević, 2014). The basic generalization is simple. When the corresponding passive participle contains *-n*, the nominalization typically contains an *-n*-based nominalizing form, as in 4. When the passive participle contains *-t*, the nominalization contains a corresponding *-t*-based form, as in 5.

- (4) a. *pis-a-ti* → *pis-a-n* →
 write-THV-INF write-THV-PASS.PTCP
 pis-a-n-je
 write-THV-PASS.PTCP-NMLZ
 ‘write’ → ‘written’ → ‘writing’ *BCMS*

- b. ps-á-t → ps-a-n-ý →
 write-THV-INF write-THV-PASS.PTCP-M.SG
 psa-n-í
 write-THV-PASS.PTCP-NMLZ
 ‘write’ → ‘written’ → ‘writing’ *Czech*
- c. pys-a-ty → pys-a-n-yj →
 write-THV-INF write-THV-PASS.PTCP-M.SG
 pys-a-nn-ja
 write-THV-PASS.PTCP-NMLZ
 ‘write’ → ‘written’ → ‘writing’ *Ukrainian*
- (5) a. u-gan-u-ti → u-gan-u-t →
 PFV-sprain-THV-INF PFV-sprain-THV-PASS.PTCP
 u-gan-u-t-je (uganuće)
 PFV-sprain-THV-PASS.PTCP-NMLZ
 ‘sprain’ → ‘sprained’ → ‘a sprain’ *BCMS*
- b. vitchn-u-c → vitchn-u-t-i →
 sprain-THV-INF sprain-THV-PASS.PTCP-M.SG
 vitchn-u-t-je (vitchnuce)
 sprain-THV-PASS.PTCP-NMLZ
 ‘sprain’ → ‘sprained’ → ‘a sprain’ *Pannonian Rusyn*
- c. zaje-ti → zaje-t →
 capture-INF capture-PASS.PTCP
 zaje-t-je
 capture-PASS.PTCP-NMLZ
 ‘capture’ → ‘captured’ → ‘capture’ *Slovenian*

At the same time, the similarity in form across Slavic does not mean that all such nominalizations have the same internal structure. Previous literature has shown substantial cross-linguistic variation in how much verbal structure they preserve. Russian perfective-derived nominalizations may occur in contexts where perfective verbs are otherwise restricted, suggesting that the nominalization does not straightforwardly preserve the aspectual syntax of the base verb (Tatevosov, 2011). Czech, by contrast, shows stronger preservation of aspectual properties in comparable nominalizations (Biskup, 2023).

BCMS is particularly relevant in this regard because it has been argued to show different nominalization types within a single

language. Imperfective-derived nominalizations are productive, semantically transparent, and prosodically faithful to the verbal base. For example, *priznavati* ‘admit.IPFV’ gives *priznavanje* ‘admitting’. Perfective-derived nominalizations behave differently. They are less productive, more prone to semantic opacity, and typically prosodically unfaithful, as in *priznati* ‘admit.PFV’ → *priznanje* ‘admittance, recognition’. Simonović and Arsenijević (2014) analyze this contrast structurally. On their account, imperfective *-nje* nominalizations inherit verbal structure, while perfective forms undergo forced lexicalization and are stored as simplex nominal items despite their morphological complexity. In other words, imperfective forms such as *priznavanje* are closer to true verbal nominalizations, whereas perfective forms such as *priznanje* are closer to lexical nouns.

One complication is that this distinction is not visible through a single categorical diagnostic. Both types reject ordinary accusative objects, as shown in 6.

- (6) a. **priznavanje krivic-u*
admitting guilt-ACC
‘admitting guilt’
b. **priznanje krivic-u*
admittance guilt-ACC
‘admittance of guilt’

This is important because accusative case is one of the clearest signs of verbal structure in English verbal gerunds. In BCMS *-nje* nominalizations, however, accusative case is generally unavailable. Therefore, accusative case cannot be used to distinguish the two types. Instead, both types can occur with genitive complements, as shown in 7.

- (7) a. *priznavanje krivic-e*
admitting guilt-GEN
‘admitting guilt’
b. *priznanje krivic-e*
admittance guilt-GEN
‘admittance of guilt’

This means that the mere possibility of a genitive complement cannot by itself prove that one type contains verbal structure and the other

does not. Both types can also occur with event-related adjectival modification.

- (8) a. iskreno priznavanje krivice
sincere admitting guilt.GEN
‘sincerely admitting guilt’
b. iskreno priznanje krivice
sincere admittance guilt.GEN
‘sincere admittance of guilt’

Finally, pluralization is also possible with both types. Both *priznavanja* and *priznanja* can occur as plural forms. This creates the same problem: if pluralization is treated as a categorical diagnostic, it does not separate imperfective and perfective *-nje* nominalizations. Both are compatible with plural morphology.

