NEOPHILOLOGICA 2025, T. 37, 1–32
ISSN 2353-088X
DOI: https://doi.org/10.31261/NEO.2025.37.07

Sergey Say
University of Potsdam
Germany
https://orcid.org/0000-0001-8066-9166

Ilja A. Seržant
University of Potsdam
Germany
https://orcid.org/0000-0002-8066-9251

A Frequency-Based Algorithm for Argument Extraction from Russian Treebanks

Abstract

Arguments, unlike adjuncts, are typically understood as verb-specific dependents, which includes the fact that the morphosyntactic devices used for argument encoding are determined by individual verbs. Building on this observation, we operationalize arguments as dependents whose encoding device occurs with a given verb at a significantly higher-than-average frequency. We apply an argument extraction algorithm to a dataset of 132,221 verb dependents from Russian treebanks available in the Universal Dependencies (UD) platform. To evaluate the algorithm's performance, we compare its results to a manually annotated subset, informed by *The Active Dictionary* and a detailed semantic understanding of argumenthood. The frequency-based algorithm achieves acceptable precision (approx. 0.83), with particularly few false positives, making it a promising tool for cross-linguistic applications in typologically diverse languages with UD treebanks. Theoretically, we argue that a quantitative distributional approach to valency—originally proposed in Ju. D. Apresjan's early pioneering work—broadly aligns with the in-depth semantic analyses of individual verbs and their meanings found in his later works, including *The Active Dictionary*.

Keywords

Russian, argument, treebank, frequency, case, preposition, dependency

1. Objectives and Approach[1]

As early as 1965, Jurij Derenikovich Apresjan proposed a hypothesis stating that "there is a regular correspondence between the syntactic properties of words and their semantic features" (Apresjan, 1965, p. 51). To the present day, this idea remains a cornerstone in the study of verbs and their valency across various approaches and perspectives (Helbig & Schenkel, 1983, pp. 61–62; Levin, 1993; Lazard, 1994, p. 133; Levin & Rappaport Hovav, 2005; Malchukov & Comrie (eds.), 2015, etc.).

Despite the apparent appeal of this claim, it immediately presents a challenge in selecting analytical tools. One extreme is to rely on the syntactic properties of verbs, such as their combinatorial potential, to infer claims about their meanings. For example, any Russian verb that can occur with v 'in' plus the accusative case might be interpreted as a kind of motion verb. This approach benefits from a potentially solid empirical foundation, including frequency data, but its drawback is that such inferences must be treated with caution and ideally verified through independent semantic analysis. The other extreme is to make meticulous judgments about verb meanings and trace the mechanisms by which syntactic patterns relate to semantic nuances. The main issue with this approach is that meanings are not directly observable and inevitably remain a theoretical construct.

While numerous insightful studies fall somewhere along the spectrum between these two extremes, Apresjan's contribution to the field is exceptional in that he made influential advances spanning the entire range of possible approaches to valency throughout his career. In his early work, he explored the extraction of semantic information from verbs' distributional properties, employing a wide array of quantitative techniques (Apresjan, 1965; 1967). Over time, he shifted toward the in-depth semantic end of the continuum, placing increasing emphasis on thought experiment (*myslennyj èksperiment*, see Apresjan & Páll, 1982, p. 39) and semantic decomposition (*tolkovanie*, Apresjan, 1974; 1995). This shift was also reflected in his extensive work on dictionaries (Mel'čuk et al., 1984; Apresjan (Ed.), 2004), culminating in *The Active Dictionary*, an endeavor launched

[1] The study reported here was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 317633480 – SFB 1287. We are indebted to the two anonymous reviewers, as well as to Peter Arkadiev, Maria Ovsjannikova, Anastasia Panova, Dmitri Sitchinava, Aigul Zakirova, and all participants of the Slavic Linguistics Department colloquium at the University of Potsdam for their comments on an earlier version of this paper, which we took into account during the revision. The usual disclaimers apply.

in 2014 and continuing to the present day, even after Jurij Derenikovich's passing (Apresjan (Ed.), 2014-).

This shift likely reflects Apresjan's growing dissatisfaction with the computational distributional approach of his early work, leading him to conclude that detailed semantic analysis is more accurate, insightful, and ultimately superior. While this may be true, lexicographic approaches to valency based on semantic decomposition have two inherent limitations (Sarkar & Zeman, 2000, p. 691). First, no dictionary can cover all verbs and their variable usage in actual texts—a limitation that has become even more apparent with the advent of large corpora. Even for well-documented languages like Russian, dictionaries inevitably have a restricted scope. Second, an approach relying on nuanced intuitions is typically feasible only for a researcher's native language and is impractical for most of the world's languages, especially those without strong lexicographic traditions.

Given these considerations, combining the quantitative distributional perspective with the semantics-oriented lexicographic approach remains essential for advancing the study of verb-dependent relationships. Our paper follows this approach to address the longstanding problem of distinguishing arguments from adjuncts — often considered central to the automatic extraction of valency frames from corpora (Sarkar & Zeman, 2000, pp. 691–692). Specifically, we propose a co-occurrence-based quantitative technique for automatically differentiating arguments from adjuncts and apply it to Russian treebanks from the Universal Dependencies (UD) project (de Marneffe et al., 2021), thereby aligning our approach with Apresjan's early methodology. We then compare the results with the treatment of argumenthood in *The Active Dictionary* (Apresjan (Ed.), 2014–), the hallmark of his later approach.

The results presented below are valuable in their own right and, hopefully, contribute to our understanding of valency from a token-based perspective. More importantly, however, the technique introduced here paves the way for its application to languages that lack detailed, semantically oriented accounts of their verbal lexica but have UD treebanks available. In this sense, our study serves as a preparatory step for a larger research project aimed at cross-linguistic comparison of valency class systems from a token-based perspective.

The paper is structured as follows. Section 2 examines the argument–adjunct distinction and key argumenthood criteria. Section 3 outlines our data and methodology, detailing the algorithm for distinguishing arguments from adjuncts based on co-occurrence frequencies in treebanks. Section 4 evaluates the algorithm's performance and theoretical implications. Finally, Section 5 provides a brief summary and outlook.

2. The argument-adjunct distinction: state of the art

Since at least Tesnière's "Éléments de syntaxe structurale" (1959), it has been widely recognized that some verb dependents are more closely associated with the verb than others (Lazard, 1994; Dixon, 2009, pp. 97–128). A textbook example of this distinction appears in (1), where *ona* 'she' and *našim otdelom* 'our department' represent participants inherent to the meaning of *rukovodit'* 'manage', while the adverbial phrase *s sentjabrja* 'since September' is not essential to the verb's meaning.

(1) *Ona rukovodit našim otdelom s sentjabrja.*
'She has been managing our department since September'.

While intuitively appealing, this distinction is far from unproblematic, even in terminology. The common English terms *arguments* and *adjuncts* are not universally accepted (see Frajzyngier et al., 2024 for an overview of alternatives), and mapping terms across linguistic traditions—e.g., *actants* vs. *circonstants* in French or *aktanty* vs. *sirkonstanty* in Russian—further complicates the picture. The real challenge, however, lies not in the terminology but in identifying suitable criteria for distinguishing verbal dependents.[2] The numerous formal and semantic criteria used to separate arguments from adjuncts often yield conflicting classifications (see the discussion of problematic patterns in Russian in Plungjan & Raxilina, 1998; Muravenko, 1998 and in other articles in the same issue of *Semiotika i informatika*). As a result, some scholars argue that a rigid two-way classification of verbal dependents is neither feasible nor necessary in typological research (Jacobs, 1994; Haspelmath, 2014, p. 9). Despite these caveats, we will use *arguments* and *adjuncts* as the standard terms. Before presenting our perspective, we briefly highlight key distinctions from the literature.

The most widely cited contrast between arguments and adjuncts, in both the Moscow Semantic School and beyond, is that arguments correspond to a predicate's inherent participants—essential to its meaning in terms of semantic decomposition (*tolkovanie*) (Apresjan, 1974; Boguslavskij, 1996, p. 23). Variations on this idea can be found in Van Valin (2001, p. 93) and Frajzyngier et al. (2024). A classic example is Apresjan's decomposition of *arendovat'*

[2] In this text, we use the term *dependents* as a cover term for arguments and adjuncts. In corpus linguistics, a similar notion is sometimes referred to more broadly as *complements* (Schulte im Walde 2009), although this usage may differ from traditional syntactic definitions.

