Illustrating the Usage of Verbs in WordNet: the Class of Self-motion Verbs

Svetlozara Leseva
Institute for Bulgarian Language
Bulgarian Academy of Sciences
Sofia, Bulgaria
zarka@dcl.bas.bg

Ivelina Stoyanova
Institute for Bulgarian Language
Bulgarian Academy of Sciences
Sofia, Bulgaria
iva@dcl.bas.bg

Abstract

The paper presents an outline of the procedures for selection and annotation of examples illustrating the usage of verbs belonging to various semantic classes (focusing on verbs of self-induced motion) in WordNet and the use of the annotated examples to validate the semantic and syntactic descriptions of the respective verbs as represented by the FrameNet valence patterns. This is done by complementing the information encoded in the synsets with conceptual information from FrameNet through the assignment of FrameNet frames and the associated valence and syntactic patterns.

The examples are collected from semantically annotated corpora for English and Bulgarian, as well as from an aligned parallel corpus. The annotation includes: assignment of a FrameNet frame to the verb in the sentence (and to the synset as a whole), annotation of the boundaries of the frame elements, their type (Self mover, Path, Source, Goal, etc.), and their syntactic category according to the types and categories defined in FrameNet. The annotated examples are then matched to the valence patterns associated with the respective verb, thus confirming their validity in Bulgarian.

The results of our work are the annotated corpus itself and the enriched representation of the verb synsets, which enables various semantic labelling and extraction tasks, parallel study of the semantic and syntactic expression, etc.

1 Introduction

We present our ongoing efforts on developing a language resource built to serve as a dataset of usage examples for the conceptual description of verbs in WordNet and their lexical, semantic and syntactic realisation in text. In this paper we focus in particular on verbs denoting self-induced motion (as a subclass of motion verbs). The annotated examples are linked to WordNet synsets and illustrate verbs in Bulgarian and English.

The complex semantic description of verbs derived from lexical semantic resources such as WordNet and FrameNet contains complementary semantic information (Baker and Fellbaum, 2009). Our work involves two main resources: (a) the Princeton WordNet, PWN (Fellbaum, 1998) and the Bulgarian WordNet (Koeva, 2021) – and for that matter, any other wordnet aligned with PWN, and (b) FrameNet (Baker et al., 1998; Ruppenhofer et al., 2016). We focus in particular on the information each of the resources provides and how it is used towards their mutual enrichment and enlargement.

We explore the transferability of the information across languages, in particular between English and Bulgarian, while also taking into account the language-specific aspects of the conceptual description in view of its syntactic realisation aiming at comprehensive study of the behaviour of verbs.

The main objective of our work is to supplement WordNet with a dataset of annotated examples illustrating the usage of verbs, in particular verbs describing self-induced motion. Moreover, we establish: (a) language-independent principles of annotation relying on the notion of universality of conceptual description using FrameNet frames; (b) cross-language alignment in terms of verb translational equivalents based on WordNet, and in terms of participants in their conceptual structure.

The principles of information transfer across languages can be beneficial for low-resourced languages such as Bulgarian. The dataset can be used to study the syntactic expression and the validity of the valence patterns across languages, thus facilitating comparative studies on conceptual structure.

2 Related work

In addition to WordNet and FrameNet which will be touched upon in Section 3 below we sketch a number of possibly interrelated resources that provide semantic and syntactic description and/or have served in various annotation initiatives.

VerbNet (Kipper-Schuler, 2005) provides good coverage of the English verb inventory and defines syntactic-semantic relations in a more explicit way by means of predicate-argument structures (combinations of thematic roles) with one-to-one linking to the syntactic category (type of phrase) and grammatical function (subject, object, etc.) of each argument expressed in a relatively small number of syntactic frames. Selectional restrictions are defined for the thematic roles assigned to a verb's arguments; they describe the semantic/ontological classes of nouns that express the arguments. However, although the verb classes describe the syntactic behaviour of verbs, many of the traditional thematic roles employed may be too general for the semantic description. Moreover, the existing mappings between WordNet synsets and VerbNet classes are very limited and do not provide sufficient data for analysis.

VerbAtlas (Fabio et al., 2019) is a lexical-semantic resource covering the verb synsets in BabelNet. BabelNet is a very large multilingual (for over 500 languages) semantic network integrating lexicographic and encyclopaedic knowledge from WordNet and Wikipedia (Navigli and Ponzetto, 2010). In VerbAtlas, each verb synset is assigned to a frame corresponding to its prototypical predicate-argument structure. Obligatory components are described using 26 semantic roles and the semantic restrictions governing their compatibility (116 types). A semantic annotation API with the frames described in it is also provided with the resource.

Predicate Matrix (de Lacalle et al., 2014) is a lexical resource resulting from the integration of several sources of predicate information: FrameNet, VerbNet, PropBank and WordNet, that have been previously aligned in Semlink[1] (Palmer, 2009). Predicate Matrix is compiled using advanced graph-based algorithms to extend the mapping coverage between resources. Additionally, by exploiting SemLink new role mappings are inferred among the different predicate schemas.

More recently, the SynSemClass lexicon[2] has marked a distinguishable effort towards combining the rich semantic description in the Vallex dictionary family with conceptual and syntactic information from external semantic resources in order to create a multilingual contextually-based verb lexicon. The aim of the lexicon is to provide a resource of classes of verbs that compares their semantic roles as well as their syntactic properties (Urešová et al., 2020a). In addition, each entry is linked to FrameNet, WordNet, VerbNet, OntoNotes and PropBank, as well as the Czech VALLEX.