The relevant contrast must therefore be stated quantitatively. If imperfective *-nje* nominalizations preserve more verbal structure, they should be more strongly associated with internal-argument realization, expressed by postnominal genitive complements. If perfective *-nje* nominalizations are more nominal or lexicalized, they should be more strongly associated with pluralization. The prediction is not that genitives occur only with imperfective forms or that plurals occur only with perfective forms. The prediction is that the two types differ systematically in frequency: imperfective nominalizations should occur more often with genitive complements, while perfective nominalizations should pluralize more readily.

3. Method

3.1. Dataset

The dataset consists of 1,000 aspectual pairs of BCMS *-nje/-će* nominalizations drawn from WeSoSlav, a Western South Slavic verbal database containing 5,300 frequent BCMS verbs annotated for a range of phonological, morphological, semantic, and syntactic properties and made available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license (Arsenijević et al., 2024). Since imperfective verbs in principle derive *-nje/-će* nominals, whereas perfective verbs do so only in a subset of cases, we started from perfective *-nje/-će* nominals marked as acceptable in WeSoSlav and identified

their imperfective counterparts. This produced a dataset of 1,000 aspectual pairs, i.e. 2,000 nominalizations.

The use of aspectual pairs is crucial. Rather than comparing unrelated lexical items, the study compares perfective and imperfective nominalizations that belong to the same aspectual pair. This makes it possible to test whether aspectual derivation correlates with different nominal distributional behavior.

3.2. Query types and operationalization

For each nominalization, three query types were generated:

1. lemma frequency;
2. plural forms;
3. postnominal genitive complements;

Since WeSoSlav contains verbs from all four BCMS varieties selected on the basis of independent frequency counts in four corpora, it annotates different variants of the same verb as independent lexical entries. This concerns not only regional variation, such as Ekavian/Ijekavian alternations and suffixes used in borrowed verbs, but also less regionally marked variation. For example, the perfective verb *iznev(j)eriti* ‘fail’ corresponds to two entries which have as their nominalizations *iznevjerjenje* and *izneverenje*. Since the searched corpora contain substantial regional variation, these variants were unified under the regular expression *iznev(e|je)renje*. The corresponding imperfective verb may be formed with either *-iva-* or *-ava-*, yielding four possible forms of the corresponding nominalization: *izneverivanje*, *izneveravanje*, *iznevjerivanje*, and *iznevjeravanje*. These were captured by the regular expression *iznev(e|je)r(a|i)vanje*.

The three query templates for this nominalization are illustrated below.

1. lemma frequency:

```
[lemma="iznev(e|je)r(a|i)vanje"]
```

2. postnominal genitive complements:

```
[lemma="iznev(e|je)r(a|i)vanje"] [tag="N.*g"]
```

3. plural forms:

```
[lemma="iznev(e|je)r(a|i)vanje" & tag="N.*p.*"]
```

The postnominal genitive complement was operationalized as a genitive-marked noun occurring immediately after the target nominalization. This restriction was deliberate. Genitive nominal arguments may also occur at a greater distance from the nominal head, with one or more words intervening. However, widening the search window increases the likelihood of retrieving genitive nouns that are not arguments of the nominalization itself, but instead participate in another syntactic relation. The immediately postnominal position was therefore treated as the most reliable diagnostic position, prioritizing precision over recall. The same query logic was applied uniformly across all aspectual pairs and all four corpora. A systematic investigation of wider windows allowing intervening non-prepositional material is left for future work.

3.3. Automated corpus querying

The study uses *Corpus Querier* (Kovačević, 2025), a Python-based system for automated large-scale corpus query execution and attestation-count extraction. The system was designed specifically for hypothesis-driven corpus linguistics, where large numbers of structurally parallel queries must be executed reproducibly across corpora while preserving an explicit correspondence between theoretical diagnostics and query templates.

The workflow proceeds through spreadsheet-based query generation. Each theoretically motivated query type is represented as a parameterized CQL template, which is automatically instantiated for every lexical item in the dataset. Regular-expression expansion is used to neutralize variation, ensuring comparability across corpora and varieties.

The generated query sets were executed automatically through the Sketch Engine API using *Corpus Querier* (Kovačević, 2025). The queries were submitted to the Sketch Engine `view` endpoint (<https://api.sketchengine.eu/bonito/run.cgi/view>). The implementation includes timeout handling, repeated query attempts, autosaving, failure detection, and resumable execution. Query results are written directly back into structured spreadsheets together with

runtime metadata and processing summaries. The system additionally tracks elapsed processing time, failed queries, and effective query throughput, allowing transparent monitoring of extraction efficiency and reproducibility.