'rent': "A rents C" means that in exchange for compensation D, person A acquires from person B the right to use property C for a period T" (Apresjan, 1995, Vol. 1, p. 120). While this approach often yields clear results, it is not without issues—for instance, the verb *arendovat'* involves five potential arguments, but not all are equally essential (e.g., property C must be identifiable, while compensation D may remain unspecified). Since semantic decomposition is a theoretical construct, criteria for determining a verb's inherent elements can vary (Testelec, 2001, p. 168–178).

Apart from the central but elusive contrast rooted in the semantic decomposition of verb meaning, argumenthood criteria and the corresponding approaches fall into two broad categories: semantic and syntactic. A common semantic strategy is to classify dependents by roles (agents, patients, experiencers, locations, etc.[3]). While some correlations exist — agents are typically arguments, while places and causes often are not—this is not a panacea. For instance, instruments can behave as either arguments or adjuncts, forming a continuum (Koenig et al., 2008; Bohnemeyer, 2007). More nuanced distinctions have been proposed, such as Jolly's (1993) view that adjuncts, as modifiers, are hybrids: they express semantic roles like arguments but also predicate information of their own. While such insights are valuable, they are difficult to operationalize, making them less useful for tasks like data annotation and cross-linguistic comparison.

Syntactic criteria might seem more reliable, with obligatoriness being an obvious candidate. While arguments are generally more obligatory than adjuncts, this is not a universal test, as strict syntactic obligatoriness is largely illusory (Helbig & Schenkel, 1983, pp. 35–36). Most languages allow omitting participants under certain conditions (see Ljaševskaja & Kaškin, 2015, p. 502 for Russian), and even non-obligatory prepositional phrases can function as arguments when they are present (Jolly, 1993, p. 283). Moreover, languages vary greatly in how freely they permit omission, further limiting obligatoriness as a cross-linguistic criterion.

Closely tied to argumenthood criteria is the distinction between *core* and *oblique* dependents. While no universal definition exists,[4] "oblique" typically

[3] A key problem with this approach is the lack of a universally accepted set of semantic roles, let alone a reliable method for classifying verbal dependents into discrete roles. See Bickel et al. (2014) for an overview of the challenges and an empirical alternative, which results in fuzzy clustering that doesn't apply to some types of dependents.

[4] In Role and Reference Grammar, "oblique" refers to core arguments marked by an adposition or "oblique case" (a morphological term that is, by the way, not typologically unproblematic), such as *to Bill* in *Harry gave the key to Bill* (Van Valin, 1993, pp. 40–41; Beavers, 2010). Others use "oblique" more broadly, to refer to adjuncts or adverbial modifiers (Dryer & Gensler, 2013).

refers to dependents that, despite being inherent to a verb's meaning, are morphosyntactically similar to adjuncts — such as prepositional phrases in *graduate from Harvard* or *apologize for the mistake*. This approach is useful for language-specific analysis but poses challenges for cross-linguistic comparison due to variation in case systems, adpositional usage, and verb indexing strategies.

Thus, defining the argument–adjunct distinction through simple formal contrasts is problematic (Haspelmath, 2014). However, deeper syntactic contrasts may be more useful for cross-linguistic studies. For instance, Helbig & Schenkel (1983, p. 38) highlight how German verbs like *wohnen* 'live, reside' and *sterben* 'die' interact differently with locative phrases. While *Er wohnte in Dresden* 'He lived in Dresden' and *Er starb in Dresden* 'He died in Dresden' appear structurally similar, their internal organization differs, as shown in (2) and (3), taken from (Helbig & Schenkel, 1983, p. 38).

(2) **Er wohnte, als er in Dresden war*, literally 'He lived when he was in Dresden'.

(3) *Er starb, als er in Dresden war* 'He died when he was in Dresden'.

Helbig & Schenkel (1983, p. 38) classify locative phrases with *wohnen* 'live' as "tightly bound verb complements" (*enge Verbergänzung*) and with *sterben* 'die' as "free verb complements" (*freie Verbergänzung*), aligning with the argument–adjunct distinction. While insightful, such syntactic tests are not universally applicable (Haspelmath, 2014).

Apart from morphosyntactic differences, arguments are inherently "verb-specific and thus have to be learned together with each verb, whereas the use of adjuncts is independent of particular verbs" (Haspelmath, 2014, p. 5; see also Beavers, 2010, p. 842). A key aspect of argument verb-specificity is that verbs typically require specific coding devices, such as cases and adpositions (Testelec, 2001, p. 187; see also Jacobs, 1994 on formal specificity as a dimension of valency). This property, known as "subcategorization" (syntactic combinability), coexists with "selection" (semantic combinability). Its lexical nature is evident in (4), where each Russian verb, despite semantic similarities, follows a distinct valency pattern: a scheme that links semantic participants to syntactic slots marked by specific encoding devices.

(4) a. *Petja simpatiziruet Maše.*
'Petja likes Masha' (the nominative + dative pattern; here and below, the coding device for the experiencer is mentioned first).

b. *Petja vosxiščaetsja Mašej.*
'Petja admires Masha' (the nominative + instrumental pattern).

c. *Petja vljubilsja v Mašu.*
'Petja fell in love with Masha' (the nominative + v 'in' accustive pattern).

d. *Petja ljubit Mašu.*
'Petja loves Masha' (the nominative + accusative pattern).

e. *Petja razočarovalsja v Maše.*
'Petja is disappointed in Masha' (the nominative + v 'in' locative pattern).

f. *Pete nadoela Maša.*
'Petja is fed up with Masha' (the dative + nominative pattern).

In contrast, the encoding of adjuncts is determined by their own meaning and lexical properties, with the head verb playing little role. This is shown in (5), where each adjunct follows a distinct coding pattern but can combine not just with *videt'sja* 'see each other' but with any Russian verb compatible with temporal adverbials (*priexat'* 'arrive', *ženit'sja* 'marry', etc.).

(5) a. *My videlis' vtorogo sentjabrja.*
'We saw each other on September 2nd' (the genitive pattern).

b. *My videlis' vo vtornik.*
'We saw each other on Tuesday' (the v 'in' + accusative pattern).

c. *My videlis' na prošloj nedele.*
'We saw each other last week' (the *na* 'on' + locative pattern).

d. *My videlis' na Pasxu.*
'We saw each other on Easter' (the *na* 'on' + accusative pattern).

e. *My videlis' prošlym letom.*
'We saw each other last summer' (the instrumental pattern).

Since arguments are verb-specific, unlike adjuncts, their documentation and analysis have largely been a lexicographic task, as seen in numerous dictionaries by Apresjan and colleagues (Apresjan & Páll, 1982; Mel'čuk et al., 1984;

Apresjan (ed.), 2004; Apresjan (ed.), 2014–). While theoretical linguistics sometimes offers gradual approaches to argumenthood with various criteria, lexicographers take a practical stance: dictionaries list verbs with a discrete set of dependents and their encoding patterns, sometimes grouped into larger sets (Zolotova 2006). This also applies to lexical databases like FrameNet and its derivatives, which focus on arguments rather than adjuncts. A case in point is FrameBank, based on Russian data (Ljaševskaja & Kaškin, 2015, p. 502).

Most approaches view the link between verbs and specific coding devices, such as cases and adpositions, as a **typical property** of arguments rather than their defining feature. Since dictionaries typically lack frequency data and do not capture usage variability, they do not emphasize this link. However, it plays a key role in the automatic identification of arguments based on their co-occurrence with verbs in corpora (see Korhonen et al., 2000; Schulte im Walde, 2009 and references cited there).

To summarize, approaches to argumenthood vary widely, and no agreed-upon procedure exists for distinguishing arguments from adjuncts. This paper does not aim to propose a new or "best" definition of arguments. Instead, we take a more modest approach: we explore argument-coding specificity as a relatively accessible, corpus-quantifiable property of verbal dependents and compare our frequency-based extraction technique to established semantics-based lexicographic approaches to Russian.

3. Data and methods

3.1. Data extraction and preannotation

The data for this study come from Universal Dependencies (UD), a collection of treebanks covering about 150 languages (Nivre et al., 2020; Zeman et al., 2022). Here, we focus on Russian, though the methodology applies to any UD treebank. All the spreadsheets mentioned in Section 3, as well as the code used to analyze the data, are available as the Supplementary materials at https://github.com/serjozhka/Algorithm_Argument_Extraction_Neophilologica.

