Multilingual Corpus of Illustrative Examples on Activity Predicates

Ivelina Stoyanova, Hristina Kukova, Maria Todorova, Tsvetana Dimitrova
Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences
{iva, hristina, maria, cvetana}@dcl.bas.bg

Abstract

The paper presents the ongoing process of compilation of a multilingual corpus of illustrative examples to supplement our work on the syntactic and semantic analysis of predicates representing activities in Bulgarian and other languages. The corpus aims to include over 1,000 illustrative examples on verbs from six semantic classes of predicates (verbs of motion, contact, consumption, creation, competition and bodily functions) which provide a basis for observations on the specificity of their realisation. The corpus of illustrative examples will be used for contrastive studies and further elaboration on the scope and behaviour of activity verbs in general, as well as its semantic subclasses.

Keywords: activity predicates, semantic frames, multilingual corpus

1 Introduction

The paper discusses the ongoing work on the compilation, analysis and annotation of a corpus of examples in several languages along with the challenges it poses. The task-specific dataset comprises examples of activity verbs and is tailored to serve as language data for contrastive conceptual analysis of verbs. Initially, verbs belonging to six semantic classes were extracted (verbs of motion, contact, consumption, creation, competition, and bodily functions) from the Bulgarian WordNet (BulNet), which were further manually filtered on account of their appurtenance to the Vendlerean aspectual class of activity verbs. Examples from monolingual (semantically annotated) and parallel corpora were then excerpted illustrating the use of the verbs in context. The verbs under observation are assigned FrameNet frames mapped to the relevant WordNet synsets.

The main objective of our work is to construct a demonstration corpus of annotated examples illustrating the usage of activity verbs. Moreover, we aim at: (a) linking various lexical, semantic and conceptual resources in order to provide comprehensive description of verbs; (b) partial (or full, in case of parallel examples) cross-language alignment in terms of verb translational equivalents based on WordNet, and in terms of participants in their semantic frames; (c) flexibility of corpus design to allow data from other languages to be added; and (d) flexibility of annotation to facilitate the expansion of the scope and variability of the examples in the corpus.

At present, the corpus includes illustrative examples for the use of activity verbs in two languages – Bulgarian and English. Further work on the corpus will include at least two more languages – Russian (a Slavic language) and Italian (a Romance language). This will provide linguistic material for observations on both closely related and more distant languages.

The methodology for constructing the corpus, the description of verbs and the annotation of examples is largely language-independent and can be applied to extract and compile datasets of various languages.

The remainder of the paper is structured as follows. Section 2 contains a theoretical overview of previous work. Section 3 gives a description of the resources used in the study. Then we discuss the linguistic data and the process of its selection (in Section 4) along with some illustrative examples (in Section 5). Section 6 involves a description of how multilingual data is represented as well as some disputable points, and Section 7 proposes an annotation schema with its main principles and steps. Section 8 summarises the main findings of the study with a view to the future work.

2 Relevant works

Activity verbs are members of Vendler's aspectual classification (Vendler, 1957, 1967), where verbs are divided into activities, states, achievements, and accomplishments. This classification has been subsequently elaborated by Dowty (1979) and Van Valin and LaPolla (1997). They propose four key semantic features which define the scope of the aspectual classes: [± static], [± dynamic], [± telic], and [± punctual]. Punctuality distinguishes achievements (which are punctual) from accomplishments (which are non-punctual).

The classes of non-stative verbs are distinguished by: dynamism; presence or absence of an internal limit – "proceed towards a terminus" (Vendler, 1957) regarding accomplishments, and time limitation – "achievements occur at a single moment" (Vendler, 1957); terminativeness (Maslov, 1982); boundedness (Paducheva, 2009). The temporal limitation is often equated with the presence of an intrinsic end-point or an instantaneous climax. Some verbs can be classified in more than one aspectual class depending on their use, for example some verbs can be both activities and accomplishments (e.g. *walk / walk to the store*), and states and achievements (e.g. the ambiguity of many mental state verbs such as *recognize, understand, know*). Thus, the aspectual classes are considered at the VP level rather than at the lexical level, which means that the aspectual properties are expressed in a complex lexical, morpho-syntactic and valence-related way by the verb and its arguments (Rappaport Hovav, 2008: 16–20).

