Chapter 4

Representation of multiword expressions in the Bulgarian integrated lexicon for language technology

Petya Osenova[a] & Kiril Simov[a]

[a]Institute of Information and Communication Technologies, Bulgarian Academy of Sciences

The chapter introduces a representation model of multiword expressions from the perspective of integrated lexicons for Bulgarian. The lexicons considered are an inflectional one, a valency one, and a wordnet. We created a joint representation entry that incorporates morphology, valency potential and lexical semantics through synonym sets. The selected mechanism for displaying all the information is catena-based since the catena allows for better modeling of idiosyncratic elements and is tree-based. Also, a general typology of multiword expressions is proposed that focuses on fixedness and (dis)continuity. We believe that providing a unified representation of multiword expressions and common lexica would improve the performance of the various natural language processing applications.

1 Introduction

This paper is based on our previous investigations on multiword expressions (MWEs) for Bulgarian (Simov & Osenova 2015a, Laskova et al. 2019). This previous research was motivated by the investigation of the most adequate representations of MWEs in treebanks, in syntax-aware lexicons like the valency ones and in lexical bases like wordnets.

Having already developed a number of language resources for Bulgarian, our current goal is to integrate them in such a way that they would allow a joint

Petya Osenova & Kiril Simov. 2024. Representation of multiword expressions in the Bulgarian integrated lexicon for language technology. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 117–146. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998637

approach to several NLP (natural language processing) tasks, including end-to-end training of neural network models.

In order to achieve this goal, we have already integrated the Bulgarian treebank (BTB) with sense annotations from the Bulgarian wordnet (BTB-WN), Bulgarian DBpedia, Bulgarian Wikipedia, Bulgarian Valency Lexicon, and a newly created small FrameNet-oriented lexicon for event annotation in the area of Digital Humanities. With respect to the integrated lexical and text resources, one of the problems is the common representation of the lemmas in the various types of lexicons, especially the representation of MWEs. Thus, one of the important requirements is that lemmas have a common representation in both – the annotated corpora and the integrated lexical resources. However, other issues appear here: what the lemma of a MWE is; how to present the syntactic potential in a lexical database including the points of flexibility and external participants; and how to map the lexical representation to the one in a corpus.

In this paper, we focus on the representation of MWEs in the framework of integrated lexical resources. In relation to that our contributions are as follows:

1. introducing the structure of the MWE lexical entry;
2. tuning the catena-based formalization to the complex structure of integrated linguistic information;
3. modeling the complexity of the entry with respect to discontinuity and fixedness.

The paper is organized as follows: in §2 related work is discussed. §3 introduces the background of our model. §4 introduces the formal definition of catena. §5 presents a model of the lexical entry. §6 suggests analyses of the specific MWE types. §7 concludes the paper.

2 Related work

The representation of MWEs in lexicons with a view to their adequate annotation in corpora has been a hot topic for quite some time. For example, Lichte et al. (2019) discuss various approaches to lexical encoding of MWEs with respect to the NLP tasks. The authors favor flexible formats like PATRII and XMG over the fixed encoding formats of a Dutch Electronic Lexicon of Multiword Expressions (Grégoire 2010), and a Polish Valency Lexicon (Przepiórkowski et al. 2014). Our current approach is somewhere between the fixed and flexible encodings. On the one hand, it uses property name sets where the main morphosyntactic, syntactic, and semantic characteristics of the MWE are given. At the same time, the notion of catena is used, which introduces a graph representation and thus falls into the tree-based approaches to MWEs. In this way, the catena ensures the flexibility of the encoding with respect to potential discontinuity or other specifics. Our approach is head-based rather than construction-based.

Dyvik et al. (2019) present the encoding of MWEs in the resource grammar NorGram which is based on the Lexical-Functional Grammar (LFG) framework. There the fixed MWEs are treated as words. For the flexible MWEs another approach is taken – namely, following the grammar apparatus of LFG, the components are presented through selection frames with a subcategorization in case of verbs and complements, and with equations for the other lexically restricted dependants – all these with their specifics. In this paper, the approach is lexico-syntactic since the representation of the MWEs combines both – the morphosyntactic and lexical specifics. Thus, through the theory mechanisms, the balance between grammar and lexicon is pertained. Our approach aims to ensure exactly such a dynamic relation between a lexicon and a grammar without the availability of a well-developed computational grammar.

Masini (2019) introduces three criteria for classifying MWEs: "(i) formal properties (degree of internal cohesion or fixity), (ii) idiomatic status […], and (iii) function, or a combination of these". In our proposed approach we focus mainly on (i) under which we also include (ii). Then we are more interested in the challenges when modeling word order than in the function of the MWE per se (see §5).

There are attempts for MWE representation in dictionaries and databases for both – humans and machines, i.e. reflecting multipurpose and multilevel aspects. For example, Vondřička (2019) uses slots for the syntagmatic information and fillers for the paradigmatic one in the entry. The author relies on the tree representation in dependency and constituency formats with the accompanying challenges. The problems come from the notion of the word and ways of spelling as well as from the not straightforward modeling of the internal elements in a MWE. In Skoumalová et al. (2024 [this volume]) the linking is described of the lexical entries in a MWE lexicon for Czech with their natural occurrences in a corpus. The relation between the lexicon and the corpus has been ensured in both directions. We aim at such an integrated resource and workflow. However, at the moment we provide a link of a MWE to its corpus occurrence only through the headwords of MWEs.

In Lion-Bouton et al. (2023) the authors propose an approach according to which the MWE identification tools consult lexicons. For this purpose, a survey has been performed on quantitative evaluation of some MWE lexicon formalisms based on the notion of observational adequacy. The suggested approach based on a generalisation of the concept of a Coarse Syntactic Structure proves to be competitive with lexicons based on a sequential representation of MWEs. Our approach is also graph/tree-based but we aim to accommodate as much information as possible in the same representation – lexical from wordnets, valency from valency dictionaries, knowledge-based from Wikipedia, etc.

Zampieri et al. (2019) show the impact of the MWE representation in the input pre-processed data as well as in two types of word embeddings (word2vec and FastText) for the task of MWE identification. They conclude that the lemma plays a positive role for all considered languages – Basque, French, and Polish. For us the most interesting part in relation to our work is the fact that the richer the information for a morphologically rich language, the better the results. We also try to represent as much integrated information about a MWE as possible.

Schneider et al. (2014) report on the annotation of MWEs in a social web corpus. They use an annotation scheme that respects the following aspects: heterogeneity (where the annotated MWEs are not restricted by syntactic construction); shallow but gappy grouping (MWEs viewed as simple groupings of tokens, which need not be contiguous in the sentence); and expression strength (where the most idiomatic MWEs are distinguished from and can belong to weaker collocations). For our work the most important focus (along the others) is the modeling of gapping, i.e. discontinuity. Authors indicate that 15% of MWEs contain at least one gap. We have to take into account that this fact is given for English as a language with a rather fixed word order. In languages like Bulgarian that have a relatively free word order, discontinuity is expected to be much higher. For that reason we are trying to find a way to model the predicted points of discontinuity within the lexical entry.

In Leseva et al. (2024 [this volume]) an elaborate bilingual model of MWEs representation is described for Bulgarian and Romanian in a uniform way. Wordnets for the two languages have been used for linking the bilingual lexicons. The focus is put on the verbal MWEs where the relations from the Universal Dependencies (UD) have been used. We also use a wordnet for Bulgarian (BTB-WN) as a linking module and UD as modeling relations within MWEs.

In the PARSEME initiative verbal MWE (VMWE) annotations, both continuous and discontinuous groups are considered (Savary et al. 2018). The annotation strategy includes the lexicalized elements, not their variations. It views the representation as a syntactic tree. However, the scheme describes also the properties for each type and provides specialized guides for each participating language, including Bulgarian. In addition to the two universal VMWE categories (light verb constructions with two subtypes and verbal idioms), our language has inherently reflexive verbs (IRV) but not verb-particle constructions (VPC). Since our task here is to show how we represent all the main types of MWEs, we focus on the variety and complexity of their modeling.