UD treebanks have several properties that make them well-suited for quantitative valency studies. First, they are based on naturalistic texts (fiction, blogs, news, etc.). Second, the Russian UD treebanks are large (see below for details) and continue to grow. Third, UD treebanks provide deep morphological and syntactic annotation, consistently applied across diverse languages, facilitating token-based typological analysis. Specifically, UD annotation includes lemmatization, allowing automatic tracing of usage patterns across morphological forms and syntactic contexts. Finally, as their name suggests, UD treebanks explicitly mark dependency relations, categorized into a concise set of universal types (e.g., "nominal subject", "indirect object", "adverbial modifier").[5]

UD treebanks also have inherent limitations. The most significant one may affect our study's planned typological extension: UD is heavily biased toward "major" languages, limiting typological diversity. "Minor" languages are underrepresented, and existing corpora vary in size, content, and quality.

Additionally, UD annotation presents technical challenges for studying verbal arguments. The framework treats function words like adpositions and auxiliaries as dependents of their associated content words (de Marneffe et al., 2021, p. 269), diverging from standard dependency grammar. For instance, in *give the toys to the children*, *give* takes *toys* as a direct object ("obj") and *children* as an oblique object ("obl"), with *to* analyzed as a dependent of *children* under a "case" relation. This approach enhances cross-linguistic comparability between case-rich languages and those relying on adpositions. However, it complicates the automatic extraction of argument-encoding devices associated with specific verbs, as discussed below.

As a first step, we extracted all dependents tagged as direct objects ("obj"), indirect objects ("iobj"), or oblique objects ("obl") of finite verbs in all Russian treebanks available as part of the UD collection ("Taiga", "SynTagRus", "GSD", and "PUD")[6]. We used UD version v2.11, published on November 15, 2022 (Zeman et al., 2022). The total size of the corpus we used is 1744K tokens. For data extraction, we used a Python script developed by Ivan Seržant. The script generates a spreadsheet where each row represents a nominal or pronominal dependent annotated for 27 parameters. The key parameters for this analysis are:[7]

[5] An important question is the exact procedures used for these annotations. Unfortunately, the UD documentation is not fully explicit, and minor inconsistencies suggest the procedures varied across treebanks.

[6] We did not extract canonical nominative subjects (tagged as, e.g., "nsubj" in UD), as they are unequivocally arguments and thus irrelevant for our analysis. Notably, UD takes a conservative approach to subjecthood: dependents often described in the literature as non-canonical subjects (e.g., accusative NPs in clauses like *menja tošnit* 'I feel nauseous') are analyzed as objects or obliques—and therefore were included in our dataset. The algorithm skipped adverbial expressions like *tuda* 'there (directional)', tagged as "advmod", even if distributionally similar to v *gorod* 'to the city'.

[7] Additional parameters related to word order, verb morphology, and animacy are not relevant here but are available in the Supplementary materials and will be used elsewhere.

- "sentence": the full sentence with the target dependent and head verb.
- "verb lemma": infinitive form of the head verb.
- "object form": inflected form of the dependent.
- "object dependency relation": UD treebank tag for the dependent, e.g. "iobj".
- "object case": case marking of the dependent (e.g., "Acc", "Dat", "Gen").
- "adposition lemma": any adposition identified as a dependent of the target dependent.

Table 1 presents the annotation of a single raw entry from our spreadsheet as an example (displayed as a column rather than a row for technical reasons).

Table 1.
Sample spreadsheet entry (8 selected parameters)

entry no	23959
corpus	ru_syntagrus-ud-dev.conllu
id	2020_Corpus2_0Khochu_byt_negrom.xml_304
sentence	Я смотрю на круглое личико, и мне кажется, что это – она.
verb lemma	смотреть
finite verb	смотрю
object dependency relation	obl
object form	личико
object case	Acc
adposition lemma	на

The spreadsheet generated by the data extraction script contained 132,221 entries, where each entry corresponds to a verb dependent token. Next, we annotated the encoding devices associated with these dependents. By default, this was a combination of case form and adposition (if present), e.g., "naACC" for the entry in Table 1. However, due to UD annotation errors—such as misdisambiguation of homonymous case forms or incorrect lemmatization of prepositions—parts of the process were manual.[8] Ultimately, we identified 92 distinct encoding devices, twice the number found in Apresjan (1965, p. 46), largely due to the presence of rare prepositions in UD treebanks, such as *posredine* 'in the middle (of)' or *vzamen* 'in exchange (for)'. The resulting spreadsheet

[8] At this stage, we either reannotated or excluded all examples tagged as nominative in the original UD annotations, as these were either annotation errors or irrelevant for our purposes (e.g. when used in citations).

(*Russian_data_with_encoding*), which served as the input for the subsequent analysis, is available in the Supplementary materials.

As the next step, we removed entries where the encoding device could not be identified (e.g., indeclinable elements), marking them with <NA> tags. We also excluded nominals that were not dependents at the verb phrase level, such as the expletive *èto* 'this' in sentences like *No èto my opjat' sil'no zabegaem* 'But here we're getting ahead of ourselves again', marking them with the "EXPL" tag. After filtering out these entries, the number of tokens available for analysis was reduced to 122,551 entries.

Table 2 presents partial annotations, including the "encoding device", for three entries from the sentence *Posle ètogo ljudi rasskazyvali zalu o rezul'tatax* 'After that, people told the audience about the results.'

Table 2.

Partial annotations of three sample entries, including "encoding device" tags

entry no	verb lemma	object dependency relation	object form	object case	adposition lemma	encoding device
2386	*рассказывать*	obl	этого	Gen	после	posleGEN
2387	*рассказывать*	iobj	залу	Dat		DAT
2388	*рассказывать*	obl	результатах	Loc	о	oLOC

As noted earlier, UD explicitly rejects the argument-adjunct distinction, opting instead for the core-oblique distinction. The UD documentation justifies this choice: "… the argument/adjunct distinction is subtle, unclear, and frequently argued over … the best practical solution is to eliminate it … The core-oblique distinction is … both more relevant and easier to apply cross-linguistically than the argument-adjunct distinction" (https://universaldependencies.org/u/overview/syntax.html; retrieved March 1, 2025). In UD, this distinction is based on the morphosyntactic encoding of dependents (de Marneffe, 2021, p. 268), partially overlapping with parameters like morphological case and disregarding whether a dependent is obligatorily selected by a specific verb.

While reluctance to provide discrete argument vs. adjunct annotation is understandable, distinguishing "obj", "iobj", and "obl" dependents is not entirely satisfactory either. The UD documentation offers only vague definitions, such as "iobj" as a "nominal core argument of a verb that is not its subject or (direct) object" or "obl" as "a nominal functioning as a non-core (oblique) modifier of a predicate" (de Marneffe, 2021, p. 266). Beyond their vagueness, these categories are inconsistently applied. The "obj" tag typically marks accusative prepositionless objects, yet 2,369 of 37,691 "obj" entries involve diverse encoding strategies with unclear rationale. Often, these cases involve verbs sometimes classified as transitive due to features like passive formation (Kamynina, 1999, p. 146; but see Fowler, 1996 for a different approach). However, annotation remains inconsistent—e.g., the instrumental *rukami* '(with) arms' as a dependent of *maxat'* 'wave' appears variably as "obj", "iobj" and "obl".

The "obj" vs. "obl" distinction is more meaningful for accusative prepositionless dependents: "obj" mainly marks clear direct objects, while "obl" (and its subtype "obl:tmod") typically indicate adverbial time or frequency modifiers, such as *vsju nedelju* 'the whole week' or *každuju minutu* 'every minute.' Though not entirely consistent, this distinction helps differentiate direct objects from adverbials, a contrast not captured by other parameters.

3.2. Argumenthood annotation algorithm

In the previous section, we showed that UD-internal tags do not distinguish arguments from adjuncts. This challenge forms the basis of our analysis, which aims to develop an algorithm for automatic argumenthood annotation.

The core principle of our algorithm, introduced in Section 2, is that argument-encoding devices—such as cases and prepositions in Russian—are verb-specific and quantitatively distinct from adjuncts, whose distribution is governed by semantic compatibility and internal structure. Based on this, we operationalize argumenthood as follows: a dependent is interpreted as an argument if the relative frequency of its encoding device with a given verb exceeds a certain baseline, as specified below. This approach treats argumenthood as a gradable phenomenon but can be converted into a discrete distinction by setting a cut-off point for the frequency difference.