The class of active predicates are broadly described in the grammars of the Bulgarian language (Maslov, 1982; Gramatika, 1983; Barkalova, 1997). Different types of verbs of states, processes, actions, activities and events have been the subject of various studies on Bulgarian. Koeva et al. (2022) offer a detailed ontological description of predicates and predicatives of state in Bulgarian, Russian and English comparatively, based on large lexical resources and corpus data. Kostova (2010) describes lexical-semantic groups of basic verbs of motion; Vateva (2005) examines lexical-semantic relations between verbs of movement in Bulgarian and their metaphorical use in different functional styles; Dekova (2006) explores particular groups of motion verbs in comparison with English verbs. A semantic description of verbs of change is proposed in connection with their representation in the semantic frames of FrameNet (Leseva and Stoyanova, 2021). A classification and semantic description of verbs of contact has been proposed by Todorova (2023).

Different features of Bulgarian verbs and classifications have been proposed. Vlahova (2004) describes systemic dependencies between predicate types, semantic roles, grammatical categories and grammatical transformations of verbs. Koeva (2006b) offers a typology of Bulgarian verbs based on their (in)ability to form diatheses and alternations.

As far as we know, there is neither a complex research on the typology, semantic and syntactic properties of activities in Bulgarian in comparison with other languages, nor a specific database with particular selection of activity predicates.

In our work we step on the analysis of Kolokovska (2005) which is close to the interpretation of Lyons (1977) and Bulygina (1982) who consider activities as a conceptual-semantic category, characterised by the following semantic features: localisation in time; change of at least one participant in the situation and activity. We also rely on the aspectual understanding of the imperfective verbs in Bulgarian, representing the activity in its progress (Stankov, 1980: 6, 43).

3 Resources

3.1 Lexical-semantic resources

The selection of lexical entries and their lexical and semantic description are extracted from two main resources – WordNet and FrameNet.

WordNet (Miller, 1995; Fellbaum, 1998) represents the lexicon in the form of a network of synonym sets (synsets) interconnected by semantic, lexical and other relations. The main structural relation is hypernymy (and its opposite relation – hyponymy), by which the vocabulary of a given semantic field is organised into a tree, the beginning (root) being the most general or abstract concept of the corresponding field.

The semantic description of verb predicates in WordNet also includes their classification into general semantic classes based on assigned semantic primitives (Miller and Fellbaum, 2007), e.g. verbs of motion, verbs of emotion, verbs of communication, verbs of change, etc. We focus on several semantic classes of verbs in WordNet that

are representative of the class of activities: verbs of contact, verbs of motion, verbs of consumption, verbs of competition, verbs of body, and verbs of creation.

WordNet ensures vast lexical coverage of the English lexicon structured and enriched with lexical and semantic information in the form of synset glosses, usage examples, notes on the usage or grammatical specificities, and a rich network of semantic relations. The Bulgarian counterpart – the Bulgarian WordNet (BulNet) (Koeva, 2006a, 2021), is linked to the Princeton WordNet through interlingual index, and serves as the main resource for the extraction of Bulgarian verbs representing activities.

However, WordNet encodes no explicit semantic information about the participants in the situations described by the predicates and only limited information about their syntactic behaviour. Moreover, WordNet does not consistently reflects the different lexical meanings of verbs that can be referred to more than one aspectual class.

FrameNet (Baker et al., 1998) is a system of semantic frames which are schematic descriptions of the conceptual structure of situations through actors, circumstances, and other conceptual roles presented as frame elements. The frames are organised using a number of relations – hierarchical (Inheritance, Use, Subframe, etc.) and other types (for example, Causation).

Lexical units in FrameNet, in particular verbs, are grouped in semantic frames based on common semantics, formalised through a common set of participants and circumstances (frame elements) and the relations between them (Fillmore, 1982, 1985, 2003; Fillmore and Baker, 2009; Ruppenhofer et al., 2016). A set of valence patterns derived from corpus evidence characterises each lexical unit. Valence patterns show the configurations of frame elements in the realisation of the verb with their respective syntactic function.

There have been efforts to construct a FrameNet-based resource for Bulgarian – BulFrameNet – a corpus-based lexicon giving an exhaustive account of the semantic and syntactic combinatorial properties of Bulgarian verbs (Koeva, 2010), while Koeva and Doychev (2022) present a web-based system for the extensive description of verbs using semantic frames offering a unified theoretical model for the formal presentation of frames and frame elements.

In our work, FrameNet description of verbs is used to design an annotation schema for the corpus which will ensure unified representation in terms of semantic frames, frame elements and syntactic function of the elements in the realisation of the frame. The description of verb semantics using FrameNet can contribute to the unified description of lexical-aspectual classes of verbs and the analysis of their specific syntactic realisation, in particular to allow comparisons between the realisation of activities with respect to other aspectual classes.

