
Approaches to Automatic Annotation of Semantic Roles and Syntactic Relations

Svetlozara Leseva¹^[0000-0001-8198-4555], Ivelina
Stoyanova¹^[0000-0003-3771-435X]

¹ Bulgarian Academy of Sciences, Institute for Bulgarian Language, Bulgaria,
{zarka|iva}@dcl.bas.bg

Scientific paper

DOI: 10.18485/judig.2026.2.ch26

Abstract

We investigate the capabilities of several openly available large language models (LLMs) for the automatic annotation of frame-semantic structure in Bulgarian, a morphologically rich, pro-drop Slavic language for which the application of LLMs to FrameNet-style annotation remains largely unexplored. The study includes a controlled pilot annotation on 350 sentences containing cleaning verbs (e.g. *чистя, мета, мия, бърша*) that evoke two closely related FrameNet frames, **Removing** and **Emptying**, which share the same core elements – AGENT, THEME, SOURCE, but differ in their syntactic profiling which distinguish frame-semantic alternations. We evaluate three models – Gemma 4 E4B-IT, BgGPT-Gemma-3-12B-IT and Qwen3-8B, on frame element assignment, with the additional evaluation of four annotation subtasks (clause-boundary identification, phrase span identification, coreference resolution, pro-drop resolution). The Bulgarian-adapted BgGPT-Gemma-3-12B-IT achieves the highest scores on all five tasks (F_1 of 58.4), followed by Qwen3-8B with chain-of-thought reasoning (50.9) and Gemma 4 E4B-IT (38.7). Bulgarian fine-tuning matters most where morphosyntax is decisive – pro-drop, clause segmentation, phrase boundaries, while reasoning partially compensates on the more conceptual tasks. The results motivate further work on retrieval-based prompt construction, iterative self-refinement, and supervised fine-tuning.

Keywords: FrameNet, frame elements, semantic annotation, large language models

1. Introduction and motivation

Research on the annotation of semantic frames and frame elements (FEs) has expanded substantially since 2020, driven by the

growing maturity of pre-trained language models and multilingual coverage, as well as the exploration of their applicability to specialised tasks.

The present work is focused on the annotation of semantic frames and their associated frame elements within a corpus of examples. Drawing on the theoretical foundations of FrameNet [Baker et al. \(1998\)](#); [Boas et al. \(2024, 2025\)](#), and related lexical-semantic resources, we are working towards developing an annotation pipeline addressing the full complexity of frame-semantic structure as it is realised in naturally occurring text. The tasks we deal with span across both the syntactic and semantic dimensions of the language, reflecting the well-established interdependence between grammatical form and semantic function in frame-based approaches to meaning.

While our experiments are performed on Bulgarian as proof-of-concept on the challenging task of semantic frame annotation of Bulgarian as a low-resource and morphologically rich language, our observations cover Bulgarian and English, and a comparative study on the approaches performance on English and Bulgarian is among our future tasks. Moreover, the pipeline is applicable to any language and the choice of the LLMs to be used can be motivated by their language coverage in training data, size, architectures, etc. in order to optimise results.

The paper proceeds as follows. Section 2. surveys related work on frame-semantic annotation, the use of LLMs as annotators, and the injection of frame-semantic knowledge into pre-trained models. Section 3. sets out the scope of the annotation task, decomposing the full workflow into interrelated stages that span the syntactic and semantic layers of analysis. Section 4. introduces the case study on Bulgarian cleaning verbs and the two related FrameNet frames – **Removing** and **Emptying**. Section 5. describes the experimental setup, the three LLMs under evaluation – Gemma 4 E4B-IT, BgGPT-Gemma-3-12B-IT and Qwen3-8B, the evaluation metrics adopted for each of the five annotation subtasks, and the results obtained on the 350-sentence pilot dataset. Section 6. draws the principal conclusions and outlines directions for further work.

2. Related Work

We have identified several overlapping directions in the growing work on automatic annotation of frames and frame elements: re-

source creation and annotation design; tools and methodological approaches; language coverage beyond English; the use of large language models (LLMs) for frame and FE annotation; and the introduction of frame-semantic knowledge into neural models during training.

2.1. The Berkeley FrameNet and its global legacy

The Berkeley FrameNet project remains the reference point for frame-semantic annotation. FrameNet (Baker et al., 1998; Baker, 2008) in the classical sense is a lexical resource that represents lexical and conceptual knowledge in the apparatus of Frame Semantics (Johnson et al., 2001; Fillmore, 2003; Ruppenhofer et al., 2016). A semantic frame is taken to be a “script-like structure of inferences, linked by linguistic convention to the meanings of linguistic units – in our case, lexical items. Each frame identifies a set of frame elements (FEs) – participants and props in the frame. A frame semantic description of a lexical item identifies the frames which underlie a given meaning and specifies the ways in which FEs, and constellations of FEs, are realized in structures headed by the word” (Johnson et al., 2001, 9). Each frame is paired with a definition that describes schematically the situation and the way in which at least the most essential FEs are involved in it. Each FE is also supplied with a definition that captures its semantics and interaction with other FEs. Frame elements have different status reflecting their role in the description of a given semantic frame: core, peripheral or extra-thematic, cf. (Ruppenhofer et al., 2016, 19–20). In what follows, we will only make reference to core FE, that is the ones that are conceptually necessary components of a frame (Ruppenhofer et al., 2016, 23). A FrameNet frame is instantiated (or evoked) by a number of linguistic items, called lexical units (LUs). A LU pairs a word and a meaning whose conceptual semantics is represented by the frame.

An indispensable part of the Berkeley project has been the FrameNet annotated corpus; the corpus consists of sentences (or clauses) annotated with the frame evoked by the predicate in the target clause and the frame elements that are syntactically expressed or non-overt but recoverable from the context or the semantics of the frame. The corpus provides valuable data as to how the frames and FEs are actually realised in configurations of FEs, along with their syntactic category and syntactic function (valence patterns).

Boas et al. (2024) provide a detailed account of the FrameNet lexicographic workflow, the definition and classification of frame elements into core, peripheral, and extra-thematic roles, and the two annotation modes—lexicographic (LU-focused) and full-text (document-level)—that together constitute the FrameNet methodology. A follow-up survey (Boas et al., 2025) charts the project’s 25-year development, cataloguing more than 1,200 frames, 10,000 FEs, and 200,000 manually annotated instances, while discussing the treatment of null instantiations (DNI, INI, CNI) as an integral part of full-text annotation.

The FrameNet model has been adopted for describing a number of languages, including Germanic – German (Burchardt et al., 2006), Dutch (Vossen et al., 2018), Danish (Pedersen et al., 2018), Swedish (Borin et al., 2010), Romance – French (Candito et al., 2014), Spanish (Subirats, 2009), Brazilian Portuguese (Torrent et al., 2018), Baltic – Latvian (Gruzitis et al., 2018), and Slavic – Bulgarian (Koeva, 2008), as well as languages typologically different from the Indo-European family, such as Chinese (You and Liu, 2005), Japanese (Ohara et al., 2004), Korean (Hahm et al., 2014). A dedicated description of *framenets*, focused on ones involved in the Multilingual FrameNet annotation initiative,¹ is offered by Gilardi and Baker (2018). FrameNet’s theoretical framework has been adopted for Bulgarian and extended into a model accounting for language-specific features, including verb aspect, semantic and syntactic diatheses and syntactic alternations. The Bulgarian FrameNet has evolved outside this initiative (Koeva, 2008, 2010, 2021) and subsequent work. One of the main venues of research for Bulgarian has been the expansion of lexical units through the alignment between WordNet synsets and FrameNet frames (Leseva and Stoyanova, 2020a,b) and the rigid description of the semantic and syntactic peculiarities of Bulgarian verbs (Koeva, 2024). While results on the annotation of linguistic data using these frames have been reported, e.g. (Leseva and Stoyanova, 2022), to the best of our knowledge, the capabilities of LLMs have not been explored with respect to the annotation of frames and frame elements in Bulgarian.