3 Background

Our work on MWEs up to now has been centred around the notion of catena. Catena (chain) was initially introduced in O'Grady (1998) as a mechanism for representing the syntactic structure of idioms. He showed that for this task a definition of syntactic patterns was needed that does not coincide with constituents. He defined the catena in the following way: "The words A, B, and C (order irrelevant) form a chain if and only if A immediately dominates B and C, or if and only if A immediately dominates B and B immediately dominates C". Some examples of catena from a dependency syntactic tree are presented in Figure 1. In our work here we convert MWEs into a representation previously defined in Simov & Osenova (2014) and in Simov & Osenova (2015b) in which the catena is depicted as a dependency tree fragment with appropriate grammatical and semantic information. The variations of the MWEs are represented through underspecification of the corresponding features, including valency frames and non-canonical basic form.

The lexical entry uses the following format: a lexicon catena (LC), semantics (SM) and valency (Frame). The lexicon catena for the MWEs is stored in its basic form. The realisation of the catena in a sentence has to obey the rules of the grammar. In this way the possible word order is managed. The semantics of a lexical entry specifies the list of elementary predicates contributed by the lexical item. When the MWE allows for some modification (including adjunction) of its elements, i.e. modifiers of a noun, the lexical entry in the lexicon needs to specify the role of these modifiers. Some first ideas in these lines are represented in the above cited works and also in Laskova et al. (2019).

We aim at an integrated and relatively flexible representation of MWE types in lexicons and their projections in corpora. We are aware that this task is not trivial and will take time. Our proposal builds on our previous modelling. Here we discuss an extended lexical entry model in order to incorporate as much linguistic information as possible. In our previous publications we already assumed that each lemma in the lexicon is represented as a catena (even when it is not a MWE). This assumption allows us to represent information in relation to analytical verb forms, to the order of the component words in the MWEs, to their morphosyntactic variations, to their syntactic and semantic behaviour, to the etymological information in cases when peculiarities of MWEs have diachronic origin. For example, in the Bulgarian expression (bg) добър вечер *dobar vecher* (lit. 'good-sg.m evening-SG.F') 'good evening', 'good' is masculine and 'evening' is feminine. The surface agreement is violated because the noun 'evening' changed its gender in contemporary language to feminine.

The model of the Valency lexicon follows our insights from the catena representation of MWEs. Such an approach allows us to introduce the integration of the necessary world knowledge to the frame elements, especially the interaction among the types of participants within a given event. Needless to say, this kind of information is not always fully compositional and the boundaries between compositional and non-compositional are not always clear. Thus, we think that the same lexicon model can be applied to the continuum from compositionality to non-compositionality in a valency-aware dictionary. We imagine that this effort will not be deterministic but incremental, since MWEs show idiosyncrasies all the time across genres, alternations, figurative meanings, etc.

Our main contribution in this paper is the structure of the lexical entry in an integrated lexicon by means of the catena notion. In the integrated resource we have included the following distinct lexicons:

Inflectional lexicon of Bulgarian (ILB): Each lemma is connected to its inflectional paradigm;

BTB Bulgarian WordNet (BTB-WN): A Bulgarian WordNet which arranges synonym sets around identical meanings. The lexical entry in BTB-WN is called SYNSET (*Synonym Set*);

Bulgarian Valency Lexicon (BVL): Complex representation of the core participants of a given event (in general sense) represented by a verb in its meaning.

The main decision we took was about the mechanism for integrating lexical entries from these three lexicons: ILB, BVL and BTB-WN. First, the initial representation of the original lexical entries is introduced. Note that we omit details that are not important for this paper. Such details, for example, include the interaction between the lexical and semantic relations in the BTB-WN.

The lexical entry of ILB includes the following main elements: *Lemma*, *Part of speech*, and *Paradigm*. The lemma is the abstract representation of the lexical entry. Each part of speech is one of the ten common parts of speech in Bulgarian (noun, adjective, numeral, adverb, pronoun, verb, preposition, conjunction, particle, interjection). For a detailed description of Bulgarian see Osenova (2010). The paradigm is a list of all the synthetic word forms related to the lemma. Bulgarian is an analytical and inflectional language. It has a rich inflectional morphology, but listing all the members of the synthetic part of the verb paradigm is still feasible, because the largest paradigm contains only 52 word forms. Each word form corresponds to a given set of grammatical features. Some word forms are analytical like part of the Bulgarian tenses. For example, the verb (bg) чета *cheta* (lit. read-1SG.PRS) 'I read' forms a future tense, second person, singular as follows: (bg) ще четеш *shte chetesh* (lit. read-2SG.FUT) 'you will read'. Such analytical word forms are formed by patterns (rules) which we consider as a part of the lexicon. They are represented using the same mechanism as the rest of the lexicon.

The Lexical entry of BTB-WN includes the following main elements: *Definition*, *Set of synonyms*, *Examples*. Each definition in BTB-WN provides a description of the meaning in Bulgarian. The set of synonyms is represented via a set of lemmas sharing the meaning of the synset. Each lemma is connected to a paradigm and a part of speech. Each example consists of one or more sentences in which the corresponding meaning is exemplified. Each example in a synset is also linked to its lemma. We usually include only one sentence, but if one sentence is not enough to disambiguate between the different meanings of the lemma, then more sentences are included. Also, the example is linked to the source from where it is taken. In this way, if necessary, we could extract more data. The current version of BTB-WN contains 53217 lemmas of which 7868 are MWEs (14.78%).

The lexical entry of BVL includes the following main elements: *Lemma*, *Definition*, *Valency frame*, and *Examples*. The lemma is the verb lemma for the lexical entry. Each definition represents a meaning of the lemma. The definition is the same as in the wordnet. The valency frame introduces a generalised representation of the core participants of the event denoted by the meaning and the syntactic behaviour of the lemma as well as by the core participants. The current version of BVL contains 6869 lemmas 1674 of which are MWEs (24.37%).

In order to integrate the lexical entries of the three lexicons we followed the following procedure:

- *Achieving a uniform representation of lemmas.* Since the three lexicons were constructed in different periods and on the basis of different machine readable sources, the lemmas of the same word could have had different representations. This holds especially for the ILB – the lexicon whose first version was created earlier.

- *Mapping of the meanings.* We have ensured that the meaning in BTB-WN and BVL are the same for the respective verbs. In this way, the verb lemmas and meanings in BTB-WN and BVL have been unified.

- *Modification of the paradigm.* Since the paradigm sometimes depends on the meaning of the lemma, the paradigm inherited from ILB had to be modified in a number of cases. For example, some nouns in some meanings are only pluralia tantum.

Thus, the lexical entry of the integrated lexicon consists of two elements: (Definition and Set of synonyms). The information about the paradigm, valency frames and examples is represented within the entry of each lemma. The record for each lemma contains also a link to its paradigm; one or more valency frames; a set of examples; and other lemma dependant classifications.

Each lemma is converted into its syntactic representation as a catena (see next section). When the lemma is a single word, the conversion to a catena is trivial. At the same time, the complexity of MWEs requires more attention to the construction of the appropriate representation. For more details see next sections. In addition to the synthetic forms, the verb paradigm contains also the analytical ones. We consider them as a special class of MWEs. The patterns for the analytical forms are represented as an addition to the main lexicon. In the lexical entry only a link to the corresponding set of patterns is given.