3.2 Corpora

Examples are extracted from a number of different corpora using a range of techniques: (i) monolingual semantically annotated corpora where the verbs are disambiguated and assigned a WordNet sense, thus making it possible to extract sentences illustrating particular verb meaning; (ii) parallel corpora which allow the extraction of parallel examples illustrating the usage of verbs in two or more languages; however, they require additional annotation, filtering and disambiguation; (iii) additional monolingual resources to collect examples for language specific usage or less frequent linguistic phenomena.

Monolingual semantically annotated corpora such as SemCor (Miller et al., 1993) and BulSemCor (Koeva et al., 2011) are used in order to extract illustrative examples for English and Bulgarian, respectively. Words are annotated with WordNet senses which enables the extraction of examples.

Parallel corpora are a useful source of examples which illustrate the use of verbs with equivalent or similar meaning in more than one language in aligned sentences.

The Bulgarian-English Sentence- and Clause-Aligned Corpus (BulEnAC)[1] (Koeva et al., 2012a) is a parallel corpus of aligned Bulgarian and English sentences and clauses with annotation of the syntactic relation between clauses. The corpus contains 366,865 tokens (176,397 tokens in Bulgarian and 190,468 tokens in English). The texts in BulEnAC cover five categories: administrative texts, fiction, journalism, science, informal texts. Texts for each language have been

[1] https://dcl.bas.bg/en/resources_list/bulenac/

annotated with sentence and clause boundaries and then semi-automatically aligned at the clause level (automatic alignment followed by manual validation). Having aligned clauses is useful for extracting parallel examples as, on one hand, the clause provides the minimal scope for realising the verb's arguments, and on the other hand, aligned clauses ensure easier matching of translational equivalents, both for the verb and for its arguments.

The Bulgarian National Corpus is the largest corpus for Bulgarian: it consists of a monolingual (Bulgarian) part and 47 parallel corpora and amounts to 5.4 billion tokens. The Bulgarian part includes about 1.2 billion tokens of running text distributed in 240,000 text samples[2] (Koeva et al., 2012b). The Bulgarian-English parallel corpus within the Bulgarian National Corpus covers over 100,000 parallel texts and 260 mln. tokens. There is also a large parallel corpus with Italian, and a small Bulgarian-Russian parallel corpus of fiction.

Additional sources of illustrative examples can also be used. For English, the corpus of examples in FrameNet[3] (Burchardt and Pennacchiotti, 2008) annotated with explicit and implicit frame elements supplies extensive empirical evidence about the syntactic realisations of semantic frames that is valuable not only for linguistic generalisations about the target language (English) but as a point of departure for cross-linguistic observations. Semantically annotated corpora exist for many languages, including for our target set of languages (Russian and Italian)[4].

In addition, examples from other parallel corpora can be added. All target languages are covered in the parallel subcorpora within the Russian National Corpus (Savchuk et al., 2024) from which additional examples can be extracted and manually validated.

4 Selection of verbs

As a first step, we extracted from the Bulgarian WordNet (which has been developing as a parallel resource to the Princeton WordNet (Koeva, 2021)) all (single) verbs from the synsets belonging to the semantic classes verb.motion, verb.creation, verb.contact, verb.competition, verb.consumption, and verb.body (Miller, 1995; Fellbaum et al., 2009). We analysed the verbs and selected only those that represent activities. The rationale behind using WordNet, and the Bulgarian WordNet in particular, as a source for verbs selection is that: (a) we are able to extract sets of verbs from the observed semantic classes; (b) we have access to the lexical and semantic description of the verbs in the Bulgarian WordNet; (c) we can also extract translational equivalents in other languages (linked to Princeton WordNet) and further collect illustrative examples for their use in corpora.

Along with the semantic information, the Bulgarian WordNet comprises also some lexicogrammatical information encoding the Bulgarian verb aspect. Using this, we filtered the extracted verbs and selected only the imperfective ones. Then we selected manually the verbal set so far leaving aside the prefixed verbs, as well as those that refer to states, accomplishments, and achievements (Vendler, 1967). The verbs that were selected refer to continuing activities, may have human or human-like volitional subjects, do not have a terminal point and a (tangible) result. More or less we step in the process of selection on the preliminary criteria offered in Koeva and Ivanova (2024).

For example, the verb *write* 'produce a literary work', which is classified as verb.creation, would refer to an activity (ongoing in the past) in Example 1a, but also to an accomplishment (with a tangible result) in Example 1b. In Bulgarian, two different verbs would be used – a non-prefixed imperfective one *пиша* (Example 1c) and the prefixed perfective verbs *напиша* (Example 1d).

