Modeling Lexicon-Syntax Interaction with Catenae

Petya Osenova and Kiril Simov

Bulgarian Academy of Sciences, Bulgaria
{petya, kivs}@bultreebank.org

The paper discusses the interaction between the lexicon and syntactic relations in texts. It considers the encoding of the language units by means of catenae. It also shows that catenae representation together with valence information can provide a good way of handling Multiword Expressions and compounds\(^1\) in the lexicon and syntax. The paper introduces a strategy for mapping noun/verb compounds with their counterpart syntactic phrases\(^2\). Thus, the investigation presents two research directions: (1) realization of lexical units from lexicons to texts, and (2) relations between identical semantic units with differing morphosyntactic properties. Although the provided examples come from Bulgarian, the proposed mechanism is language independent.

Keywords: Catenae, Multi-word Expressions, Compounds, Syntax, Lexicon, Modeling

---

\(^1\) In Bulgarian the notion of compound refers to one-word expressions with two roots and additional affixes/inflections. For example, the compound word бilkолечение ‘bilkolechenie, herb-[linking vowel]-cur-ing’, curing with herbs, has two roots: билк- (herb) and леч- (cure), which are connected by a linking vowel (-о-). It has also affixes (-ен-и) and an ending (-е) for a neuter, singular noun: билк-о-леч-ен-и-е.

\(^2\) The counterpart syntactic phrases are syntactic paraphrases of the one-word compound expressions (which are mostly nouns). For example, the compound word бilkолечение ‘bilkolechenie, herb-curing’ can be expressed syntactically in the following way: лекувам с билки ‘lekuvam s bilki’, cure with herbs.

Date submitted: 03/01/15   Date reviewed: 05/01/15
Date confirmed for publication: 09/08/15
©2015 Institute for Cognitive Science, Seoul National University
1. Introduction

The modeling of interaction between lexicon and syntax is not trivial, since their relation is asymmetrical at least in the following three senses: (1) the types of information that is encoded in the lexicon and in the text; (2) the ways of mapping from lexicon to text; (3) idiosyncrasies of various kinds in the lexicon and in the text. In order to show the interaction between the lexicon and the syntax, we rely on resources, such as valency lexicons and treebanks. Although the language examples in the paper are from Bulgarian (and English), the proposed mechanism is language independent and can be applied to other languages as well. From the lexicon perspective the valency dictionary of Bulgarian is considered (Osenova et al. 2012). From the text perspective the syntactic annotation in BulTreeBank (Simov et al. 2004) is taken into account. As crossing paths between lexis and grammar the following cases are discussed with a special focus in the paper: Multiword Expressions (MWEs) and deverbal compounds with or without respective syntactic counterparts.

BulTreeBank exists in two formats: HPSG-based (original – constituent-based with head annotation and grammatical relations) and dependency-based. In both of them the representations of the various kinds of MWEs is an important problem. We need a mechanism for connecting the MWEs in the lexicon with their actual usages within sentences. As an interesting case of MWEs at the interface of morphology and syntax we consider compounds. They are usually derived from several lexical units and exhibit an internal structure with a specified derivation model and semantics. Compounds should be viewed as transforming MWEs or syntactic phrases into one-word units. Here we are especially interested in the mapping among deverbal compounds and their counterpart syntactic phrases, when available.

There is no broadly accepted standard for Multiword Expressions (see the various classifications in Villavicencio and Kordoni (2012)). For the task of automatic recognition of MWEs in Bulgarian Stoyanova (2010) adopts the classification of Baldwin et al. (2003). It is a more semantically oriented division, since the MWEs are classified as: non-decomposable by meaning; idiosyncratically decomposable and simple decomposable. However, we
adopt the well known Multiword Expressions classification, presented in Sag et al. (2002), since we are more interested in the morphosyntactic behavior of the MWEs. They are divided into two: lexicalized phrases and institutionalized phrases. Here we do not consider institutionalized phrases (being semantically and syntactically compositional, but statistically idiosyncratic) as a distinct group. Thus, the former are further subdivided into fixed-expressions, semi-fixed expressions and syntactically-flexible expressions. Fixed expressions are said to be fully lexicalized and undergoing neither morphosyntactic variation nor internal modification. Semi-fixed expressions have a fixed word order, but “undergo some degree of lexical variation, e.g. in the form of inflection, variation in reflexive form, and determiner selection” (non-decomposable idioms, proper names). Syntactically-flexible expressions show more variation in their word order (light verb constructions, decomposable idioms).

We follow the understanding of O’Grady (1998) that MWEs have their own internal syntactic structure which needs to be represented in the lexicon as well as in the sentence analysis. Such a mapping would provide a mechanism for accessing the literal meaning (if any) of MWEs. The provisional inclusion of the compounds into the MWE classification raises additional challenges. As it was mentioned, an important question is the prediction of the compound semantics formed on the basis of the related phrases containing verb + dependents. In this paper we discuss the usage of the same formal instrument – catenae – for the representation and analysis of MWEs and compounds as linguistic cross-paths at lexicon-syntax interface. Catenae are a path in the syntactic or morphemic analysis that is continuous in the vertical (dominance) dimension. It operates on the surface level. Its full potential is discussed further in the text.