Efforts on mappings lexical semantic resources are also relevant to our work. Correlation of WordNet with FrameNet is proposed for different languages, e.g. Danish (Pedersen et al., 2018), Dutch (Horak et al., 2008), Korean (Gilardi and Baker, 2018), Bulgarian (Leseva and Stoyanova, 2020). One of the challenges in aligning resources based on different methodologies is the alignment between the units that are represented in them. When aligning lexical units evoking particular frames from FrameNet and literals from synonym sets in WordNet, a coverage of 30.5% was achieved (Leseva and Stoyanova, 2019). New methods have been proposed to increase the coverage by discovering suitable literals based on semantic relations with literals already described in semantic frames (Burchardt et al., 2005) or on the basis of the inheritance of conceptual features in hypernym–hyponym trees, i.e., by assigning frames from hypernyms to hyponyms where possible and implementing a number of validation procedures based on the structural properties of the two resources, primarily the relations encoded in them (Leseva and Stoyanova, 2020).

Combining the semantic description of verbs from different resources is proposed by Urešová et al. (2020a,b). The result is a multilingual dictionary with the comprehensive description of the semantic classes of verbs and the semantic roles and syntactic properties of their arguments. The project is also aimed at creating an ontology of events, processes and states, and for this purpose each dictionary entry is linked to its correspondences in FrameNet, WordNet, VerbNet, Ontonotes and PropBank, as well as the Valence Dictionary of Czech Verbs (Lopatková et al., 2016), which presents the predicate-argument structure of each verb, its semantic class and the syntactic transformations (diatheses) in which it participates.

3 Resources

Here, we outline the lexical-semantic, conceptual and corpus resources employed in the study.

[1] https://verbs.colorado.edu/semlink/

[2] https://ufal.mff.cuni.cz/synsemclass

3.1 WordNet

WordNet[3] (Miller, 1995; Fellbaum, 1998) provides extensive lexical coverage; the verbs represented in it are organised in 14,103 synsets (including verb synsets specific for Bulgarian). We use both Princeton WordNet and the Bulgarian WordNet (Koeva, 2021), which are aligned at the synset level using unique synset identifiers. WordNet provides the most coarsely-grained semantic division in terms of a set of language-independent semantic primitives (semantic classes) assigned to all the nouns and verbs in the resource. The verbs fall into 15 groups, such as verb.change (verbs describing change in terms of size, temperature, intensity, etc.), verb.cognition (verbs of mental activities or processes), verb.motion (verbs of change in the spacial position), verb.communication (verbs describing communication and information exchange), etc.[4]

Verb synsets are interrelated and form a hierarchical structure by a troponymy relation (a manner relation analogous to hypernymy in nouns); for example, in *talk – whisper* the second member of the pair refers to a particular, semantically more specified, manner of performing the action referred to by the first verb (Fellbaum, 1999).

3.2 FrameNet

FrameNet[5] (Baker et al., 1998; Baker, 2008) is a lexical semantic resource that couches lexical and conceptual knowledge using the apparatus of frame semantics. Frames are conceptual structures that describe types of objects, situations, or events along with their components (frame elements) (Baker et al., 1998; Ruppenhofer et al., 2016). Depending on their status, the frame elements (FE) can be core, peripheral, or extra-thematic (Ruppenhofer et al., 2016). In terms of the conceptual description, we deal primarily with core FEs, which instantiate conceptually necessary components of a frame and which in their particular configuration make a frame unique and different from other frames.

FrameNet frames are organised into a hierarchical network, using a number of frame-to-frame relations (Ruppenhofer et al., 2016, 81–84). Here we list the hierarchical relations that bear most relevance to the internal structure of thematic verb classes. These are: Inheritance – a relationship between a parent frame and a more specific (child) frame, such that the child frame elaborates the parent frame; Uses (also called ‘weak inheritance’) – a relationship between two frames where the first one makes reference in a very general kind of way to the structure of a more abstract, schematic frame; Perspective – a relation indicating that a situation viewed as neutral may be specified by means of perspectivised frames that represent different possible points-of-view on the neutral state-of-affairs; Subframe – a relation between a complex frame referring to sequences of states and transitions (each of which can be separately described as a frame), and the frames denoting these states or transitions.

FrameNet has been employed in various initiatives, most notably ones focused on: (i) the creation of FrameNet-like resources for other languages such as the efforts undertaken within the Multilingual FrameNet initiative (Gilardi and Baker, 2018); (ii) the annotation of data using these resources, which has been carried out within the Multilingual FrameNet Shared Annotation Task (Torrent et al., 2020). Both research venues have confirmed empirically the applicability of the FrameNet descriptions across languages.

3.3 Combining WordNet and FrameNet

The combination of the resources helps redeem some of their shortcomings regarding conceptual description. A particular deficiency to the optimal use of the rich semantic information provided by FrameNet is its relatively small coverage in terms of lexical units. One way to alleviate this is to expand the coverage of FrameNet against the WordNet sense inventory through procedures for mapping WordNet synsets whose members evoke an existing frame but have not been matched with one yet, as well as through defining new frames to describe parts of the lexicon that have not been described yet.

We use a mapping between WordNet and FrameNet obtained from several already available ones (Shi and Mihalcea, 2005; Tonelli and Pighin, 2009; Leseva and Stoyanova, 2020). The task essentially involves manual validation of the accuracy of the proposed automatic mapping for the lexical units selected for the study (i.e. lexemes describing motion and self-induced motion in particular) and correction of the frames assigned to them if necessary. The validation consists in checking whether the proposed definition for the frame, the configuration of frame elements and their syntactic expres-

[3] `https://wordnet.princeton.edu/`

[4] The list of semantic primes along with short definitions is available at: `https://wordnet.princeton.edu/documentation/lexnames5wn`.

