
Modeling Expression of Eventuality Types in Bulgarian with Large Language Models

Svetla Koeva¹[0000-0001-5947-8736],

¹ Bulgarian Academy of Sciences, Institute for Bulgarian Language, Bulgaria,
svetla@dcl.bas.bg

Scientific paper

DOI: 10.18485/judig.2026.2.ch25

Abstract

The paper examines the relationship between Bulgarian verbal aspect and eventuality types. It adopts a semantic framework based on two privative oppositions, intervality and terminativity, to distinguish eventuality types, and applies these features to Bulgarian perfective and imperfective verbs. The proposed features provide a systematic account of how Bulgarian imperfective and perfective verbs are distributed across eventuality types, while also capturing contextual coercion. A dataset of 254 Bulgarian verbs (the *Aspect:BG* dataset) was annotated with appropriate derivational and semantic features for each verb. Multiple experimental runs investigated the use of Large Language Models (LLMs) to evaluate human classification and to further support semi-automated classification. The analysis shows that different LLMs can be used individually or in combination to generate annotations for an expanded version of the dataset, with different models applied to different semantic or derivational features; however, expert verification is still recommended.

Keywords: eventuality types, Bulgarian verb aspect, Large Language Models

1. Introduction

The study assumes that the semantic category of aspect and the expression of eventuality types constitute a single semantic domain that, cross-linguistically, may or may not be realised at the lexical level, while grammatical and compositional levels may also play a significant role. In Bulgarian, verbal aspect (вид на глагола) provides a derivational means of expressing eventuality types at the lexical level. Bulgarian therefore exhibits both purely lexical means

of expression, in primary imperfective and primary perfective verbs, and derivational means, in secondary imperfective verbs and in perfective verbs formed through prefixation and suffixation.¹ The paper briefly discusses grammatical and compositional means of expressing the semantic category of aspect and focuses on the correspondences between lexically and derivationally expressed verbal aspect in Bulgarian and eventuality types.

The approach that defines the privative semantic oppositions underlying the differentiation of eventuality types is briefly presented in (Koeva, 2025a). These oppositions involve intervality versus the absence of interval structure in eventualities, and open-endedness versus terminativity in eventualities. The distribution of these features across eventuality types characterises basic properties of eventualities. When these features are applied to Bulgarian perfective and imperfective verbs, a clear distribution of verbs from different aspectual classes across eventuality types emerges, together with contextual coercion between types arising at the grammatical and compositional levels. The theoretical assumptions are illustrated using a dataset of 254 Bulgarian verbs classified according to both their derivational and semantic characteristics. Furthermore, Large Language Models (LLMs) are employed, first to evaluate the human-generated classification and, second, to investigate whether they can be used for semi-automatic classification and expansion of the dataset.

The main research questions addressed in the paper are:

- Can a classification of Bulgarian verbs based on semantic and derivational features distinguish verb semantic classes that systematically share properties related to eventuality types?
- Can LLMs be successfully used to evaluate human classifications of Bulgarian verbs and to support the semi-automatic development of aspect-related datasets?

The paper is organised as follows. First, we briefly present the theory of lexical aspect (Section 2.) and our approach to the complex notion of semantic aspect (Section 3.). We then describe the

¹ Following tradition, we call **primary imperfective** and **primary perfective** verbs those that are not the product of derivation, or whose derivation is opaque, **secondary imperfective** verbs those formed by imperfective suffixation, and **perfective** verbs those formed by affixation. There are also **biaspectual** verbs with the form of imperfective verbs, for which verb aspect and eventuality type are interpreted in context.

Bulgarian dataset, *Aspect:BG* (Section 4.). Several experimental runs were conducted. The first evaluates the manually annotated dataset (Section 5.) and, together with the next two, demonstrates the ability of LLMs to classify Bulgarian verbs by eventuality type, as discussed in Section 6. Section 7. presents the conclusions.

The main contributions of our work are as follows. **First**, the paper proposes a unified privative, feature-based representation of eventuality types. **Second**, it presents the *Aspect:BG* dataset of 254 Bulgarian verbs annotated for derivational and semantic features. **Third**, it evaluates the feasibility of using LLMs to support the semi-automatic development of linguistic datasets.

2. Lexical Aspect

The **lexical aspect** (Aktionsart, situation aspect) covers “distinctions between properties of event types denoted by verbal expressions ... classifying verbs into verb classes” (Rothstein, 2004, 1). Aktionsart denotes “kinds of action” such as terminative, resultative, delimitative, iterative, semelfactive, etc. (Filip, 2011, 1187), and is related to the Bulgarian category “manner of the verb action” (начин на глаголното действие) (Ivanova, 1983, 282–286), which comprises more than fifty identified manners, mainly determined by prefixes and suffixes.