The paper is structured as follows: In the next section an overview of previous work on catenae is presented. In Section 3 a typology of the most frequent Multiword Expressions in BulTreeBank is outlined. Section 4 considers possible approaches to consistent analyses of MWEs. Section 5 introduces formalization of the catenae. Section 6 describes the structure of the lexical entry and the representation of MWEs. Section 7 outlines the relation of the syntactic expressions with compound morphology. Section 8 concludes the paper.
2. Related Work on Catenae

The notion of catenae (chain) was introduced in O’Grady (1998) as a mechanism for representing the syntactic structure of idioms. He showed that for this task there is a need for a definition of syntactic patterns that do not coincide with constituents. He defines the catenae 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.* In recent years the notion of catenae revived again and it was applied also to dependency representations. Catenae are used successfully for modeling of various problematic language phenomena.

Osborne et al. (2012) give extensive argumentation on why to consider catenae as a basic syntactic unit instead of constituent. They define the catenae in the following way: *a word or a combination of words that is continuous with respect to dominance.* Also, catenae are viewed as “any tree or any subtree of a tree” (p. 359). The gain in this proposal is that constituents are always catenae, while there are catenae that are not constituents. This applies to idioms, elided expressions, etc. that are hard to be explained through the notion of constituent only. Thus, the explanatory coverage of catenae is better than the one of the constituent. For example, through the differentiation of the notions of head and governor, the discontinuous material can be explained in a straightforward way via certain principles, such as: either the head of a raised catena or the raised catena itself dominates its governor (p. 362). Another advantage of the proposed approach is that despite the surface-oriented analysis, catenae ensure also a direct connection to semantics. This means that all semantically related units are always stored as catenae in the lexicon. However, in real texts the MWEs might stop being continuous catenae due to the inclusion of various tense forms or other language material. For example, the MWEs ‘pull strings’ and ‘the devil is in the details’ are considered continuous catenae. However, in the following sentences their parts become separated by tense forms: ‘Which strings have they been pulling?’ or ‘I believe the devil will be in the details’. Thus, the catenae license the language material from the lexicon to the real usage in syntactic contexts, requiring
means for presenting discontinuous catenae. The authors apply declarative mechanisms, not procedural, to phenomena like raising and ellipsis. It makes them compatible with all the constraint-based approaches. Although the catenae approach is applicable in constituent-based frameworks, its full potential is shown in dependency-based theories.

The catenae-based approach acknowledges not only the interword dependencies, but also the intraword ones. The latter are discussed in more detail in Gross (2011b) and Gross (2011c). For example, the clitics are viewed as borderline morphs between free and bound morphs, which express meanings of free morphs but fail to be prosodic words (Gross 2011b: 49), and the compounds – as words that contain at least two lexical morphs (p. 50). These observations are of great importance to our paper, since Bulgarian has clitics in its grammar system, and they are often part of MWEs. Also, compounds are considered in this paper. Our idea is that we can profit from the combination of the catenae approach and valency information when handling the phenomena at lexicon-syntax interface.

Gross (2010) presents problems in syntax and morphology that have led to the introduction of the subconstituent catenae level. Constituency-based analysis faces non-constituent structures in ellipsis, idioms, verbal complexes. In morphology the constituent-oriented bracketing paradoxes have been also introduced ([moral] [philosoph -er] vs. [moral philosoph]-er). Catena is viewed as a dependency grammar unit. At the morphological level morphemes (affixes) receive their own nodes forming chains with the roots (such as tenses: has...(be)en; be...(be)ing, etc.). In Gross (2011) the author again advocates his approach on providing a surface-based account of the non-constituent phenomena via the contribution of catena. Here he introduces a notion at the morphological level — morph catena. Also, he presents the morphological analysis in the Meaning-Text Theory framework, where (due to its strata) there is no problem like the one present in constituency. Apart from linguistic modeling of language phenomena, catena was used in a number of NLP applications. Maxwell et al. (2013) present an approach to Information retrieval based on catenae. The authors consider catena as a mechanism for semantic encoding which overcomes the problems of long-distance paths and elliptical sentences. The employment of catenae in NLP applications is additional motivation for us
to use them in the modeling of an interface between the lexicon and syntax.

In this paper we consider catenae as both — unit of syntax and unit of morphology. In a syntactic or morphological tree (constituent or dependency) catena is: Any element (word) or any combination of elements that are continuous in the vertical dimension (y-axis).

In syntax it is applied to the idiosyncratic meaning of all sorts, to the syntax of ellipsis mechanisms (e.g. gapping, stripping, VP-ellipsis, pseudogapping, sluicing, answer ellipsis, comparative deletion), to the syntax of predicate-argument structures, and to the syntax of discontinuities (topicalization, wh-fronting, scrambling, extraposition, etc.). In morphology it is applied to the bracketing paradox problem, clitics and compounds. It provides a mechanism for a (partial) set of interconnected syntactic or morphological relations. The set is partial in cases when the elements of the catenae can be extended with additional syntactic or morphological relations to elements outside of the catenae. The relations within the catenae cannot be changed.

These characteristics of catena make it a good candidate for representing the various types of Multiword Expressions in lexicons and treebanks. In the lexicon each MWE represented as a catena can specify the potential extension of each element of this catena. As part of the morphemic analysis of compounds, catena is also a good candidate for mapping the elements of the syntactic paraphrase of the compound to its morphemic structure.