[5] `https://framenet.icsi.berkeley.edu/fndrupal/`

sion are reflected by the semantics and syntactic properties of the respective verb.

We first inspect the verb senses that have counterparts in both FrameNet and WordNet, i.e. verbs that have been encoded in both resources and have been mapped to each other. If the alignment is correct, the two lexemes will describe (near-)identical senses. Compare, for instance, the verb synset {walk:1} ("use one's feet to advance; advance by steps") and the lexical unit *walk.v* ("move at a regular and fairly slow pace by lifting and setting down each foot in turn"). Although the phrasing differs, the two definitions clearly describe the same sense, as additionally confirmed by the usage examples and the verbs' place in the overall internal structure of the respective resource: {walk:1} is a hyponym of the synset {travel:1; move:1; go:1; locomote:1}, the root of the subtree which contains most of the self-induced motion verbs, and *walk.v* evokes the Self_motion frame (one of the principal frames describing motion); it is also a descendant of the Motion frame, the prototypical representative of this semantic domain (Johnson et al., 2001, 16).

At the next stage we move on to validating the assignment of frames to WordNet verbs that do not have a counterpart in FrameNet. We implement this step through exploring the system of semantic relations in the two resources, in particular the inheritance of semantic information between frames. For instance, the verb synset {gallop:4} ("go at galloping speed") does not have a correspondence in FrameNet, but its hypernym {pace:5} ("go at a pace") is assigned the Self_motion frame. After inspecting {gallop:4}, we are able to confirm the validity of the automatic assignment of the frame of its hypernym. Other procedures involving the internal structure of the resources are also applied in the process.

As a result of the validation of the synset-to-frame alignment of the verbs belonging to the domain of motion, we obtain a list of pairs of verb senses and FrameNet frames which describe the semantics of the verbs in the respective synsets. This list represents the collection of senses and the pertaining semantic descriptions derived from both resources that serves as an inventory for which to supply examples.

3.4 Corpora

In order to explore the syntactic expression of the verbs and their participants we study the use examples from various corpora. First, we rely on semantically annotated corpora – the English SemCor and its counterpart BulSemCor, both of which are annotated with WordNet senses.

SemCor (current version 3.0) (Miller et al., 1993, 1994; Landes et al., 1998) is compiled by the Princeton WordNet team and covers texts excerpted from the Brown Corpus. SemCor is supplied with POS and grammatical tagging and all open-class words (both single words and multiword expressions, as well as named entities) are semantically annotated by assigning each word a unique WordNet sense (synset ID). SemCor is the largest manually annotated corpus of this kind and amounts to 226,040 sense annotations.

BulSemCor (Koeva et al., 2006, 2011) is designed according to the general methodology of the original SemCor and criteria for ensuring an appropriate coverage of contemporary general lexis. In addition to open-class words, BulSemCor includes annotation of closed-class words such as preposition, conjunctions, particles, etc.; for that purpose the Bulgarian WordNet has been expanded with closed-class words (Koeva et al., 2011). The size of the corpus is close to 100,000 annotated units.

In addition, we employ parallel resources to extract bilingual examples that would be annotated and analysed in juxtaposition. In particular, we use the Bulgarian-English Sentence- and Clause-Aligned Corpus (**BulEnAC**)[6], a parallel corpus aligned at sentence- and clause level and containing annotations of the syntactic relations between the pairs of clauses and the lexical or other elements realising this relation (conjunctions, complementisers, punctuation). The corpus contains 366,865 tokens altogether – 176,397 tokens in Bulgarian and 190,468 tokens in English (Koeva et al., 2012a). BulEnAC is particularly suitable for both mono- and bilingual semantic annotation tasks as it provides aligned translation equivalents at sentence- and clause level, i.e. the context in which a predicate's semantic and argument structure is realised.

When the above corpora do not provide sufficient data, we could supplement the dataset with examples from the **Bulgarian National Corpus**, which consists of a monolingual (Bulgarian) part and 47 parallel corpora. The Bulgarian part amounts to 1.2 billion words of running text distributed in 240,000 samples, which reflect the language predominantly in its written modality from the mid-20th century (1945) until the present day (Koeva et al., 2012b).

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

4 Selection and annotation of examples

Below we outline the steps involved in the selection and annotation of examples.

Selection of verbs and verb senses. We focus on verbs expressing self-induced non-directed translational motion, in particular verbs that evoke the FrameNet frame Self_motion and their counterparts in WordNet.

In total, the class of motion verbs in WordNet covers 1,463 synsets. Out of this number, we have identified 248 verb synsets representing the subclass of self-motion evoking the Self_motion frame. There are 140 synsets assigned the Self_motion frame in the Bulgarian WordNet, including 6 language-specific synsets with no counterpart in English.

Automatic collection of examples from corpora. For each literal from the selected synset inventory, we perform automatic collection of usage examples in English and Bulgarian from the corpora described in Section 3.4.

We start by extracting sentences from SemCor and BulSemCor as the verbs in these corpora are assigned WordNet senses and can be used for the annotation task in a straightforward manner. As a result, we obtained 824 examples in English and 186 in Bulgarian.

In order to increase the number of examples and to provide more representative data in terms of the valence patterns covered and the variation in the syntactic expression of the frame elements, we supplement the collection of examples with ones from the Bulgarian National Corpus and the Bulgarian-English Clause-aligned Corpus. As these two resources are not word sense disambiguated, we apply additional manual filtering to make sure that the automatically collected sentences contain at least one of the verb senses selected for the study. As a result of this procedure, we were able to increase the data by 745 parallel Bulgarian–English sentence pairs. The bilingual examples are especially valuable as they allow for a direct comparison between the ways of expressing similar or equivalent linguistic content in the two languages.