(1) **a.** En: He WROTE novels, short stories, lyrics, essays, plays for almost 70 years.
b. En: He did the research, and he WROTE the book.
c. Bg: Той ПИШЕ романи, кратки разкази, поезия, есета и пиеси в продължение на почти 70 години.
d. Bg: Той направи проучването и НАПИСА книгата. 'He did the research and WROTE the book.'

Idiomatic and phrasal verbs (as in *ходя на лов* 'hunt; run; hunt down; track down' with a definition 'pursue for food or sport (as of wild animals)'), light verbs (as in *търси отговор* 'looking for an answer') and English verbs with no lexicalisation in Bulgarian (e.g., *краде база* 'steal'

[2] https://search.dcl.bas.bg/
[3] http://framenet.icsi.berkeley.edu/
[4] http://globalwordnet.org/resources/wordnet-annotated-corpora/

with a definition 'steal a base' in the domain of basketball) were also discarded (*ловувам* 'hunt' which is the synonym of *ходя на лов*, however, would be left in the verb dataset).

semantic class	Synsets	Literals
verb.contact	265	334
verb.motion	199	296
verb.creation	89	109
verb.consumption	49	69
verb.competition	70	100
verb.body	24	33

Table 1: Selected verbs according to their semantic class

The resulting set consists of 941 verbs, most of which are members of synsets classified as verb.contact (334 (single) verbs, e.g. *чистя* 'clean, make clean' with a definition 'make clean by removing dirt, filth, or unwanted substances from') and verb.motion (296 verbs, e.g. *плувам* 'swim' with a definition 'travel through water'); among the selected verbs there are 100 verbs classified as verb.competition (e.g. *воювам* 'war' defined as 'make or wage war'); 109 verbs are classified as verb.creation (e.g. *свиря* 'play' with a definition 'play on an instrument'); relatively smaller are the class verb.consumption – 69 verbs (e.g. *ям* 'eat' defined as 'eat a meal; take a meal'), and verb.body – 33 verbs (e.g. *душа* 'choke; strangle' defined as 'constrict (someone's) throat and keep from breathing').

5 Selection of illustrative examples

The verbs are used to automatically extract examples from: (i) semantically annotated corpora and (ii) parallel corpora with no semantic annotation.

From semantically annotated corpora we have extracted automatically examples using the predetermined verb set. The verbs are disambiguated, and assigned a particular WordNet sense (see section 4). Example 2 shows two sentences from semantically annotated corpora showing two words belonging to the same WordNet sense.

From parallel corpora with no semantic annotation, we extract Bulgarian examples containing verbs from the predetermined set with all their possible WordNet senses, which are then aligned to a sentence in the other language (English) in which the verb is also identified. Example 3 shows two aligned parallel sentences; all possible senses of the identified Bulgarian verb are given; for each of the possible senses it is checked whether the English verb can be found among the literals of that synset, and in this way the most likely candidate sense(s) are identified.

(2) Examples from SemCor and BulSemCor for verbs of the same synset (sentences are not parallel, only verbs are linked to WordNet synset: eng-30-01698271-v / пиша 'write', verb.creation, 'produce a literary work').
a. En: Mr. Sansom WRITES of foreign parts with a dedication to decoration worthy of a pastry chef creating a wedding cake. (SemCor)
b. Bg: Баща ѝ ПИШЕШЕ криминални романи, които се славеха с особен успех и тя ги четеше с удоволствие. (BulSemCor)

(3) Examples from BulEnAC for verbs in aligned parallel sentences.
a. **Bg:** Колко пъти съм СЛАГАЛА камъни да варя, за да не разберат съседките, че дни наред нямаме какво да сложим в тенджерата!
En: How many times have I PUT stones to boil, so that the neighbours won't know that days after days we have nothing to put in the pot!
Possible WordNet senses for *put*:
eng-30-01493380-v / verb.contact / 'place temporarily'
eng-30-01494310-v / verb.contact / 'put into a certain place or abstract location'
eng-30-00050652-v / verb.body / 'put clothing on one's body'
eng-30-01500372-v / verb.contact / 'cause to sit or seat or be in a settled position or place'
eng-30-01465921-v / verb.contact / 'arrange or fix in the desired order'

b. Several times I've had to PUT stones on to boil so the neighbors wouldn't know that we often go for many days without putting on the pot.
Then the possible verb senses / WordNet synsets were aligned to their BulNet correspondences:

eng-30-01493380-v / Not aligned
eng-30-01494310-v / **Aligned**
eng-30-00050652-v / Not aligned
eng-30-01500372-v / Not aligned
eng-30-01465921-v / Not aligned

Next, we selected manually the appropriate illustrative examples for Bulgarian and for English. We disregard examples with verbs used in figurative (metaphorical) contexts (as in *The world 's best golfer, shooting below par, came to the last hole of the opening round*…), in multiword expressions (as part of light verb constructions or idioms, as in *eat humble pie*), verbs in passive constructions, or in other uses which do not refer to ongoing activities in the present or in the past (thus, certain verb forms can be excluded, such as past simple, past perfect, etc.).