3. Multiword Expressions in BulTreeBank

In its inception and development phase, the HPSG-based BulTreeBank adopted the following principles: When the MWE is fixed, which means inseparable, with fixed order and can be viewed as a part-of-speech, it receives lexical treatment. This group concerns mainly the multiword closed class parts-of-speech: multiword prepositions, conjunctions and pronouns, but it also contains some adverbs. There are 1081 occurrences of such multiword closed class parts-of-speech (POS) in the treebank, which makes around 1.9% of the token occurrences in the text. Thus, this group is not problematic. Of course, there are also exceptions. For example, one of the multiword indefinite pronouns in Bulgarian allows for three alternatives.
in the auxiliary verb part, when the pronoun is in its plural form: каквито и да е/са/бил (‘what.PL and to is/are/been’ whatever). The alternatives are: a 3rd-person-singular-present-tense-auxiliary (is), its 3rd-person-neuter-singular-past-participle (been) or 3rd-person-plural-present-tense-auxiliary (are). The semi-fixed expressions (mainly proper names) have been interpreted as Multiword Expressions.

However, all the idioms, light verb constructions, etc. have been treated only syntactically, without any reference to their lexical semantics in a lexicon. This means that in the annotations there is no difference between the literal and idiomatic meaning of the expression: kick the bucket (= kick some object) and kick the bucket (= die). In both cases we indicated that the verb kick takes its nominal complement.

After some exploration of the treebank, such as the extraction of the valency frames and training of statistical parsers, we discovered that the present annotations of Multiword Expressions are not the most useful ones. In both applications the corresponding generalizations are overloaded with specific cases which are not easy to incorporate in more consistent classifications. The group of lexically treated POS remained stable. However, the other two groups were reconsidered. Proper names, as semi-fixed, are treated separately, i.e. as non-Multiword Expressions, since we need to coreference the single occurrence of the name with the occurrence of two or more parts of the name. Light verb constructions have to be marked as such explicitly in order to differentiate their specific semantics from the semantics of the verbal phrases with semantically non-vacuous verbs. The same holds for the idioms.

Here we outline only three illustrative frequent multiword types in our syntactic resource with respect to the syntactic relations (adjunction and complementation) among their elements. In the next sections also their modeling is presented with examples in the treebank and the lexicon. The adjunction and complementation types do not affect the formalization, which generalizes over both of them. However, it shows differences at the syntax-lexical interface.
The adjunction is expressed in the following MWE types:

1. **Noun phrases of type Adjective – Noun**
   Example: вътрешен министър ‘interior minister’ (Minister for Internal Affairs); снежен човек ‘snowy man’ (snowman)
   These patterns allow inflection in both elements for number. The first element can get a definite article. The noun phrase can be further modified: ‘our interior minister’; ‘a nice snowy man’, etc. Semantically, the first example is a metonymical synthetic form of the phrase ‘Minister for Internal Affairs’. The second one conveys its literal meaning of: (1) a man-like sculpture from snow or (2) hypothetical man living in Himalayas or some other regions.

2. **Noun phrases of type Noun – Prepositional Phrase**
   Example: среща на върха ‘meeting at peak.DEF’ (summit)
   Here среща ‘meeting’ can inflect in all its forms and allows for some modifications: ‘past meetings’, etc.

   The complementation is expressed in the following MWE type:

3. **Verb phrases of type Verb – Complement**
   Example: знае си работата ‘knows.3PERSON.SG his business.DEF’ (one knows one’s own business); затварям си очите ‘close one’s own eyes.DEF’ (to ignore the facts).
   Here ‘business’ allows for only reflexive possessive forms (one knows their-own business), but the nominal phrase always has to be definite, singular. The verb ‘know’ can vary in all its wordforms and it allows for modification: one knows their-own business well. The same holds for the second example.

4. **Some Approaches for Encoding Multiword Expressions**

There are a number of possible approaches to handling idioms, light verb constructions and collocations. The approaches are not necessarily conflicting with each other.
**The first approach is selection-based.** This approach is appropriate for Multiword Expressions in which there is a word that can play the role of a head which subcategorizes for its dependent. For example, a verb subcategorizes for only one lexical item or a very constrained set of lexical items. When combined with nouns, such as время (time), форма (shape), надежда (hope), the verb губя ‘lose’ forms idioms: губя время ‘lose time’ (waste one’s time); губя форма ‘lose shape’ (to be unfit); губя надежда ‘lose hope’ (lose one’s hope). However, when combined with other nouns, such as портфейл (wallet) or роднина (relative), the verb takes canonical complements. In the former cases, verbs like обръщам ‘to turn’ pay – take only the noun внимание ‘attention’, for making an idiom: обръщам внимание на някого ‘to turn attention to somebody’ (pay attention to somebody). Another example is the verb вземам ‘to take’, which combines in such cases with дума ‘word’: вземам думата ‘to take the word’ (take the floor). However, light verb expressions with de-semantisized verbs, such as имам (have) or става (happens) (имам думата ‘I have the word’ (to have the floor) or става дума за нещо ‘it happens word’ (something refers to something)) can take numerous semantic classes as dependents. In this case we mark the information only on the head of these MWE.

In this approach the assumption is that the verb posits its requirements on its dependents. However, a very detailed valency lexicon is required. One problem with this approach is when the dependent elements allow for modifications.

**The second approach is construction-based.** In this case there is no head, and thus the MWE is considered a construction without head-dependent relations among its parts. Multiword Expressions are with fixed order and inseparable parts. They are annotated via brackets at the lexical level. One example is the idiom от игла до конец ‘from needle to thread’ (from the beginning to the end). This approach is appropriate for the fixed MWEs, but problematic for the syntactically flexible ones.