The resulting query sets were executed across four BCMS web corpora: srWaC (Ljubešić and Klubička, 2016c), hrWaC (Ljubešić and Klubička, 2016b), bsWaC (Ljubešić and Klubička, 2016a), and meWaC (Ljubešić and Erjavec, 2021). For each variety, 6,000 queries were executed: three query types for each of 2,000 nominalizations. Across all four varieties, this yielded a total of 24,000 corpus queries.

To assess the reliability of the morphosyntactic classifications underlying the corpus tagging, two random samples of 1,000 concordance hits from meWaC were manually verified (Simonović, 2026), one for genitives and one for plural forms. For the plural query, 950 of the 1,000 sampled hits (95.0%) were manually classified as plural forms. For the genitive query, 998 of the 1,000 sampled hits (99.8%) were manually confirmed as genitive. In both samples, the annotator was required to make a binary decision even in ambiguous cases, reflecting the categorical analysis assigned by the automatic morphosyntactic tagger.

The system processes approximately 300–400 corpus queries per hour under ordinary server conditions while requiring minimal researcher intervention after initialization. By contrast, equivalent manual querying requires a sequence of repetitive operations for each datapoint, including copying queries, executing searches, retrieving counts, and recording results. Under ideal conditions, manual querying typically requires approximately 15–20 seconds per query while demanding uninterrupted researcher attention and introducing numerous potential failure points, including transcription errors, skipped datapoints, accidental duplication, and spreadsheet misalignment. At the scale of the present study, fully manual extraction would therefore require many dozens of hours of highly repetitive labor.

The automated workflow substantially improves reproducibility and scalability. Because query generation is template-based and fully explicit, the entire extraction procedure is transparent, auditable, and directly replicable across datasets, corpora, and languages. This makes it possible to evaluate theoretically motivated grammatical hypotheses over datasets of a scale that would be im-

practical to construct manually within ordinary linguistic research workflows.

Statistical analyses were conducted in R (R Core Team, 2026) using mixed-effects logistic regression models implemented in the `lme4` package (Bates et al., 2015). Data manipulation and visualization were carried out using the `tidyverse` ecosystem (Wickham et al., 2019), and model summaries were extracted using `broom.mixed` (Robinson et al., 2023).

4. Results

The results reveal a stable contrast between imperfective and perfective *-nje* nominalizations across all four BCMS varieties. The percentages in Table 1 are pooled corpus-level proportions rather than averages across individual nominalizations. Within each corpus, counts were aggregated separately for imperfective and perfective nominalizations. The pluralization rate was calculated as the total number of plural-form occurrences divided by the total number of occurrences of nominalizations of the corresponding aspect, and the genitive-complement rate as the total number of occurrences immediately followed by a genitive-marked noun divided by that same total occurrence count.

Table 1 presents the overall shares of plural forms and occurrences with postnominal genitive complements for imperfective and perfective NTN across the four corpora. These values are calculated directly from the aggregated token counts within each aspect–corpus combination, by dividing the total number of plural or genitive-complement tokens by the total frequency of the corresponding NTNs. Thus, unlike the lemma-level averages presented in the following table, the rates in Table 1 are weighted by lemma frequency.

Table 1: Pluralization and postnominal genitive-complement rates by aspect and corpus.

Corpus	IPFV PL (%)	PFV PL (%)	IPFV GEN (%)	PFV GEN (%)
srWaC	5.32	14.64	32.44	21.05
hrWaC	5.03	9.71	32.14	21.53
bsWaC	4.91	10.88	33.90	22.62
meWaC	4.23	13.41	35.79	25.34

Table 2 presents the same measures calculated at the level of individual lemmas. For each lemma, the proportion of plural and genitive-complement occurrences was first calculated by dividing the respective number of attestations by the lemma’s overall frequency, after which these proportions were averaged within each aspect–corpus combination. Unlike the pooled token-level rates in Table 1, this procedure gives equal weight to each lemma irrespective of its corpus frequency. The close correspondence between the two sets of values indicates that the observed aspectual differences are not driven primarily by a small number of highly frequent NTNs.

Table 2: Pluralization and postnominal genitive-complement rates by aspect and corpus.