A set of corpus examples was selected with the distribution shown in Table 2. So far there are 245 examples selected for Bulgarian and 257 for English. The aim is to achieve a dataset of over 1000 examples for each language.

	Bulgarian		English	
Semantic class	**All**	**Selected**	**All**	**Selected**
verb.contact	116	50	3485	54
verb.motion	153	96	2867	81
verb.creation	52	35	1060	40
verb.consumption	50	31	1005	37
verb.competition	24	14	576	20
verb.body	38	19	604	25

Table 2: Selected examples from BulSemCor (Bulgarian) and SemCor (English)

6 Challenges in the process of selection

The main challenge was, first, to differentiate between the different senses of the verb, which in some cases are very close (Example 4), and second, to distinguish activities and accomplishments and achievements with verbs that select direct configuration of arguments or arguments with particular semantic characteristics (Example 5). Some verbs may be categorised differently with respect to the intentional elements – such as *плача* 'to cry', which can refer to a non-intentional, as well as to an intentional act.

(4) Examples from BulEnAC for verbs in aligned parallel sentences with several possible closely related WordNet senses.
a. Bg: През целия си живот СИ ТЪРСИЛ това съкровище, само за да получиш уважението на историците.
b. En: You've spent your entire life SEARCHING for this treasure, only to have the respect of the historical community.
Possible WordNet senses:
eng-30-01317533-v / търся / verb.contact / 'go in search of or hunt for'
eng-30-01315613-v / търся / verb.contact / 'try to locate or discover, or try to establish the existence of'

(5) Examples from BulEnAC for verbs in aligned parallel sentences with several possible closely related WordNet senses.
a. Since he couldn't sleep anyway, he might as well stand their watches for them or WRITE their reports. (telic, thus classified as an accomplishment)
b. All his life her father WAS WRITING poems and novels. (generalised atelic; or telic + iterative)
c. In his poems he WROTE about the beauty of the countryside.
WordNet synset:
eng-30-01698271-v / write / verb.creation / 'produce a literary work'

In addition, mismatches are observed in the extracted parallel examples on various levels as illustrated in the examples below.

(6) Example from BulEnAC for verbs in aligned parallel sentences with mismatch in translation.
a. We're not even out.
b. Дори не СЕ РАЗХОЖДАМЕ.
'We do not even walk around.'

(7) Example from the Russian-Bulgarian parallel corpus within RusNC[5] for verbs in aligned parallel sentences with different verbs. The example was extracted manually.
a. Ru: Внизу проплывали игрушки-парусники… (Alexander Belyaev. The Ruler of the World. Russian edition: 1940.
b. Bg: Отдолу ПЛАВАХА играчки платноходки. (Bulgarian translation:

[5]The examples are taken from The Russian National Corpus (https://ruscorpora.ru).

1988, trans. by Assen Trayanov)
'Toy sailboats floated from below...'

This initial work on data selection served as a starting point for laying out the theoretical basis in determining the scope of **activity predicates**, their specific features and the possible approaches for the distinction between different senses of a verb with a view to its realisation in text.

7 Proposed annotation schema

The corpora used for extracting examples are supplied with basic annotation such as sentence splitting, POS tagging, lemmatisation, performed for both English and Bulgarian (Koeva et al., 2020).

The annotation aims at identification and description of the following syntactic components: (a) the verb, its WordNet sense, and the semantic frame it evokes; (b) noun phrases matched to frame elements and serving as external argument (NP.ext) or direct object (NP.Obj); (c) prepositional phrases (PP) matched to frame elements; (d) subordinate clauses marked with different conjunctions, direct quotes usually marked using punctuation, and other lexical elements that realise frame elements. In particular, as a minimum we aim to identify and annotate core frame elements, but in some cases non-core frame elements which are essential for the interpretation of the verb, are also annotated (e.g., when an element is essential to distinguish a verb as an activity rather than other aspectual class).

Here we present the main principles and steps for the annotation of the illustrative corpus of examples which aim to ensure the consistency of the annotation as well as the flexibility allowing for its expansion in terms of including more languages, more examples and more levels of annotation. The annotation is ongoing.