¹ <https://www.globalframenet.org/>

2.2. Frame Semantics Annotation

Beyond the Berkeley FrameNet and the FrameNet-inspired annotated corpora developed and/or expanded under the Multilingual FrameNet annotation initiative, cf. (Gilardi and Baker, 2018), several projects have extended the FrameNet model to specialised corpora and new annotation granularities. Belcavello et al. (2024) present Frame2, a Brazilian Portuguese multimodal corpus of 230 minutes of travel television annotated with frames and FEs at both the sentence and video-frame level, combining text annotation sets with bounding-box FE annotations over images—one of the first resources to operate simultaneously below and above the sentence boundary. Timponi Torrent et al. (2022) propose extending the FrameNet Brasil annotation scheme with qualia-structure relations between FEs and with a multimodal layer, arguing that contextual and encyclopaedic information must be encoded alongside the standard FE tiers to support downstream semantic tasks.

Tonelli and Menini (2021) push FE annotation into an entirely new perceptual domain, defining a FrameNet-inspired schema for olfactory events whose roles (Smell_source, Location, Carrier, Quality) differ systematically from general-purpose FEs, and applying it to historical travel writings. This schema is scaled to six European languages in a diachronic benchmark (Menini et al., 2022).

At the biomedical end of the spectrum, Ferreira and Pinheiro (2021) introduce SpiNet, a domain-specific FrameNet for spinal medicine whose four frames and associated FEs are linked to the MeSH ontology, enabling automatic semantic annotation of clinical research papers.

2.3. Annotation Tools and Approaches

A persistent challenge in frame-semantic annotation is the high cost of manual effort and the difficulty of achieving reliable inter-annotator agreement on the boundaries and labels of FEs. Remijne et al. (2022) address both concerns simultaneously by combining referential annotation (grounding mentions to real-world entities via a structured knowledge base) with FrameNet FE annotation. Their Dutch annotation tool produces discourse-level annotations that exceed sentence-level FE consistency, reporting high inter-annotator agreement scores—an outcome they attribute to the refer-

ential grounding acting as a pre-structuring aid for human annotators.

The problem of insufficient annotated data has prompted several data-augmentation strategies. [Pancholy et al. \(2021\)](#) exploit the frame hierarchy by transferring existing FE annotations from annotated lexical units to unannotated “sister” lexical units that evoke the same frame, thereby expanding coverage without additional manual effort. [Anwar et al. \(2023\)](#) extend this idea through lexical substitution: they replace frame-evoking words with semantically compatible alternatives, generating new annotated sentences while preserving FE structure, and demonstrate downstream gains in frame induction and parsing. [Cui and Swayamdipta \(2024\)](#) go further by using LLMs to generate entirely new sentences conditioned on a target frame’s structural description, showing that generated data is effective for data augmentation in low-resource FE labelling settings, though not in high-resource ones.

On the tooling side, [Marşan and Yıldız \(2021\)](#) present FrameForm, an open-source web interface for predicate-level frame and FE annotation used in building the Turkish FrameNet, designed to be easily adapted to any language. [Belcavello et al. \(2022\)](#) describe Charon, a semi-automatic annotation pipeline for multimodal corpora that annotates static images and video sequences with FrameNet frames and FEs, interfacing directly with existing FrameNet database software.

The LLM-assisted annotation pipeline of [Belcavello et al. \(2025\)](#) provides the most comprehensive evaluation of automation’s impact on FrameNet annotation to date, comparing annotation time, FE coverage, and lexical diversity between human annotators and an LLM-based semantic role labeller under a perspectivized full-text methodology.

Extending frame-semantic annotation to languages other than English requires decisions about whether to reuse Berkeley FrameNet frames, adapt them, or build new inventories from scratch. [Remijnse et al. \(2022\)](#) develop a Dutch annotation resource and tool that operates over language-independent corpora, demonstrating that discourse-level FE annotation with referential grounding is feasible for a non-English language. [Bick \(2022\)](#) take a coverage-first approach for German, constructing PFN-DE, a parsing-oriented FrameNet with almost 15,000 frames covering 11,300 verb lemmas, specifically designed for robust annotation of German social media

texts; they achieve an F-score of 93.6% for frame sense tagging on Twitter data using a syntax-driven annotator.

The Global FrameNet initiative addresses cross-lingual annotation directly. [Czulo et al. \(2020\)](#) report on the Global FrameNet Shared Annotation Task, in which multiple national FrameNet teams jointly annotate a TED Talk in their respective languages, revealing typological differences in frame-bearing word density and the need for pragmatic frames to handle phenomena such as greetings and tag questions that are not lexically evoked in standard FrameNet. [Gamonal et al. \(2024\)](#) evaluate the LOME multilingual semantic parser on Brazilian Portuguese texts annotated with FrameNet Brasil frames and FEs, finding that over 80% of frame-evoking words can be assigned Berkeley FrameNet frames, but that register differences substantially affect parser coverage. [Marşan and Yıldız \(2021\)](#) further contribute the Turkish FrameNet as a practical output of a language-agnostic annotation interface.

Multilingual coverage also motivates the olfactory benchmark of [Menini et al. \(2022\)](#), which annotates FEs in English, French, German, Italian, Dutch, and Slovenian historical texts, demonstrating that a FrameNet-inspired schema can be applied consistently across typologically diverse languages when the domain is sufficiently constrained.

The use of LLMs as annotators—rather than merely as components within a supervised pipeline—represents a recent and rapidly developing strand of research. [Cui and Swayamdipta \(2024\)](#) are among the first to deploy LLMs specifically for generating FrameNet-annotated data: they condition sentence generation on a frame’s structural definition and show that the resulting sentences preserve FE assignments more faithfully than prior augmentation approaches, as judged by both automatic metrics and human evaluation. [Belcavello et al. \(2025\)](#) conduct an extensive human-vs-LLM comparison in the context of full-text FrameNet annotation of Brazilian Portuguese, finding that while LLMs offer substantial speed advantages, they differ from human annotators in FE coverage and in the handling of perspectivized or culturally specific framings.

[Devasier et al. \(2025\)](#) provide the most systematic benchmark of LLM capability on frame element extraction to date, evaluating GPT-4o, GPT-4o-mini, DeepSeek V3, Llama, Qwen, and Phi-4 under both in-context learning and fine-tuning conditions on FrameNet 1.5

and 1.7. They find that models excelling at multi-step reasoning (as measured by the MUSR benchmark) also perform best at FE extraction, and that fine-tuning substantially closes the gap between LLMs and dedicated parsers.

Gamonal et al. (2024) evaluate the LOME parser—built on XLM-RoBERTa—for automatic frame and FE assignment in Brazilian Portuguese, characterising the types of post-annotation correction required and the extent to which automated pre-annotation reduces overall annotation effort.