This approach closely resembles Apresjan's concept of "government strength" (*sila upravlenija*), defined as "the proportion of cases in which a given verb occurs with a specific (prepositional-)case form out of the total occurrences of the verb" (Apresjan, 1965, pp. 51–52), where adjuncts are dependents with very low government strength. Comparing observed frequency to a baseline is a standard method in subcategorization frame acquisition (Korhonen et al., 2000; Schulte im Walde, 2009). The main challenge is determining this baseline. A simple approach sets a fixed threshold (e.g., >20% of a verb's occurrences), while we use a more refined method, comparing observed frequency to expected frequency, which is calculated under the assumption that the encoding device occurs equally often with all verbs in the corpus (see Sarkar & Zeman, 2000 for a similar idea as applied to Czech data and Schulte im Walde 2009, p. 957 for a brief overview of available methods). Our approach avoids "penalizing" low-frequency coding devices, aiming instead to detect all cases where individual verbs attract specific constructions. For example, the combination *nad* 'above' with the instrumental case occurs 38 times with the verb *rabotat'* 'work' (as in *rabotat' nad proektom* 'work on a project') — just 8% of the 467 clauses with this verb. Yet this is high relative to the baseline frequency of *nad* + instrumental, which appears with only about 0.4% of all verb tokens. This discrepancy reflects the intuitively correct idea that *nad* + instrumental is a lexically specified pattern for expressing the domain or object of work.

This concept was implemented in R (R Core Team, 2021). The procedure was as follows. Based on the raw dataset from Section 3.1 (*Russian_data_with_encoding*), we created a spreadsheet (*valency_frames*) where each row represents a finite verb token in the treebank, and columns correspond to the 92 attested encoding devices (e.g., "ACC", "INS", "naACC"). Cells contain "0" or "1" depending on whether the specific verb token has a dependent in the specified (prepositional-) case form. For example, the three entries in Table 2 are merged into one row with "1" in the columns for "posleGEN", "DAT", and "oLOC". The *valency_frames* spreadsheet includes 87,797 entries, matching the number of verb tokens in the raw data. It was also used to calculate verb lemma prevalence; for instance, *rasskazyvat'* 'tell (IPFV)', shown in Table 2, appears 153 times in the dataset.

Next, we created a summary spreadsheet (*verbs_and_encoding_devices*) grouping co-occurrences with encoding devices by verb lemmas. The 7,538 rows represent distinct verb lemmas, while the 92 columns indicate the prevalence of encoding devices co-occurring with each verb. For example, for *rasskazyvat'* 'tell (IPFV)', which appears 153 times in the data, 76 instances involve a dependent encoded by o 'about' + locative, 58 by the dative case, and only 2 by *posle* 'after' + genitive (as seen in Table 2), along with additional dependent types.

The figures in this spreadsheet represent observed frequencies. Expected frequencies were calculated by determining the overall relative prevalence of each encoding device across all verb tokens in the corpus and multiplying it by the total number of tokens for each verb lemma. For example, dependents in the prepositionless dative case appear in 7,929 out of 87,797 clauses, yielding a ratio of 0.09. Under the null hypothesis that *rasskazyvat'* 'tell (IPFV)' selects the dative case at the same rate as other verbs, we would expect about 13.8 occurrences (= 153 × 0.09). The observed count (58) far exceeds this, indicating a stronger-than-average preference for the dative case with *rasskazyvat'*.

As the final stage in the analysis, we needed a procedure to infer argumenthood status from the differences between observed and expected frequencies. We used a two-step algorithm:

- If the expected frequency of a verb-encoding combination was ≥5, we applied the χ^2-test. The combination was classified as an argument relation if the observed frequency exceeded the expected frequency and the p-value was <0.05.
- Otherwise, we used Fisher's exact test with the same p-value threshold (<0.05) and an additional condition that the observed count was at least 3.

This procedure is largely similar to two subcategorization frame acquisition methods discussed by Korhonen et al. (2000). The key difference is the use of Fisher's exact test to address issues with low-frequency items, which make the χ^2-test unreliable (Korhonen et al., 2000, p. 205).[9] The additional requirement that the observed frequency exceed 2 helps prevent premature argumenthood classification for rare combinations.

Using this procedure, we assigned a binary argument vs. non-argument status to each verb-encoding combination, following the standard discrete approach to argumenthood (Levin & Rappaport Hovav, 2005). For example, *rasskazyvat′* 'tell (IPFV)' was classified as selecting arguments marked by the dative case, o 'about' + locative, and *pro* 'about' + accusative.

Argumenthood judgments are summarized in a spreadsheet (*argumenthood_df*), where "1" denotes arguments and "0" adjuncts. Like *verbs_and_encoding_devices*, it contains 7,538 rows (verb lemmas) and 92 columns (encoding devices). However, this format is not very clear visually. Instead of presenting it directly, Table 3 shows an excerpt combining data from two spreadsheets: it shows observed frequencies of some verb-encoding combinations from *verbs_and_encoding_devices*, with boldface indicating "1"s from *argumenthood_df*, marking algorithmically identified arguments.

Table 3.
Selected verb – encoding device combinations frequencies and their argumenthood

		INS	**ACC**	**DAT**	**vLOC**	**sGEN**	**nadINS**	**dljaGEN**
представлять	'represent'	33	**244**	28	22	1	0	**14**
иметь	'have'	7	**794**	1	102	10	1	**24**
посоветовать	'advise'	0	**22**	**15**	1	0	0	1
называть	'call'	**109**	**273**	0	33	2	0	0

[9] The main downside of Fisher's exact test is that it is computationally demanding, which makes it poorly suited for across-the-board application.

Table 3 (Continuation)

		INS	ACC	DAT	vLOC	sGEN	nadINS	dljaGEN
подвергнуться	‘undergo’	0	1	**21**	3	1	0	0
вынудить	‘force’	0	**13**	0	1	1	0	0
закончить	‘finish’	5	**50**	0	14	0	1	0
работать	‘work’	38[10]	19	0	**193**	**22**	**38**	5
становиться	‘become’	**256**	2	1	34	5	0	**20**

The spreadsheet *argumenthood_df*, which annotates each verb-encoding combination based on the quantitative argumenthood test, is valuable in its own right and can serve as a reference for future analyses.

However, at this stage, we encountered a systematic complication, exemplified by the verb *rasskazyvat'* ‘tell’ discussed above. Intuitively, this verb is transitive, with accusative dependents like *istorii* ‘stories’ functioning as clear arguments.[11] Indeed, *rasskazyvat'* appears with an accusative dependent 28 times in our data. Yet, under our frequency-based algorithm, accusative dependents of *rasskazyvat'* did not meet the quantitative argumenthood criteria: their frequency was not significantly higher than expected, reflecting the verb’s diverse valency patterns (see Table 2). This negative result is purely mathematical—since transitive verbs are common, the expected co-occurrence rate with an accusative object is high, complicating algorithmic identification (Korhonen et al., 2000, p. 204).

This issue is even more pronounced in languages with many labile verbs, where overt direct objects appear less frequently than in strictly transitive verbs. To address this, we made an exception when integrating algorithmic argumenthood judgments into our raw data: every accusative dependent was initially classified as an argument (see Section 4.3 for an additional correction). This decision reflects the special status of the basic transitive pattern in most languages, encompassing a broad range of verbs centered on action (Lazard, 1994, pp. 134–135). Many

[10] The frequency of *rabotat'* with an instrumental-case dependent did not surpass the argumenthood threshold, contrary to standard views: many such combinations involve professions (e.g., *rabotat' povarom* ‘work as a cook’) and would count as arguments in decomposition-based approaches. In principle, the algorithm could be refined by factoring in grammatical animacy — *rabotat'* with animate instrumentals is unusually frequent — but we deliberately avoided overfitting, opting for a simpler, cross-linguistically more transferable model.

[11] The issue of temporal adverbials expressed by prepositionless accusatives, such as *vsju nedelju* ‘the whole week’, will be discussed separately in Section 4.3.

linguists argue that case assignment in canonical transitives is structural rather than lexical (Yip et al., 1987, p. 222; see also de Marneffe et al., 2021, p. 267).

In any event, our binary annotation remained fully automatic and content-agnostic—aside from accusative dependents, which were assigned "1" by default, all other combinations were tagged "1" or "0" based on the frequency-based algorithmic test.

The resulting spreadsheet (*data*) integrates UD treebank annotations, Python-generated annotations, semi-manual encoding device annotations, and algorithmic binary argumenthood classifications. As discussed in Section 3.1, the number of verb-dependent tokens suitable for further analysis is 122,551. In Section 4, we evaluate the quality of our argumenthood classification.