Corpus	IPFV PL (%)	PFV PL (%)	IPFV GEN (%)	PFV GEN (%)
srWaC	5.32	14.60	32.40	21.10
hrWaC	5.03	9.71	32.10	21.50
bsWaC	4.91	10.90	33.90	22.60
meWaC	4.23	13.40	35.80	25.30

As shown in Tables 1 and 2, imperfective nominalizations are less likely to occur in plural forms and more likely to occur with postnominal genitive complements. Perfective nominalizations show the opposite tendency: they pluralize more readily and occur less frequently with genitive complements.

To test the effect of aspect on genitive-complement realization, we fitted a mixed-effects logistic regression model predicting the proportion of genitive complements relative to overall lemma frequency, with aspect as a fixed effect and aspectual pair included as a random intercept. The model reveals a highly significant effect of aspect ($\beta = -0.81$, $z = -309.75$, $p < .001$). Perfective nominalizations are substantially less likely to occur with postnominal genitive complements than imperfective nominalizations. The random intercept structure further indicates that the effect remains robust across aspectual pairs rather than being driven by a small number of high-frequency lexical items. The magnitude of the effect suggests that aspect is not merely a marginal correlate of genitive-complement realization but a major predictor of nominal distributional behavior in BCMS *-nje* nominalizations.

To examine potential microvariation across BCMS varieties, we fitted an additional mixed-effects logistic regression model in-

cluding the interaction between aspect and variety. The model again reveals a highly significant main effect of aspect ($\beta = -0.87$, $z = -198.62$, $p < .001$), confirming that perfective nominalizations are systematically less likely to occur with genitive complements. At the same time, the interaction terms between aspect and variety are also significant, indicating limited but detectable cross-variety differences in the magnitude of the effect. Crucially, however, the direction of the asymmetry remains stable across all four varieties: imperfective nominalizations consistently favor genitive complements more strongly than perfective nominalizations. These results suggest that the underlying structural contrast is robust throughout BCMS, while its corpus realization is subject to gradient and variety-sensitive distributional variation.

To test the effect of aspect on pluralization, we fitted a mixed-effects logistic regression model predicting plural frequency relative to overall lemma frequency, with aspect as a fixed effect and aspectual pair included as a random intercept. The model reveals a highly significant effect of aspect ($\beta = 1.87$, $z = 310.9$, $p < .001$). Perfective nominalizations are substantially more likely to occur in plural forms than imperfective nominalizations. The size of the effect indicates that aspect is a major predictor of pluralization behavior in BCMS-*nje* nominalizations. The random intercept structure further shows that the effect is stable across aspectual pairs and not reducible to a limited set of lexical items.

A parallel mixed-effects logistic regression model predicting pluralization likewise reveals a highly significant main effect of aspect ($\beta = 1.94$, $z = 226.56$, $p < .001$). Perfective nominalizations are systematically more likely to occur in plural forms than imperfective nominalizations. Interaction terms between aspect and variety are again significant, indicating limited cross-variety differences in the strength of the effect. Nevertheless, the direction of the asymmetry remains stable throughout BCMS: perfective nominalizations consistently show higher pluralization rates across all four corpora. Together with the genitive-complement models, these results support the view that the perfective/imperfective distinction corresponds to a robust structural contrast reflected in large-scale distributional behavior.

5. Discussion

5.1. Structural interpretation of the contrast

The findings strongly support the structural contrast proposed by [Simonović and Arsenijević \(2014\)](#). Across all four BCMS varieties, imperfective *-nje* nominalizations systematically pattern as structurally more verbal, while perfective *-nje* nominalizations pattern as more lexicalized and more strongly noun-like. Importantly, this contrast does not emerge only in isolated examples or in a restricted lexical domain. Rather, it appears robustly across thousands of nominalizations and tens of thousands of corpus queries, suggesting that the asymmetry reflects a stable grammatical generalization rather than accidental lexical tendencies.

The quantitative profile of imperfective nominalizations is particularly revealing. Their lower pluralization rates and substantially higher frequencies of postnominal genitive complements are precisely what would be expected if these forms preserve more internal verbal structure than their perfective counterparts. The results therefore provide large-scale corpus support for analyses according to which imperfective *-nje* nominalizations retain verbal projections absent from more lexicalized perfective forms.

At the same time, the findings also clarify the limits of the contrast. The distinction is not categorical in the sense that one class categorically permits a construction while the other categorically excludes it. Both perfective and imperfective nominalizations may occur with genitive complements, and both may occur in plural forms. The contrast instead emerges as a stable quantitative asymmetry distributed across large numbers of lexical items. This point is theoretically important because it suggests that the relevant grammatical distinction is gradient and distributional rather than reducible to binary grammaticality contrasts.