The question of what LLMs already “know” about frame semantics is probed by Jayanth et al. (2025), who evaluate Llama-3.1-8B under zero- and few-shot prompting for frame identification on FrameNet 1.5 and 1.7, and additionally test whether models can generate semantically coherent frame definitions from frame names alone. Their results suggest that instruction-tuned LLMs have internalised substantial knowledge of FrameNet’s conceptual structure without task-specific training.

A complementary strand asks not what LLMs already know about frames, but how frame-semantic knowledge can be systematically injected during or after pre-training to improve downstream tasks.

Zheng et al. (2023) propose AGED (ArGument Extractor with Definitions in FrameNet), a query-based framework that feeds FE definitions from FrameNet as natural-language queries into a cross-encoder, guiding the model to learn label semantics and argument interactions jointly. Compared to naïve multi-class classification of roles, AGED improves F1 by up to 1.3 points on FrameNet 1.5 and 1.7, and shows markedly better generalisation in zero-shot and few-shot settings.

Ai and Tu (2024) extend the modelling of argument interactions by proposing arbitrary-order conditional random field inference over a neural encoder, surpassing models that only capture pairwise dependencies. Although not an LLM fine-tuning method per se, this work directly exploits FrameNet’s FE co-occurrence structure to inform a structured prediction layer.

Liu et al. (2025) take a parameter-efficient fine-tuning approach, adapting LLMs with LoRA using instructional templates derived from FrameNet FE definitions and annotated examples. The resulting Frame-Aware LLM achieves zero-shot semantic role labelling

performance competitive with supervised baselines on CoNLL-2005 and FrameNet benchmarks, demonstrating that frame-semantic structure can be encoded as lightweight adapter weights rather than requiring full fine-tuning.

Devasier et al. (2025) show that fine-tuned small LLMs (e.g., Qwen 2.5 14B) rival or exceed GPT-4o on FrameNet FE extraction when trained on structured JSON representations of frame-annotated sentences, providing a practical recipe for injecting FrameNet supervision into open-weight models.

Finally, Remijnse and Minnema (2020) lay an early conceptual foundation for this line of work by arguing that grounding FE annotations referentially—linking them to real-world entities—provides richer training signal for models than purely span-level FE labels, a hypothesis increasingly borne out by the entity-grounded pre-training strategies adopted in more recent work.

3. Scope of the Task

Frame-semantic annotation requires a principled treatment of a broad range of linguistic phenomena. Unlike shallow semantic role labelling, full frame annotation demands that each predicate be situated within a precise conceptual frame, and that each participant and circumstantial element be identified and categorised with respect to that frame. This entails not only the recognition of surface syntactic constituents, but also the resolution of discourse-level dependencies that bear on the interpretation of frame elements.

The present effort is motivated by the need for high-quality, theoretically grounded annotated data to support both computational modelling and linguistic investigation of semantic structure, with a particular focus on Bulgarian as a morphologically rich Slavic language.

The annotation workflow encompasses a set of interrelated tasks, each targeting a distinct layer of linguistic structure. These tasks are designed to proceed in a principled sequence, with later stages building upon the outputs of earlier ones.

The full set of tasks is as follows:

1. **Identification of the verb and verb form.** Each predicate is identified in full, including all auxiliary and modal elements that form part of the verbal complex. This encompasses pe-

riphrastic tense, aspect, and voice constructions, ensuring that the complete predicate is delimited prior to further analysis.

2. **Word sense disambiguation (WSD) of the verb.** The target verb is disambiguated with respect to its sense inventory. Accurate sense assignment is a prerequisite for all subsequent frame-semantic decisions, as the frame evoked by a verb is sense-dependent. Automatic WSD is validated manually.
3. **Frame assignment.** On the basis of the disambiguated verb sense, the appropriate semantic frame is assigned from the frame inventory. This step uses the mapping between WordNet (for the sense inventory) and FrameNet (for frame inventory), to select the frame that most precisely captures the situation described by the verb.
4. **Clause structure and clause boundary identification.** As a supplementary task, the clause structure of each sentence is established, mainly focusing on clause boundaries but also including the identification of main and subordinate clauses in order to identify elements that are expressed in other clauses within the same sentence. This provides the structural scaffolding within which frame elements are subsequently located.
5. **Identification of noun phrases, prepositional phrases, and adverbial phrases.** All noun phrases, prepositional phrases, and adverbial phrases that potentially express frame elements are identified and their true constituent boundaries established. Particular care is taken to avoid both under- and over-segmentation, especially in cases involving complex or nested phrasal structure.
6. **Coreference and anaphora resolution.** Coreference chains and anaphoric relations are resolved across the discourse, enabling the correct identification of the referents of frame element fillers. This step is essential for cases in which a frame element is realised by a pronominal or other anaphoric expression, and for the subsequent investigation of the semantic properties of frame element fillers.
7. **Pro-drop resolution.** Bulgarian, as a pro-drop Slavic language, permits the systematic omission of pronominal subjects

and, in certain constructions, objects. The annotation explicitly identifies and recovers dropped arguments, specifying their referents where recoverable from context. This is particularly consequential for frame-semantic annotation, as dropped arguments frequently correspond to core frame elements.

8. **Frame element assignment.** Finally, the identified phrases are mapped to the frame element roles defined by the assigned frame. This step determines each syntactic constituent that expresses a participant or circumstantial role within the evoked frame, completing the annotation of the predicate–argument structure at the semantic level.

The tasks outlined above are not independent. Rather, they form an interrelated pipeline in which each layer of annotation informs and constrains the next. Clause boundary determination, for instance, restricts the search space for frame element fillers, whilst coreference and pro-drop resolution are indispensable for ensuring that the full set of frame participants is identified, including those not overtly expressed in the immediate syntactic environment of the predicate.

The joint treatment of syntactic and semantic relations is thus a defining feature of the present annotation scheme, and reflects a commitment to theoretical adequacy alongside practical utility for natural language processing tasks.

Our work described below involves an experimental setup, analysis and evaluation of the results of applying LLMs for each of the described annotation subtasks (except the first three, which are predetermined in the present setting). The fine-tuning is based on the manual correction of errors and the elaboration of the prompts. With respect to frame element assignment, we have selected verbs exhibiting variable syntactic behaviour aligned with two different related frames **Emptying** and **Removing**. The two frames share the same inventory of core frame elements (Agent / Cause, Theme and Source), but these elements are profiled in a different way, resulting in a different syntactic expression for the Theme and the Source as either a direct or PP-object depending on the frame. This allows us to test the correct assignment of frame elements to the appropriate syntactic group in very similar contexts.

4. Case Study: Verbs of Cleaning

The so-called verbs of cleaning, e.g. verbs such as *чистя* (*chistya*) ‘clear, clean’, *бърша* (*barsha*) ‘wipe’, *мета* (*meta*) ‘sweep’, etc. contain the notion of an inherent scale in their meaning. The progress of the event of cleaning may be interpreted, and measured out, in terms of the change in the degree of a scalar property lexicalised by the verb. For deadjectival predicates (such as *chistya*) the property is encoded by the adjectival root. Thus in the case of *чистя* (*chistya*), the property is cleanliness and the upper bound of the scale is the state of being clean. Verbs with these properties are subsumed in the class of the so-called ‘degree achievements’ to be identified and studied in the dedicated literature (Hay et al., 1999; Rappaport Hovav, 2008; Kennedy, 2012), among many others.