3.3. Incorporating lexicographic annotations

To assess the algorithm's performance (Section 3.2), we compared its results with annotations based on a qualitative item-by-item analysis of raw entries, distinguishing arguments from adjuncts (this is a standard evaluation procedure in automatic acquisition of verb frames from corpora, see Schulte im Walde, 2009, p. 958). For this, we relied on two dictionaries: Rozental's *Upravlenie v russkom jazyke* (*Government in the Russian Language*) (Rozental', 1986) and *The Active Dictionary*, initiated by Ju.D. Apresjan (Apresjan (Ed.), 2014–).[12] These sources differ in scope, target audience, and approach to argumenthood.

Rozental's dictionary is a practical, prescriptive reference that lacks detailed analysis of argumenthood. It focuses on verbs with variable or problematic valency, omitting frequent, prototypically transitive verbs like *ubivat'* 'kill', while including more complex derivatives like *ubivat'sja* 'grieve, mourn bitterly.' For covered verbs, it provides typical valency patterns, often with brief explanations of relevant meanings.

In contrast, *The Active Dictionary* is a comprehensive academic work that reflects Apresjan and colleagues' views on argumenthood, semantic decomposition, and combinatorial potential. Each entry provides detailed information on meaning, syntax, and usage, including a standardized representation of government (*upravlenie*). Importantly, government patterns are provided separately for each of the verb's meanings. To compare the two sources, we included one sample dictionary entry from each of them in the Appendix at the end of the article.

[12] The annotations discussed in this section were manually prepared by Vera Arbieva Pais, to whom we express our sincere gratitude.

The dictionary-based annotations relied on whether the dictionaries listed the relevant encoding devices for the specific meaning of the given verb. Since this process was manual and required semantic analysis, it was time-consuming and impractical for the entire dataset. Ultimately, we obtained annotations for the first 5,423 entries in our *data* spreadsheet, ordered alphabetically by verb lemma. Unlike the algorithmic approach, this manual procedure was token-based — meaning that the same combination of a verb lemma and an encoding device could be classified as an argument in one entry and as an adjunct in another, depending on the meanings of the verb and its dependent in the specific sentence.

The annotations distinguish between "1" (the usage pattern is listed in the verb's government pattern and considered an argument), "0" (not listed and not an argument), and <NA> (verb not included in the dictionary). Additional tags, detailed in the Supplementary materials, appear in separate columns. The most important is "s" (for *semantic argument*), used when a syntactic dependent fulfills a semantic valency in the verb's decomposition but the specific form is not listed in its government pattern. An example is entry 44925, shown in (6).

(6) *V otvet Irina Vasil'evna rasstegnula vorotnik u Semëna Petroviča i bryznula* ***na nego*** *vodoj.*
'In response, Irina Vasilievna unbuttoned Semyon Petrovich's collar and sprayed him with water.'

The Active Dictionary states that in this meaning, *bryzgat'* 'spray, sprinkle' has three semantic arguments, one of which expresses the endpoint of spraying. We infer that *na nego*, literally 'onto him' functions as an argument in Apresjan's framework (hence the main tag "1"). However, the exact syntactic form (*na* 'on, onto' + accusative) is not specified in the verb's government pattern (*model' upravlenija*), warranting the additional tag "s" in our dataset, which reflects the morphosyntactic flexibility of the encoding pattern.

Manual dictionary-based argumenthood annotations are included in the final spreadsheet (*data_with_dictionary_annotations*).

4. Results

In this section, we compare the performance of our argument extraction algorithm with a manually annotated data subset based on dictionaries. Our goals are twofold: first, to identify weaknesses in the algorithm with a view to future improvements; second, to highlight quantitative insights that may be overlooked in lexicography, which typically ignores text frequency. That said, we do not seek to replace the standard, semantically grounded concept of argumenthood with a simplistic, black-box algorithm. Rather, we aim to compare the two perspectives for their mutual benefit.

4.1. Rozental's dictionary

Before evaluating our algorithm, we quantitatively compare two lexicographic approaches to argumenthood—those of Rozental's dictionary (Rozental', 1986) and *The Active Dictionary* (Apresjan (Ed.), 2014–)—as shown in Table 4. The counts reflect manual annotations of a subset of UD entries, as discussed in Section 3.3.[13]

Table 4.
Argumenthood: Rozental's dictionary vs. The Active Dictionary[14]

		The Active Dictionary			
		yes	no	<NA>	total
Rozental's dictionary	yes	**984**	47	65	1,096
	no	696	**510**	37	1,243
	<NA>	2,267	646	171	3,084
	total	3,947	1,203	273	5,423

As the number of <NA>s in Table 4 shows, *The Active Dictionary* covers 95% (5,150 of 5,423) of the manually annotated dataset, far surpassing Rozental's 43%. Overall, the two sources align, with matching annotations (boldfaced in the table) more common than discrepancies. However, Apresjan's approach is more inclusive, cf. 696 tokens classified as arguments in *The Active Distionary* but not in Rozental's dictionary as opposed to just 47 tokens of the opposite type.

[13] Apart from directly identifying the necessary verbs in dictionaries, we sometimes inferred them from aspectual pairs or reflexive counterparts (see the discussion of additional tags in the Supplementary materials). We used <NA> only when none of these approaches yielded a satisfactory result.

[14] In this and the subsequent tables, "yes" indicates dependents mentioned in the government patterns in the dictionaries (≈ arguments), while "no" refers to all other dependents (≈ adjuncts).

This difference underscores the lack of a principled consensus on argumenthood in naturalistic data as compared to preselected examples. In Rozental’s approach, fewer than half of (pro)nominal dependents in the text are arguments, compared to about 77% in Apresjan’s. This discrepancy should be taken into account when evaluating our algorithm’s performance. Table 5 cross-tabulates our algorithmized annotations with those based on Rozental’s dictionary.

Table 5.
Argumenthood: Rozental’s dictionary vs. frequency-based algorithm

		frequency-based algorithm		
		yes	no	total
Rozental’s dictionary	yes	1,007	89	1,096
	no	682	561	1,243
	<NA>	2,193	891	3,084
	total	3,882	1,541	5,423

As noted earlier, more than half of UD entries contain verbs missing from Rozental’s dictionary, making algorithmic approaches inherently superior for corpus analysis—even aside from the time-consuming nature of dictionary-based annotations. More importantly, the two approaches handle arguments asymmetrically: nearly all dependents listed in Rozental’s government patterns (92%, see Table 4, first row) are also classified as arguments by our algorithm. The reverse does not hold — Rozental’s non-arguments are more often identified as arguments than as adjuncts by our algorithm. This again underscores the restricted and conservative nature of Rozental’s approach to argumenthood. Manual inspection shows that these additional arguments identified by the algorithm, such as the dative NP in (7), typically do meet standard semantics-based criteria.

(7) ***Vot ètim parnjam*** *ja verju!*
‘**These guys** I do trust!’

Rozental’s dictionary does not list human dative arguments as part of the verb *verit'* ‘believe’ government pattern, likely because it focuses on variable cases, while the dative use in (7) is regular and predictable from the verb’s meaning. We conclude that our algorithm is better suited for detecting argumenthood in corpora than Rozental’s dictionary. This is not a critique of Rozental’—his dictionary was intended for contexts unrelated to automatic annotation. The rest of the paper compares our results with *The Active Dictionary*, which has a broader scope and a strong theoretical foundation.

4.2. Apresjan's dictionary: an overview

Table 6 summarizes our algorithm's performance relative to *The Active Dictionary*, as Table 5 did for Rozental's dictionary.

Table 6.
Argumenthood: The Active Dictionary vs. frequency-based algorithm

		frequency-based algorithm		
		yes	no	total
The Active Dictionary	yes	3,409	538	3,947
	no	329	874	1,203
	<NA>	144	129	273
	total	3,882	1,541	

Based on Table 6, our algorithm performs sufficiently well if the published volumes of *The Active Dictionary* are taken as the gold standard. Of the 3,738 entries identified as arguments by the algorithm, 3,409 are also classified as such in *The Active Dictionary* (excluding verbs missing from the dictionary), which yields a precision of approximately 0.91 for the algorithmic positives (see Korhonen et al., 2000 for the definition of precision). Precision for negatives is lower (≈0.62),[15] resulting in an overall precision of ≈0.83. We refer to these two kinds of mismatches as "false positives" (Section 4.3) and "false negatives" (Section 4.4). This standard terminology does not imply that the dictionary is always correct and the algorithm is wrong; as we show below, the two approaches sometimes capture different aspects of argumenthood.