Vendler’s four aspectual verb classes, with Bulgarian examples, are (Vendler, 1957):

- **States** — static, durative, atelic, homogeneous: *помня* ‘remember’, *мразя* ‘hate’.
- **Activities** — dynamic, durative, atelic, homogeneous: *плувам* ‘swim’, *пия* ‘drink’.
- **Accomplishments** — dynamic, durative, telic, non-homogeneous: *да напиша писмо* ‘write a letter’, *да построя къща* ‘to build a house’.
- **Achievements** — dynamic, punctual, telic, non-homogeneous: *да счупя чашата* ‘break the cup’, *да скоча* ‘jump’.

Activities and accomplishments freely take the English progressive, reflecting the presence or absence of a terminal point: activities “go on in time in a homogeneous way,” while accomplishments “proceed toward a terminus” (Vendler, 1957, 146).

Carlson distinguishes individual-level (permanent, e.g. *She knows French*) from stage-level (temporary, e.g. *She sleeps*) predi-

cates (Carlson, 1980). Dowty (1979) classifies situations according to whether they are momentary or interval-based, whether they involve change, and, where relevant, the type of change involved; his interval predicates broadly correspond to Carlson’s stage-level predicates.

- **Individual-level states** — static, durative, atelic, permanent, homogeneous: *помня* ‘remember’, *мразя* ‘hate’.
- **Stage-level states** — static, durative, atelic, temporary, homogeneous: *спя* ‘sleep’, *лежа* ‘lie’.

Smith adds a fifth class, semelfactives: punctual, repeatable, but atelic – a subtype of activity (Smith, 1991, 49–53).

- **Semelfactives** — dynamic, punctual, repeatable, atelic, non-homogeneous: *мигам* ‘blink’, *чукам* (*на вратата*) ‘knock’.

Filip (2012, 721) summarises: lexical aspect concerns “properties of eventualities... the presence of some end, limit or boundary in the lexical structure of certain classes of verbs and its lack in others.” Mourelatos suggests that perfective/imperfective morphology parallels these aspectual classes, motivating a single underlying dimension linking grammatical aspect and aspectual class (Mourelatos, 1978, 418).

3. Complex Notion of Aspect

Aspect has been analysed within numerous theoretical frameworks, resulting in considerable terminological variation and differences in the claims made about its nature. Aspect is famously defined as “different ways of viewing the internal temporal constituency of a situation” (Comrie, 1976, 3). Jespersen distinguished four sources of aspectual meaning (lexical, contextual, derivational, tense-based) and several oppositions (aorist/imperfect, conclusive/non-conclusive, durative/punctual, etc.) (Jespersen, 1924, 286–288). Verkuyl treats Aktionsart and aspect as a single phenomenon, “aspectuality,” which is encoded differently across languages (Verkuyl, 1999, 125–126).

The global category of aspect is considered a semantic category that can be expressed at one of the following levels, or by a combination of some or all of them (lexical, morphological, syntactic). It represents eventualities either as a homogeneous, non-interval continuum or as interval-based structures, the latter further classified as terminative or non-terminative (Koeva, 2025a).

The theory adopts a privative (rather than equipollent) feature approach, following (Olsen, 1994, 363), who showed that treating

features such as *telic*, *dynamic* and *durative* as privative allows both restricted compositional possibilities and broader pragmatic effects for unmarked members.

Two privative oppositions are proposed, rather than the usual three (Koeva, 2025a, 213–216):

- **Intervality** – marked member: interval structure (an event represented as having internal temporal structure, regardless of whether that structure corresponds to an extended duration); unmarked member: an event with non-interval structure.
- **Terminativity** – marked member: terminative (reaches a natural, inherent endpoint); unmarked: open-ended (no inherent endpoint).

The two privative features do not provide a complete taxonomy of the six eventuality types; rather, they capture two dimensions along which the types cluster.

In Bulgarian, lexical aspect² is often overtly marked through derivation, unlike in English, although English phrasal verbs frequently parallel Slavic perfectives. The Bulgarian derivational system is rich, comprising affixless verbs, prefixed perfectives derived from imperfectives, suffixed imperfectives derived from perfectives, and cases where the direction of derivation is unclear.

Key derivational findings for Bulgarian imperfective and perfective verbs can be summarised as follows (Koeva, 2021, 2025b):

- **Imperfective verbs** without affixes, or with prefixes or certain suffixes, may express states (*горчи* ‘is bitter’, *зная* ‘know’) or activities (*пия* ‘drink’, *ям* ‘eat’), depending on their lexical meaning.
- **Primary perfective verbs** and prefixed perfectives (*изпека* ‘bake up’) are accomplishments or achievements, depending on durativity; suffixed perfectives with *-н* (*бодна* ‘prick up’) are only achievements.
- **Secondary imperfectives** (*давам* ‘give’, *купувам* ‘buy’) derived from perfectives can be activities, accomplishments, or achievements, depending on lexical meaning and context.
- **Biaspectual verbs**, imperfective in form but usually interpreted as perfective or imperfective in context, are treated as aspectless at the lexical/derivational level, with aspect assigned only com-

² In this paper, lexical refers to the aspectual value (eventuality type) associated with a verb, whereas derivational refers to the morphological process through which that value is established or changed.

positionally (Kabakchiev, 2020, 118–121). Some of these verbs, however, irrespective of their verb aspect, are only lexically imperfective or only lexically perfective (Koeva, 2026, 404).