**The third approach marks all the parts of the Multiword Expressions.** It is based on the notion of catena as introduced above, which means that the catena is stored in the lexicon first and then realized in the text. Here is
an example of this annotation in the treebank:\(^3\):

\[
\begin{align*}
(VPS & \text{Той} (VPC-C (V-C ритна) (N-C камбаната)))) \\
(VPS & \text{He} (VPC-C (V-C kicked) (N-C bell.DEF)))
\end{align*}
\]

\begin{quote}
He kicked the bucket
\end{quote}

where the suffix “-C”, attached to the node label, marks the catena. This approach maybe adds some spurious compositionality to the idioms, but it would be indispensable for handling idiosyncratic cases, such as separable MWEs.

However, in order to model the various MWEs and to ensure mappings among compounds and related syntactic phrases\(^4\), the combination of catenae with selection-based approach is needed. Note that in this phrase-based representation the verb dominates the noun forming a catena of type verb-complement.

In case the MWE does not allow for any modifications, for each element of the catena it is specified that the element does not allow any modifications. Thus, catena plus the selection-based approach is an adequate means for linguistically challenging analyses, since it combines the lexical representation (catenae) with syntactic behavior (valency). The construction-based approach does not make any difference for the strict idioms, since there is no lexical variation envisaged there. In Fig. 1 and Fig. 2 we present two sentences from BulTreeBank in which the same verb \textit{затварям} ‘to close’ is used in its literal meaning (Fig. 1) and as a part of idiomatic expression (Fig. 2). The catena is highlighted.

Note that this idiom has a possessive clitic in its structure, which sometimes comes in a raised position. However, it satisfies the clitic constraints (as

\(^3\) Here the example is in the original HPSG-based annotation which uses a constituent tree representation and head-dependent relational labels. In the examples, VPC stands for verbal phrase of type head-complement, VPS is for verbal phrase of type head-subject, and VPA is for verbal phrase of type head-adject. Later in the paper we switch to a dependency representation which exploits these head-dependent relational labels for converting the trees into dependency.

\(^4\) This kind of paraphrases is discussed further in the text. We have in mind the following situation: ‘heartbreaking’ (compound) and ‘to break one’s heart’ (syntactic paraphrase).
stated in Gross (2011c)), namely – that despite the prosodic position of the clitic (to the verb or the noun), the verb always governs either the clitic...

Figure 1. HPSG-based tree for the sentence “Гениите си затварят очите при свирене” (‘Geniuses REFL.POSS.SHORT close eyes when playing’ *Geniuses close their eyes when playing some instrument*).

Figure 2. HPSG-based tree for the sentence “Гениите си затварят очите пред дребните неща” (‘Geniuses REFL.POSS.SHORT close eyes.DEF before minor things’ *Geniuses ignore minor issues*).
itself together with its nominal head (затваря̀м си очите ‘close-I REFL. POSS.SHORT eyes.DEF’), or the nominal head which licenses the clitic (затваря̀м очите си ‘close-I eyes.DEF REFL.POSS.SHORT’).

Here is the dependency representation of the example from Fig. 2. The catena is represented by the relations - clitic and obj (red arcs).

In the lexicon we would like to represent each MWE in a way similar to single words. On the other hand, some of the wordforms in MWEs are not allowed to be changed and they have to be stored as they appear. In order to achieve this we have to store the catena in the lexical unit. Additionally, the valency of MWE is expressed either for the whole catena or for its parts. When the MWE allows for some modification of its elements – i.e. modifiers of a noun, the lexical unit in the lexicon needs to specify the role of these modifiers. For example, the canonical form of the MWE in Fig. 2 is затваря̀м си очите. Having all this in mind we define our desired representation in the lexicon as follows:

```
lexicon-catena:
   (VPC-C (V-C (V-C затваря̀м) (Pron-C си)) (N-C очите) )
semantics: ignore-the-facts_rel(e, [1]fact)
valency: < indobj; (PP (P x) (N [1]y)) : x ∈ \{ пред, за \} >
```

The specification above shows that the catena includes the elements ‘close one’s eyes’ in the sense of ‘ignore the facts’, which is presented in the semantics part as a relation. In this part the noun ‘fact’ is indicated via a structure-sharing mechanism – [1]. This is necessary, because in the valency part of the lexical unit the noun within the subcategorized PP by the catena ‘close one’s eyes’ reproduces some fact from the world. Also, if more than one preposition is possible, they are presented as a set of x-values. In
the rest of the paper we give a more precise definition of lexical entries which unifies the representation of single words (compounds), MWEs and derivational morphology.

5. Formal Definition of Catenae

In this section we define the formal presentation of catena as it is used in syntax and in 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 a syntactic tree. This definition of catena for dependency trees is equivalent to a subtree definition. Following Osborne et al. (2012) we prefer to use the notion of catena to that of dependency subtree, because some of the dependency subtrees (complete subtrees) are considered as constituents. We have to utilize the notion of catena for several purposes: representation of words and multiword expressions in lexicon, their realization in the actual trees expressing the analysis of sentences as well as for representation of derivational structure of compounds in the lexicon.