4 Formal definition of catena

In this section we define the formal presentation of the catena as it is used in syntax and in the lexicon. Here we follow the definition of catena provided by O'Grady (1998) and Groß (2010): a CATENA is a word or a combination of words directly connected in the dominance dimension. In reality, this definition of catena for dependency trees is equivalent to a subtree definition. Figure 1 depicts a complete dependency tree and some of its catenae. Notice that the complete tree is also a catena itself. With "root$_C$" we mark the root of the catena. It might be the same as the root of the complete tree, but also different as in the cases of "John" and "apple". Following Osborne et al. (2012) we prefer to use the notion of catena to that of dependency subtree or treelet. We aim to utilize the notion of catena for several purposes: representation of words and MWEs in the lexicon, their realization in the actual trees that present the sentence analysis, as well as for the representation of the derivational structure of compounds in the lexicon.

In order to model the variety of phenomena and characteristics encoded in a dependency grammar we extend the catena with partial arc and node labels. We follow the approach taken in CoNLL shared tasks on dependency parsing (Buchholz & Marsi 2006) representing for each node its word form, lemma, part of speech, extended part of speech, grammatical features (and later – semantics). This provides a flexible mechanism for expressing the combinatorial potential of

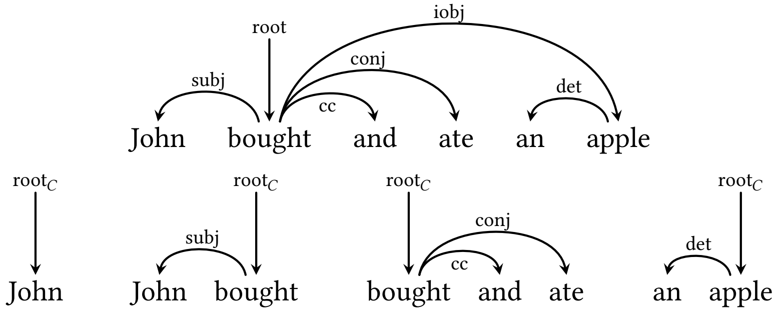


Figure 1: A complete dependency tree and some of its catenae. The complete list of catenae of the complete tree is too large to be presented here.

lexical items. In the following definition all grammatical features are represented as part-of-speech (POS) tags.[1]

Let us have the sets: LA – a set of POS tags,[2] LE – a set of lemmas, WF – a set of word forms, and a set of dependency tags D ($root \in D$). Let us have a sentence $x = w_1, ..., w_n$. A TAGGED DEPENDENCY TREE is a directed tree $T = (V, A, \pi, \lambda, \omega, \delta)$ where:

1. $V = \{0, 1, ..., n\}$ is an ordered set of nodes that corresponds to an enumeration of the words in the sentence (the root of the tree has an index 0);

2. $A \subseteq V \times V$ is a set of arcs. For each node i, $1 \leq i \leq n$, there is exactly one arc in A: $\langle i, j \rangle \in A$, $0 \leq j \leq n$, $i \neq j$. There is exactly one arc $\langle i, 0 \rangle \in A$;

3. $\pi : V - \{0\} \rightarrow$ LA is a total labelling function from nodes to POS tags.[3] π is not defined for the root;

4. $\lambda : V - \{0\} \rightarrow$ LE is a total labelling function from nodes to lemmas. λ is not defined for the root;

5. $\omega : V - \{0\} \rightarrow$ WF is a total labelling function from nodes to word forms. ω is not defined for the root;

[1] In fact, our tagset encodes all the morphosyntactic tags related to each part-of-speech, but here we use the notion of POS tag as a more common term. The tagset is described here: http://bultreebank.org/wp-content/uploads/2017/04/BTB-TR03.pdf.

[2] In the formal definitions here we use tags as entities, but in practice they are sets of grammatical features like values for gender, number, etc.

[3] In case we are interested in part of the grammatical features encoded in a POS tag we could consider π as a set of different mappings for the different grammatical features. It is easy to extend the definition in this respect, but we do not do this here.

6. $\delta : A \rightarrow D$ is a total labelling function for arcs corresponding to the dependency label. Only the arc $\langle i, 0\rangle$ is mapped to the label *root*;

7. 0 is the root of the tree.

We will hereafter refer to this structure as a parse tree for the sentence x. Node 0 does not correspond to a word form in the sentence, but plays the role of a root of the tree.

Let $T = (V, A, \pi, \lambda, \omega, \delta)$ be a tagged dependency tree. A directed tree $G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G)$ is called DEPENDENCY CATENA OF T if and only if there exists a mapping $\psi : V_G \rightarrow V$[4] such that:

1. $A_G \subseteq A$, the set of arcs of G;

2. $\pi_G \subseteq \pi$ is a partial labelling function from nodes of G to POS tags;

3. $\lambda_G \subseteq \lambda$ is a partial labelling function from nodes of G to lemmas;

4. $\omega_G \subseteq \omega$ is a partial labelling function from nodes of G to word forms;

5. $\delta_G \subseteq \delta$ is a partial labelling function for arcs of G to dependency labels.

A directed tree $G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G)$ is a DEPENDENCY CATENA if and only if there exists a dependency tree T such that G is a dependency catena of T. We mark the root catena with *root*$_C$ arc in graphical representation.

The partial functions for assigning POS tags, dependency labels, word forms and lemmas allow us to construct arbitrary abstractions over the structure of a catena. Thus, the catena could be underspecified for some of the node labels, like grammatical features, lemmas and also some dependency labels. In this way the catena could be a dependency catena of dependency trees which differ with respect to labels of different kinds. Thus, catenae are a good choice for encoding variability of lexical representation of MWEs.

Thus mapping ψ parameterizes the catena with respect to different dependency trees. Using the mapping, there is a possibility to realize different word orders of the catena nodes, for instance. The omission of node 0 from the range of the mapping ψ excludes the external root of the tagged dependency tree from each catena. The catena could be a word or an arbitrary subtree.

[4] This mapping allows for embedding of G in different tagged dependency trees and thus different word order realizations of the catena nodes (corresponding to word forms in T). The mapping ψ is specific for G and T. It allows also the image of G in T not to be a subtree of T, but several subtrees of T. A special case is discussed below – partition and extension operations.

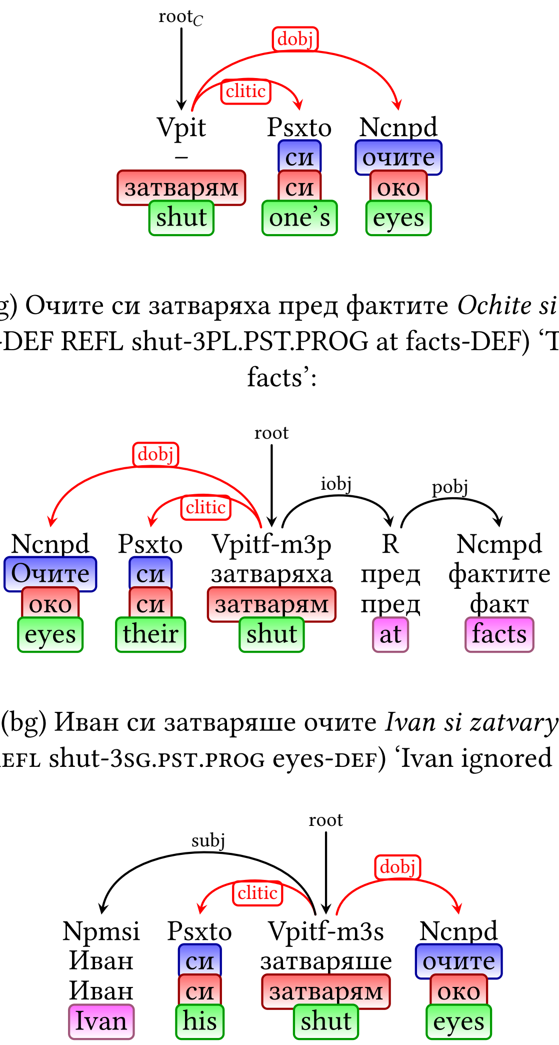
We call the mapping of a catena into a given dependency tree the REALIZATION OF THE CATENA IN THE TREE. We consider the realization of the catena as a fully specified subtree including all node and arc labels. For example, the catena for "to spill the beans" will allow for any realization of the verb form like in: "they spilled the beans" and "he spills the beans". Thus, the catena in the lexicon will be underspecified with respect to the grammatical features and word forms for the verb.