5.2. BCMS nominalizations and the problem of categorical diagnostics

The results also clarify why BCMS *-nje* nominalizations resist a simple mapping onto the English-based typology of verbal gerunds, process nominals, and result nominals developed in work such as [Alexiadou \(2001\)](#). English verbal gerunds preserve extensive verbal structure, including accusative object licensing and adverbial modification. Result nominals, by contrast, behave as ordinary lexical

nouns. BCMS *-nje* nominalizations occupy a more complex intermediate position.

Imperfective *-nje* nominalizations are clearly not English-type verbal gerunds. They do not normally license accusative objects and therefore do not preserve the full range of verbal behavior associated with canonical gerundive structures. However, they are also not ordinary result nominals. Their semantic transparency, prosodic faithfulness, high productivity, and strong association with genitive complements all point toward the preservation of at least some verbal structure.

One possible interpretation of this intermediate status is that imperfective *-nje* nominalizations involve nominalization over reduced verbal structure, plausibly including passive-like verbal material (see [Kovačević and Simonović 2026](#) for elaboration). On such an analysis, the absence of accusative case follows naturally if the relevant structure lacks an accusative-assigning active Voice/v head. Genitive complements can then be understood as internal arguments licensed in the nominal domain, for example by a categorizing *n* head, rather than by verbal accusative assignment ([McIntyre, 2014](#)). This analysis captures the hybrid nature of the constructions: they preserve important traces of verbal structure while simultaneously behaving as morphosyntactic nominals.

Perfective-derived *-nje* nominalizations, by contrast, are more plausibly analyzed as root-based or lexicalized nominal formations. On this view, genitive complements associated with perfective forms are not inherited verbal arguments in the same sense as those associated with imperfective nominalizations. Rather, they are nominal dependents introduced after nominalization. This explains why genitive complements remain possible while also predicting that they should occur less systematically and less productively than with imperfective-derived forms.

The results therefore support a view in which BCMS *-nje* nominalizations do not instantiate a single uniform structural type. Instead, they appear distributed across a continuum of verbal and nominal behavior, with imperfective forms clustering closer to structurally verbal nominalizations and perfective forms clustering closer to lexicalized nominal nouns.

5.3. Quantitative asymmetries and gradient grammatical structure

One of the broader theoretical implications of the study concerns the status of grammatical diagnostics themselves. In much theoretical work, diagnostics such as pluralization and internal-argument realization are often treated categorically: either a construction allows them or it does not. The BCMS data suggest that such a perspective may obscure important structural generalizations.

Both pluralization and genitive complements are grammatically possible with both nominalization types. If the analysis were restricted to isolated examples, the contrast between perfective and imperfective nominalizations would therefore appear considerably weaker and potentially unstable. The large-scale corpus evidence presented here shows that the relevant distinction emerges not through categorical possibility but through systematic differences in frequency and distribution.

This point is particularly important because it shifts the theoretical focus from isolated grammaticality judgments toward distributional structure across the lexicon. The results suggest that grammatical architecture may manifest itself probabilistically in corpus behavior even when categorical contrasts are weak or absent. In this respect, the study aligns with a broader movement toward integrating quantitative evidence into formally oriented grammatical research without abandoning structural explanation.

The interaction models further reinforce this interpretation. While the direction of the contrast remains stable across all four BCMS varieties, the magnitude of the effect varies across corpora. This indicates that the underlying grammatical asymmetry is robust but that its usage realization is gradient and sensitive to variety-specific distributional tendencies. Such findings are difficult to accommodate within rigid categorical models but fit naturally within approaches that allow grammatical structure to interact with lexicalization, usage frequency, and distributional conventionalization.

5.4. Methodological implications

The study also has broader methodological implications for corpus-based theoretical linguistics. The relevant contrast investigated here is not whether a form is merely possible, but how systemat-

ically it occurs across thousands of lexical items and multiple corpora. Testing such hypotheses using manual corpus methods would require extremely large amounts of repetitive annotation and would be highly vulnerable to transcription errors and researcher inconsistency.

The automated querying workflow introduced here makes this type of research practically feasible. Because query generation is template-based and reproducible, theoretically motivated hypotheses can be tested systematically across large corpora while preserving explicit correspondence between grammatical diagnostics and corpus queries. This substantially changes the empirical scale on which theoretically oriented linguistic work can operate.

More broadly, the study demonstrates that automated corpus querying can function not merely as a convenience tool for corpus linguistics, but as a methodological infrastructure for formal grammatical research. The combination of explicit structural hypotheses, reproducible query generation, and large-scale quantitative testing makes it possible to evaluate grammatical generalizations over datasets far larger than those traditionally used in theoretically oriented studies.