Let GF is a set of grammatical features, including parts-of-speech, POS tag is a non-empty set of grammatical features from GF. Let LA be a set of POS tags, LA sub be the set of all subset of all tag in LA (LA ⊆ LA sub), LE be a set of lemmas, WF be a set of wordforms and D be a set of dependency tags (ROOT ∈ D). For a given wf ∈ WF, a given lemma ∈ LE, and a given tag ∈ LA sub we say that they agree with each other if and only if wf is a wordform of lemma with grammatical features in tag. Similarly, wf agrees with lemma if and only if wf is a wordform of lemma; tag agrees withwf if and only if grammatical features in tag are grammatical features of wf; tag agrees with lemma if and only if there is a wordform of lemma that agrees with tag. Let x = w_1 … w_n be a sentence where w_i, 1 ≤ i ≤ n are the wordforms in the sentence (w_i ∈ WF).

A tagged dependency tree for sentence x is a directed tree T = (V, A, π, λ, ω, δ) where:

1. V = {0,1,…, n} is an ordered set of nodes, where nodes from 1 to n correspond to the wordforms w_i in the sentence x;
2. A ⊆ V × V is a set of arcs. For each node i, 1 ≤ i ≤ n, there is exactly
one arc in \( A \): \(<i, j> \in A, 0 \leq j \leq n, i \neq j \). There is exactly one arc \(<i, 0> \) in \( A \);

3. \( \pi : V \setminus \{0\} \to LA \) is a total labeling function from nodes to POS tags;

4. \( \lambda : V \setminus \{0\} \to LE \) is a total labeling function from nodes to lemmas;

5. \( \omega : V \setminus \{0\} \to WF \) is a total labeling function from nodes to wordforms of the sentence \( x \), such that \( \omega(i) = w_i \);

6. \( \pi, \lambda, \) and \( \omega \) agree with each other in the sense that their values for each node in \( V \setminus \{0\} \) agree with each other;

7. \( \delta : A \to D \) is a total labeling function for arcs. Only the arc \(<i, 0> \) is mapped to the label \( \text{ROOT} \);

8. \( 0 \) is the root of the tree.

The node \( 0 \) does not correspond to a wordform 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. Let \( G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G) \) be a directed tree such that:

1. \( V_G = \{1, \ldots, m\} \) is an ordered set of nodes;

2. \( A_G \subseteq V_G \times V_G \) is a set of arcs. For each node \( i, 1 \leq i \leq m \), there is exactly one arc in \( A_G \): \(<i, j> \in A_G, 0 \leq j \leq m, i \neq j \);

3. \( \pi_G : V_G \to LA_{\text{sub}} \) is a partial labeling function from nodes to POS tags in \( LA \), their subsets, or an empty set;

4. \( \lambda_G : V_G \to LE \) is a partial labeling function from nodes to lemmas. Some nodes might have no lemma assigned to them;

5. \( \omega_G : V_G \to WF \) is a partial labeling function from nodes to wordforms;

6. \( \pi_G, \lambda_G, \) and \( \omega_G \) agree with each other in the sense that their values for each node in \( V_G \) agree with each other;

7. \( \delta_G : A_G \to D \) is a partial labeling function for arcs;

8. \( \text{CatR} \in V_G \) is the root of \( G \).

\( G \) is a **dependency catena of \( T \)** if and only if there exists a mapping \( \psi : V_G \to V \setminus \{0\} \) such that:

---

5 In case when we are interested in part of the grammatical features encoded in a POS tag we could consider \( p \) 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 This mapping allows for embedding of \( G \) in different tagged dependency trees and thus different word order realizations of the catena nodes (corresponding to wordforms in \( T \)). The mapping \( \psi \) is specific for \( G \) and \( T \). It allows also the image of \( G \) in \( T \) not to be a subtree of \( T \). A special case is discussed below.
1. For each arc \(<n_i, n_j> \in A_G \rightarrow <\psi(n_i), \psi(n_j)> \in A\). The image of \(G\) in \(T\) under \(\psi\) cannot introduce new arcs to \(G\), or delete arcs from \(G\);

2. For each node \(n \in V_G \rightarrow \pi_G(n) \subseteq \pi(\psi(n))\). The grammatical features assigned to nodes in \(G\) form a subset of complete POS tags assigned to the corresponding nodes in \(T\). \(\pi_G(n)\) could be empty;

3. For each node \(n \in V_G \rightarrow \lambda_G(n) = \lambda(\psi(n))\). The lemmas presented in \(G\) are the same as lemmas in \(T\). Some nodes in \(G\) could be without assigned lemmas;

4. For each node \(n \in V_G \rightarrow \omega_G(n) = \omega(\psi(n))\). The wordforms presented in \(G\) are the same as wordforms in \(T\). Some nodes in \(G\) could be without assigned wordforms;

5. For each arc \(<n_i, n_j> \in A_G \rightarrow \delta_G(<n_i, n_j>) \text{ is defined then } \delta_G(<n_i, n_j>) = \delta(<\psi(n_i), \psi(n_j)>)\). The arc labels in \(G\) have to agree with the labels in \(T\). Some arcs in \(G\) could be without assigned arc label.

The subgraph of \(T\) determined by the nodes \(\psi(n)\), such that \(n \in V_G\) is called an image of \(G\) under \(\psi\) in \(T\), denoted as \(\psi(G)\). The image of \(G\) in \(T\) will be called realization of \(G\) in \(T\).