In general, Bulgarian imperfective verbs denote states and activities, and perfective verbs denote accomplishments and achievements – the traditional view also applied to Russian (Braginsky and Rothstein, 2008, 8) – although a strong competing tendency explains aspect compositionally, mainly for languages lacking derivational aspect (Verkuyl, 1993, 1999, 2022). Table 1 shows the main distribution of Bulgarian imperfective verbs across eventuality types, irrespective of particular features that may arise from their subgroups based on aspectual derivation.

Table 1: Imperfective verbs: distribution across six eventuality types

Type	Non-Interval	Interval	Terminative	Open-ended
individual-level states	+	–		
stage-level states	–	+		
activities	–	+	–	+
accomplishments	–	+	+	–
achievements	–	+	+	–
semelfactives	–	+	–	+

There are not many examples of imperfective verbs that express accomplishments or achievements that are not a result of coercion. These include, for example, some of the so-called bi-aspectual verbs, which are imperfective in form but lexically perfective (Koeva, 2026, 404): *ликвидирам* ‘liquidate’, *парафирам* ‘initial’, *катастрофирам* ‘crash’, *фотографирам* ‘photograph’.

Eventuality types other than individual-level states can be analysed as having temporal interval structure, whether durative or momentary. Achievements and semelfactives are typically interpreted as momentary predicates, corresponding to events occurring within a very short interval. Imperfective Bulgarian verbs that derivationally correspond to perfective verbs expressing semelfactive events usually denote repeatable, very short, momentary events. Although

each individual semelfactive is bounded and momentary, an imperfective verb can present a repetition of such events as an unbounded, potentially open-ended series; therefore, intervality does not correspond directly here to the traditional durative feature.

Under certain contextual or interpretive conditions, predicates belonging to eventuality types with temporal interval structure may receive a property-denoting or habitual interpretation, thereby approaching the interpretation characteristic of individual-level states ('having the property').

activities → individual-level state

Старецът кѳѳа към кѳѳата. 'The old man limps towards the house' [activity]

Старецът кѳѳа от много години. 'The old man has been limping for many years (is lame)' [individual-level state]

On the other hand, activity predicates may receive accomplishment or achievement interpretations.

activities → accomplishments/achievements

Момчето изпиѳа водата бавно. 'drinks up the water slowly' [activity]

Момчето изпиѳа водата за една минута. 'The boy drinks up the water in one minute' [accomplishment]

What is observed is coercion, understood as a reduction of the opposition in which one member occurs in contexts typically associated with the other. In this case, predicates with an open-ended temporal structure may, under appropriate contextual conditions, receive a terminative interpretation.

Table 2 shows the distribution of Bulgarian perfective verbs across the eventuality types.

Table 2: Perfective verbs: limited distribution, no coercion

Type	Interval	Terminative	Open-ended
Accomplishments	+	+	–
Achievements	+	+	–

Within the inventory represented in the *Aspect:BG* dataset, perfective verbs are classified as accomplishments or achievements. A more precise classification can take into account factors such as the lexical semantics of certain perfective verbs, which can involve

delimitative meaning and denote the eventuality type of the source root, i.e. activity (*нея* ‘sing’, *noneя* ‘sing a little’).

Coercion is not permitted with perfective verbs, which can also be explained by the perfective derivational meaning, as it contradicts both structures in which the intervals cannot be recognised and those that do not presuppose a termination.

Grammatical aspect is distinguished here from derivational aspect and refers to the contribution of temporal categories (chiefly tense) to a verb’s lexical aspect. For example, in Bulgarian, the imperfect/aorist distinction is considered aspectual rather than temporal (Maslov, 1959, 165): continuity is inherent in the present and imperfect tenses and in imperfective verbs (Marovska, 2016, 156–157), while the aorist implies discontinuity and typically combines with perfective verbs. In Bulgarian, perfective/imperfective marking alone often suffices to distinguish an activity from an accomplishment; however, this part of the investigation lies outside the scope of the paper.

In addition, **compositional aspect** concerns how argument structure and its lexical/grammatical realisation jointly determine the aspectual meaning of the verb phrase or clause. Several scholars (Verkuyt, 1993, 1999, 2022; Krifka, 1989; Rothstein, 2004) locate aspectual class at the VP level rather than in the verb alone. Telicity in English often depends on whether the object NP is cumulative (atelic: *Mary ate apples for five minutes*) or quantised (telic: *Mary ate three apples in five minutes*) (Krifka, 1989, 75). Compositional aspect has also been demonstrated for Bulgarian (Kabakchiev, 2020). Compositional aspect is also not a concern in this paper.