In this sense, the contribution of the present study is not only empirical but also methodological. It shows that large-scale corpus evidence can be integrated directly into formal argumentation about grammatical structure while maintaining explicit theoretical interpretability of the relevant diagnostics and constructions.

5.5. Limitations and future directions

Several limitations of the present study should be acknowledged. At the same time, many of these limitations follow directly from the decision to prioritize reproducible large-scale corpus querying over manually curated small-scale annotation. The trade-off between empirical scale, annotation depth, and reproducibility is therefore central to interpreting the present results.

A first limitation concerns the operationalization of post-nominal genitive complements. The extraction procedure identifies genitive-marked nominal material occurring in a defined postnominal context relative to the target nominalization. While this captures the overwhelming majority of relevant constructions, the automatically extracted counts inevitably include some degree of noise typical of large web corpora. In particular, the extraction procedure does not

fully distinguish inherited verbal arguments from all possible nominal dependents, adjunct-like modifiers, or parsing ambiguities. The present study therefore treats the extracted frequencies as large-scale distributional proxies for argument realization rather than as manually validated token-level syntactic annotations.

At the same time, several factors mitigate this limitation. Most importantly, the same query logic was applied uniformly across all aspectual pairs and all corpora. The central results therefore depend not on perfect extraction accuracy for individual tokens, but on systematic distributional asymmetries emerging across thousands of nominalizations and tens of thousands of automatically generated corpus queries. Moreover, manual verification of random samples of 1,000 meWaC concordance hits for each diagnostic showed reassuring agreement with the automatic morphosyntactic classifications (Simonović, 2026): 95.0% of hits returned by the plural query were manually classified as plural, while 99.8% of hits returned by the genitive query were manually confirmed as genitive. These results provide additional support for the reliability of the morphosyntactic information underlying the large-scale extraction, although the genitive verification concerns case identification rather than the syntactic relation of the genitive to the nominalization.

An important avenue for extending the present methodology is the use of dependency-parsed corpora and syntactic extraction tools. The token- and morphosyntactic-tag-based CQL approach adopted here enables efficient and reproducible querying at scale, but it cannot directly identify syntactic dependencies between non-adjacent elements. Tools developed for dependency-based corpus exploration offer a promising way of overcoming this limitation. In particular, STARK (Krsnik and Dobrovoljc, 2025), a toolkit for dependency subtree extraction and analysis, could enable genitive complements to be identified on the basis of their syntactic relation to the nominalization rather than linear adjacency. Related work on dependency-treebank exploration and the Drevesnik environment (Dobrovoljc, 2024) further demonstrates the potential of syntactically informed querying for this type of research. Integrating such approaches with the scalable workflow developed here would be a valuable direction for future work, allowing the present diagnostics to be extended beyond linear CQL patterns while retaining systematic large-scale extraction.

A second limitation concerns the interpretation of frequency-based diagnostics themselves. The present study does not claim that corpus frequencies directly reveal syntactic structure in any mechanical sense. Rather, the argument is that large-scale quantitative asymmetries may provide indirect evidence compatible with particular structural analyses when interpreted together with independent theoretical considerations. The claim is therefore probabilistic and inferential rather than categorical. Frequent association with genitive complements does not by itself prove the presence of preserved verbal structure, nor does frequent pluralization automatically demonstrate lexical nominal status. Instead, the significance of the results lies in the convergence of multiple diagnostics across large lexical domains and multiple corpora.

Relatedly, the present study focuses on broad aspectual contrasts across nominalization types and does not yet systematically model additional semantic or lexical factors that may influence distributional behavior. Certain subclasses of nominalizations may plausibly differ with respect to event structure, lexicalization, semantic transparency, register sensitivity, or degree of conventionalization. Likewise, particular verbal classes may favor or disfavor specific nominal patterns independently of aspect. Future research can therefore refine the present models by incorporating semantic subclassification, lexical frequency effects, collocational behavior, and additional morphosyntactic diagnostics.

Another important limitation concerns the nature of the corpora themselves. The study relies on large web corpora, which provide substantial empirical coverage but also contain heterogeneous material with varying levels of normalization and annotation quality. Web corpora are particularly valuable for studying large-scale distributional tendencies because they capture broad usage across many lexical items and genres. However, they also inevitably introduce variability associated with noise, duplication, inconsistent metadata, and annotation imperfections. Future work could therefore compare the present findings with results derived from more carefully balanced corpora or dependency-parsed datasets.

More broadly, the present study should be understood as an initial demonstration of a scalable methodology rather than as a definitive account of all dimensions of BCMS nominal structure. The workflow introduced here is readily extendable to ad-

ditional diagnostics, additional Slavic languages, and more complex query architectures. In particular, future work could integrate dependency-parsed corpora, contextual semantic annotation, acceptability-judgment data, or embedding-based lexical representations in order to investigate how distributional tendencies interact with grammatical structure across larger typological domains.