An important feature of cleaning verbs is that the scale measuring out the progress of the event may be construed with respect to either the *THEME* – here understood in a narrow sense as the thing(s) or substance that is removed from a location or entity so that it becomes clean, or the location *SOURCE* – the container, space, etc. undergoing the process of cleaning. This results in an object alternation, the so-called clear-alternation discussed by Levin (1993, 51–52) and much subsequent literature. The possibility to associate the scale with properties of different participants gives rise to two related senses: (i) a predicate describing the removal of a *THEME* out of a *SOURCE*, e.g., *cleaning the snow from the deck*, in which case the *THEME* is the direct object complement (ex. 1), and (ii) a predicate denoting the emptying of a *SOURCE* from the *THEME*, e.g., *cleaning the deck from the snow*, where the *SOURCE* is expressed as the direct object NP (ex. 2).

In the first interpretation, the *THEME* is syntactically expressed, while the *SOURCE* PP may remain implicit (in brackets).

- (1) Те_{AGENT} чистят снега_{THEME} (от палубата_{SOURCE}).
They clean-PRS snow-DEF (from deck-DEF).
‘They are cleaning the snow (from the deck).’

In the second case (ex. 2), the *SOURCE* is understood to be “completely”, or holistically affected by the event, which needs not be the case when it is expressed as the object of a preposition Levin (1993, 50). (Note that in sentences involving an ongoing, not yet

completed situation, this is to be construed as ‘intended’, rather than as available holistic result). The PP may be left syntactically non-overt.

- (2) Те_{AGENT} чистят палубата_{SOURCE} (от снега)_{THEME}.
 They clean-PRS deck-DEF (of snow-DEF).
 ‘They are cleaning the deck (of the snow)’

In FrameNet, the semantic distinction motivating this object alternation is captured in terms of two related frames, **Removing** (1) and **Emptying** (2), which describe the Theme-focused and the holistically-affected-Source-focused scenarios, respectively.

For the sake of this study, we use contexts (sentences) representing either of these two frames. We have filtered the examples that may represent a different frame, e.g. the **Grooming** sense of чистят. At the current stage of frame element annotation, our goal is to identify correctly the AGENT, CAUSE, THEME and SOURCE among the constituents. The fine-tuning stage is aimed particularly at: (i) perfecting the identification of pro-drop pronouns, which are quite common among AGENTS (or CAUSE); (ii) perfecting the distinction of SOURCES and THEMES, which, especially in the cases where the third core frame element is non-overt, may be difficult.

5. Experiments

The experiments reported in this section assess the capabilities of several openly available LLMs on the annotation tasks introduced in Section 3., with a particular focus on the five tasks related to frame-semantic analysis: (i) clause structure and clause boundary identification; (ii) identification of noun, prepositional and adverbial phrases as potential frame element fillers; (iii) coreference and anaphora resolution; (iv) pro-drop resolution; and (v) frame element assignment.

The present study is set as a *controlled pilot* in order to explore the capabilities of LLMs towards building a complete annotation solution. The empirical material is restricted to Bulgarian sentences in which the target predicate is one of the cleaning verbs discussed in Section 4. and evokes one of two closely related frames – **Removing** and **Emptying**. This narrow scope serves three purposes: it explores the application of frame-specific content, errors can be examined in detail, and it is feasible to explore the complete workflow.

Consequently, the results reported below should be read as a first, focused characterisation of LLM behaviour on a well-defined slice of the annotation problem before launching experiments on a wider Bulgarian FrameNet inventory observing how patterns generalise over other frames and predicate classes.

5.1. Experiments setup

The experiments were performed in Google Colab, a cloud-based Jupyter notebook environment providing access to NVIDIA A100 GPU hardware via Google’s infrastructure.

The notebook was organised into a pipeline of sequential steps covering dependency installation and model loading, data retrieval, prompt execution, and output, allowing individual stages to be re-run independently for different models and when sessions were interrupted. The code will be released for replication purposes.

The models were imported directly from HuggingFace, using the `transformers` library installed from its development branch on GitHub. Model weights were downloaded automatically from the HuggingFace Hub on first use and cached in the Colab session’s local system for the duration of the session.

The structure of the initial experimental prompts is shown in the Appendix. It contains: a) the name and definition of the FrameNet frame; b) the list and definitions of the frame elements; c) description of the required tasks.

5.2. Models

Gemma 4 E4B-IT² (Google DeepMind, 2025) is released by Google DeepMind in April 2025. Gemma 4 E4B-IT is an instruction-tuned model from the Gemma 4 family, with ‘effective’ 4 billion parameters, achieved through Per-Layer Embeddings (PLE) that maximise representational capacity for on-device and edge deployments. The model employs a hybrid local/global attention mechanism and supports a context window of 128K tokens. Pre-training covered a diverse corpus of over 140 languages, with Bulgarian being represented in the base model as part of broad web-document coverage, though without dedicated language-specific fine-tuning. Benchmark evaluations place it at the Pareto frontier for its parameter class, with

² <https://huggingface.co/google/gemma-4-E4B-it>

strong results in reasoning, instruction following, and structured output generation.

BgGPT-Gemma-3-12B-IT³ (INSAIT Institute, 2025) is a Bulgarian-adapted LLM produced by the INSAIT Institute in 2025, forming part of the BgGPT 3.0 series. Built on Google’s Gemma 3-12B backbone and comprising 12 billion parameters, it has been further pre-trained and instruction-tuned on Bulgarian and bilingual Bulgarian–English data. The model supports a context window of 131K tokens. BgGPT 3.0 demonstrates substantial gains over its predecessor on Bulgarian generative benchmarks including TriviaQA, GSM8k, IFEval, and BigBenchHard.

Qwen3-8B⁴ (Qwen Team, 2025) is an 8-billion-parameter dense language model released by the Qwen Team (Alibaba Cloud) in April 2025 (Qwen Team, 2025). A distinctive architectural feature is the *hybrid thinking mode*: the model can seamlessly switch between a chain-of-thought reasoning mode, suited to complex logical, mathematical, and coding tasks, and a fast non-thinking mode for efficient general-purpose dialogue. Pre-training used approximately 36 trillion tokens covering 119 languages and dialects and a corpus that explicitly includes Slavic languages, providing a reasonable baseline for Bulgarian even without dedicated fine-tuning. The pre-training mix also incorporated high-quality multilingual web documents, books, and synthetic reasoning data. Qwen3-8B outperforms Qwen2.5-14B on the majority of benchmarks, including several multilingual evaluation tasks, and supports a native context window of 32K tokens, extendable to 131K via YaRN scaling.

5.3. Evaluation metrics

Each of the five annotation tasks is reported with a single primary simplified metric. The main purpose of the pilot evaluation is: 1) to establish whether LLMs perform reliable annotation of components required for semantic annotation (see tasks above in Section 3.); 2) to establish whether LLMs perform reliable semantic annotation in general; and 3) to investigate whether the performance of various LLMs differ, in particular to see the performance of LLMs focused on Bulgarian.

³ <https://huggingface.co/INSAIT-Institute/BgGPT-Gemma-3-12B-IT>

⁴ <https://huggingface.co/Qwen/Qwen3-8B>

Clause-boundary identification. The clause-boundary task is scored with token-overlap (partial) F_1 , which credits a predicted clause whenever its span intersects a gold standard clause. Partial F_1 is the harmonic mean of precision (predicted spans that overlap a gold span) and recall (gold spans recovered by any predicted span).