Here we defined catena with respect to a tagged dependency tree \(T\). In this way catena corresponds to a subtree of \(T\). In this way the graph \(G\) has a linguistic meaning in the sense of corresponding to a subset of a syntactic tree. In the next definition we define catena independently from a concrete tagged dependency tree.

A directed tree \(G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G)\) as defined above is a dependency catena (or just catena) if and only if there exists a dependency tree \(T = (V, A, \pi, \lambda, \omega, \delta)\) such that \(G\) is a dependency catena of \(T\).

Having partial functions for assigning POS tags, dependency labels, wordforms and lemmas allows us to construct arbitrary abstractions over the structure of catena. Thus, the catena could be underspecified for some of the node labels like grammatical features, lemmas and also some dependency labels. The mapping \(\psi\) parameterizes the catena with respect to different dependency trees. This freedom provides mechanisms for using catenae on different levels of abstraction. Such catenae will encode only the linguistic information that is obligatory for the encoded unit — a word or a phrase. Additional constraints could be added in the grammar. Different mappings provide a possibility for different word order realizations of the
same catena, for example. The excluding of node 0 from the range of the mapping \( \psi \) excludes the external root of the tagged dependency tree from each catena.

![Diagram](image)

**Figure 3:** Two realizations of the lexicon catena “затварям си очите” (ignore the facts) in two sentences: “Очите си затваряха пред фактите” (They ignored the facts) and “Иван си затваряше очите” (Ivan ignored [the facts]). Here we show also the root nodes of the two sentences. In the next examples we will not represent them when they are not crucial for these examples.

As it was defined above, the image of a catena in a given dependency tree is called **realization of the catena in the tree**. The realization of the catena is a fully specified subtree including all node and arc labels. For example, the catena for “spill the beans” will allow for any realization of the verb form like in: “they spilled the beans” and “he spills the beans”. Fig. 3 depicts two realizations (with different word orders) of the catena for the Bulgarian idiom затварям си очите (lit. close one’s eyes).

Let \( G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G) \) and \( G' = (V_{G'}, A_{G'}, \pi_{G'}, \lambda_{G'}, \omega_{G'}, \delta_{G'}) \) be two dependency catenae. We will say that \( G \) is a subcatena of \( G' \) if and only if there exists a mapping \( \psi : V_G \rightarrow V_{G'} \) such that:

1. For each arc \( <n_i, n_j> \in A_G \rightarrow <\psi(n_i), \psi(n_j)> \in A_{G'}; \)
2. For each node \( n \in V_G \rightarrow \pi_G(n) \subseteq \pi_{G'}(\psi(n)). \pi_G(n) \) could be empty;

---

7 We do not represent all the information for all the nodes in the lexicon catena, but only wordforms and lemma for the root. The wordform for the root of the catena will be different in different realizations.
3. For each node $n \in V_G \rightarrow$ if $\lambda_G(n)$ is defined, then $\lambda_G(n) = \lambda_{G'}(\psi(n))$;
4. For each node $n \in V_G \rightarrow$ if $\omega_G(n)$ is defined, then $\omega_G(n) = \omega_{G'}(\psi(n))$;
5. For each arc $<n_i, n_j> \in A_G \rightarrow$ if $\delta_G(<n_i, n_j>)$ is defined, then $\delta_G(<n_i, n_j>) = \delta_{G'}(<\psi(n_i), \psi(n_j)>)$.

The image of $G$ into $G'$ under $\psi$ is called realization of $G$ in $G'$.

Each catena might play a different linguistic role. The catena in the lexicon is underspecified with respect to the grammatical features and wordforms of the verb. This underspecified catena will be called a lexicon catena, since we assume that lexicon entries contain catenae for the defined lexical units. Some catenae play role in grammar for forming verbal forms. Such type of catena will be called for convenience auxiliary catena in this paper although it could cover linguistic constructions of different kinds (auxiliary, modal, control, etc.) depending on the verb forms. Thus under auxiliary catena we understand a catena that is part of the verbal complex and contains nodes for the auxiliary verbs. In the grammars for different languages different kinds of catenae could be defined on the basis of their role in the grammar.

As it was shown above in the examples of idioms with an explicit subject like: the devil is in the details, the realizations from the lexicon catena to syntax often are accompanied by intervening material – see the discussion in Osborne et al. (2012). Our insight, supported by the examples, is that the intervening material forms an auxiliary catena of a certain type. In order to implement this idea we need some additional notions, which follow.

Let $G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G)$ be a catena and $n \in V_G$ then $\{G_1, G_2, ..., G_m\}$ is a partition of $G$ on node $n$ if and only if:

1. for each $i, 1 \leq i \leq m$, $G_i$ is catena which is a subcatena of $G$;
2. for each node $n' \in V_G$, such that $n' \neq n$, there is exactly one catena $G_i, 1 \leq i \leq m$, such that $n'$ is an image of a node in $G_i$ under the corresponding

---

8 There are also similar kinds of MWEs in Bulgarian like рибата се вмирисва откъм главата ‘fish-DEF REFL rots from head-DEF’ (Its meaning is that corruption starts from people in power).