This underspecified catena will be called a LEXICON CATENA (LC), because it will be stored in the lexical entries. Figure 2 depicts two realizations (with different word orders) of the catena for the idiom (bg) затварям си очите *zatvaryam si ochite* (lit. shut-1SG.PRS REFL eyes-DEF) 'I ignore the facts'. The upper part of the image represents the lexicon catena for the idiom. It determines the fixed elements of the catena: the arcs, their labels, the nodes and their labels: extended part of speech (first row), word forms (second row), lemmas (third row), and gloss in English (fourth row).[5] The dash (–) in the word form row means that the word form is not defined for the verbal node. In this way the word form could be different in the different realization of the catena. Also, the POS tag in the catena is underspecified with respect to features of the different word forms. In the two realizations, the verbal forms received their specific tags. Also, fixed elements of the catena are represented as in the image of the catena. The word order in the two realizations is different. Thus, catenae with different underspecified elements define different levels of freedom in the realization of the MWEs.

Let G_1 and G_2 be two catenae. A COMPOSITION of G_1 and G_2 is a catena G_c, such that

1. the catenae G_1 and G_2 are realized in catena G_c,
2. each node in catena G_c is an image of a node from G_1 or G_2, or both,
3. the root of catena G_c is an image of the root of catena G_1,
4. if a node i in catena G_c is an image of node i_1 in catena G_1 and i_2 in G_2, then all the information assigned to these nodes is compatible and fully represented in the node i,
5. if an arc $\langle i, j\rangle$ in catena G_c is an image of arc $\langle i_1, j_1\rangle$ in catena G_1 and $\langle i_2, j_2\rangle$ in G_2, then the label of $\langle i, j\rangle$ if it exists, has to be compatible with the labels of the arc $\langle i_1, j_1\rangle$ in G_1 and $\langle i_2, j_2\rangle$ in G_2.

[5] In the next examples we present only the important information, thus, some of these rows will be missing. In other cases new rows will be used to represent additional information.



Lexicon catena:

Realization 1: (bg) Очите си затваряха пред фактите *Ochite si zatvaryaha pred faktite* (lit. eyes-DEF REFL shut-3PL.PST.PROG at facts-DEF) 'They ignored the facts':

Realization 2: (bg) Иван си затваряше очите *Ivan si zatvaryashe ochite* (lit. Ivan-SG REFL shut-3SG.PST.PROG eyes-DEF) 'Ivan ignored the facts':

Figure 2: Two realizations of the lexicon catena for the idiom (bg) затварям си очите *zatvaryam si ochite* (lit. shut-1SG.PRS REFL eyes-DEF) 'I ignore the facts'.

The lemma information for two nodes i_1 in G_1 and i_2 in G_2 is compatible if at least one of the nodes does not have an assigned lemma, or if both nodes have the same assigned lemma. It is similar for word forms. For POS tags the compatibility is defined as a tag representation that contains the information of tags defined for both nodes. For example, if we have partial POS tag specifications 'Vpit' and 'Vp–m2s', the compatible specification is 'Vpit–m2s'. The arc labels are compatible if and only if they are the same, or at least one of them is not defined. If for both arcs the labels are not defined, then the label for the image arc is also not defined. Similar definitions could be stated for any other information added to the nodes and arcs such as semantic information, etc.

Using the composition operation we could realize the selectional restrictions of a given lexical unit with respect to a catena in a sentence.

For example, let us assume that the verb 'to read' requires the subject to be a human and the object to be an information object. In Figure 3 we present how the catena for 'I read' is combined with the catena 'a book' in order to form the catena 'I read a book'. The figure represents only the level of word forms and a level of semantics (specified only for the node on which the composition is performed). The catena for 'I read …' specifies that the unknown direct object has the semantics of an *Information Object (InfObj)*. The catena for 'a book' represents the fact that the book is an Information Object. Thus the two catenae could be composed on the two nodes marked as InfObj. The result is represented at the bottom of Figure 3.[6]

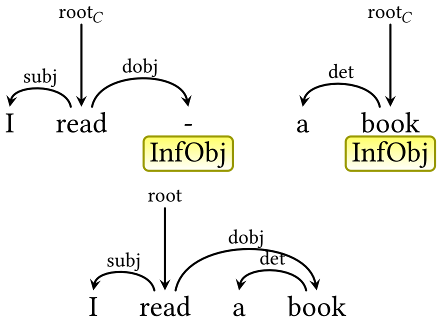


Figure 3: Composition of catenae.

[6] In this representation many details like lemmas and grammatical features are not presented because they are not important for the example.

Some MWEs require more complex operations over catenae. Such a class of MWEs are idioms with a lexicalized subject, such as "the devil is in the details"; the realizations of catenae from the lexicon into dependency trees are often accompanied by intervening material – see the discussion in Osborne et al. (2012). For example, the above-mentioned idiom allows realizations such as: "the devil will be in the details", "the devil seems to be in the details", etc. Thus we need to modify the internal structure of the lexicon catena.

Our insight, supported by the examples, is that the intervening material forms a catena of a certain type. Such a type of catena will be called an AUXILIARY CATENA[7] in this paper, although it could be of different kinds (auxiliary, modal, control, etc.), depending on the verb forms. In order to implement this idea we need some additional notions.

Let $G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G)$ be a catena and $k \in V_G$ and m is integer and $m > 1$, then $G_1, G_2, ..., G_l$ is a partition of G on node k if and only if:

1. each G_i for for $1 \leq i \leq m$ is a catena which is a subtree of G;
2. one or more subcatenae G_i for each $1 \leq i \leq m$ have k as a root node;
3. the only common node for all subcatenae G_i is k;
4. the mappings $\pi_{Gi}, \lambda_{Gi}, \omega_{Gi}, \delta_{Gi}$ are the same as for the whole catena G, except for the node k where the mappings $\pi_{Gi}, \lambda_{Gi}, \omega_{Gi}$ could be partial with respect to the original mappings.

An example of the operation **partition** of *the devil is in the details* is given in Figure 4.

After the partition of the catena, we need a mechanism to connect the different catenae of the partition with the auxiliary catena.

Let G be a catena and for $n \in V_G, G_1, G_2, ..., G_n$ be a partition of G and G_a be an auxiliary catena. An EXTENSION of G on partition $G_1, G_2, ..., G_n$ with catena G_a is a catena G_e such that each catena $G_1, G_2, ..., G_n$ and the auxiliary catena G_a are realized in G_e in such a way that the node n_i in G_i (corresponding to the original node n) is mapped to a node in G_e to which a node of G_a is mapped. Each node in G_e is an image of a node from $G_1, G_2, ..., G_n$ or G_a.

An example of the operation **extension** is presented in Figure 5.[8]

[7] Under auxiliary catena we assume a catena that is part of the verbal complex (i.e. an analytical tense of a verb, where elements such as clitics can be inserted between components) and contains nodes for the auxiliary verbs. In the grammars for the different languages different kinds of catena could be defined on the basis of their role in the grammar. In this respect, the definition of extension here is restricted to the verbal complex, but could be easily adapted for other cases when necessary.

[8] Note that there are alternative analyses in which the auxiliary verb is not a head of the sentence, but a dependent of the copula.