4. The Dataset *Aspect:BG*

The dataset *Aspect:BG* contains a list of 254 Bulgarian verbs (lexical units associated with exactly one meaning),³ selected to cover each of the following features: imperfective verbs distributed across six eventuality types and perfective verbs distributed across two eventuality types. Each verb was classified according to the following features: (P-PFV: primary perfective; D-PFV: derivationally derived perfective), (P-IPFV: primary imperfective; D-IPFV: secondary, derivationally derived imperfective). This classification may

³ For simplicity, the meanings of the verbs are not provided in the paper.

be further refined according to different means of derivation, which would provide a more detailed classification of verb classes. Each verb was further classified into one of six main eventuality types: I-L state: individual-level state; S-L state: stage-level state; activity; ACC: accomplishment; ACH: achievement; and semelfactive. For imperfective activities, the option of coercing activities into accomplishments or achievements, and semelfactives into achievements, is indicated. Verbs are also classified as non-interval or interval, and as terminative or open-ended.

The verbs were selected from relatively high-frequency verbs using a procedure designed to cover the relevant feature combinations.

Table 3 shows the annotation of ten verbs from the dataset, which is available under the CC BY 4.0 licence.⁴

Table 3: Example of the dataset *Aspect:BG*

Verb	Verb aspect	Eventuality type	Contextual coercion	Intervality	Terminativity
<i>бера</i> 'pick'	P-IPFV	Activity	—	+	Open-ended
<i>бия</i> 'beat'	P-IPFV	Activity	—	+	Open-ended
<i>блестя</i> 'shine'	P-IPFV	I-L state	—	—	—
<i>взема</i> 'take'	P-PFV	ACH	—	+	Terminative
<i>вземам</i> 'is taking'	D-IPFV	Activity	→ ACH	+	Open-ended
<i>виждам</i> 'see'	P-IPFV	S-L state	—	+	—
<i>давам</i> 'is giving'	D-IPFV	Activity	→ ACH	+	Open-ended
<i>дам</i> 'give'	P-PFV	ACH	—	+	Terminative
<i>донасям</i> 'is bringing'	D-IPFV	Activity	→ ACC	+	Open-ended
<i>донеса</i> 'bring up'	D-PFV	ACC	—	+	Terminative

⁴ <https://zenodo.org/records/21982545>

During annotation, the principle is applied that secondary (derived) imperfective verbs denoting activity are subject to coercion into accomplishments or achievements.

Double human annotation is not envisaged because one of the goals is to develop a model for collaborative annotation between humans and LLMs, ensuring human expertise while avoiding the resources required for duplicating human annotation. The expertise arising from the knowledge of different experts might be provided by annotation of different parts of the dataset by different experts and by arbitrary evaluation by various experts of data generated solely by models.

5. Experiments with Large Language Models

An initial experimental run was conducted with the three LLMs using a zero-shot prompt and the list of verbs from the dataset. Its aim was to identify potential inconsistencies in the expert annotation, which, after minor revisions, was then used as the ground truth to assess the models' ability to classify Bulgarian verbs according to selected semantic and derivational properties.

5.1. Brief description of LLMs selected for the experiments

Three major proprietary LLMs were selected for the experimental runs. The aim is not to establish a general ranking of the models, but to assess their suitability for the specific classification task investigated here.

Gemini 3.5 Flash is a multimodal LLM developed by Google DeepMind and released in 2026.⁵ The model supports text, image, audio, video, and PDF inputs, offers a one-million-token context window, and is available through the Gemini API, Google AI Studio, and Vertex AI.

GPT-5.5 is a general-purpose LLM developed by OpenAI and released in 2026.⁶ It is designed for reasoning, instruction following, coding, and agentic tasks, and supports a 1.05-million-token context window. The model is available through the OpenAI API.

Claude Sonnet 5 is an instruction-tuned LLM developed by Anthropic and released on 30 June 2026.⁷ It is designed for reasoning,

⁵ <https://deepmind.google/models/gemini/flash> (accessed on 15.07.2026)

⁶ <https://openai.com/index/introducing-gpt-5-5/> (accessed on 15.07.2026)

⁷ <https://www.anthropic.com/news/claude-sonnet-5> (accessed on 15.07.2026)

coding, and agentic workflows, supports a context window of up to one million tokens, and is available through the Anthropic API.

Experiments conducted with models from different families demonstrate some variability in performance. Consequently, our analysis focuses on broader performance trends and patterns rather than on conclusions drawn from the results of any single model or model release. Experiments with open-weight models (including those specifically designed for Bulgarian) are not reported, as the goal is to establish a framework for human–model interaction to achieve the best possible annotation, which for the moment is not supported by the initial results of open-weight models for the tasks investigated.