Finally, the study points toward a broader methodological shift in theoretically oriented corpus linguistics. Traditionally, formally explicit grammatical analyses have often relied on relatively small datasets because large-scale manual querying and annotation impose severe practical limitations. The present workflow demonstrates that automated hypothesis-driven querying makes it possible to evaluate structural hypotheses across empirical scales that would be difficult to achieve manually while still preserving explicit theoretical interpretability of the relevant diagnostics. The broader challenge for future work is therefore not merely to increase corpus scale, but to integrate scalability, reproducibility, and formal grammatical reasoning into a unified empirical research infrastructure.

6. Conclusion

This paper has presented a reproducible method for hypothesis-driven corpus querying at scale and applied it to the study of BCMS *-nje* nominalizations. Combining automated corpus querying with large-scale quantitative analysis, the study investigated whether perfective and imperfective *-nje* nominalizations differ systematically with respect to two distributional diagnostics associated with internal verbal structure: pluralization and postnominal genitive complements.

The empirical results reveal a highly stable asymmetry across all four BCMS varieties examined in the study. Imperfective nominalizations are systematically less likely to occur in plural forms and systematically more likely to occur with postnominal genitive complements, while perfective nominalizations show the opposite tendency. Mixed-effects logistic regression models demonstrate that these effects are extremely robust statistically and remain stable across aspectual pairs and corpora. At the same time, interaction models reveal limited but detectable cross-variety differences in the magnitude of the effects, indicating that the underlying grammatical contrast is

stable throughout BCMS while its corpus realization remains gradient and variety-sensitive.

These findings provide large-scale quantitative support for analyses such as [Simonović and Arsenijević \(2014\)](#), according to which imperfective *-nje* nominalizations preserve more verbal structure than perfective-derived forms. More broadly, the results suggest that BCMS *-nje* nominalizations cannot be adequately described using simple categorical diagnostics inherited from English-based nominalization typologies. Both perfective and imperfective nominalizations permit pluralization and genitive complements to some extent. The relevant contrast instead emerges distributionally, through systematic quantitative asymmetries visible only at sufficiently large empirical scale.

The study therefore has broader implications for the relationship between grammatical structure and corpus distribution. The findings suggest that structurally meaningful distinctions may manifest probabilistically rather than categorically, particularly in domains involving lexicalization and mixed verbal-nominal behavior. In this respect, the results support approaches that integrate quantitative corpus evidence into formally oriented grammatical analysis while retaining explicit structural interpretation of the relevant diagnostics.

Methodologically, the paper demonstrates that theoretically motivated corpus hypotheses can be tested reproducibly over datasets substantially larger than those typically employed in formally oriented research on nominal structure. Using template-based automated query generation and execution, the study carried out 24,000 corpus queries across four BCMS web corpora and thousands of lexical items. This makes it possible to evaluate structural hypotheses not only through isolated examples or small manually collected datasets, but through large-scale, systematically constructed bodies of evidence.

More generally, the workflow introduced here suggests a broader role for automated corpus querying in theoretical linguistics. Rather than functioning merely as a technical convenience for corpus extraction, reproducible large-scale querying can serve as methodological infrastructure for testing formally explicit grammatical hypotheses across large lexical domains and multiple varieties. The present study therefore contributes not only to the analysis of Slavic nomi-

nalization, but also to the development of scalable empirical methodologies for theoretically oriented linguistic research.

Acknowledgements

Marko Simonović's work on this paper was supported by the Austrian Science Fund (FWF), project Multifunctionality in Morphology (Grant DOI: 10.55776/I6258; PI: Marko Simonović; University of Graz).

Language Resources

Arsenijević, Boban and Gomboc Čeh, Katarina and Marušič, Franc Lanko and Milosavljević, Stefan and Mišmaš, Petra and Simić, Jelena and Simonović, Marko and Žaucer, Rok. 2024. *Database of the Western South Slavic Verb HyperVerb 2.0 – WeSoSlav*. Slovenian language resource repository CLARIN.SI. PID <http://hdl.handle.net/11356/1846>. ISSN 2820-4042; CC BY 4.0.

Ljubešić, Nikola and Erjavec, Tomaž. 2021. *Montenegrin web corpus meWaC 1.0*. Slovenian language resource repository CLARIN.SI. PID <http://hdl.handle.net/11356/1429>. ISSN 2820-4042.