[15] Sarkar & Zeman (2000, p. 696) observe a similar disparity in the performance of the best-performing algorithm among the three they tested for automatic verb frame extraction in Czech.

4.3. False positives

As seen in Table 6, false positives are rare, accounting for 329 out of 5,423 entries. One possible source is our decision (see Section 3.2.) to classify all prepositionless accusatives as arguments, including cases like (8).

(8) *Želudok bolit ne prekraščaja pjatyj den'.*
‘The stomach has been hurting nonstop for the fifth day’.

The noun phrase *pjatyj den'* ‘for the fifth day’ is clearly an adjunct, traditionally analyzed as a circumstantial (*obstojatel'stvo*) in Russian grammar. Thus, classifying all prepositionless accusatives as arguments inevitably leads to errors in such cases. The UD distinction between dependents bearing “obj” and “obl” (or occasionally “obl:tmod”) relations might seem helpful here. Table 7 cross-tabulates this distinction with the manual annotation based on *The Active Dictionary* for all prepositionless accusatives.

Table 7.
Prepositionless accusative dependents: UD-internal annotations vs. The Active Dictionary

		UD-internal annotations			
		“obj”	“obl”	“obl:tmod”	total
The Active dictionary	yes	1,068	53	6	1,127
	no	6	19	7	32
	<NA>	73	0	1	74
	total	1,147	72	14	1,233

As seen in Table 7, the “obl” and especially “obl:tmod” UD tags are slightly more common among adjuncts (“no” in the table) than arguments (“yes” in the table) when distinguished based on *The Active Dictionary*. To refine our algorithm, we adjusted the final annotation (the “argumenthood_corrected” column in the *data* spreadsheet) to rely on UD-inherent tags for accusative dependents: “obj” was classified as an argument, while all others were adjuncts. This adjustment may seem questionable for Russian, as accusative dependents tagged as “obl” often correspond to arguments in *The Active Dictionary*. However, we applied it anyway, since it helps with languages where object-like dependents frequently express adverbial meanings and had little effect on our algorithm’s accuracy for Russian. In any case, Russian accusative dependents contribute minimally to false positives. Since we applied a special rule for prepositionless accusatives, we exclude them from further discussion of our algorithm's performance, focusing on the remaining 4,190 entries.

As noted above, false positives are generally rare, and there are few consistent patterns behind their occurrence. However, a few scenarios can still be illustrated — one involves polysemous encoding devices, as shown in examples (9) and (10).

(9) *S ego pomošč'ju možno uznat', čto* ***varitsja v staleplavil'noj peči****.*
‘With its help, one can find out what is **being brewed in the steelmaking furnace**.’

(10) ***V drevnem Kitae varilos'*** *pivo iz prorosšego risa.*
‘**In ancient China**, beer **was brewed** from sprouted rice.’

Our algorithm detected a statistical association between *varit'sja* ‘be brewed’ and the encoding device v ‘in’ with the locative case. In (9), this correctly classifies the boldfaced phrase as an argument, aligning with *The Active Dictionary*. However, the algorithm generalizes this pattern to all occurrences of the same encoding device with *varit'sja*, leading to an error in (10), where the phrase denotes general localization rather than the vessel or container inherent in the verb's semantics.

While cases like (10) reflect clear errors in the algorithm's output, it occasionally captures patterns that are overlooked even in the detailed analysis of *The Active Dictionary*. One such case is shown in (11).

(11) ***Dlja pozitivista est'*** *tol'ko fakty i različnye sposoby ix vzaimouvjazki.*
‘**For a positivist, there are** only facts and various ways to connect them’.

The algorithm classified *dlja* ‘for’ phrases as arguments of *byt'* ‘be’ based on their frequent co-occurrence, though *The Active Dictionary* does not list this among the verb's many meanings and government patterns. While the boldfaced phrase in (11) is not a typical argument, it is not a standard adjunct either. Adjuncts typically narrow a clause's truth-conditional reference —for instance, *She baked a cake for him* necessarily entails *She baked a cake*. In (11), however, there is no such entailment that only facts exist and can be connected in various ways. Instead, the *dlja*-phrase effectively introduces a new meaning of *byt'* akin to ‘seem’ or ‘be considered,’ where it functions as an argument.

Another case where the algorithm offers semantic insights is shown in (12).

(12) *No kogda oni gus'kom* ***breli po lesnoj tropinke*** *… on uže tverdo znal, čto ničego ne budet.*
'But when they **trudged** in single file **along the forest path** … he already knew for sure that nothing would happen'.

Based on frequent co-occurrence, the algorithm classified *po* 'along, by' phrases and instrumental-case dependents as arguments of *bresti* 'trudge'. While *The Active Dictionary* does not include such phrases in the verb's government pattern, their frequent corpus occurrence aligns with its semantic decomposition: 'A person (A1) slowly moves toward place (A2) from place (A3), struggling to lift their feet due to weakness, fatigue, or difficult conditions—water, deep snow, or mud' (Apresjan (ed.), 2014, p. 351). Though the path is not treated as a variable like A1, A2, or A3, which defines arguments, its typical properties ("water, deep snow, or mud") are explicitly mentioned, making it more integral to the verb's meaning than a standard adjunct.

To summarize, cases we conventionally label as "false positives" partly result from the algorithm's inability to distinguish different senses of the same encoding device. However, some of these so-called false positives actually reveal insights that go beyond standard lexicographic perspectives, showing that they are not always truly false.

4.4. False negatives

False negatives arise when the algorithm fails to classify a verb's dependent as an argument, even though the dictionary does. In our manually annotated dataset, they are more frequent (538 entries) than false positives (329). Two main causes account for false negatives: data sparsity and verb polysemy.

The first scenario involves rare verbs. The algorithm requires statistically significant differences and sets an absolute threshold of at least three occurrences for a verb-plus-encoding-device combination. Under these conditions, verbs that occur in the dataset only once or twice cannot, in principle, be classified as taking arguments by the algorithm, even when the dependent in question is clearly an argument in terms of content, as illustrated by (13).

(13) *Novikov uže ballotirovalsja v mèry.*
'Novikov has already run for mayor'.

In (13), *ballotirovat'sja* ‘run for office’ appears with v (‘in’) and the rare “2nd accusative”.[16] This encoding device, specific to verbs denoting a change in social status, clearly marks arguments. However, with only one occurrence of *ballotirovat'sja* in our dataset, the algorithm could not extract its argument-encoding pattern.[17]

Thus, data sparsity is the primary cause of false negatives—an issue with no perfect solution. While our algorithm currently produces binary annotations (argument vs. adjunct), a more realistic approach might be to filter out low-frequency verbs and assign <NA>s in unclear cases instead of (potentially false) negatives.

The second major source of false negatives occurs with highly frequent, syntactically versatile verbs (see Tao et al., 2024 on the correlation between verb frequency and valency diversity). Verbs like *bit'* ‘beat’, and *vesti* ‘lead’ appear in numerous valency patterns, some of which are relatively infrequent. As a result, the algorithm fails to identify certain argument structures. This issue was already noted in Apresjan (1965), where polysemy was disregarded when extracting semantic information from frequency distributions.

Apart from the two main causes of false negatives, we also expected them to arise with verbs exhibiting “flexible” government patterns compatible with the same meaning. As noted in Apresjan and Páll (1982), many Russian verbs do not require a fixed preposition but can combine with various preposition-case pairs expressing a shared semantic role. For instance, *vernut'sja* ‘return’ can take different prepositions depending on the noun rather than the verb, e.g., *vernut'sja iz otpuska* ‘return from vacation’, *ot druga* ‘from a friend’s place’, s *koncerta* ‘from a concert’. In Apresjan and Páll’s framework, flexible patterns are categorized as P1 (source), as exemplified by examples with *vernut'sja* ‘return’, as well as P2 (goal), P3 (static location), and P4 (path). Similar observations appear in Helbig & Schenkel (1983, p. 43), Ljaševskaja & Kaškin (2015, p. 482), and *The Active Dictionary* (Apresjan (Ed.), 2014–).

We accounted for flexible government patterns in our manual annotations based on *The Active Dictionary* (see Section 3.3), marking verbs with such patterns with an additional “s” tag, as discussed in Section 3.3. This distinction is illustrated in (14), which follows a rigid government pattern, and (15), which

[16] The second accusative is a form of Russian animate nouns whose exponent coincides with the nominative (not the usual accusative), but which appears after prepositions that normally require the accusative.