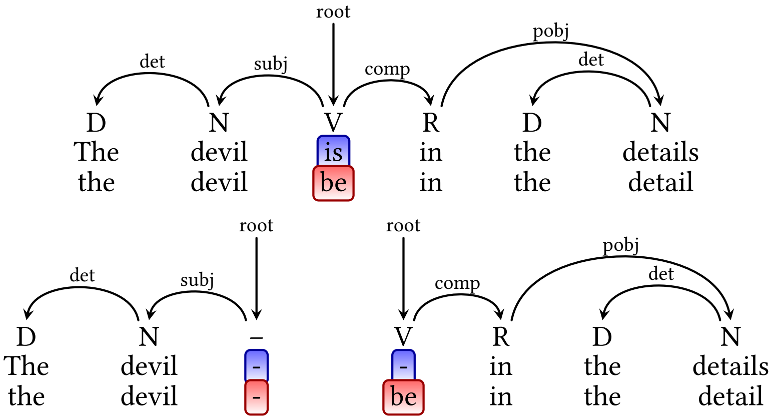

root
pobj
det
subj
comp
det
D N V R D N
The devil is in the details
the devil be in the detail
root
root
pobj
det
subj
comp
det
D N – V R D N
The devil - - in the details
the devil - be in the detail

Figure 4: Partition of the catena for “the devil is in the details”.

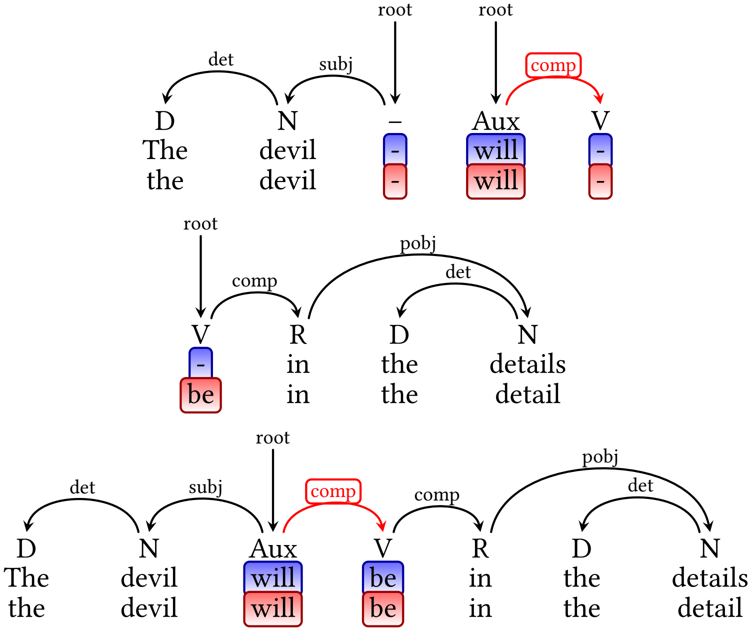


Figure 5: Extension.

Two catenae G_1 and G_2 could have the same set of realizations. In this case, we will say that G_1 and G_2 are EQUIVALENT. Representing the nodes via paths in the dependency tree from root to the corresponding node and imposing a linear order over this representation of nodes facilitates the selection of a unique representative of each equivalent class of catenae. Thus, in the rest of the paper we assume that each catena is representative of its class of equivalence. This representation of a catena will be called CANONICAL FORM.

5 A model of a lexical entry

In this section we use the notion of catena already introduced in Section 4, to define in greater detail the structure of a lexical entry as presented above. Through the operations of *composition*, *partition* and *extension* it becomes possible to compose the different parts of this structure and thus manage the actual realization of the lexical items in text. In this paper we represent the syntactic information in terms of the dependency grammar, but it can be done in a similar way within phrase-based grammars.

For each node in a catena or a dependency tree we present the following information: POS, Grammatical Features, Word Form, Lemma, Node identifier (the position of a word form in a catena or a sentence). Each piece of information is depicted in the node representation at a different row.

In Figure 6 a model of the lexical entry is presented. Each lexical entry for a synset includes (minimally): *Synset* which defines the synset information and *SynsetID* which identifies the synset in a unique way; *Definition* which expresses the content of the meaning of the synset; *Lemma list* which contains the representation of each lemma that shares the meaning of the synset. Each lemma is represented by the following elements: *LemmaID* which introduces the lemma in a unique way in the whole lexicon; *Basic Form* is a selected word form from the paradigm of the lemma; *Paradigm* is a list of pairs consisting of a word form, represented as a catena, and a tag, encoding the grammatical features of the word form. Each word form is a catena; *Valency Frame* represents the selectional restrictions of the lemma. The valency frame is represented as a catena. *Examples* is a list of example sentences or short texts. The realization of a lemma in a text requires the selection of the appropriate word form from the paradigm, represented as a Word Form Catena (WFC), composed with the Valency Frame Catena (VFC).

In Figure 7 we give an example lexical entry for the verb (bg) бягам *byagam* (lit. run-1SG.PRS) ‘to run’. The most important information is presented in the following sections: *Paradigm*, where we could see two catenae for *present tense, first person, singular*, and *present tense, second person, singular*, and in *Valency frame* (V. Frame) where a catena for the valency restrictions is given.

<table>
<tr><td colspan="2"><b>Synset:</b> <i>Example entry</i> Synset ID: <i>SynsetID</i></td></tr>
<tr><td colspan="2"><b>Definition:</b> <i>Text of the definition</i></td></tr>
<tr><td><b>Lemma list:</b></td><td>
<table>
<tr><td><b>LemmaID:</b></td><td><i>Lemma-ID1</i></td></tr>
<tr><td><b>Basic Form:</b></td><td><i>BasicForm-Lemma-ID1</i></td></tr>
<tr><td><b>Paradigm:</b></td><td>WordForm_{11} : $\text{GrammaticalTag}_{11}$
WordForm_{12} : $\text{GrammaticalTag}_{12}$
…
WordForm_{1n} : $\text{GrammaticalTag}_{1n}$</td></tr>
<tr><td><b>Valency Frame:</b></td><td><i>Valency Frame Description</i></td></tr>
<tr><td><b>Examples:</b></td><td><i>List of examples for this lemma</i></td></tr>
<tr><td><b>Analytical Class:</b></td><td><i>Pattern Class</i></td></tr>
</table>
…
<table>
<tr><td><b>LemmaID:</b></td><td><i>Lemma-IDK</i></td></tr>
<tr><td><b>Basic Form:</b></td><td><i>BasicForm-Lemma-IDK</i></td></tr>
<tr><td><b>Paradigm:</b></td><td>WordForm_{K1} : $\text{GrammaticalTag}_{K1}$
WordForm_{K2} : $\text{GrammaticalTag}_{K2}$
…
WordForm_{Kn} : $\text{GrammaticalTag}_{Kn}$</td></tr>
<tr><td><b>Valency Frame:</b></td><td><i>Valency Frame Description</i></td></tr>
<tr><td><b>Examples:</b></td><td><i>List of examples for this lemma</i></td></tr>
<tr><td><b>Analytical Class:</b></td><td><i>Pattern Class</i></td></tr>
</table>
</td></tr>
</table>

Figure 6: Lexical entry model.