Ljubešić, Nikola and Klubička, Filip. 2016a. *Bosnian web corpus bsWaC 1.1*. Slovenian language resource repository CLARIN.SI. PID <http://hdl.handle.net/11356/1062>. ISSN 2820-4042.

Ljubešić, Nikola and Klubička, Filip. 2016b. *Croatian web corpus hrWaC 2.1*. Slovenian language resource repository CLARIN.SI. PID <http://hdl.handle.net/11356/1064>. ISSN 2820-4042.

Ljubešić, Nikola and Klubička, Filip. 2016c. *Serbian web corpus srWaC 1.1*. Slovenian language resource repository CLARIN.SI. PID <http://hdl.handle.net/11356/1063>. ISSN 2820-4042.

Simonović, Marko. 2026. *Manual Verification Dataset for Plural and Genitive Identification in BCMS Web Corpora*. Zenodo. PID <https://doi.org/10.5281/zenodo.22280756>. Dataset.

References

- Artemis Alexiadou. 2001. *Functional Structure in Nominals: Nominalization and Ergativity*, volume 42 of *Linguistik Aktuell/Linguistics Today*. John Benjamins, Amsterdam and Philadelphia.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. *Fitting Linear Mixed-Effects Models Using lme4*. *Journal of Statistical Software*, 67(1):1–48.
- Petr Biskup. 2023. *Aspect Separated from Aspectual Markers in Russian and Czech*. In Petr Biskup, Marcel Börner, Olav Mueller-Reichau, and Iuliia Shcherbina, editors, *Advances in Formal Slavic Linguistics 2021*, volume 8 of *Open Slavic Linguistics*, pages 61–98. Language Science Press, Berlin.
- Kaja Dobrovoljc. 2024. *Can’t see the Forest for the Trees: Tools and Services for Investigating Slovene Dependency Treebanks*. In *CLARIN Annual Conference Proceedings*, pages 107–111.
- Predrag Kovačević. 2025. *Corpus querier: A python script for automated corpus querying with fair use compliance*. GitHub repository, <https://github.com/pedja-kovacevic/Corpus-Querier>. Version 1.0. Computer software.
- Predrag Kovačević and Marko Simonović. 2026. *Typology of nominalizations and the NP/DP parameter: Evidence from minimal pairs of native and latinate nominalizations in BCMS and English*. Manuscript under review.
- Luka Krsnik and Kaja Dobrovoljc. 2025. *STARK: A Toolkit for Dependency (Sub)Tree Extraction and Analysis*. In *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, pages 44–51, Ljubljana, Slovenia. Association for Computational Linguistics.
- Andrew McIntyre. 2014. *Constraining argument structure in nominalizations: The case of english -er*. *Lingua*, 141:121–138.
- R Core Team. 2026. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- David Robinson, Alex Hayes, and Simon Couch. 2023. *broom.mixed: Tidying Methods for Mixed Models*. *R package version 0.2.9.5*.

- Marko Simonović and Boban Arsenijević. 2014. [Regular and Honorary Membership: On Two Kinds of Deverbal Nouns in Serbo-Croatian](#). *Lingue e Linguaggio*, 13(2):185–210.
- Sergei Tatevosov. 2011. [Severing Perfectivity from the Verb](#). *Scandoslavica*, 57(2):216–244.
- Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemond, Alex Hayes, Lionel Henry, Jim Hester, et al. 2019. [Welcome to the Tidyverse](#). *Journal of Open Source Software*, 4(43):1686.

Korpusno pretraživanje velikog obima zasnovano na hipotezi: Metod i njegova primena na glagolske imenice na *-nje* u bchs-u

Predrag Kovačević, Marko Simonović, Marija Đokić Petrović

Sažetak

Rad predstavlja reproducibilan metod za korpusno pretraživanje velikog obima zasnovano na hipotezama te njegovu primenu na nominalizacije na *-nje* u bosanskom/crnogorskom/hrvatskom/srpskom jeziku. Polazeći od pluralizacije i genitivnih dopuna kao dijagnostika unutrašnje glagolske strukture, rad ispituje razlike između perfektivnih i imperfektivnih glagolskih imenica. Rezultati pokazuju stabilne asimetrije: imperfektivne nominalizacije ređe se pluralizuju i češće se javljaju sa genitivnim dopunama, dok perfektivne nominalizacije pokazuju suprotnu tendenciju. Time se potvrđuju ranije teorijske pretpostavke i pokazuje da automatizovano korpusno pretraživanje omogućava pouzdano i reproducibilno testiranje gramatičkih hipoteza.

Cljučne reči: nominalizacija, slovenska lingvistika, bchs, glagolski vid, korpusna lingvistika, reproducibilnost

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