[17] The algorithm extracted this pattern only for three verbs: *pojti* ‘become X’ in such contexts, *vyjti* ‘make one’s way to X’, and *vybit'sja* ‘rise to be an X’.

exhibits a flexible one—despite both involving the same preposition-case combination.

(14) *Ušakov* ***vgljadyvalsja v temnotu*** *trezvymi glazami.*
'Ushakov peered into the darkness with sober eyes'.

(15) *Timorcy prinosjat v žertvu byka – zakalyvajut i s pesnjami* ***brosajut v more****.*
'The Timorese sacrifice a bull—slaughter it and throw it into the sea with songs'.

Since our argument extraction algorithm relies on the frequency of specific encoding devices rather than their groups, verbs with flexible government patterns were expected to yield more diffuse distributions, reducing the algorithm's efficiency. Table 8 presents the relevant data: a breakdown of 2820 entries identified as arguments in *The Active Dictionary*, excluding those marked by prepositionless accusative (handled separately; see Section 4.3). As noted above, 538 of these were not recognized as arguments by the frequency-based algorithm ("false negatives"). Table 8 tests whether the false negative rate correlates with the rigidity or flexibility of the government pattern in *The Active Dictionary*.

Table 8.
Frequency-based argument extraction algorithm: rigid vs. flexible government patterns

	frequency-based algorithm		
	yes	no	total
rigid	837	221	1,058
flexible	1,445	317	1,762
	2,282	538	2,820

An unexpected result emerges from the data in Table 8: false negatives occur at similar rates for rigid (21%) and flexible (18%) government patterns identified in *The Active Dictionary*. This suggests that, despite their theoretical compatibility with various encoding devices, verbs with flexible government patterns tend to favor specific ones in actual usage, making their distributional behavior resemble that of verbs with rigid patterns. For example, *vernut'sja* 'return' favors *iz* 'from' plus the genitive (as in *vernut'sja iz komandirovki* 'return from a business trip'), although other ablative prepositions can also occur with this verb (e.g., *vernut'sja s zanjatij* 'return from class' or *vernut'sja ot vsenoščnoj* 'return from an all-night vigil'). The frequency-based algorithm captures this asymmetry by classifying dependents introduced by *ot* plus the genitive, but not those with the other two prepositions, as arguments of *vernut'sja* 'return'.

5. Summary and outlook

The argument–adjunct distinction is complex, spanning formal and semantic dimensions that create interrelated but not identical contrasts. This complexity is evident in the evolution of Ju. D. Apresjan's views on argumenthood, from his early quantitative distributional approach to the in-depth semantic analyses of his later work.

This paper aims to reconcile different approaches to argumenthood using Russian data. Our primary goal is not to uncover new facts about Russian grammar and lexicon, but to establish a new methodological avenue for token-based typology, using Russian as a starting point. Specifically, we propose an algorithm for extracting arguments from UD treebanks and evaluate its performance against semantically nuanced lexicographic sources, primarily *The Active Dictionary* (Apresjan (ed.), 2014–). The key idea is that verb arguments typically require specific encoding forms, such as prepositions and cases in Russian, leading to frequency peaks in the distribution of encoding devices across verb lemmas. In contrast, adjuncts are not lexically determined and show flatter distributions. Based on this principle, the algorithm identifies verb–encoding device combinations that occur significantly more often than expected under a flat zero hypothesis.

Evaluated against *The Active Dictionary*, the algorithm performed well, achieving an overall precision of about 0.83. False positives were particularly rare, indicating that dependents identified as arguments by the algorithm generally align with those recognized in semantic-based lexicography. The asymmetry between infrequent false positives and more common false negatives likely reflects an inherent distinction between semantic valency and formal government: while formal government presupposes semantic valency, the reverse is not necessarily true (Helbig & Schenkel, 1983, p. 44).

The low rate of false positives is promising for the practical goals of this study. This paper is part of a broader effort to analyze argumenthood and valency cross-linguistically in typologically diverse languages. For Russian, such analysis—though time-consuming—can, in principle, rely on semantic pattern analysis, given the availability of high-quality lexicographic resources. However, for most languages, such resources are missing and unlikely to emerge soon. In this context, an algorithm that reliably extracts arguments from UD treebanks, even if incomplete, provides a valuable starting point for a quantitative, token-based study of valency patterns. We hope that the UD platform will continue to expand in favor of lesser-documented languages and become increasingly suitable for full-scale typological research.

While the current algorithm achieves a reasonably good accuracy rate as a starting point, further improvement is likely necessary. Two main avenues appear promising in this regard. First, the algorithm should be adapted to better handle rare verbs. Even a seemingly defeatist solution—such as refraining from assigning argumenthood labels to rare verbs altogether—may be more useful in practice than producing false negatives. Second, in its current form, the algorithm processes each verb-dependent token in isolation. In reality, a verb's arguments typically occur as part of a larger frame, with dependents interacting with one another. Such interaction is crucial both theoretically—being a cornerstone of Construction Grammar (Goldberg, 1995), for instance—and practically, as it can significantly improve the efficiency of verb frame extraction algorithms (Sarkar & Zeman, 2000).

Beyond its practical applications, the frequency-based argument extraction algorithm also yields insights into the theoretical understanding of argumenthood. (i) Verbs can exhibit co-occurrence frequency peaks with dependents not traditionally considered full-fledged arguments, yet their presence can alter verb meaning and affect argument structure (e.g., *byt' dlja* + Genitive, 'be for someone'; see Section 4.3). (ii) While canonical arguments correspond to variables in a verb's semantic decomposition, frequency peaks also occur with dependents that represent key semantic components, revealing complex links between syntactic and semantic structure (e.g., *bresti* 'trudge'; see Section 4.3). (iii) A quantitative, token-based perspective suggests that distinctions between rigid government patterns, which require a specific encoding form, and more flexible patterns, where verbs allow multiple competing encoding devices with similar spatial meanings, may not be entirely clear-cut (see Section 4.4).

References

Apresjan, Ju. (1965). Opyt opisanija značenij glagolov po ix sintaksičeskim priznakam (tipam upravlenija). *Voprosy jazykoznanija, 5*, 51–66.

Apresjan, Ju. (1967). *Èksperimental'noe issledovanie semantiki russkogo glagola*. Nauka, Moskva.

Apresjan, Ju., & Páll, E. (1982). *Russkij glagol — vengerskij glagol. Upravlenie i sočetaemost'* (Vols. 1–2). Tankönyvkiadó, Budapest.

Apresjan, Ju. (1974). *Leksičeskaja semantika: Sinonimičeskie sredstva jazyka*. Nauka, Moskva.

Apresjan, Ju. (1995). *Izbrannye trudy* (Vols. 1–2). Jazyki russkoj kul'tury, Moskva.

Apresjan, Ju. (Ed.). (2004). *Novyj ob''jasnitel'nyj slovar' sinonimov russkogo jazyka* (2nd edition). Jazyki russkoj kul'tury, Moskva.

Apresjan, Ju. (Ed.). (2014–). *Aktivnyj slovar' russkogo jazyka* (Vol. 1, 2014; Vol. 2, 2014; Vol. 3, 2017; Vol. 4, Part 1, 2023; Vol. 4, Part 2, 2024). Jazyki slavjanskoj kul'tury Moskva.

Barðdal, J. (2011). Lexical vs. structural case: A false dichotomy. *Morphology, 21*, 619–654. https://doi.org/10.1007/s11525-010-9174-1

Beavers, J. (2010). The structure of lexical meaning: Why semantics really matters. *Language, 86*(4), 821–864.

Bickel, B., Zakharko, T., Bierkandt, L., & Witzlack-Makarevich, L. (2014). Semantic role clustering: An empirical assessment of semantic role types in non-default case assignment. *Studies in language*, *38(3)*, 485–511.

Boguslavskij, I. (1996). *Sfera dejstvija leksičeskix edinic.* Jazyki russkoj kul'tury, Moskva.

Bohnemeyer, J. (2007). Morpholexical transparency and the argument structure of verbs of cutting and breaking. *Cognitive Linguistics, 18(2)*, 153–177.

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics, 47(2)*, 255–308.

Dixon, R. M. W. (2009). *Basic linguistic theory. Volume 1: Methodology.* Oxford University Press, Oxford.