9 In this respect the definition of extension here is restricted to verbal complex, but easy could be adapted for other cases when necessary.
mapping $\psi_i$;

3. at most one subcatena $G_i$, $1 \leq i \leq m$, has a leaf node whose image in $G$ is $n$;

4. the root of each subcatena $G_i$, $1 \leq i \leq m$, is mapped to the node $n$, except for one catena for which a leaf node is mapped to $n$, if there is such a catena;

5. the mappings $\pi_{Gi}, \lambda_{Gi}, \omega_{Gi}, \delta_{Gi}$ are the same as the mappings for catena $G$ for all nodes without the nodes mapped to the node $n$ where the mappings $\pi_{Gi}, \lambda_{Gi}, \omega_{Gi}$ could be partial with respect to the original mappings.

**Figure 4.** Partition of the catena for “The devil is in the details” into two catenae: one for the subject catena and one for the indirect object catena. All the labeling mappings are the same as in the original catena except for the node $n^{10}$. In the example, the node $n$ is for the copula. In the original catena the node is connected to the lemma ‘be’, the wordform ‘is’ and part of speech ‘V’. In the subject catena the node $n$ mappings are made underspecified with respect to wordform and lemma. Thus, the mapping is on the level of part of speech — V. This fact is denoted by a box marked by V. In the indirect object catena the node $n$ mappings are the same as in the original catena except for the wordform and the grammatical features. This fact is denoted by a box marked by the lemma ‘be’.

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 $G_i$ 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

---

$^{10}$ The node on which the partition is performed.
node of $G_a$ is mapped. Each node in $G_c$ is an image of a node from $G_1, G_2, \ldots, G_n$ or $G_a$. An example is presented in Fig. 5.

Let $G_1$ and $G_2$ be two catenae. A composition of $G_1$ and $G_2$ is a catena $G_c$ such that the catenae $G_1$ and $G_2$ are realized in $G_c$ in such a way that the root node CatR$_2$ of $G_2$ is mapped to a node in $G_c$ to which a node of $G_1$ is mapped. Each node in $G_c$ is an image of a node from $G_1$ or $G_2$. An example is presented in Fig. 6.

**Figure 5.** In the figure the extension of the catena “The devil is in the details” with the auxiliary catena for ‘will’ is shown. Here the root nodes of the catenae are not marked explicitly, but they could be easily identified.

**Figure 6.** In the figure the concatenation of the catena “I read” and the catena “a book” as an object of a verb is shown. The fact that the catena “a book” is an object is represented via an underspecified node.

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. If for a node in a catena the lemma and the wordform are not specified, the number of the catena realizations could be infinite in number. In this sense the definition for equivalence of catena is not applicable in practice for determining the
The alternative definition is based on the realization of a catena into another catena. Two catenae $G_1$ and $G_2$ are **equivalent** if there is a realization of $G_1$ into $G_2$ and there is a realization of $G_2$ into $G_1$.

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 for its class of equivalence.

Using the notions presented in this section we define the structure of lexical items in the lexicon and how they are realized in actual sentence analyses.

In the rest of the paper we represent dependency trees in CoNLL 2006 shared task format with the necessary changes. This format is a table format where each node in the dependency tree (except for the root node 0) is represented as a row; the cells in a row are separated by a tabulation symbol. The fields are: Number, WordForm, Lemma, POS, ExtendedPOS, GrammaticalFeatures (in a form of attribute value pairs, attr=v, separated by a vertical bar), parent node, and dependency relation. In the paper we do not use columns 9 and 10 in the way they were used in the CoNLL 2006 format. Here column 9 is used for annotation of the node as being part of the catena for a given MWE. The rows that represent the nodes belonging to a catena are marked with the same identifier. If a node is not part of a catena for MWE, column 9 of the corresponding row contains an underscore symbol.

Since a sentence might contain more than one MWE, the catena for each one is numbered in a different way. We do not allow any MWE overlapping.

Let $T = (V, A, \pi, \lambda, \omega, \delta)$ be a tagged dependency tree:

- The nodes of $V - \{0\}$ are represented in the first cell of each row in the table. This is column No in the table;
- For each arc $<d, h> \in A$, the head node $h$ is represented in cell 7 of the row for node $d$. This is column Head in the table;
- For each node $n \in V - \{0\}$, the value $\pi(n)$ is represented in cells 4, 5, and
6 of the row for node n. These are columns POS, ExPOS, and GramFeat in the table;

• For each node \( n \in V - \{0\} \) the value \( \lambda(n) \) is represented in cell 3 of the row for node n. This is column Le in the table;

• For each node \( n \in V - \{0\} \) the value \( \omega(n) \) is represented in cell 2 of the row for node n. This is column Wf in the table;

• For each arc \( < d, h > \in A \) the label \( \delta(< d, h >) \) is represented in cell 8 of the row for node d. This is column Rel in the table.

• The root 0 is not represented in the table.