Phrase span identification. Noun, prepositional and adverbial phrases are evaluated with exact-match span F_1 , computed jointly over the three categories. F_1 is the harmonic mean of span precision (predicted spans exactly matching a gold span) and span recall (gold spans recovered exactly).

Coreference resolution. Coreference is evaluated with the standard F_1 score restricted to mentions within the discourse window of the sentence.

Pro-drop resolution. Pro-drop resolution is evaluated on the subject slot of the target verb only. For each target verb, the model must decide whether the subject is syntactically explicit or left out, in which case it inserts an empty element ($[\]$) immediately before the verb. We report on F_1 where precision is the proportion of correctly predicted pro-drop cases out of all predicted pro-drop cases, and recall is the proportion of correctly predicted pro-drop subjects that the model identifies out of all true pro-drop cases in the dataset.

Frame element assignment. Frame element assignment is evaluated with weighted F_1 over all core roles for the assigned frame: AGENT, THEME, SOURCE and CAUSE for both **Removing** and **Emptying**.

5.4. Results

Table 1 reports the primary evaluation scores for each of the five annotation tasks across the three LLMs, computed on the 350 sentences pilot dataset. BgGPT-Gemma-3-12B-IT attains the highest score on all five metrics, followed by Qwen3-8B in the thinking-mode configuration, with Gemma 4 E4B-IT consistently last. The overall macro F_1 across the five tasks is 58.4 for BgGPT, 50.9 for Qwen3-8B and 38.7 for Gemma 4 E4B-IT, an absolute gap of 19.7 points between the strongest and the weakest model.

BgGPT’s lead over Qwen3-8B is largest precisely on the tasks that depend on Bulgarian morphosyntactic competence: pro-drop identification, phrase span identification and clause-boundary iden-

tification. On the more abstract, reasoning-heavy tasks of frame element assignment, the gap narrows. This is consistent with the interpretation that chain-of-thought reasoning in Qwen3 partially compensates for the absence of dedicated language adaptation when the task is more conceptual than morphological.

Even the strongest model attains scores in the range 51–65 for any of the annotation tasks, a level below the threshold of practical usability. Further in-depth observations reveal the interdependence between the performance on the semantic annotation tasks and the syntactic complexity of the sentence in terms of clause structure, phrase boundaries, etc.

Table 1: Pilot results on 350 Bulgarian sentences with cleaning verbs. Each cell reports the primary evaluation score (%) for the task, computed on all 350 sentences. Best per row in **bold**.

<i>Task (primary metric)</i>	<i>Gemma-4 E4B-IT</i>	<i>BgGPT-3- 12B</i>	<i>Qwen3-8B (thinking)</i>
Clause boundaries (partial F_1)	46.2	64.8	57.2
Phrase spans, NP/PP/AdvP (span F_1)	41.5	58.1	52.6
Coreference (F_1)	27.7	51.2	49.4
Pro-drop resolution (F_1)	36.3	59.5	42.1
Frame element assignment (weighted F_1)	41.9	58.2	53.4

Figure 1 shows relatively consistent results across all four core frame elements: AGENT is the easiest for every model, followed by THEME and SOURCE, which show similar results, and the CAUSE is systematically the hardest. BgGPT-Gemma-3-12B-IT leads on every role, but its margin over Qwen3-8B is largest on AGENT (+5.5 points), where Bulgarian morphology and pro-drop recovery are particularly relevant.

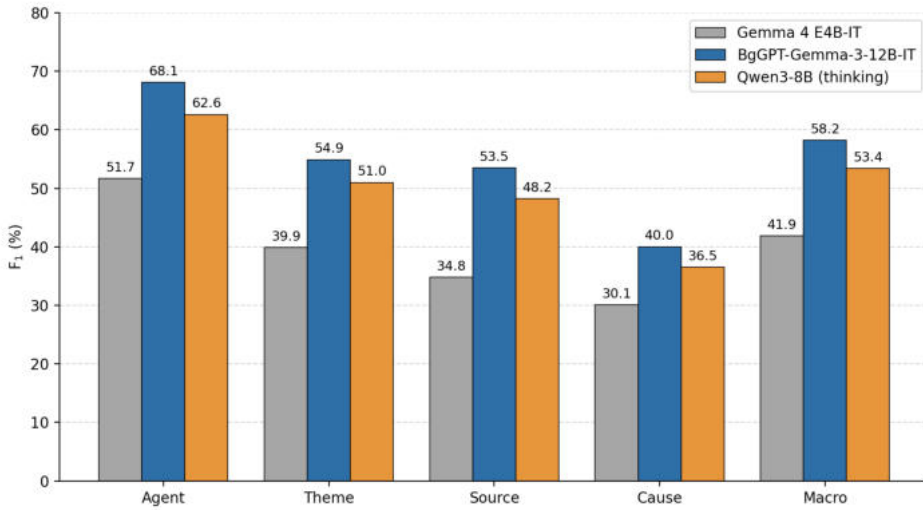


Figure 1: Results of the semantic frame element assignment task. Comparison between the models.

5.5. Discussion

Even though the models are best at identifying the AGENT argument as compared with the other frame elements under discussion, the results are far from impressive. The models are prone to systematic errors, in particular – failure to identify the non-overt (pro-drop) subjects of active clauses and to a much lesser extent – failure to recognise and annotate overt AGENTS. A typical instance of the former is illustrated by Example 3:

- (3) Двамата_{AGENT} почистват масата, []_{AGENT}
 Two-DEF clean up-PRS table-DEF, pro-drop
 метат [пода на верандата]_{SOURCE}.
 sweep-PRS floor-DEF of porch-DEF.
 ‘The two of them clean the table and sweep the porch floor.’

The sentence comprises several coordinated clauses with coreferential AGENT subjects. Except for the first one where an explicit AGENT occurs, the other clauses have non-overt subjects which, along with the coreference chain formed by them, remain unidentified. In such instances, the annotation is evaluated as incorrect.

A related issue arises from the annotation of pro-drop subjects of passive clauses as AGENTS, although the passivised argument projected in the subject position must be a SOURCE or a GOAL depending on the meaning of the sentence and respectively, the frame evoked (see Section 4.). An illustration of this is provided in Example 4, where the non-overt argument licensed for the subject position of the passive clause is the SOURCE.

- (4) Той_{SOURCE} изглеждаше сякаш []_{AGENT} не беше
 It look-PST as if pro-drop not be-PST
 метен [от години]_{TIME}.
 sweep-PTCP from years.
 ‘It looked as if it had not been swept in years.’

These errors – non-identifying a pro-drop AGENT and assigning this element to the subject in passive clauses – account for the major part of wrong annotations, while the assignment of the frame element AGENT to a wrong constituent is much rarer. These conclusions indicate that the greatest effort in identifying AGENTS should be on refining the pro-drop rules and enhancing them with instructions regarding passive subjects.

The frame element recognition task yields similar results for the THEME and the SOURCE. A closer look at the data shows that the two are not well discriminated by the models, especially in the cases where an entity may be either a THEME or a SOURCE depending on the context in which it occurs. Thus *feathers* is incorrectly identified as a THEME in Example 5 instead as a SOURCE.

- (5) Патиците_{AGENT} чистеха [перата си]_{THEME}.
 Ducks-DEF clean-PST.PROG feathers-DEF their.
 ‘The ducks were cleaning their feathers.’