7.1 Matching a verb to a WordNet sense

Each verb is matched to a WordNet sense in order to: (1) provide cross-language linking to translational equivalents of the verb in different languages by linking them through the inter-lingual index of WordNet; and (2) provide linking to FrameNet and assign a FrameNet semantic frame to the verb, so that we can investigate the syntactic realisation of the verb and the frame elements in its evoked semantic frame.

As seen above, in semantically annotated corpora verbs have already been disambiguated and assigned a WordNet sense. However, these corpora are limited in size and coverage, and other more general corpora are also used for deriving examples. Additional semi-automatic procedures are applied to identify the WordNet verb sense. For example, in parallel corpora the two verbs within the aligned sentences can be used for additional automatic validation whenever possible. For monolingual corpora and other cases of verb ambiguity, manual validation has been performed.

7.2 Identification of the FrameNet frame evoked by a verb

After the verb has been matched to a particular WordNet sense, it can be assigned the FrameNet frame that characterises the verbs of the respective synset. For this purpose we rely on the mapping between WordNet synsets and FrameNet frames ((Shi and Mihalcea, 2005), (Tonelli and Pighin, 2009), (Palmer et al., 2014), (Leseva et al., 2018), among others).

By identifying the frame evoked by the verb, we are able to analyse the configurations of frame elements in the example sentences and to make observations on the verb class based on its realisation.

7.3 Identification of the syntactic components corresponding to core frame elements

We take as a point of departure the valence patterns as a collective set for all lexical units evoking a given frame. The generalised valence patterns show the possible configurations of frame elements for the evoked semantic frame and their corresponding syntactic realisations.

(8) Valence patterns for the FrameNet frame **Text creation** evoked by *write* 'produce a literary work'.
$[\text{NP.Ext}]_{\text{Author}}$ $[\text{NP.Obj}]_{\text{Text}}$
$[\text{NP.Ext}]_{\text{Author}}$ $[\text{NP.Obj}]_{\text{Text}}$ $[\text{PP}]_{\text{Time}}$
$[\text{NP.Ext}]_{\text{Author}}$ $[\text{NP.Obj}]_{\text{Text}}$ $[\text{ADVP}]_{\text{Time}}$
$[\text{NP.Ext}]_{\text{Author}}$ $[\text{NP.Obj}]_{\text{Text}}$ $[\text{PP}]_{\text{Manner}}$
$[\text{NP.Ext}]_{\text{Author}}$ $[\text{NP.Obj}]_{\text{Text}}$ $[\text{ADVP}]_{\text{Manner}}$

7.4 Identification of the valence pattern associated with the example

The FrameNet valence patterns describe all the co-occurrence combinations of frame elements (both core and non-core) attested for each annotated lexical unit in the FrameNet annotated corpus. The set of the identified and annotated frame elements is matched against the set of possible valence patterns associated with the semantic frame of the verb. Priority is given to valence patterns containing only core frame elements than to more elaborate patterns, as well as to more frequent patterns (frequency is extracted from the dataset of annotated examples in English from FrameNet).

There may be a mismatch in the syntactic category across languages, e.g., a certain frame element may be a direct object in one language and a prepositional object in another. Languages may also differ in terms of the overtness of syntactic information, i.e. the possibility to leave an obligatory element non-explicit (null instantiations retrievable from the context or the grammatical construction); language-specific diatheses, constructions, word order, morphosyntactic features, etc. The inventory of means that introduce certain frame elements such as prepositions, conjunctions, wh-words, etc. are also language-specific.

In annotating the data we pay attention to the cases of null instantiation, where the frame element is not overt – definite (e.g. pro-drop in Bulgarian), indefinite when the frame element represents a generalised non-specific entity (e.g. with communication verbs that are not directed to an addressee but such is implied), constructional when the lexical omission is licensed by certain constructions (e.g. imperative), and incorporated frame element – where the meaning of the frame element is incorporated in the meaning of the verb, and thus not expressed in the sentence.

Original patterns from FrameNet are generalised in order to allow cross-language match with the Bulgarian data. Particular attention is paid to examples which are not matched to a pattern in order to identify patterns characteristic for Bulgarian that do not appear in FrameNet or for English in general.

8 Applications and future work

The corpus aims at providing illustrative examples for the usage of activity verbs – a large and diverse class of verb predicates which shows various specific characteristics in contrast to other aspectual classes. Further, the aspectual properties of many verbs are not realised on the lexical level (in the lexical meaning) but within the larger unit (as in VP). Thus, the corpus will be a useful source of examples for studying the syntactic realisation of activity verbs.

The work on the corpus demonstrates the principles of information transfer: (a) by linking different resources in terms of scope, coverage, description layers, granularity of semantic categories, to provide a basis for comprehensive description of verb semantics; and (b) across languages to facilitate the development of resources and language processing tools for low-resourced languages such as Bulgarian.