The examples in both languages are POS-tagged, morphosyntactically annotated and lemmatised.

Assignment of valence patterns to English and Bulgarian synsets. FrameNet describes the semantic and syntactic properties of lexical units evoking a given semantic frame in terms of valence patterns: co-occurring combinations of frame elements attested in the FrameNet corpus, i.e. the actual realisations of a lexical unit in context. As these patterns are derived from the annotated data, they may not be exhaustive in the sense that they may not cover all the possible combinations of frame elements and different syntactic realisations, or may not be the most representative ones (i.e. the most frequent ones found in the language).

Semantic frames are relatively universal and language-independent by design as they are grounded in human cognition and experience. This assumption, while not explored here (but cf. (Boas, 2020)), has been implicitly taken for granted in previous and ongoing work, thus providing the motivation for adopting the FrameNet methodology in the creation of framenets for a number of typologically diverse languages where their cross-lingual application has been tested empirically in a satisfactory way (Tiago Torrent and Matos, 2018). Our own experience with Bulgarian has shown that the frames are comprehensive enough to enable a detailed description of the Bulgarian lexical units studied so far, and sufficiently general to allow for further refinements, if needed. As valence patterns describe the combinations of co-occurring frame elements in actual data, they are also quite applicable across languages. The observed variations in the attested configurations cross-linguistically or among same-language verbs may point to important contrasts and are thus all the more interesting to study.

The greatest differences are found at the level of syntactic expression as different languages have different inventories of grammatical and lexical devices. While being more language-specific, syntactic expression in one language may also be used as a point of departure for analysis and comparison in another language (at least in the case of English and Bulgarian), especially in the scenario where annotated data are lacking or scarce as is the case for Bulgarian. We have thus started with the syntactic descriptions provided for English through the FrameNet system of frames and annotated examples and have confirmed, rejected, modified or elaborated on them if necessary.

In certain cases, the original patterns attested in FrameNet have been generalised in order to match the Bulgarian data. For example, patterns involving finite and non-finite clauses have been clustered together and labelled as `Clause` to account for the fact that Bulgarian lacks non-finite clauses and such clauses will have as counterparts finite clauses or will be rendered in another way. Prepositional phrases realising the same frame element with PPs headed by different prepositions (e.g. PP[of], PP[from] when used to introduce the frame element COMPONENTS in the frame `Building`) have also been grouped together.

Particular attention is paid to examples which cannot be matched to any available pattern as this might signal that the respective pattern is specific to Bulgarian.

Below we illustrate some of the patterns attested for the lexical units evoking the frame Self_motion as identified in the FrameNet data. For the sake of easier understanding, we give only English examples adapted from the FrameNet corpus.

[NP.Ext]$_{\text{SELF\_MOVER}}$ [PP]$_{\text{PATH}}$
[She]$_{\text{SELF\_MOVER}}$ **walked** [along the beach]$_{\text{PATH}}$.

[NP.Ext]$_{\text{SELF\_MOVER}}$ [PP]$_{\text{AREA}}$
[He]$_{\text{SELF\_MOVER}}$ **ran** [about the room]$_{\text{AREA}}$.

[NP.Ext]$_{\text{SELF\_MOVER}}$ [PP]$_{\text{GOAL}}$
[They]$_{\text{SELF\_MOVER}}$ **walked** [to the entrance]$_{\text{GOAL}}$

[NP.Ext]$_{\text{SELF\_MOVER}}$ [AdvP]$_{\text{AREA}}$
[Pelicans]$_{\text{SELF\_MOVER}}$ **were flying** [about]$_{\text{AREA}}$.

[NP.Ext]$_{\text{SELF\_MOVER}}$ [AdvP]$_{\text{GOAL}}$
[The boy]$_{\text{SELF\_MOVER}}$ **sneaked** [home]$_{\text{GOAL}}$.

[NP.Ext]$_{\text{SELF\_MOVER}}$ [AdvP]$_{\text{MANNER}}$ [PP]$_{\text{PATH}}$
[The two guards]$_{\text{SELF\_MOVER}}$ **were strolling** [leisurely]$_{\text{MANNER}}$ [around the fence]$_{\text{PATH}}$.

[NP.Ext]$_{\text{SELF\_MOVER}}$ [PP]$_{\text{SOURCE}}$ [PP]$_{\text{GOAL}}$
[The toddler]$_{\text{SELF\_MOVER}}$ **jumped** [from the boulder]$_{\text{SOURCE}}$ [into the shallow water]$_{\text{GOAL}}$.

[NP.Ext]$_{\text{SELF\_MOVER}}$ [PP]$_{\text{PATH}}$ [PP]$_{\text{GOAL}}$
[Jenny]$_{\text{SELF\_MOVER}}$ **dashed** [down the bank]$_{\text{PATH}}$ [to the river]$_{\text{GOAL}}$.

[NP.Ext]$_{\text{SELF\_MOVER}}$ [AdvP]$_{\text{MANNER}}$ [AdvP]$_{\text{AREA}}$
[The men]$_{\text{SELF\_MOVER}}$ **danced** [merrily]$_{\text{MANNER}}$ [around]$_{\text{AREA}}$.

Annotation of the frame elements. At this stage, the annotation of frame elements is performed predominantly manually in order to ensure better precision and analysis. For a part of the syntactic components, more specifically the the subject, some preliminary annotation has been performed automatically, followed by manual post-editing.