The information related to the nodes in the catena is represented on different layers as follows: the bottom row contains the names of the corresponding nodes: CNo1, CNo2, etc. (in many examples in the paper this information is not presented, because it is redundant to a certain extent); the next row up contains the translation of the word form in English; the next two rows up are for the lemma of the node and for the word form. If the word form row contains "–" then the node is underspecified for a word form and it is determined by another catena during the composition operation. The last two rows up represent the grammatical features for the corresponding word forms. The first row contains information for each word form in its own lexical entry. The second row (the top one) contains grammatical information for the node when it is realized in the complete word form. When the word form is a single word, then the value in the two rows coincides. The difference could appear when in MWEs (including

Synset: *бягам от отговорност*	Synset ID: *SID-003592*
Definition: *Отбягвам да поема отговорност*	
Lemma list:	

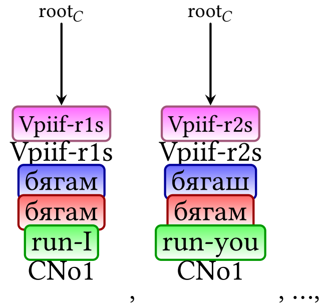
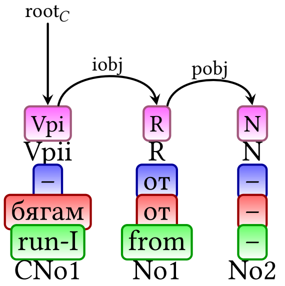
LemmaID:	*btbwn-041000447-v*
B. Form:	*бягам*
Paradigm:	 root$_C$ root$_C$ Vpiif-r1s Vpiif-r1s бягам бягам run-I CNo1 , Vpiif-r2s Vpiif-r2s бягаш бягам run-you CNo1 , …,
V. Frame:	 root$_C$ iobj pobj Vpi Vpii – бягам run-I CNo1 R R от от from No1 N N – – – No2
Examples:	*List of examples for this lemma*
Analytical Class:	*PatternClassVp*

Figure 7: Lexical entry for the verb (bg) бягам (от отговорност) *byagam (ot otgovornost)* (lit. run-1SG.PRS (from responsibility-SG.F)) 'to run away from one's responsibility'.

analytical forms) some of the grammatical features are modified. In the example above, the word form for future tense is composed of the auxiliary particle (bg) ще *ste* (lit. will-FUT) ‘will’ and the verb form for *present tense, second person, singular*. The whole word form is in future tense. In the example, the morphosyntactic tag Vpiif-r2s (tag for present tense) becomes Vpiif-f2s (tag for future tense in an analytical verb form). In the text realization we perform composition of one catena from the paradigm and the catena from the valency frame. Thus, the result from this operation between the analytical word given above and the valency catena results in the following catena – see Figure 8.

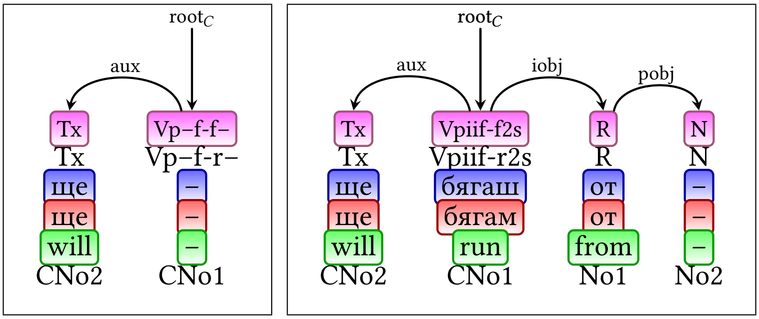


Figure 8: On the left, the auxiliary catena for future tense is given. As can be seen, the head node for the verb is unspecified for lemma and word form. It is also unspecified for the grammatical features of the main verb which has to be in present tense. The auxiliary and the main verb together build an analytical word form that is in future tense. On the right side, the following information is given: the result from the composition of the auxiliary catena, the word form catena and the valency frame catena. The resulting verb catena is for the string (bg) ще бягаш от отговорност *ste byagash ot otgovornost* (lit. will-FUT run-2SG.PRS (from responsibility-SG.F)) ‘you will run from responsibility’.

Coming back to modeling MWEs and their representation in the lexicon and their realization in the text, we model them in the lexicon as described above assigning an appropriate catena for the forms of the MWE in the paradigm and catena for the valency frame. The realization in the text is performed by the operations defined in the section above. We also represent the grammatical features over two layers: one for the components of the MWE as they appeared in the lexicon, and one for the realization in the text. In the next section we present a classification of the different types of MWEs included in the final integrated lexicon.

6 Analyses of MWE types

In our previous research we gave credit to the most frequent head-based types of MWEs (this means that the MWE is analysed according to its syntactic head – noun, verb, etc.) as presented in BTB-WN. The influence of BTB-WN mapping to the English wordnet also played a big role. When transferred from English, the resulted MWEs in Bulgarian might include free phrases, collocations, etc. to ensure the correct relation to the English notion.

Here we would like to present our model with respect to the complexity of the MWE representation. We view complexity in the following way: a) from fixedness towards flexibility. Here several options are considered: morphological flexibility, syntactic flexibility, semantic flexibility, and combination of two or all of them; b) from continuity to discontinuity. We consider MWEs with at least two words. Please note that the named entities are not discussed. We assume that the more words constitute the MWE, the more complex this MWE is. Idiomaticity is hidden in fixedness. Here are the types we consider: fixed, continuous; fixed, discontinuous; semi-fixed, continuous; semi-fixed, discontinuous; flexible, continuous; flexible, discontinuous.

It can be seen that the fixed, continuous type is mainly nominal or prepositional while the fixed, discontinuous type is rare. The most frequent type is the semi-fixed one. In the continuous subtype noun phrases prevail while in the discontinuous one verbal MWEs are typical. We build on the representation described in Simov & Osenova (2015a,b). Let us consider them in order below. In the graphical representations below we present the main word forms in the paradigm, instead of complete lexical entries.

6.1 Fixed, continuous

Here three main structural variants are detected. They are all idiomatic.

(1) Noun Conj Noun: (bg) живот и здраве *zhivot i zdrave* (lit. life-sg.m and health-sg.n) ‘some day’ – see Figure 9

(2) Prep NP:

a. (bg) за вечни времена *za vechni vremena* (lit. for eternal-PL times-PL) ‘forever’;

b. (bg) между другото *mezhdu drugoto* (lit. between other-SG.DEF) ‘by the way’;

c. (bg) на легло *na leglo* (lit. on bed-sg.n) ‘ill’

(3) Adjective Noun: (bg) добро утро *dobro utro* (lit. good-sg.n morning-sg.n) ‘good morning’

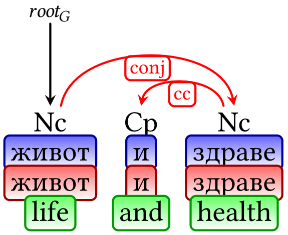


Figure 9: Catena for fixed, continuous expressions: (bg) живот и здраве *zhivot i zdrave* (lit. life-SG.M and health-SG.N) ‘some day’.

The new additions to the catena representation in comparison to our previous work are: the incorporation of the synonyms to the idioms as in examples 1 and 2, and the handling of pragmatic formulae in example 3.

A challenge that appears in this group are the boundaries of the MWEs. For example, (bg) на легло *na leglo* (lit. on bed-SG.N) ‘ill’ might be extended also to the inclusion of a copula: (bg) на легло съм *na leglo sam* (lit. on bed-SG.N am-1SG) ‘to be ill’. The question is whether the copula element should be represented as a component of the MWE or not. According to our suggestion the catena (bg) на легло *na leglo* (lit. on bed-SG.N) ‘ill’ can combine with the catena of the auxiliary and form another catena.

6.2 Fixed, discontinuous

This class is a speaker strategy rather than a distinct type of its own. The strategy can contextualize a fixed MWE and thus add to it more elements. For example, the MWE (bg) без капка разум *bez kapka razum* (lit. without drop-SG.F sense-SG.M) ‘without an iota of sense’ can be extended with a modifier to the noun ‘sense’ such as (bg) без капка медицински разум *bez kapka meditsinski razum* (lit. without drop-SG.F medical-SG.M sense-SG.M) ‘without an iota of medical sense’ in a specific context. These cases are rare and non-systematic.

6.3 Semi-fixed, continuous

This predominantly nominal group contains terms, idiomatic expressions as well as every-day-life expressions. However, its main specificity is the fact that they do exhibit morphosyntactic varieties such as changes in definiteness and number but on the head word only. The dependant remains unchanged.