Moreover, the corpus can be used to study the syntax and valence patterns across languages, thus facilitating comparative studies on conceptual structure. The corpus will provide empirical material for the comparative study between languages with lexical aspect (such as Bulgarian, Russian and other Slavic languages) and those without lexical aspect (such as English). The flexible structure and the annotation scheme allow the corpus to be expanded with more examples, languages and annotation.

The collection of parallel data of activity predicates is aimed at abstract ontological description and will allow the comparison of the features of conceptualisation, lexicalisation and grammaticalisation of activities in Bulgarian and other languages. The parallel collection will be used for theoretical comparison in the conceptualisation of different types of activities that correlate with the grammatical structure in individual languages. This will make possible the typological description of activity predicates and the highlighting of language-specific and universal features at the semantic and syntactic level.

Acknowledgments

This research is carried out as part of the project *Ontology of Activity Predicates – Linguistic Modelling with a Focus on Bulgarian* funded by the Bulgarian National Science Fund, Grant Agreement No КП–06–Н80/9 from 8.12.2023.

References

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.

Petia Barkalova. 1997. *Balgarskiat sintaksis – poznat i nepoznat*. Plovdivsko Universitetsko Izdatelstvo.

Tatiana V. Bulygina. 1982. K postrojeniju tipologiji predikatov v russkom jazuke [to build typologies predicates in russian]/Tv buligina. *Semanticheskiji tipu predikatov/[otv. red. ON Seliverstova].–M.: Nauka*, pages 7–85.

Aljoscha Burchardt and Marco Pennacchiotti. 2008. FATE: a FrameNet-Annotated Corpus for Textual Entailment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco.

Rositsa Dekova. 2006. *Lexical encoding of verbs in English and Bulgarian*. Det historisk-filosofiske fakultet.

David R. Dowty. 1979. *The Semantics of Aspectual Classes of Verbs in English*, pages 37–132. Springer Netherlands, Dordrecht.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.

Christiane Fellbaum, Anne Osherson, and Peter E. Clark. 2009. Putting Semantics into WordNet's "Morphosemantic"Links. In *Proceedings of the Third Language and Technology Conference, Poznan, Poland. [Reprinted in: Responding to Information Society Challenges: New Advances in Human Language Technologies. Springer Lecture Notes in Informatics]*, volume 5603, pages 350–358.

Charles J. Fillmore. 1982. Frame Semantics. In *Linguistics in the Morning Calm (Ed. by The Linguistic Society of Korea)*, pages 111 – 137. Soeul: Hanshin.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni Di Semantica*, 6:222–254.

Charles J. Fillmore. 2003. Valency and semantic roles: the concept of deep structure case. In Vilmos Ágel, Ludwig M. Eichinger, Hans Werner Eroms, Peter Hellwig, Hans Jürgen Heringer, and Henning Lobin, editors, *Dependenz und Valenz: Ein internationales Handbuch der zeitgenössischen Forschung*, pages 457 — 475. Walter de Gruyter.

Charles J. Fillmore and Collin F. Baker. 2009. A frames approach to semantic analysis. In B. Heine and H. Narrog, editors, *The Oxford handbook of linguistic analysis*, pages 313—340. Oxford: Oxford University Press.

Gramatika. 1983. *Gramatika na savremenen balgarski knijoven ezik T3. Sintaksis*. Sofia: Akademichno izdatelstvo „Prof. Marin Drinov".

Svetla Koeva. 2006a. BulNet (leksikalno-semantichna mrezha na balgarskiya ezik) — chast ot svetovnata leksikalno-semantichna mrezha. *Balgarski ezik*, pages 19 – 32.

Svetla Koeva. 2006b. *Sintaktichni transformatsii*, pages 106–138. SemaRSh, Sofia.

Svetla Koeva. 2010. *Balgarskiyat Freymnet*. Institute for Bulgarian Language, Sofia.

Svetla Koeva. 2021. The Bulgarian WordNet: Structure and specific features. *Papers of Bulgarian Academy of Sciences*, 8(1):47–70.

Svetla Koeva and Emil Doychev. 2022. Ontology supported frame classification. *Proceedings of the Fifth International Conference Computational Linguistics in Bulgaria*, pages 203–214.

Svetla Koeva and Elena Ivanova. 2024. Izsledvane na lingvistichni testove za razgranichavane na aspektualnite klasove (s fokus varhu balagrski i ruski). In *Derzhavinski cheteniya, Moskva*.