We have adopted the Berkeley FrameNet approach to annotation. The process consists of the identification and labelling of the syntactic constituents that realise each frame element. Hence, the projection of frame elements into syntactic positions is implemented in a straightforward manner by associating each frame element with a syntactic category that may be further specified for its grammatical function – specifically for subject (NP.Ext) and object (NP.Obj) phrases. Object and adverbial PPs are not explicitly distinguished, but this information is recoverable from the semantics of the respective frame element; for instance, PLACE, TIME, SPEED, FREQUENCY, etc. are adverbial PPs, while other frame elements qualify as prepositional objects. This declarative linking enables the direct observation of the syntactic properties and behaviour of lexical units.

The aggregation of the examples annotated for each target Lexical Unit provides empirical data about the attested valence patterns in terms of the combinations of overtly expressed frame elements (and possibly non-overt elements understood from the context, see next paragraph) and the specific ways in which they are realised syntactically.

An important feature of the FrameNet methodology and by extension of the annotation adopted in our corpus, is the labelling of syntactically non-overt but semantically obligatory frame elements, the so-called null instantiations (NIs) cf. Ruppenhofer et al. (2016, 28–30). Null instantiations have different status depending on whether the referent of the respective frame element is retrievable from the previous context. A definite null instantiation (DNI) stands for a non-expressed frame element that has a definite reference, e.g. the non-overt subject in the case of pro-drop languages such as Bulgarian. An indefinite null instantiation (INI) is observed where a frame element represents a generalised non-specific entity understood from the broader context by virtue of some convention or habitual interpretation: for instance, the frame element INGESTIBLE in the following sentence is not expressed, but is understood to be some kind of food or meal: *[She]*$_{\textit{INGESTOR}}$ ***ate*** *[hastily]*$_{\textit{MANNER}}$ *[_]*$_{\textit{INGESTIBLE:INI}}$. A constructional null instantiation (CNI) is observed when the lexical omission is licensed by the grammatical construction in which the frame element is found, e.g. the subject of an imperative sentence in both Bulgarian and English.

In such a way the annotated data provide information about the regularities and dependencies

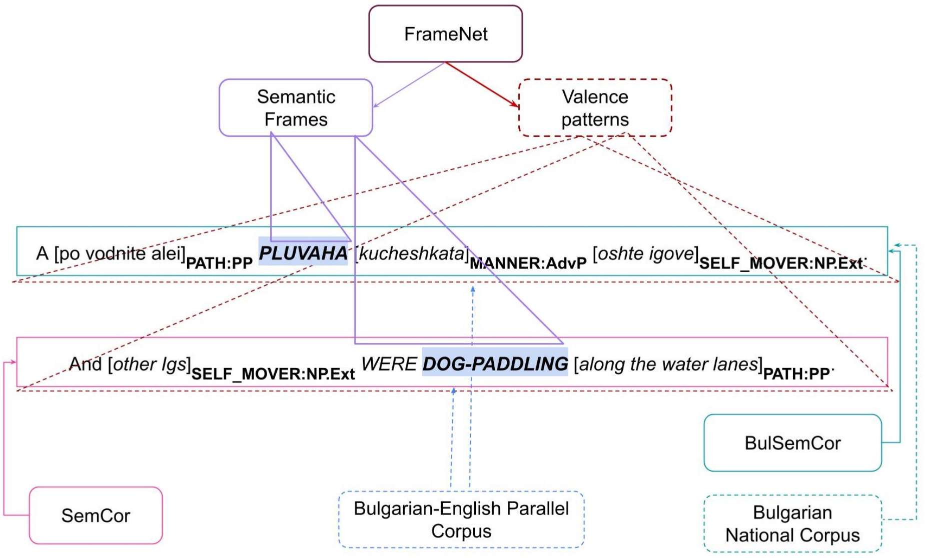

FrameNet
Semantic Frames
Valence patterns
A [po vodnite alei]PATH:PP PLUVAHA [kucheshkata]MANNER:AdvP [oshte igove]SELF_MOVER:NP.Ext.
And [other Igs]SELF_MOVER:NP.Ext WERE DOG-PADDLING [along the water lanes]PATH:PP.
BulSemCor
SemCor
Bulgarian-English Parallel Corpus
Bulgarian National Corpus

Figure 1: Interaction between the various resources

existing with respect to the co-occurrence or, oppositely, the competition between certain frame elements, including the possibility to leave some of them implicit under certain conditions.

Figure 1 summarises the interaction among the different resources and the information encoded in them. The dotted lines stand for data that need to be validated manually (examples from corpora lacking sense annotation and requiring additional filtering), while the solid lines denote verified examples, such as sentences extracted from sense-annotated corpora. Solid lines are also used to mark semantic frames since they provide relatively universal descriptions, while dotted lines are reserved for valence patterns (and syntactic realisations), which need to be verified for each language and for each verb individually.

The Self_motion frame. Self_motion describes the self-induced and self-controlled motion of an entity along a trajectory. More precisely, a Self_mover, a living being (or by metaphorical extension a self-directed entity such as a vehicle) moves under its own direction along a Path – any trajectory of motion confined between a starting point, the Source, and an end point, the Goal. An Area covered may be mentioned when the motion does not occur along a single linear trajectory, as can be the Direction – i.e. the general spatial orientation of the motion from the deictic centre towards a (possibly implicit) reference point.

Not all core-elements need to be expressed simultaneously. In particular, due to the fact that Direction, Goal, Path and Source define a linear trajectory together, they form a single coreness frame element set (Ruppenhofer et al., 2016, 25–26), meaning that it is usually sufficient to realise only one of them in order to satisfy the semantic valence of the verb; in other terms, each of them on its own evokes the entire notion of motion along a trajectory. In addition, as the Area defines motion that cannot be described along a single coherent route, it can only be expressed when none of the frame elements in the above core set is realised.

In addition to the core frame elements, a number of others may also be expressed, specifying various circumstances or aspects of the situation, such as Time, Duration, Speed, Manner, Place, etc.