Dryer, M. S. (with Gensler, O. D.). (2013). Order of object, oblique, and verb. In M. S. Dryer & M. Haspelmath (Eds.), *WALS Online* (v2020.4) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13950591 (Available online at http://wals.info/chapter/84, accessed on 2025-02-25.)

Fowler, G. (1996). Oblique passivization in Russian. *The Slavic and East European Journal, 40(3)*, 519–545.

Frajzyngier, Z., Gurian, N., & Karpenko, S. (2024). Minimal participant structure and the emergence of the argument/adjunct distinction. *Studies in Language, 48(1)*, 181–227. https://doi.org/10.1075/sl.22029.fra.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, Chicago.

Haspelmath, M. (2014). Arguments and adjuncts as language-particular syntactic categories and as comparative concepts. *Linguistic Discovery, 12*(2), 3–11.

Helbig, G., & Schenkel, W. (1983). *Wörterbuch zur Valenz und Distribution deutscher Verben.* Tübingen: Max Niemeyer.

Jacobs, J. (1994). *Kontra Valenz.* Trier: Wissenschaftlicher Verlag Trier.

Jolly, J. A. (1993). Preposition assignment in English. In R. D. Van Valin Jr. (Ed.), *Advances in role and reference grammar* (pp. 275–310). Benjamins, Amsterdam.

Kamynina, A. (1999). *Sovremennyj russkij jazyk. Morfologija.* MGU, Moskva.

Koenig, J.-P., Mauner, G., Bienvenue, B., & Conklin, K. (2008). What with? The anatomy of a role. *Journal of Semantics, 25(2)*, 175–220.

Korhonen, A., Gorrell, G., & McCarthy, D. (2000). Statistical filtering and subcategorization frame acquisition. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (pp. 199–205). Hong Kong, China.

Lazard, G. (1994). *L'actance.* Presses Universitaire de France, Paris.

Ljaševskaja, O., & Kaškin, E. (2015). Tipy informacii o leksičeskix konstrukcijax v sisteme FrejmBank. *Trudy Instituta russkogo jazyka im. V. V. Vinogradova, 6*, 464–555.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* University of Chicago Press, Chicago.

Levin, B., & Rappaport Hovav, M. (2005). *Argument realization*. Cambridge University Press, Cambridge.

Malchukov, A., & Comrie, B. (Eds.). (2015). *Valency classes in the world's languages* (Vols. 1-2). De Gruyter Mouton, Berlin, Boston.

Mel'čuk, I., Žolkovskij, A., & Apresjan, Ju. (1984). *Tolkovo-kombinatornyj slovar' sovremennogo russkogo jazyka: Opyty semantiko-sintaktičeskogo opisanija russkoj leksiki.* Wiener Slavistischer Almanach, Wien.

Muravenko, E. (1998). *O slučajax netrivial'nogo sootvetstvija semantičeskix i sintaksičeskix valentnostej glagola. Semiotika i informatika, 36*, 71–81. Jazyki russkoj kul'tury, Moskva.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal dependencies v2: An ever-growing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 4034–4043). European Language Resources Association, Marseille.

Plungjan, V., & Raxilina, E. (1998). Paradoksy valentnostej. *Semiotika i informatika, 36*, 108–119. Jazyki russkoj kul'tury, Moskva.

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna. Retrieved from https://www.R-project.org/ (accessed May 11, 2024).

Rozental', D. (1986). *Upravlenie v russkom jazyke* (2nd edition). Kniga, Moskva.

Sarkar, A., & Zeman, D. (2000). Automatic extraction of subcategorization frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics* (pp. 691–697). Saarbrücken, Germany.

Schulte im Walde, S. (2009). The induction of verb frames and verb classes from corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 952–972). Walter de Gruyter, Berlin.

Tao, S., Donatelli, L., & Hahn, M. (2024). More frequent verbs are associated with more diverse valency frames: Efficient principles at the lexicon-grammar interface. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 11795–11810.

Testelec, Ja. (2001). *Vvedenie v obščij sintaksis.* RGGU, Moskva.

Tesnière, L. (1959). *Éléments de syntaxe structurale.* Klincksieck, Paris.

Van Valin, R. D. Jr. (1993). A synopsis of Role and Reference Grammar. In R. D. Van Valin Jr. (Ed.), *Advances in Role and Reference Grammar* (pp. 1–166). Benjamins, Amsterdam.

Van Valin, R. D. Jr. (2001). *An introduction to syntax.* Cambridge University Press, Cambridge.

Yip, M., Maling, J., & Jackendoff, R. (1987). Case in tiers. *Language, 63*(2), 217–250.

Zeman, D., et al. (2022). *Universal dependencies 2.11.* LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Retrieved from http://hdl.handle.net/11234/1-4923.

Zolotova, G. (2006). *Sintaksičeskij slovar'. Repertuar èlementarnyx edinic russkogo jazyka.* 3rd edition. URSS, Moskva.

Appendix

The entries for *bazirovat'sja* 'base oneself, to be based' in (I) Rozental's dictionary (1986) and in (II) *The Active Dictionary* (Apresjan (Ed.), 2014–)

(I) базироваться *на чём* и *на что*. 1. *на чём* (основываться на чем-л. в своих суждениях или действиях). Базироваться на данных опыта. Нравственное воспитание коммунистического человека, прежде всего, базируется на воспитании его способностей, на развитии его сил, его созидательного, творческого актива (Макаренко). 2. *на что* и *на чём* (опираться на что, иметь что-л. в качестве базы). Самолёты базировались на новый аэродром (на новом аэродроме).

(II) БАЗИ́РОВАТЬСЯ, ГЛАГ.; -руюсь, -руется; НЕСОВ; СОВ нет.

базироваться 1

Базироваться на фактах <на опыте>.

ЗНАЧЕНИЕ. *А1 базируется на А2* ‘В своих действиях или деятельности А3 человек А1 использует информацию А2 или информацию о явлении А2, считая, что она необходима для успешности А3’.

УПРАВЛЕНИЕ 1.

А1 * редк. ИМ: *(В своих прогнозах) метеорологи базируются (на анализе спутниковых данных*).

А2 * *на* ПР: *базироваться на анализе спутниковых данных.*

А3 * *в* ПР: *базироваться в своих выводах <в своих прогнозах>.*

УПРАВЛЕНИЕ 2.

А3 * ИМ: *Прогнозы экономистов базируются (на оптимистической оценке роста цен на углеводороды).*

А2 * *на* ПР: *базироваться на опти*

СОЧЕТАЕМОСТЬ. *Базироваться на других принципах* <на теории Дарвина, на новой системе подсчёта голосов>; *базироваться на чьих-л. оценках* <на устаревших представлениях>.

🕮 *Экономическое обоснование такой стратегии базируется на отчётах аналитиков* («Computerworld», № 25, 2004). *На этом и будут базироваться наши дальнейшие рассуждения* («Информационные технологии», № 8, 2004). *А она [власть] базируется не столько на силе, богатстве, благостности, мудрости etc. властителя, сколько* […] *на готовности подданных к повиновению* (М. Соколов). *Все эти программы эффективнее предшествующих, так как базируются на новых объективных данных о путях оптимизации работы мозга взрослого и ребёнка* («Вопросы психологии», 2004.10.12). *Ведь в англосаксонском праве всё базируется на прецеденте* (К. Амелин).

СИН: *основываться (на чем-л.), опираться (на что-л.);* **АНА:** *исходить из чего-л.;* **КОНВ:** *базировать, основывать;* **ДЕР.:** *база.*

базироваться 2

На этом аэродроме базировалось несколько эскадрилий.

ЗНАЧЕНИЕ. *А1 базируется на А2*: ‘Организация или войска А1 расположены на территории или в месте А2, где они обеспечиваются всем необходимым для их нормального функционирования’.

УПРАВЛЕНИЕ.

А1 * ИМ: *Фирма базируется (в Лондоне).*

А2 * ГДЕ: *базироваться в Токио <на Кольском полуострове, под Москвой>.*

🕮 *Попал я морчасти погранвойск, базирующиеся в дальневосточной Находке* (Л. Вертинская). *Огрызкову сообщили, куда они должны прибыть,*

[…] – *урочище Боговизна, где базировалось руководство партизанской зоны* (В. Быков). *Организация экономического сотрудничества и развития* […] базируется в Париже (В. Ключников). На самом деле там базировалась разведшкола одной из спецслужб (В. Морозов). ДЕР: *базирование*; *база (эскадрилий)*. [Ю.А.]