Svetla Koeva, Elena Ivanova, Yovka Tisheva, and Anton Zimmerling, editors. 2022. *Ontologiya na situatsiite za sastoyanie – lingvistichno modelirane. Sapostavitelno izsledvane za balgarski i ruski*. Prof. Marin Drinov Publishing House of Bulgarian Academy of Sciences.

Svetla Koeva, Svetlozara Leseva, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Hristina Kukova, and Maria Todorova. 2011. Design and development of the Bulgarian sense-annotated corpus. In *Information and communications technologies: present and future in corpus analysis: Proceedings of the III International Congress of Corpus Linguistics*, pages 143 – 150.

Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. Natural language processing pipeline to annotate Bulgarian legislative documents. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6988–6994, Marseille, France. European Language Resources Association.

Svetla Koeva, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Rositsa Dekova, Ivelina Stoyanova, Svetlozara Leseva, Hristina Kukova, and Angel Genov. 2012a. Bulgarian-English Sentence- and Clause-Aligned Corpus. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, page 51–62. Lisboa: Colibri.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012b. The Bulgarian National Corpus: theory and practice in corpus design. *Journal of Language Modelling*, 0(1):65–110.

Siya Kolokovska. 2005. Semantika i motiviranost na terminite za protsesi v savremennia balgarski knizhoven ezik. *Elektronno spisanie LiterNet 11*.

Nadezhda Kostova. 2010. *Osnovni glagoli za dvizhenie v balgarskiya ezik*. Avangard Prima, Sofia.

Svetlozara Leseva and Ivelina Stoyanova. 2021. Semantichno opisanie na glagoli za promyana i yerarhichna organizatsiya na kontseptualnite freymove. In *Proceedings from the Annual Conference of the Institute for Bulgarian Language*, volume 2.

Svetlozara Leseva, Ivelina Stoyanova, and Maria Todorova. 2018. Classifying verbs in WordNet by harnessing semantic resources. In *Proceedings of CLIB 2018, Sofia, Bulgaria*.

John Lyons. 1977. *Semantics*. Cambridge University Press.

Yuriy Maslov. 1982. *Gramatika na balgarskiya ezik*. Nauka i izkustvo, Sofia.

George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.

George A. Miller and Christiane Fellbaum. 2007. WordNet Then and Now. *Language Resources and Evaluation*, 41:209 – 214.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A Semantic Concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Elena V Paducheva. 2009. Leksicheskaya aspektualynosty i klassifikatsia predikatov po maslovu-vendleru. *Voprosi yazikoznania*, (6):3–20.

Martha Palmer, Claire Bonial, and Diana McCarthy. 2014. SemLink+: FrameNet, VerbNet and event ontologies. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014), Baltimore, Maryland USA, June 27, 2014*, pages 13–17. Association for Computational Linguistics.

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 Publishing Company.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher. R. Johnson, Collin. F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: extended theory and practice*. International Computer Science Institute, Berkeley, California.

Svetlana O. Savchuk, Timofey Arkhangelskiy, Anastasiya A. Bonch-Osmolovskaya, Ol'ga V. Donina, Yuliya N. Kuznetsova, Ol'ga N. Lyashevskaya, Boris V. Orekhov, and Mariya V. Podryadchikova. 2024. Nacional'nyj korpus russkogo jazyka 2.0: novye vozmožnosti i perspektivy razvitija [russian national corpus 2.0: new opportunities and development prospects. *Voprosy Jazykoznanija*, 2:7–34.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Computational Linguistics and Intelligent Text Processing: 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005. Proceedings 6*, pages 100–111. Springer.

Valentin Stankov. 1980. *Glagolniyat vid v balgarskiya ezik*. Nauka i izkustvo, Sofia.

Maria Todorova. 2023. Semantic annotation of common lexis verbs of contact in Bulgarian. In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, pages 11–17, Nancy, France. Association for Computational Linguistics.

Sara Tonelli and Daniele Pighin. 2009. New Features for FrameNet – WordNet Mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09), Boulder, USA*.

Robert D. Van Valin and Randy J. LaPolla. 1997. *Syntax: Structure, meaning and function*. Cambridge: Cambridge University Press.

Veselina Vateva. 2005. *Balgarskite glagoli za dvizhenie v semantichen i stilistichen aspekt*. Burgas: Diamant.

Zeno Vendler. 1957. Verbs and Times. *Philosophical Review*, pages 143 – 160.

Zeno Vendler. 1967. *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press.

Radka Vlahova. 2004. Kam strukturno-semantichnata harakteristika na nyakoi prefigirani glagoli v savremennia balgarski ezik. *Godishnik na SU. T. Ezikoznanie*, 89:5 – 29.