5.2. Description of experiments

For each model, the list of verbs from the dataset was provided with the following zero-shot prompt:

There is a list of Bulgarian verbs. Create a table using the exact category labels given below for each verb.

1st column **Bulgarian verb**: one **verb** from the list.

2nd column: **Verb aspect**: **P-IPFV** (primary imperfective, with no clear derivation), **P-PFV** (primary perfective, with no clear derivation), **D-IPFV** (secondary imperfective, with clear derivation), **D-PFV** (perfective, with clear derivation).

3rd column **Eventuality type**: **individual-level state**, **stage-level state**, **activity**, **accomplishment**, **achievement**, **semelfactive**.

4th column **Contextual coercion** – for some activity verbs, does a plausible context support reinterpretation as an accomplishment (**ACC**), an achievement (**ACH**) or neither of the two? Choose exactly one or leave **blank**.

5th column **Intervality**: **+Intervality** (eventuality with temporal interval structure), **–Intervality** (eventuality with no temporal interval structure).

6th column **Terminativity**: **Terminative** (an event reaches a natural, inherent endpoint), **Open-ended** (an event has no inherent endpoint). Choose exactly one or leave **blank**.

Do not provide any additional information.

The number and order of verbs were identical, and no context was provided for each verb. The human annotations and theoretical discussion were not provided to the models, which nevertheless had access to the web and were not required to explain their answers.

Next, all instances in which all three models gave identical answers that differed from the human decision were identified. This yielded 122 feature-level decisions for which all three models agreed with one another but disagreed with the human annotation, as shown in Table 4. After manual analysis of these disagreements, 48 inconsistencies in the original annotations, mostly technical and some classification errors, were identified and corrected.

Table 4: Dataset corrections after comparison with the LLMs’ results

Category	Contradicting identical answers	Number of corrections
Verb	0	0
Verb aspect	7	7
Eventuality type	9	6
Contextual coercion	21	0
Intervality	78	35
Terminativity	7	0
Total	122	48

The improved human annotation in the dataset *Aspect:BG* is the ground truth against which the LLMs’ responses are compared. The responses contained missing (Models 1 and 3), duplicated (Model 1), and substituted verbs (Model 1): as shown in the first row of Table 5.

Table 5: LLMs’ correct answers compared with the ground truth

Category	Model 1	Model 2	Model 3
Verb coverage	66.3%	100.0%	96.6%
Verb aspect	40.2%	55.9%	86.2%
Eventuality	27.2%	85.8%	90.6%
Coercion	38.2%	54.7%	10.6%
Intervality	39.8%	70.9%	78.0%
Terminativity	37.8%	90.9%	92.9%
Average	41.6%	76.3%	75.8%

For matching purposes, capitalisation and arrows are treated as irrelevant; transliterations of verbs provided by Model 1 are also treated as irrelevant. Only exact semantic labels are counted. The average score is the arithmetic mean of the models’ accuracy across the five linguistic features and the verb coverage accuracy.

Model 1 performs worse because all verbs are supplied with transliterations (which are sometimes incorrect), and many verbs are replaced or omitted, degrading performance. Models 2 and 3 achieve similar overall performance, but Model 3 outperforms Model 2 except when predicting coercion. The custom F1-like score (the harmonic mean of verb coverage and per-feature accuracy) for the three LLMs is shown in Table 6.

Table 6: F1 scores for the three LLMs

Feature	Model 1	Model 2	Model 3
Verb aspect	50.0%	71.7%	91.1%
Eventuality	38.6%	92.4%	93.5%
Coercion	48.5%	70.7%	19.1%
Intervality	49.7%	83.0%	86.3%
Terminativity	48.2%	95.2%	94.7%
Average	47.0%	82.6%	77.0%

6. Discussion of Results

Two tests, McNemar’s test and Cochran’s Q test (Agresti, 2013), are designed for paired or related categorical data on the same items: here, the same 254 verbs are classified repeatedly by different models. McNemar’s test is intended specifically for two related conditions on the same items (e.g. Model A vs Model B), comparing only the discordant pairs – items where the two conditions disagree – to assess whether one condition performs significantly better than the other. Cochran’s Q test generalises this to three or more related conditions (e.g. three repeated runs of the same model), testing whether the proportion of “correct” outcomes differs across all conditions simultaneously, without inflating the false-positive rate as multiple separate McNemar tests would. To apply Cochran’s Q test, two additional runs of the experiment with the same prompt and models were carried out.