The training should thus be focused on improving the model’s discriminating ability. This will also have effect on the recognition task as well, as in a number of cases the models do not identify an NP or a PP as a realisation of either of the two frame elements when it actually should be.

6. Conclusions

The pilot reported in this paper supports the conclusion that the openly available LLMs evaluated here are not reliable annotators of Bulgarian frame-semantic structure when used without any preliminary fine-tuning. The best-performing model, BgGPT-Gemma-3-12B-IT, achieved an overall macro of 58.4 across the tasks, which is below reasonable reliability.

Considerably more work is required before the pipeline can be deployed, particularly experiments in several directions: implementing an example-selection strategy that retrieves the most syntactically similar gold-annotated sentences for each input with focus on the structurally non-typical sentences; iterative re-prompting with the model’s own intermediate output and an explicit feedback (instructions) and reduction of cross-task error propagation; supervised fine-tuning.

The error analysis further shows that the annotation tasks are closely intertwined: errors in the structural layers propagate into frame element assignment, so that treating the tasks jointly, rather than as isolated steps, is likely to yield more reliable annotation.

An important empirical result of the pilot is that dedicated Bulgarian fine-tuning is decisive, in particular for tasks involving analysis of Bulgarian morphosyntax – pro-drop, clause segmentation, phrase boundaries. This suggests on possible experimentations on architectures that combine a language-specialised model for the structural pre-processing stages with a reasoning-augmented model for the frame-semantic decisions, rather than a single model handling the entire pipeline.

7. Limitations

The present study is a controlled pilot and its scope is limited in three respects: 1) the training and evaluation set is restricted to 350 Bulgarian sentences instantiating two related frames – **Removing** and **Emptying**, evoked by a small inventory of verbs describing cleaning; 2) it is unclear whether the observations generalise to other frames, predicate classes, or text registers; 3) the annotation of the dataset was manually validated which is expensive to perform on a large scale; 4) results reflect a single prompt formulation per task, and prompt-engineering variance was not systematically explored; 5)

due to the small size of the dataset, small differences between models should be interpreted with caution, as no significance testing was performed.

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

- Chaoyi Ai and Kewei Tu. 2024. [Frame Semantic Role Labeling Using Arbitrary-Order Conditional Random Fields](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17638–17646.
- Saba Anwar, Artem Shelmanov, Nikolay Arefyev, Alexander Panchenko, and Christian Biemann. 2023. [Text Augmentation for Semantic Frame Induction and Parsing](#). *Language Resources and Evaluation*, pages 1–46.
- Collin F. Baker. 2008. [FrameNet, Present and Future](#). In Jonathan Webster, Nancy Ide, and Alex Chengyu Fang, editors, *Proceedings of the First International Conference on Global Interoperability for Language Resources*. City University, Hong Kong.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. [The Berkeley FrameNet Project](#). In *COLING-ACL '98: Proceedings of the Conference*. Montreal, Canada, pages 86–90.
- Frederico Belcavello, Ely E. Matos, Arthur Lorenzi, et al. 2025. [Perspectivized setting: The case of FrameNet annotation](#). In *arXiv preprint*. ArXiv:2510.25904.
- Frederico Belcavello, Tiago Timponi Torrent, Ely E. Matos, Adriana S. Pagano, Maucha Gamonal, et al. 2024. [Frame2: A FrameNet-based Multimodal Dataset for Tackling Text-image Interactions in Video](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Re-*

- sources and Evaluation (LREC-COLING 2024)*, pages 7429–7437, Torino, Italia. ELRA and ICCL.
- Frederico Belcavello, Tiago Timponi Torrent, Ely E. Matos, et al. 2022. [Charon: A FrameNet Annotation Tool for Multimodal Corpora](#). In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) at LREC 2022*, pages 91–96, Marseille, France. European Language Resources Association.
- Eckhard Bick. 2022. [A Framenet and Frame Annotator for German Social Media](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 3942–3949, Marseille, France. European Language Resources Association.
- Hans Christian Boas, Josef Ruppenhofer, and Collin F. Baker. 2024. [FrameNet at 25: Results and Applications](#). *International Journal of Lexicography*, 37(3):1–22.
- Hans Christian Boas, Josef Ruppenhofer, and Collin F. Baker. 2025. [FrameNet at 25: Results and Applications \(follow-up\)](#). *International Journal of Lexicography*. Advance article. <https://doi.org/10.1093/ijl/ecaf011>.
- Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2010. [The Past Meets the Present in Swedish FrameNet++](#). In Anne Dykstra and Tanneke Schoonheim, editors, *Proceedings of the 14th EURALEX International Congress*, pages 269–281. Fryske Akademy, Leeuwarden/Ljouwert, The Netherlands.
- Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. [The SALSA Corpus: A German Corpus Resource for Lexical Semantics](#). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. European Language Resources Association (ELRA), Genoa, Italy.
- Marie Candito, Pascal Amsili, Lucie Barque, Farah Benamara, Gaël de Chalendar, Marianne Djemaa, Pauline Haas, Richard Huyghe, Yvette Yannick Mathieu, Philippe Muller, Benoît Sagot, and Laure Vieu. 2014. [Developing a French FrameNet: Methodology and First results](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1372–

1379. European Language Resources Association (ELRA), Reykjavik, Iceland.
- Xinyue Cui and Swabha Swayamdipta. 2024. [Annotating FrameNet via Structure-Conditioned Language Generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 681–692, Bangkok, Thailand. Association for Computational Linguistics.
- Oliver Czulo, Alexander Ziem, and Tiago Timponi Torrent. 2020. [Beyond Lexical Semantics: Notes on Pragmatic Frames](#). In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 1–7, Marseille, France. European Language Resources Association.
- Jacob Devasier, Rishabh Mediratta, and Chengkai Li. 2025. [Can LLMs Extract Frame-Semantic Arguments?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics. ArXiv:2502.12516.
- Vanessa C. Ferreira and Vladia Pinheiro. 2021. [SpiNet: A FrameNet-like Schema for Automatic Information Extraction about Spine from Scientific Papers](#). In *AMIA Annual Symposium Proceedings*, pages 452–461.
- Charles J. Fillmore. 2003. [Valency and Semantic Roles: The Concept of Deep Structure Case](#). In Vilmos Ágel, Ludwig Eichinger, Hans Werner Eroms, Peter Hellwig, Hans Jürgen Heringer, and Henning Lobin, editors, *Dependenz und Valenz: Ein internationales Handbuch der zeitgenössischen Forschung*, volume 1, pages 457–475. Walter de Gruyter, Berlin.
- Maucha Gamonal, Adriana Pagano, Tiago Timponi Torrent, Ely E. Matos, and Arthur Lorenzi. 2024. [Automated Semantic Frame Annotation](#). In *Proceedings of the XXI EURALEX International Congress*, pages 67–81, Zagreb, Croatia. Institute for the Croatian Language.
- Luca Gilardi and Collin F. Baker. 2018. [Learning to Align across Languages: Toward Multilingual FrameNet](#). In Tiago T. Torrent, editor, *Proceedings of the International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons*, pages 13–22. European Language Resources Association (ELRA), Miyazaki, Japan.