We annotate both core and non-core frame elements. Examples 1–3 include annotated sentences along with the patterns to which they are matched. Note that Example 1 illustrates a mismatch in the patterns for each language since the English verb *dog-paddle* incorporates a component describing the manner of swimming, while the Bulgarian sentence employs a Manner frame element realised as an adverbial – *kucheshkata* (doggy-style) – to render the same meaning in conjunction with the verb *pluvam* (swim). In fact, this would be the conventional way of translating the English verb as it does not have a straightforward equivalent in Bulgarian. Example 3 illustrates a difference in the syntactic expression of the frame element Direction in the two languages, resulting in different valence patterns, although consisting of the same configuration of frame elements.

Example 1. FrameNet frame: `Self_motion`

BG: [NP.Ext]$_{\text{SELF\_MOVER}}$ [AdvP]$_{\text{MANNER}}$ [PP]$_{\text{PATH}}$

EN: [NP.Ext]$_{\text{SELF\_MOVER}}$ [PP]$_{\text{PATH}}$

A [*po vodnite alei*]$_{\text{PATH:PP}}$ ***pluvaha*** [*kucheshkata*]$_{\text{MANNER:AdvP}}$ [*oshte igove*]$_{\text{SELF\_MOVER:NP.Ext}}$.

And [*other Igs*]$_{\text{SELF\_MOVER:NP.Ext}}$ ***were dog-paddling*** [*along the water lanes*]$_{\text{PATH:PP}}$.

Example 2. FrameNet frame: `Self_motion`

[NP.Ext]$_{\text{SELF\_MOVER}}$ [PP]$_{\text{PATH}}$

[_]$_{\text{SELF\_MOVER:DNI:NP.Ext}}$ ***varvyahme*** [*v neshto kato ledena zala*]$_{\text{PATH:PP}}$.

[*We*]$_{\text{SELF\_MOVER:NP.Ext}}$ ***were walking*** [*through a kind of ice hall*]$_{\text{PATH:PP}}$.

Example 3. FrameNet frame: `Self_motion`

BG: [NP.Ext]$_{\text{SELF\_MOVER}}$ [PP]$_{\text{DIRECTION}}$ [PP]$_{\text{PATH}}$

EN: [NP.Ext]$_{\text{SELF\_MOVER}}$ [AdvP]$_{\text{DIRECTION}}$ [PP]$_{\text{PATH}}$

[*Toy*]$_{\text{SELF\_MOVER:NP.Ext}}$ ***tichashe*** [*na zapad*]$_{\text{DIRECTION:PP}}$ [*prez lozyata*]$_{\text{PATH:PP}}$.

[*He*]$_{\text{SELF\_MOVER:NP.Ext}}$ ***was running*** [*west*]$_{\text{DIRECTION:PP}}$ [*through the vineyards*]$_{\text{PATH:PP}}$.

5 Results

While in this paper we focus on self-motion verbs, the principles and methodology adopted here are applied to the description of verbs belonging to other semantic classes as well.

There are two principle results from our work: (i) a corpus of examples illustrating the use of a given class of verbs in Bulgarian annotated according to the methodology proposed by the Berkeley FrameNet project, in which some of the sentences are paired with their annotated English counterparts if the examples are extracted from the Bulgarian-English Sentence- and Clause-Aligned Corpus; (ii) a collection of verb synsets from the Princeton WordNet and the Bulgarian WordNet aligned with a number of FrameNet frames relevant for the studied class of verbs and the semantic and syntactic information that can be derived from the frame's description and the annotated examples.

More specifically, for each verb that has a counterpart in FrameNet, we list the patterns attested in the FrameNet Corpus that meet several criteria: appear in three or more examples; contain at least one core frame element; appear in their canonical form (and not in alternations, e.g. preferring active-voice rather than passive-voice examples).

For the verbs in WordNet which are assigned a given frame but do not have a correspondence denoting a (near-)equivalent sense in FrameNet, we assign the aggregate of valence patterns attested for all the verbs evoking the relevant frame. As part of them may not be relevant for the particular verb, the need for providing examples confirming the valence patterns is tantamount.

As a result each of the verbs in the studied inventory is supplied with a list of valence patterns. While for the verbs in the Princeton WordNet the patterns are confirmed by the FrameNet corpus examples, they are not necessarily valid for the equivalent Bulgarian verbs and need to be validated against corpus evidence.

At the next stage, for each annotated sentence in our corpus we extract the configuration of frame elements in order to identify the valence pattern realised in it and match the pattern to the identical one in the FrameNet frame. The patterns confirmed by examples are marked in bold.