6.1. McNemar test

The results of the McNemar test applied to the data shown in Table 5 reveal substantial differences between the three classifications. For **Verb aspect**, Models 1 and 2 do not differ ($p = 1.000$), whereas Model 3 significantly outperforms both Model 1 and Model 2 ($p < .001$ for both comparisons). For **Eventuality type**, Model 2 significantly outperforms Model 1 ($p < .001$), and Model 3 also significantly outperforms Model 1 ($p < .001$). The difference between Models 2 and 3, however, does not reach statistical significance ($p = .073$). The results for **Contextual coercion** show a different pattern. There is no significant difference between Models 1 and 2 ($p = .725$). In contrast, both comparisons involving Model 3 are highly significant ($p < .001$). The direction of the discordant pairs indicates that Model 3 performs substantially worse on this feature than either Model 1 or Model 2. Its accuracy is only 10.6%, compared with 38.2% for Model 1 and 54.7% for Model 2. For **Intervality**, all three pairwise comparisons are significant. Model 2 significantly outperforms Model 1 ($p < .001$), while Model 3 significantly outperforms both Model 1 and Model 2 ($p < .001$). A similar pattern is observed for **Terminativity**: both Model 2 and Model 3 significantly outperform Model 1 ($p < .001$), whereas the difference between Models 2 and 3 is not significant ($p = .227$).

Overall, the results indicate that Model 3 performs particularly well in the classification of Verb aspect, Eventuality type, Intervality, and Terminativity, but poorly in the classification of Contextual coercion. Model 2 provides the best overall performance when all five features must be classified correctly. This suggests that accurate classification of individual semantic features and the ability to produce a fully consistent multi-feature classification do not necessarily coincide.

6.2. Cochran’s Q test

The second and third runs of the experiment were conducted to provide data for the second statistical analysis. Cochran’s Q test showed that classification accuracy differed significantly across the three runs for most model–feature combinations, but the pattern was not uniform across models. Model 1 showed pronounced and significant instability on every feature (all $p < .001$), with Total accuracy

(the percentage of verbs, out of 254, for which all five features were classified correctly simultaneously) ranging from 7.5% to 48.8%, depending on the run. Run-to-run accuracy of each model across the three repeated runs (Run 1, Run 2, Run 3), evaluated with Cochran’s Q test and Fleiss’ κ , is shown in Table 7, in which M stands for Model, F for Feature, and T for Total accuracy.

Table 7: Run-to-run accuracy of each model, evaluated with Cochran’s Q test and Fleiss’ κ .

M	F	Run 1	Run 2	Run 3	Q	df	p	κ
M1	F1	40.2%	90.2%	83.5%	169.75	2	<.001***	-0.08
M1	F2	27.2%	89.8%	67.3%	211.60	2	<.001***	-0.02
M1	F3	38.2%	78.7%	61.8%	122.58	2	<.001***	0.29
M1	F4	39.8%	76.8%	67.7%	119.06	2	<.001***	0.33
M1	F5	37.8%	94.1%	72.8%	218.78	2	<.001***	0.13
M1	T	7.5%	48.8%	26.4%	124.65	2	<.001***	0.13
M2	F1	55.9%	83.5%	89.4%	90.15	2	<.001***	0.01
M2	F2	85.8%	67.3%	61.8%	54.70	2	<.001***	0.28
M2	F3	54.7%	61.8%	60.2%	4.32	2	.115	0.33
M2	F4	70.9%	67.7%	70.1%	1.73	2	.420	0.63
M2	F5	90.9%	72.8%	72.8%	56.43	2	<.001***	0.41
M2	T	25.2%	26.4%	27.2%	0.35	2	.839	0.27
M3	F1	86.2%	82.3%	82.7%	3.50	2	.174	0.50
M3	F2	90.6%	75.6%	76.0%	56.28	2	<.001***	0.58
M3	F3	10.6%	63.0%	66.5%	263.43	2	<.001***	0.24
M3	F4	78.0%	79.1%	79.1%	0.44	2	.803	0.68
M3	F5	92.9%	92.9%	93.3%	0.50	2	.779	0.92
M3	T	7.9%	26.0%	28.7%	93.85	2	<.001***	0.58

Model 2 was significantly unstable on **Verb aspect**, **Eventuality type**, and **Terminativity** (all $p < .001$), but showed no significant run-to-run differences on **Contextual coercion** ($Q(2) = 4.32$, $p = .115$), **Intervality** ($Q(2) = 1.73$, $p = .420$), or **Total accuracy** ($Q(2) = 0.35$, $p = .839$). Model 3 showed the most selectively stable profile of the three: **Verb aspect** ($Q(2) = 3.50$, $p = .174$), **Intervality** ($Q(2) = 0.44$, $p = .803$), and **Terminativity** ($Q(2) = 0.50$, $p = .779$, $\kappa = .92$, almost perfect agreement) did not differ significantly across runs, indicating that Model 3’s classification of these three features is reproducible. However, **Eventuality type** ($Q(2) = 56.28$, $p < .001$)

and, especially, Contextual coercion ($Q(2) = 263.43$, $p < .001$) varied significantly, with Coercion accuracy rising from 10.6% in Run 1 to 63.0% and 66.5% in Runs 2 and 3. Because Total accuracy requires all five features to be correct simultaneously, it inherits this instability ($Q(2) = 93.85$, $p < .001$; $7.9\% \rightarrow 26.0\% \rightarrow 28.7\%$), even though three of the five underlying features were themselves stable.