- Google DeepMind. 2025. [Gemma 4: Frontier Multimodal Intelligence on Device](#). Google DeepMind Blog. Model available at <https://huggingface.co/google/gemma-4-E4B-it>. Released April 2025 under Apache 2.0 licence.
- Normunds Gruzitis, Gunta Nespore-Berzkalne, and Baiba Saulite. 2018. [Creation of Latvian FrameNet based on Universal Dependencies](#). In *Proceedings of the International FrameNet Workshop (IFNW)*, pages 23–27.
- Young Gyun Hahm, Youngsik Kim, Yousung Won, Jongsung Woo, Jiwoo Seo, Jiseong Kim, Seongbae Park, Dosam Hwang, and Key-Sun Choi. 2014. [Toward matching the relation instantiation from DBpedia ontology to Wikipedia text: Fusing FrameNet to Korean](#). In Harald Sack, Agata Filipowska, Jens Lehmann, and Sebastian Hellmann, editors, *Proceedings of the 10th International Conference on Semantic Systems, SEMANTiCS 2014, Leipzig, Germany, September 4-5, 2014*, pages 13–19. ACM.
- Jennifer Hay, Christopher Kennedy, and Beth Levin. 1999. [Scalar Structure Underlies Telicity in Degree Achievements](#). In Tanya Matthews and Devon Strolovitch, editors, *Proceedings of the 9th Semantics and Linguistic Theory (SALT 9), Ithaca, NY*, pages 127–144. CLC Publications: Cornell University.
- INSAIT Institute. 2025. [BgGPT 3.0 Release](#). INSAIT Blog. Model available at <https://huggingface.co/INSAIT-Institute/BgGPT-Gemma-3-12B-IT>.
- Adithya Jayanth et al. 2025. [Do LLMs Encode Frame Semantics? Evidence from Frame Identification](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. ArXiv:2509.19540.
- Christopher R. Johnson, Charles J. Fillmore, Esther J. Wood, Margaret Urban, Miriam R. L. Petruck, and Collin F. Baker. 2001. The FrameNet Project: Tools for Lexicon Building. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, pages 86–90. Stroudsburg: Association for Computational Linguistics.
- C. Kennedy. 2012. [The composition of incremental change](#). In Violeta Demonte and Louise McNally, editors, *Telicity, Change, and State*:

- A Cross-Categorial View of Event Structure*, pages 103–121. Oxford University Press.
- Svetla Koeva. 2008. *Balgarskiyat FreymNet. Semantiko-sintaktichen rechnik na balgarskiya ezik – kontseptualen model [Bulgarian FrameNet. Semantic-Syntactic dictionary of Bulgarian – Conceptual Model]*, pages 5–51. Institute for Bulgarian Language Prof. Lyubomir Andreychin, Sofia.
- Svetla Koeva. 2010. *Balgarskiyat FreymNet [Bulgarian FrameNet]*. Institute for Bulgarian Language Prof. Lyubomir Andreychin, Sofia.
- Svetla Koeva. 2021. [Towards Expanding WordNet with Conceptual Frames](#). In Sonja Bosch, Christiane Fellbaum, Marissa Griesel, Alexandre Rademaker, and Piek Vossen, editors, *Proceedings of the 11th Global WordNet Conference*, pages 182–191. Global Wordnet Association.
- Svetla Koeva, editor. 2024. *Semantichni studii. Sintaktichno i semantichno opisanie na balgarski glagoli [Semantic Studies. Syntactic and semantic description of Bulgarian verbs]*. Publishing House of the Bulgarian Academy of Sciences, Sofia.
- Svetlozara Leseva and Ivelina Stoyanova. 2020a. Beyond Lexical and Semantic Resources: Linking WordNet with FrameNet and Enhancing Synsets with Conceptual Frames. In Svetla Koeva, editor, *Towards a Semantic Network Enriched with a Variety of Semantic Relations*, pages 21–48. Prof. Marin Drinov Publishing House of the Bulgarian Academy of Sciences, Sofia.
- Svetlozara Leseva and Ivelina Stoyanova. 2020b. [Towards a Conceptual Description of Verbs](#). *Cybernetics and Information Technologies*, 20:108–124.
- Svetlozara Leseva and Ivelina Stoyanova. 2022. [Linked Resources towards Enhancing the Conceptual Description of General Lexis Verbs Using Syntactic Information](#). In *Proceedings of the Fifth International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, pages 214–223. Department of Computational Linguistics, IBL – BAS.
- Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago and London: The University of Chicago Press.

- Xinpeng Liu et al. 2025. Injecting Frame Semantics into Large Language Models for Role Labeling. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*. Association for Computational Linguistics.
- Büşra Marşan and Olcay Taner Yıldız. 2021. [FrameForm: An Open-source Annotation Interface for FrameNet](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 257–262, Online. Association for Computational Linguistics.
- Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, et al. 2022. [A Multilingual Benchmark to Capture Olfactory Situations over Time](#). In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change (LChange 2022)*, pages 1–10, Dublin, Ireland. Association for Computational Linguistics.
- Kyoko Ohara, Seiko Yamaguchi Fujii, Toshio Ohori, Ryoko Suzuki, Hiroaki Saito, and Shun Ishizaki. 2004. [The Japanese FrameNet Project: An Introduction](#). In *Proceedings of the Satellite Workshop on Building Lexical Resources from Semantically Annotated Corpora*, pages 9–11. European Language Resources Association (ELRA), Lisbon, Portugal.
- Ayush Pancholy, Miriam R. L. Petruck, and Swabha Swayamdipta. 2021. [Sister Help: Data Augmentation for Frame-Semantic Role Labeling](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Bolette Pedersen, Sanni Nimb, Anders Søgaard, Mareike Hartmann, and Sussi Olsen. 2018. [A Danish FrameNet Lexicon and an Annotated Corpus Used for Training and Evaluating a Semantic Frame Classifier](#). In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.
- Qwen Team. 2025. [Qwen3 Technical Report](#). Model available at <https://huggingface.co/Qwen/Qwen3-8B>.

- Malka Rappaport Hovav. 2008. [Lexicalized meaning and the internal temporal structure of events](#). In Susan Rothstein, editor, *Theoretical and Crosslinguistic Approaches to the Semantics of Aspect*, pages 13–42. John Benjamins, Amsterdam.
- Levi Remijnse and Gosse Minnema. 2020. [Towards Reference-Aware FrameNet Annotation](#). In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 13–19, Marseille, France. European Language Resources Association.
- Levi Remijnse, Piek Vossen, Antske Fokkens, and Sam Titarsolej. 2022. [Introducing Frege to Fillmore: A FrameNet Dataset that Captures both Sense and Reference](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 39–50, Marseille, France. European Language Resources Association.
- J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson, C. F. Baker, and J. Scheffczyk. 2016. [FrameNet II: Extended Theory and Practice](#). International Computer Science Institute, Berkeley, California.
- Carlos Subirats. 2009. [Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon](#). In Hans C. Boas, editor, *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, pages 135–162. De Gruyter Mouton, Berlin.
- Tiago Timponi Torrent, Ely E. Matos, Frederico Belcavello, Marcelo Viridiano, Maucha Gamonal, Arthur D. Costa, and Mariana C. Marim. 2022. [Representing Context in FrameNet: A Multidimensional, Multimodal Approach](#). *Frontiers in Psychology*, 13:838441.
- Sara Tonelli and Stefano Menini. 2021. [FrameNet-like Annotation of Olfactory Information in Texts](#). In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCHCLfL 2021)*, pages 11–20, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.
- Tiago T. Torrent, Ely Matos, Ludmila Lage, Adrieli Laviola, Tatiane Tavares, Vânia Almeida, and Natália Sigiliano. 2018. [Towards Continuity between the Lexicon and the Constructicon in FrameNet Brasil](#). In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara,