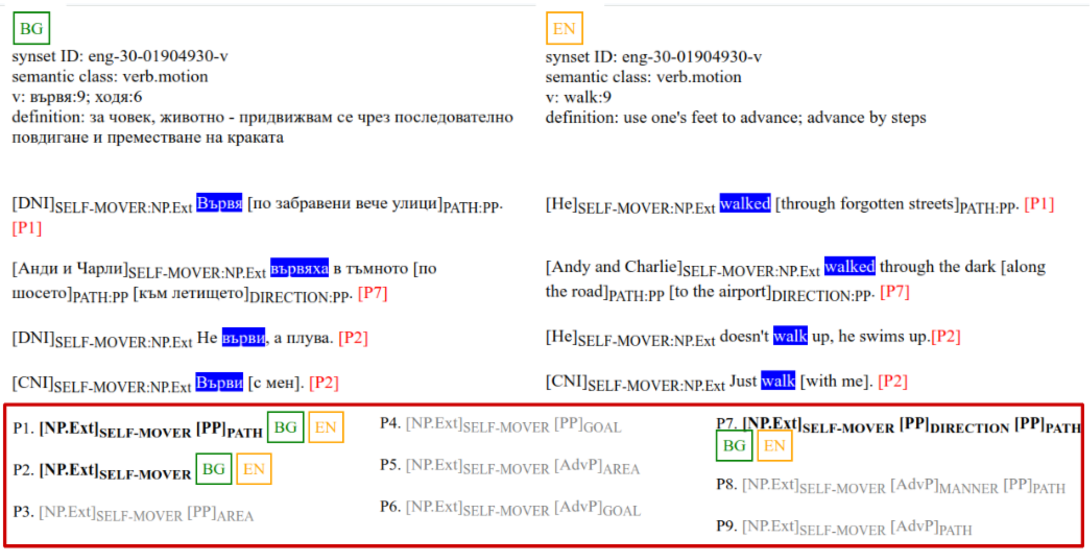
As some patterns are more frequent than others, the annotated examples would help to obtain a more comprehensive and accurate picture of the combinatorial properties of verbs and the typical syntactic realisation of their frame elements. While we focus on Bulgarian and English, the valence patterns should be applicable to other languages.

Language	EN	BG	Aligned
# Verbs	65	32	26
# WordNet Synsets	31	16	15
# Valence patterns	40	32	30
# Sentences	254	228	50
# Annotated FEs	541	508	–

Table 1: Distribution of annotated examples for self-motion verbs.

Table 1 shows the distribution of the annotated examples across synsets and patterns. The English dataset covers 254 fully annotated examples of self-motion verbs, while the Bulgarian dataset contains 228 examples. The total number includes 50 parallel pairs of sentences.

The self-motion subset is part of a larger corpus of annotated examples of verbs in WordNet, which covers several semantic classes involving motion and includes so far over 1,200 examples for



BG

synset ID: eng-30-01904930-v
semantic class: verb.motion
v: вървя:9; ходя:6
definition: за човек, животно - придвижвам се чрез последователно повдигане и преместване на краката

[DNI]$_{\text{SELF-MOVER:NP.Ext}}$ Вървя [по забравени вече улици]$_{\text{PATH:PP}}$. [P1]

[Анди и Чарли]$_{\text{SELF-MOVER:NP.Ext}}$ вървяха в тъмното [по шосето]$_{\text{PATH:PP}}$ [към летището]$_{\text{DIRECTION:PP}}$. [P7]

[DNI]$_{\text{SELF-MOVER:NP.Ext}}$ Не върви, а плува. [P2]

[CNI]$_{\text{SELF-MOVER:NP.Ext}}$ Върви [с мен]. [P2]

EN

synset ID: eng-30-01904930-v
semantic class: verb.motion
v: walk:9
definition: use one's feet to advance; advance by steps

[He]$_{\text{SELF-MOVER:NP.Ext}}$ walked [through forgotten streets]$_{\text{PATH:PP}}$. [P1]

[Andy and Charlie]$_{\text{SELF-MOVER:NP.Ext}}$ walked through the dark [along the road]$_{\text{PATH:PP}}$ [to the airport]$_{\text{DIRECTION:PP}}$. [P7]

[He]$_{\text{SELF-MOVER:NP.Ext}}$ doesn't walk up, he swims up.[P2]

[CNI]$_{\text{SELF-MOVER:NP.Ext}}$ Just walk [with me]. [P2]

P1. **[NP.Ext]$_{\text{SELF-MOVER}}$ [PP]$_{\text{PATH}}$** BG EN

P2. **[NP.Ext]$_{\text{SELF-MOVER}}$** BG EN

P3. [NP.Ext]$_{\text{SELF-MOVER}}$ [PP]$_{\text{AREA}}$

P4. [NP.Ext]$_{\text{SELF-MOVER}}$ [PP]$_{\text{GOAL}}$

P5. [NP.Ext]$_{\text{SELF-MOVER}}$ [AdvP]$_{\text{AREA}}$

P6. [NP.Ext]$_{\text{SELF-MOVER}}$ [AdvP]$_{\text{GOAL}}$

P7. **[NP.Ext]$_{\text{SELF-MOVER}}$ [PP]$_{\text{DIRECTION}}$ [PP]$_{\text{PATH}}$** BG EN

P8. [NP.Ext]$_{\text{SELF-MOVER}}$ [AdvP]$_{\text{MANNER}}$ [PP]$_{\text{PATH}}$

P9. [NP.Ext]$_{\text{SELF-MOVER}}$ [AdvP]$_{\text{PATH}}$

Figure 2: Visualisation of annotated examples for the verbs in a synset in Bulgarian and English.

Bulgarian and over 1,500 examples in English.

A possible application of the corpus is the cross-lingual analysis aiming to match the pairs of literals within corresponding synsets for Bulgarian and English that exhibit the same set of valence configurations and syntactic patterns. On this basis we can identify closer translational pairs as opposed to verbs that share only part of the valence patterns or differ significantly in their syntactic realisation despite their similar meaning.

In the current version of the dataset, all patterns attested for Bulgarian are matched to patterns attested in FrameNet. Moreover, the data show a considerably low degree of variation in terms of the syntactic realisation of the frame elements in the two languages.

The Bulgarian dataset needs to be further extended to provide sufficient data that would enable us to make reliable conclusions on the pattern correspondences. This includes the annotation of examples exhibiting language-specific syntactic patterns that are not found (or are rare) in English.