Taken together, these results indicate that accuracy is feature-dependent rather than a single property of a model: Model 3, in particular, combines near-perfect run-to-run reproducibility on Verb aspect, Intervality, and Terminativity with substantial volatility specifically on Contextual coercion, the same feature that was also the hardest for all three models to classify accurately in the first-run comparison (Table 5).

6.3. Limitations

Among the limitations, the following must be mentioned. The dataset is relatively small; however, it serves as both a case study and a methodological instrument. Only three models are used, and LLM outputs can vary by model, version, and the timing of the experiment. As the intention is to model how LLMs can be involved as producers of dataset annotation, these deviations cannot be avoided, but they can be reduced through human post-editing. Another limitation is that no human inter-annotator agreement can be reported; however, the paper demonstrates how LLMs can be used to correct human annotations.

Previous research has shown how multilingual LLMs can be used to annotate text datasets for low-resource languages (such as Indic languages) (Bhat and Varma, 2023); to investigate the alignment between LLM-generated codes and labels and human-generated ground truth in text data analysis (Kristensen-McLachlan et al., 2025); and to test LLMs as span-level annotators (for example, in evaluating data-to-text generation errors, translation mistakes, and propaganda detection) by comparing them directly with skilled human annotators (Kasner et al., 2026), among other applications.

In summary, different models can be used individually or in combination to generate annotations for an extended version of the dataset, with different models applied to different semantic or derivational features. Using all models together provides a useful check against a fully manual annotation of the dataset. When automati-

cally generating annotations for new verbs, a manual check should be performed.

7. Conclusion

This paper has examined the relationship between Bulgarian verbal aspect and eventuality types within a privative feature-based framework centred on intervality and open-endedness. The analysis shows that the proposed features capture two dimensions along which the six eventuality types are systematically distributed, while also accounting for contextual coercion.

The study also demonstrates the potential of LLMs for evaluating and extending an expert-annotated dataset of Bulgarian verbs. The results show that the performance of these models varies considerably across semantic features: some models achieve high accuracy for aspect, eventuality type, intervality, and terminativity, while contextual coercion remains substantially more difficult. Overall, the findings suggest that LLMs can support the semi-automatic development of linguistic datasets, but that expert annotation and theoretical guidance remain essential, particularly for semantically complex and context-dependent classifications.

Acknowledgements

The present study is carried out within the project Ontology of Activity Predicates – Linguistic Modelling with a Focus on Bulgarian funded by the Bulgarian National Science Fund, Grant Agreement No KP-06-H80/9 from 8.12.2023.

References

- Alan Agresti. 2013. *Categorical Data Analysis*, 3rd edition. Wiley, Hoboken, NJ.
- Savita Bhat and Vasudeva Varma. 2023. *Large Language Models As Annotators: A Preliminary Evaluation For Annotating Low-Resource Language Content*. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, page 100–107, Bali, Indonesia. Association for Computational Linguistics.

- Pavel Braginsky and Susan Rothstein. 2008. *Vendlerian Classes and the Russian Aspectual System*. *Journal of Slavic Linguistics*, 16(1):3–55.
- Gregory N. Carlson. 1980. *Reference to Kinds in English*. Number 150 in *Outstanding Dissertations in Linguistics*. Garland Publishing, New York.
- Bernard Comrie. 1976. *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Number 2 in *Cambridge Textbooks in Linguistics*. Cambridge University Press, Cambridge.
- David Dowty. 1979. *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Number 7 in *Studies in Linguistics and Philosophy*. Reidel, Dordrecht.
- Hana Filip. 2011. *Aspectual class and Aktionsart*. In *Semantics*, pages 1186–1217. de Gruyter, Berlin and New York.
- Kalina Ivanova. 1983. Vid na glagola. In *Gramatika na savremenniya balgarski knizhoven ezik*, volume 2, pages 258–287. BAN, Sofia.
- Otto Jespersen. 1924. *The Philosophy of Grammar*. George Allen & Unwin, London.
- Krassimir Kabakchiev. 2020. *Two Major Manifestations of Compositional Aspect in Bulgarian*. *Studia Philologica Universitatis Velikotarnovensius*, 39(1):115–125.
- Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtové, Ivan Kartác, Kristýna Onderková, Ondrej Plátek, Dimitra Gkatzia, Saad Mahamood, Ondrej Dusek, and Simone Balloccu. 2026. *LLMs as Span Annotators: A Comparative Study of LLMs and Humans*. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 1–22, Rabat, Morocco. Association for Computational Linguistics.
- Svetla Koeva. 2021. *Predikati za sastoyanie: svoystva i lingvistichni testove*. *Zeszyty Cyrylo-Metodiańskie*, 10:137–157.
- Svetla Koeva. 2025a. *Bulgarian Aspect in the Context of the Theories of Aspect*. In *Ezik. Teoriya. Interpretatsiya. Yubileen sbornik v chest na prof. d.f.n. Stefana Dimitrova*, pages 204–219. Institute for Bulgarian Language, Bulgarian Academy of Sciences, Sofia.
- Svetla Koeva. 2025b. *Vidat na glagola v balgarski v konteksta na situacionnite tipove*. *Orbis Linguarum*, 23(3):22–30.