- and Tiago T. Torrent, editors, *Constructicography: Constructicon Development Across Languages*, volume 22 of *Constructional Approaches to Language*, pages 107–140. Benjamins.
- Piek Vossen, Antske Fokkens, Isa Maks, and Chantal van Son. 2018. [Towards an Open Dutch FrameNet Lexicon and Corpus](#). In Tiago T. Torrent, Lars Borin, and Collin F. Baker, editors, *Proceedings of the International FrameNet Workshop 2018: Multilingual Framenets and Constructicons (LREC 2018)*, pages 75–80. European Language Resources Association (ELRA).
- Liping You and Kaiying Liu. 2005. [Building Chinese FrameNet database](#). In *International Conference on Natural Language Processing and Knowledge Engineering*, pages 301–306.
- Ce Zheng, Yiming Wang, and Baobao Chang. 2023. [Query Your Model with Definitions in FrameNet: An Effective Method for Frame Semantic Role Labeling](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14029–14037.

Pristupi automatskom označavanju semantičkih uloga i sintaksičkih odnosa

Svetlozara Leseva, Ivelina Stojanova

Sažetak

U radu se ispituju mogućnosti nekoliko javno dostupnih velikih jezičkih modela (LLM) za automatsko obeležavanje semantičkih uloga u bugarskom, morfološki bogatom, slovenskom pro-drop jeziku. Primena velikih jezičkih modela za ovu vrstu anotacije u bugarskom i dalje je nedovoljno istražena. Studija obuhvata kontrolisanu pilotsku anotaciju 350 rečenica koje sadrže glagole čišćenja (npr. *чистя, мета, мия, бърша*). Ovi glagoli prizivaju dva blisko povezana okvira u FrameNet-u, **Uklanjanje** (Removing) i **Pražnjenje** (Emptying), koji dele iste centralne elemente – AGENS, TEMU i IZVOR – ali ih različito sintaksički profilisu, na osnovu čega se uočavaju alternacije okvira. U radu se ocenjuju tri modela — Gemma 4 E4B-IT, BgGPT-Gemma-3-12B-IT i Qwen3-8B — u zadatku pripisivanja elemenata okvira, uz dodatnu evaluaciju za četiri podzadatka anotacije: određivanju granica klauza, identifikovanju raspona sintagmi, razrešavanju koreferencije i određivanju izostavljenog subjekta. Model BgGPT-Gemma-3-12B-IT, prilagođen bugarskom jeziku, postiže najbolje rezultate u svih pet zadataka ($F1 = 58,4$), a slede ga Qwen3-8B uz lančano zaključivanje (50,9) i Gemma 4 E4B-IT (38,7). Fino podešavanje (*fine-tuning*) za bugarski jezik najznačajnije je tamo gde je morfosintaksa presudna – kod izostavljanja zameničkog subjekta (pro-drop), segmentacije rečenica i granica sintagmi – dok sposobnost zaključivanja delimično kompenzuje nedostatke na zadacima koji su više konceptualne prirode. Dobijeni rezultati daju podsticaj za dalji rad na kreiranju upita zasnovanim na pretraživanju (*retrieval-based*), iterativnom samousavršavanju i nadgledanom finom podešavanju (*supervised fine-tuning*).

Ključne reči: FrameNet, elementi okvira, semantička anotacija, veliki jezički modeli

Appendices

A Prompt

```
prompt = (  
    "You are a linguistic annotator for Bulgarian."  
    "Annotate sentences containing forms of the verb"  
    " ' ' (to clean).\n\n"
```

```
"CRITICAL RULES:\n"
"- Copy the FULL original sentence (only from the "
"third column of the tsv) exactly, word for word\n"
"- Do NOT shorten, summarize or omit any part of "
"the sentence\n"
"- Only ADD annotation markers, never remove or "
"change words\n"
"- Identify explicit Agents in the sentence; "
"if there is pro-drop and not an explicit subject, "
"use empty brackets [] \n\n"
"FRAME DEFINITION:\n"
"Emptying\n"
"These are words relating to emptying containers "
"and clearing areas of some substance or items. "
"The area or container can appear as the direct "
"object with all these verbs, and is designated Source "
"because it is the source of motion of the Theme. "
"Corresponding to its nuclear argument status, it is "
"also affected in some crucial way, unlike Source in "
"other frames. Some words in this frame, such as "
"empty, may also occur with Path or Goal expressions.\n\n"
"FRAME ELEMENTS TO ANNOTATE:\n\n"
"[1] AGENT\n"
"The Agent is the entity that does the emptying.\n"
"-> Mark the FULL noun phrase as [phrase]_AGENT_i\n"
"-> Only use []_AGENT_i (pro-drop) if there is NO explicit"
"subject in the sentence\n\n"
"[2] CAUSE\n"
"The Cause is the non-sentient force which instigates "
"the emptying. Excludes Agent.\n"
"-> A basic washing is the CAUSE\n\n"
"[3] SOURCE\n"
"The Source is the region or container which is emptied "
"of something. In this frame, the Source occurs as an "
"NP Object.\n"
"-> the table is the SOURCE\n\n"
"[4] THEME\n"
"The Theme is the substance or items removed.\n"
"-> dishes is the THEME\n\n"
"Additional elements to annotate if present:\n"
"[5] GOAL -- destination of the removed items\n"
"[6] PATH -- path of movement of the removed items\n"
"[7] INSTRUMENT -- tool used to do the emptying\n"
"[8] LOCATION -- location where the emptying takes place\n\n"
"ANNOTATION RULES:\n"
"1. Mark the verb form with curly braces: {verb_form}\n"
"   - The main part of the verb form is in the "
"first column of the tsv\n"
```

```

" - Must be a whole word, never part of a word\n"
" - Can be multi-word (e.g. preceding auxiliary + main verb)\n\n"
" - Must annotate one verb in each sentence!\n\n"
" - Definitely mark as verb      ,      , "
"      ,      ,      \n\n"
"2. Mark each frame element as [phrase]_ELEMENT_i\n"
" - Use the FULL phrase, not just the head noun\n"
" - The phrase can be a noun phrase or a prepositional "
"phrase, sometimes an adverbial phrase\n"
" - If Agent is pro-drop (missing), insert []_AGENT_i "
"immediately before the verb\n"
" - i is a unique index per sentence, each frame element "
"gets its own index\n\n"
"3. Mark coreferential phrases with [phrase]_ELEMENT_coref_i\n"
" where i matches the index of the phrase it corefers with\n\n"
"OUTPUT FORMAT:\n"
"- Output exactly the same number of sentences as input\n"
"- One fully annotated sentence per line, no numbering\n"
"- Preserve the full original sentence with annotations added\n"
"- No explanations, no commentary\n\n"
"CORRECT EXAMPLES TO FOLLOW EXACTLY:\n"
"Below are correctly annotated sentences. Follow this "
"style of annotation:\n"
+ examples_text + "\n\n"
"Now annotate the sentences below. "
"Copy each sentence in full and exactly and add ONLY "
"annotations, no new words, in or out of brackets. "
"Output only the annotated sentences, one per line:\n\n"
+ batch
)

```

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

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