6 Conclusions and future work

The compiled dataset of annotated examples is part of an ongoing effort on the semantic, syntactic and aspectual analysis of several large semantic classes of verbs – verbs of motion, verbs of communication, and verbs of change. Our next task will be to expand the data further by covering more verbs, verb classes and peculiarities of the semantics and syntactic behaviour of the studied predicates. Extending the scope of annotation systematically beyond core frame elements is also a research venue to be pursued, especially as the expression of some frame elements such as GOAL, MANNER or PURPOSE, among others, may be correlated to changes in the aspectual interpretation of verbs.

The empirical data enable the study of the two languages under discussion individually, as well in comparative or contrastive terms. The linking of the annotated examples to lexical resources such as WordNet and FrameNet facilitates the applicability of the corpus for various research tasks.

The dataset can be employed in the training of semantic role labelling, semantic disambiguation, syntactic pattern analysis, as well as in extracting parallel valence patterns, translation equivalents of verb phrases, etc. The proposed approach can also be extended to other languages, in particular to ones that have their own wordnets linked to PWN, thus resulting in the creation of a multilingual and more universally applicable resource.

The resources created as part of this work are made available to the community under the Creative Commons Attribution 4.0 International license.[7]

Acknowledgements

This research is carried out as part of the project *An Ontology of Activity Predicates – Linguistic Modelling with a Focus on Bulgarian* funded by the Bulgarian National Science Fund, Grant Agreement No KP–06–N80/9 from 8.12.2023.

[7] https://dcl.bas.bg/corpus-data-semantic-frames-2024/

References

C. F. Baker and C. Fellbaum. 2009. WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP '09), Association for Computational Linguistics, Stroudsburg, PA, USA*, pages 125–129.

Collin F. Baker. 2008. FrameNetPresent and Future. In *The First International Conference on Global Interoperability for Language Resources*, Hong Kong. City University, City University.

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.

Hans C. Boas. 2020. *A roadmap towards determining the universal status of semantic frames*, pages 21–52. De Gruyter Mouton, Berlin, Boston.

Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A WordNet detour to FrameNet. In *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, volume 8 of *Computer Studies in Language and Speech*. Lang, Frankfurt, Germany.

Maddalen Lopez de Lacalle, Egoitz Laparra, and German Rigau. 2014. Predicate Matrix: extending SemLink through WordNet mappings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 903–909, Reykjavik, Iceland. European Language Resources Association (ELRA).

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China. Association for Computational Linguistics.

C. Fellbaum. 1999. The Organization of Verbs and Verb Concepts in a Semantic Net. In P. Saint-Dizier, editor, *Predicative Forms in Natural Language and in Lexical Knowledge Bases*, volume 6 of *Text, Speech and Language Technology*, pages 93 – 110. Springer, Dordrecht.

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

Luca Gilardi and Collin F. Baker. 2018. Learning to Align across Languages: Toward Multilingual FrameNet. In *Proceedings of the International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons*, pages 13–22.

Ales Horak, Piek Vossen, and Adam Rambousek. 2008. The Development of a Complex-Structured Lexicon based on WordNet. In *Proceedings of the Fourth International Global WordNet Conference (GWC 2008), Szeged*, pages 200–208, Szeged, Hungary. University of Szeged, Department of Informatics.

Christopher R. Johnson, Charles J. Fillmore, Esther J. Wood, Margaret Urban, Miriam R. L. Petruck, Collin F. Baker, and et al. Charles J. Fillmore. 2001. The FrameNet Project: Tools for Lexicon Building. `https://citeseerx.ist.psu.edu/pdf/0ece390b6f4e6b38c5733248992ff73f846d91aa`.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon. PhD Thesis*. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.

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

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, Svetlozara Leseva, and Maria Todorova. 2006. Bulgarian sense tagged corpus. In *Proceedings of LREC 2006*, pages 79 – 86.

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.

Shari Landes, Claudia Leacock, and R. Tengi. 1998. Building Semantic Concordances. In *WordNet: An Electronic Lexical Database*.

Svetlozara Leseva and Ivelina Stoyanova. 2019. Enhancing conceptual description through resource linking and exploration of semantic relations. In *Proceedings of 10th Global WordNet Conference, 23 – 27 July 2019, Wroclaw, Poland*, pages 229–238.

Svetlozara Leseva and Ivelina Stoyanova. 2020. Beyond lexical and semantic resources: linking WordNet with FrameNet and enhancing synsets with conceptual frames. In *Towards a Semantic Network Enriched with a Variety of Semantic Relations*. Prof. Marin Drinov Academic Publishing House of the Bulgarian Academy of Sciences.

Markéta Lopatková, Václava Kettnerová, Eduard Bejček, Anna Vernerová, and Zdeněk Žabokrtský. 2016. *Valenční slovník českých sloves VALLEX*. Karolinum, Praha.

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

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a Semantic Concordance for Sense Identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

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*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Martha Palmer. 2009. Semlink: linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*. 9–15.

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 *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: combining FrameNet, VerbNet and WordNet for robust semantic parsing. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science*, volume 3406. Springe, Berlin, Heidelbergr.

Collin Baker Tiago Torrent, Michael Ellsworth and Ely Matos. 2018. The multilingual framenet shared annotation task: a preliminary report. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

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*.

Tiago T. Torrent, Collin F. Baker, Oliver Czulo, Kyoko Ohara, and Miriam R. L. Petruck, editors. 2020. *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*. European Language Resources Association, Marseille, France.

Zdenka Urešová, Eva Fucíková, Eva Hajičová, and Jan Hajič. 2020a. SynSemClass Linked Lexicon: Mapping Synonymy between Languages. In *Proceedings of the Globalex Workshop on Linked Lexicography, Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020*, pages 10 – 19.

Zdenka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2020b. Syntactic-Semantic Classes of Context-Sensitive Synonyms Based on a Bilingual Corpus. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 242–255. Springer International Publishing.