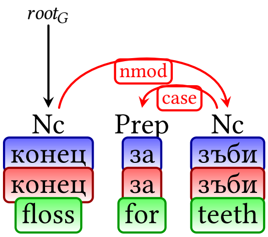


Figure 10: Catena for semi-fixed, continuous expressions: (bg) конец за зъби *konets za zabi* (lit. floss-SG.M for teeth-PL) ‘dental floss’.

1. Noun Noun: (bg) муха цеце *muha tsetse* (lit. fly-SG.F tsetse) ‘tsetse fly’; (bg) ангел хранител *angel hranitel* (lit. angel-SG.M guardian-SG.M) ‘guardian angel’

2. Noun prep Noun: (bg) конец за зъби *konets za zabi* (lit. floss-SG.M for teeth-PL) ‘dental floss’ – see Figure 10; (bg) лак за нокти *lak za nokti* (lit. polish-SG.M for nails-PL) ‘nail polish’; (bg) яйце на очи *yaytse na ochi* (lit. egg-SG.N on eyes-PL) ‘a fried egg’

6.4 Semi-fixed, discontinuous

This group contains mainly verbal MWEs. These are: the quasi-reflexive verbs (the so-called middle verbs where the participating reflexive has no semantics but only a derivational function), and the light verb constructions.

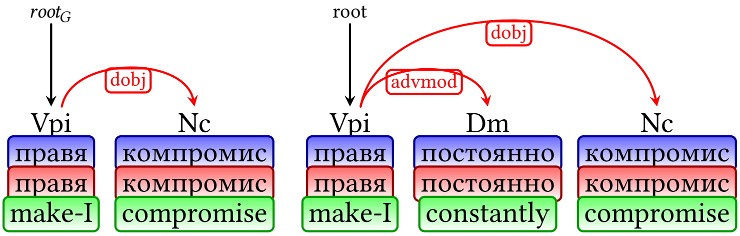


Figure 11: Catena for a light verb construction (semi-fixed, discontinuous expressions): (bg) правя компромис *pravya kompromis* (lit. do-1SG.PRS compromise-SG.M) ‘to make a compromise’. On the left side is the lexical catena. On the right side is a modification with an adverb, which is realized between the two parts of the MWE.

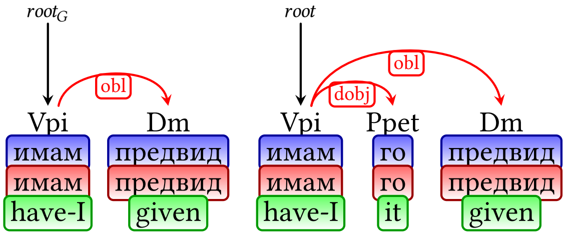


Figure 12: Catena for a light verb construction (semi-fixed, discontinuous expressions): (bg) имам предвид *imam predvid* (lit. have-1SG.PRS given) 'to have in mind'. This is similar to the previous example, but the intervening material is a pronoun.

1. Quasi-reflexive verbs: (bg) адаптирам се *adaptiram se* (lit. adapt-1SG.PRS REFL) 'to adapt'; (bg) вкисвам се *vkisvam se* (lit. get-sour-1SG.PRS REFL) 'to feel bad'

2. Light verb constructions: (bg) правя компромис *pravya kompromis* (lit. do-1SG.PRS compromise-SG.M) 'to make a compromise' – see Figure 11; (bg) правя почивка *pravya pochivka* (lit. do-1SG.PRS rest-SG.F) 'to take a break'; (bg) давам обещание *davam obeshtanie* (lit. give-1SG.PRS promise-SG.N) 'to make a promise'; (bg) вкарвам в употреба *vkarvam v upotreba* (lit. implement-1SG.PRS in usage-SG.F) 'to put into use'; (bg) имам предвид *imam predvid* (lit. have-1SG.PRS given) 'to have in mind' – see Figure 12; (bg) давам под наем *davam pod naem* (lit. give-1SG.PRS under rent-SG.M) 'to rent out'

The two parts of the quasi-reflexive verbs can be discontinued by the auxiliary in some forms in the verb paradigm ((bg) адаптирал съм се *adaptiral sam se* (lit. adapt-PTCP.PST am-1SG.PRS REFL) 'I have adapted'). Most of the light verbs have single verbs as synonyms. For example, (bg) давам обещание *davam obeshtanie* (lit. give-1SG.PRS promise-SG.N) 'to make a promise' has a synonym (bg) обещавам *obeshtavam* (lit. promise-1SG.PRS) 'to promise'. They also can often be discontinued by a modifier on the noun element ((bg) давам голямо обещание *davam golyamo obeshtanie* (lit. give-1SG.PRS big-SG.N promise-SG.N) 'to make a big promise') or by another participant in the sentence (bg) давам насила обещание *davam nasila obeshtanie* (lit. give-1SG.PRS reluctantly promise-SG.N) 'to make a promise reluctantly'). The variant (bg) давам под наем *davam pod naem* (lit. give-1SG.PRS under rent-SG.M) 'to rent out' allows for an object coming after the verb (bg) давам *davam* (lit. give-1SG.PRS) ‘to give’: (bg) давам стаята под наем *davam stayata pod naem* (lit. give-1SG.PRS room-SG.F.DEF under rent-SG.M) ‘to rent out the room’.

6.5 Flexible, continuous

This group consists of just one nominal type which is “Adjective Noun”. Some of the MWEs are literal, and some are figurative. In the examples below the last one is figurative.

(4) Adjective Noun

a. (bg) бежански лагер *bezhanski lager* (lit. refugee-SG.M camp-SG.M) ‘a refugee camp’ – see Figure 13;

b. (bg) гол охлюв *gol ohlyuv* (lit. naked-SG.M snail-SG.M) ‘a slug’;

c. (bg) домашна работа *domashna rabota* (lit. home-SG.F work-SG.F) ‘homework’;

d. (bg) ахилесова пета *ahilesova peta* (lit. Achilles’-SG.F heel-SG.F) ‘Achilles’ heel’

Here the MWEs are mostly terms or near-terms. Both elements form a concept, so they cannot be discontinued but they are flexible with respect to their morphosyntactic behaviour. They can be used with an article or in a plural form. The article occurs only once in a phrase but both elements in the MWE can inflect in number. Also, the idiomatic expressions like (bg) ахилесова пета *ahilesova peta* (lit. Achilles’-SG.F heel-SG.F) ‘Achilles’ heel’ have synonyms, in this case – weakness.

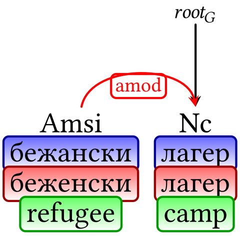


Figure 13: Catena for flexible, continuous expressions: (bg) бежански лагер *bezhanski lager* (lit. refugee-SG.M camp-SG.M) ‘a refugee camp’.

6.6 Flexible, discontinuous

Here some verbal expressions are listed which are flexible with respect to morphosyntax. This means that the verb can inflect in all verb tenses and other verb forms.

(5) Verb NP

a. (bg) развързвам кесията *razvarzvam kesiyata* (lit. untie-1SG.PRS purse-SG.F.DET) 'I pay generously' – see Figure 14;

b. (bg) играя открито *igraya otkrito* (lit. play-1SG.PRS openly) 'I play fair';

c. (bg) избирам страна *izbiram strana* (lit. choose-1SG.PRS side-SG.F) 'to take side';

d. (bg) тегля един бой *teglya edin boy* (lit. drag-1SG.PRS one fight-SG.M) 'to draw a fight', etc.