- Svetla Koeva. 2026. Otnoshenieto mezhdu leksikalnia aspekt i vida na glagola v balgarski s fokus varhu dvuvidovite glagoli. *Balgarski ezik, Prilozhenie*, 73:396–409.
- Manfred Krifka. 1989. Nominal Reference, Temporal Constitution and Quantification in Event Semantics. In Renate Bartsch, Johan van Benthem, and Peter van Emde Boas, editors, *Semantics and Contextual Expression*, pages 75–116. De Gruyter Mouton, Berlin and Boston.
- Ross Deans Kristensen-McLachlan, Miceal Canavan, Márton Kárdos, Mia Jacobsen, and Lene Aarøe. 2025. Are chatbots reliable text annotators? Sometimes. *PNAS Nexus*, 4(4):pgaf069.
- Vera Marovska. 2016. Leksikalno-gramatichnata kategoriya vid na glagola. *Filologiya*, 54(1A):155–192.
- Jurij S. Maslov. 1959. *Glagol'nyj vid v sovremennom bolgarskom literaturnom jazyke*. AN SSSR, Moskva.
- Alexander P. D. Mourelatos. 1978. Events, Processes, and States. *Linguistics and Philosophy*, 2(3):415–434.
- Mari B. Olsen. 1994. The Semantics and Pragmatics of Lexical Aspect Features. *Studies in the Linguistic Sciences*, 24:361–375.
- Susan Rothstein. 2004. *Structuring Events: A Study in the Semantics of Lexical Aspect*. Blackwell, Malden.
- Carlota S. Smith. 1991. *The Parameter of Aspect*. Number 43 in Studies in Linguistics and Philosophy. Kluwer Academic Publishers, Dordrecht.
- Zeno Vendler. 1957. Verbs and Times. *Philosophical Review*, 66(2):143–160.
- Henk J. Verkuyl. 1993. *A Theory of Aspectuality: The Interaction between Temporal and Atemporal Structure*. Number 64 in Cambridge Studies in Linguistics. Cambridge University Press, Cambridge.
- Henk J. Verkuyl. 1999. Tense, Aspect and Aspectual Composition. In Lars Hellan and Mila Dimitrova-Vulchanova, editors, *Topics in South-Slavic Syntax and Semantics*, pages 123–162. John Benjamins, Amsterdam.
- Henk J. Verkuyl. 2022. *The Compositional Nature of Tense, Mood and Aspect*. Cambridge University Press, Cambridge.

Modelovanje izražavanja tipova eventualnosti u bugarskom jeziku pomoću velikih jezičkih modela

Svetla Koeva

Sažetak

Rad ispituje odnos između bugarskog verbalnog aspekta i tipova eventualnosti. Usvaja semantički okvir zasnovan na dve privativne opozicije, intervalnosti i terminativnosti, kako bi se razlikovali tipovi eventualnosti, i primenjuje ove karakteristike na bugarske perfektivne i imperpektivne glagole. Predložene karakteristike pružaju sistematski prikaz načina na koji se bugarski imperpektivni i perfektivni glagoli raspoređuju po tipovima eventualnosti, a istovremeno obuhvataju kontekstualnu koerciju. Skup podataka od 254 bugarska glagola (skup podataka *Aspect:BG*) anotiran je odgovarajućim derivacionim i semantičkim karakteristikama za svaki glagol. Upotreba velikih jezičkih modela (LLM) za procenu ljudske klasifikacije i za podršku poluautomatskoj klasifikaciji i proširenju skupa podataka istražena je u nekoliko eksperimenata. Analiza pokazuje da se različiti LLM-ovi mogu koristiti pojedinačno ili u kombinaciji za generisanje anotacija za proširenu verziju skupa podataka, pri čemu se različiti modeli primenjuju na različite semantičke ili derivacione karakteristike; međutim, i dalje se preporučuje stručna verifikacija.

Ključne reči: tipovi situacija, bugarski glagolski vid, veliki jezički modeli

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

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