The MWE can be used also without the reflexive particle. At the moment we view both possibilities as synonyms. These expressions also allow for some discontinuous material. For example, an adverbial of manner can come between the verb and the object in the first listed MWE above – (bg) развързвам си сериозно кесията *razvarzvam si seriozno kesiyata* (lit. untie-1SG.PRS REFL seriously purse-SG.F.DET) 'I pay very generously' – the second tree in Figure 13.

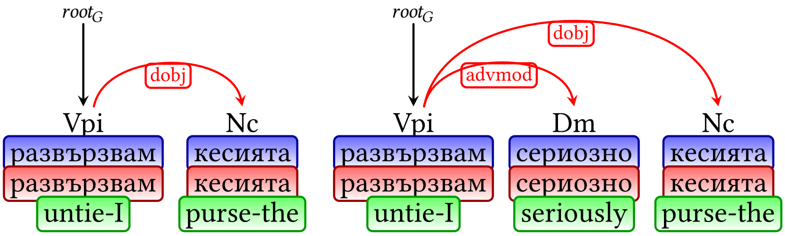


Figure 14: Catena for flexible discontinuous expressions: (bg) развързвам кесията *razvarzvam kesiyata* (lit. untie-1SG.PRS purse-SG.F.DET) 'I pay generously'.

In this section various examples were outlined according to a proposed classification that respects the complexity of the MWEs. The catena illustrations follow the Universal Dependencies guide.[9] The fixed, discontinuous type turned out to be a strategy where the speaker can personalize fixedness and thus legitimate the addition of new elements in a specific context.

7 Conclusions and future work

The representation of MWEs in an integrated model has never been a trivial task. Our proposal is to use the catena notion since it allows for a graph-based realization where all the characteristics of interest can be added: the internal structure specifics as well as the external ones, if needed. In addition, the interaction among morphology, syntax (including valency potential and a vanilla mechanism[10] for word order) as well as semantics can be illustrated. We are aware of the fact that our model is similar in many aspects to the other tree-based approaches. At the same time, our representation model is put in the context of an integrated resource and we believe that here come the main novel directions in our work.

It has become clear for quite some time that MWEs are a phenomenon that is not always trivial to define, classify, annotate, analyse and integrate. For that reason, we view our work as a bottom-top effort that would gradually cover specific lemmas, meanings and cases.

Our future work is envisaged in several directions: to fully implement the suggested mechanism, to evaluate it on downstream tasks, and also in the backward direction – to identify the problematic places and repair them in the lexicon. Some already identified problematic places are the MWE boundaries and the degree of granularity in their representation.

Abbreviations

BTB	Bultreebank	IRV	inherently reflexive verbs
BTB-WN	Bultreebank Wordnet	LC	lexicon catena
BVL	Bulgarian Valency Lexicon	LFG	Lexical-Functional Grammar
ID	identifier	MWE	multiword expressions
ILB	Inflectional lexicon of Bulgarian	NLP	Natural Language Processing
		POS	part-of-speech

[9] https://universaldependencies.org/guidelines.html

[10] This means that our approach is very standard and basic, initially predicting the clear places of discontinuity on the encountered examples without ensuring that all cases are covered appropriately.

SM	semantics	VPC	verb-particle constructions
VMWE	verbal multiword expressions	WFC	Word Form Catena
VFC	Valency Frame Catena		

Acknowledgements

The reported work has been partially supported by CLaDA-BG, *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH*, funded by the Ministry of Education and Science of Bulgaria (support for the Bulgarian National Roadmap for Research Infrastructure). We would also like to thank the reviewers for their valuable and insightful comments.

References

Buchholz, Sabine & Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*. https://api.semanticscholar.org/CorpusID:13075323.

Dyvik, Helge, Gyri Smørdal Losnegaard & Victoria Rosén. 2019. Multiword expressions in an LFG grammar for Norwegian. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 69–108. Language Science Press. DOI: 10.5281/zenodo.2579037.

Grégoire, Nicole. 2010. DuELME: A Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44(4). 23–39. DOI: 10.1007/s10579-009-9094-z.

Groß, Thomas. 2010. Chains in syntax and morphology. In Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto & Yasunari Harada (eds.), *Proceedings of the 24th Pacific Asia conference on language, information and computation*, 143–152. Tohoku University, Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University. https://aclanthology.org/Y10-1018.

Laskova, Laska, Petya Osenova, Kiril Simov, Ivajlo Radev & Zara Kancheva. 2019. Modeling MWEs in BTB-WN. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović & Verginica Barbu Mititelu (eds.), *Proceedings of the joint workshop on multiword expressions and WordNet (MWE-WN 2019)*, 70–78. Florence, Italy: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5109.

Leseva, Svetlozara, Verginica Barbu Mititelu, Ivelina Stoyanova & Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 73–116. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998635.

Lichte, Timm, Simon Petitjean, Agata Savary & Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of "irregular" regularities. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions*, 1–33. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579033.

Lion-Bouton, Adam, Agata Savary & Jean-Yves Antoine. 2023. A MWE lexicon formalism optimised for observational adequacy. In Archna Bhatia, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han & Shiva Taslimipoor (eds.), *Proceedings of the 19th workshop on multiword expressions (MWE 2023)*, 121–130. Dubrovnik, Croatia: Association for Computational Linguistics. https://aclanthology.org/2023.mwe-1.16.

Masini, Francesca. 2019. *Multi-word expressions and morphology*. Oxford: Oxford University Press.

O'Grady, William. 1998. The syntax of idioms. *Natural Language and Linguistic Theory* 16. 279–312.

Osborne, Timothy, Michael Putnam & Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax* 15(4). 354–396. DOI: 10.1111/j.1467-9612.2012.00172.x.

Osenova, Petya. 2010. Bulgarian. In *The languages of the new EU member states*, vol. 88 (Revue Belge de Phologie et D'Historie 3), 643–668.

Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski & Marek Świdziński. 2014. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14)*, 2785–2792. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/279_Paper.pdf.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten Van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke Van Der Plas, Behrang Qasemizadeh, Carlos Ramisch, Federico Sangati, Ivelina

Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI: 10.5281/zenodo.1471590.

Schneider, Nathan, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad & Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14)*, 455–461. Reykjavik, Iceland: European Language Resources Association (ELRA). https://aclanthology.org/L14-1433/.

Simov, Kiril & Petya Osenova. 2014. Formalizing MultiWords as catenae in a treebank and in a lexicon. In Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova & Adam Przepiórkowski (eds.), *Proceedings of the thirteenth international workshop on Treebanks and Linguistic Theories (TLT13)*, 198–207. Tübingen: University of Tübingen.

Simov, Kiril & Petya Osenova. 2015a. Catena operations for unified dependency analysis. In Joakim Nivre & Eva Hajičová (eds.), *Proceedings of the third international conference on dependency linguistics (depling 2015)*, 320–329. Uppsala, Sweden: Uppsala University. https://aclanthology.org/W15-2135.

Simov, Kiril & Petya Osenova. 2015b. Modeling lexicon-syntax interaction with catenae. *Journal of Cognitive Science* 16(3). 287–322. DOI: 10.17791/jcs.2015.16.3.287.

Skoumalová, Hana, Marie Kopřivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondřička & Milena Hnátková. 2024. LEMUR: A lexicon of Czech multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 1–37. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998631.

Vondřička, Pavel. 2019. Design of a multiword expressions database. *The Prague Bulletin of Mathematical Linguistics* 112. 83–101. https://ufal.mff.cuni.cz/pbml/112/art-vondricka.pdf.

Zampieri, Nicolas, Carlos Ramisch & Geraldine Damnati. 2019. The impact of word representations on sequential neural MWE identification. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović & Verginica Barbu Mititelu (eds.), *Proceedings of the joint workshop on multiword expressions and*

WordNet (MWE-WN 2019), 169–175. Florence, Italy: Association for Computational Linguistics. https://aclanthology.org/W19-5121.