<table>
<thead>
<tr>
<th>No</th>
<th>Wf</th>
<th>Le</th>
<th>POS</th>
<th>ExPOS</th>
<th>GramFeat</th>
<th>Head</th>
<th>Rel</th>
<th>Catena</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>те</td>
<td>те</td>
<td>P</td>
<td>Pp</td>
<td>number=pl</td>
<td>case=nom</td>
<td>3</td>
<td>subj</td>
</tr>
<tr>
<td>2</td>
<td>си</td>
<td>си</td>
<td>P</td>
<td>Pp</td>
<td>form=possessive</td>
<td>3</td>
<td>clitic</td>
<td>el</td>
</tr>
<tr>
<td>3</td>
<td>затварят</td>
<td>затварям</td>
<td>V</td>
<td>Vpi</td>
<td>number=pl</td>
<td>person=3</td>
<td>0</td>
<td>Root</td>
</tr>
<tr>
<td>4</td>
<td>очите</td>
<td>око</td>
<td>N</td>
<td>Nc</td>
<td>number=pl</td>
<td>definiteness=y</td>
<td>3</td>
<td>obj</td>
</tr>
<tr>
<td>5</td>
<td>пред</td>
<td>пред</td>
<td>R</td>
<td>R</td>
<td>_</td>
<td>3</td>
<td>indobj</td>
<td>_</td>
</tr>
<tr>
<td>6</td>
<td>истината</td>
<td>истина</td>
<td>N</td>
<td>Nc</td>
<td>number=sg</td>
<td>definiteness=y</td>
<td>5</td>
<td>prepobi</td>
</tr>
</tbody>
</table>

In the table it can be seen that three elements are part of the catena: си затварят очите (positions 2, 3 and 4) “they close their eyes” (they ignore). In this way, the idiomatic meaning of the expression is ensured. Thus, each MWE in a dependency tree is represented via its realization. This representation of MWEs is convenient for dependency trees in dependency treebanks on analytical (or surface) level of dependency analysis.

In order to model the behavior of lexical units in a better way we need to add semantics to the dependency representation. We do not do this exhaustively in this paper, but we only suggest a direction to go. In order to represent the MWEs in the lexicon, we assume a semantic analysis based on Minimal Recursion Semantics (MRS) (see Copestake et al. (2005)). MRS is used to support semantic analyses in HPSG English grammar – ERG (Copestake and Flickinger, 2000), but also in other grammar formalisms like LFG. The main idea is the formalism to rule out spurious analyses resulting from the representation of logical operators and the scope of quantifiers. Here we will present only basic definitions from (Copestake et
al., 2005). For more details the cited publication should be consulted. An MRS structure is a tuple $<GT, R, C>$, where $GT$ is the top handle, $R$ is a bag of EPs (elementary predicates) and $C$ is a bag of handle constraints, such that there is no handle $h$ that outscopes $GT$. Each elementary predication contains exactly four components: (1) a handle which is the label of the EP; (2) a relation; (3) a list of zero or more ordinary variable arguments of the relation; and (4) a list of zero or more handles corresponding to scopal arguments of the relation (i.e., holes). Here is an example of an MRS structure for the sentence “Every dog chases some white cat.”

$$<h0, \{h1: every(x,h2,h3), h2: dog(x), h4: chase(x, y), h5: some(y,h6,h7), h6: white(y), h6: cat(y)\}, {}>$$

The top handle is $h0$. The two quantifiers are represented as relations $every(x, y, z)$ and $some(x, y, z)$ where $x$ is the bound variable, $y$ and $z$ are handles determining the restriction and the body of the quantifier. The conjunction of two or more relations is represented by sharing the same handle ($h6$ above). The outscope relation is defined as a transitive closure of the immediate outscope relation between two elementary predications – EP immediately outscopes EP’ iff one of the scopal arguments of EP is the label of EP’. In this example the set of handle constraints is empty, which means that the representation is underspecified with respect to the scope of both quantifiers. Here we finish with the brief introduction of the MRS formalism. In the rest of the paper we exploit only the set of elementary predicates of the MRS structure introduced by the corresponding lexical entry. Another simplification is that we do not use handles in the examples in the paper. Although they are important for the complete formalization, in our view, they only will complicate the examples without providing better understanding. For dependency analyses the MRS structures are constructed in a way similar to the one presented in Simov and Osenova (2011). In this work, the root of a subtree of a given dependency tree is associated with the MRS structure corresponding to the whole subtree. This means that for the semantic interpretation of MWEs we use the root of the corresponding catena. In the dependency tree for the corresponding sentence the catena root provides the interpretation of the MWE and its dependent elements, if any. In the lexicon we provide the corresponding structure to model the idiosyncratic semantic content of MWE.
A similar representation is used for the dependency structure of the morphological elements of words. The main difference is that the nodes of the tree are morphs and the lemmas are substituted by morphemes. Below the representation of the deverbal noun подписване (signing) is given.

<table>
<thead>
<tr>
<th>No</th>
<th>Morph</th>
<th>Morpheme</th>
<th>ExPOS</th>
<th>MorphT</th>
<th>GramFeat</th>
<th>Head</th>
<th>Rel</th>
<th>Catena</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>под</td>
<td>_</td>
<td>_</td>
<td>pref</td>
<td>_</td>
<td>2</td>
<td>pref</td>
<td>_</td>
</tr>
<tr>
<td>2</td>
<td>пис</td>
<td>_</td>
<td>_</td>
<td>vroot</td>
<td>_</td>
<td>3</td>
<td>root</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>в</td>
<td>_</td>
<td>_</td>
<td>imsufs</td>
<td>aspect=imp</td>
<td>4</td>
<td>suf</td>
<td>_</td>
</tr>
<tr>
<td>4</td>
<td>ан</td>
<td>_</td>
<td>_</td>
<td>dnsuff</td>
<td>_</td>
<td>5</td>
<td>suf</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>е</td>
<td>e</td>
<td>Nc</td>
<td>ending</td>
<td>gender=n</td>
<td>number=sg</td>
<td>definiteness=n</td>
<td>0</td>
</tr>
</tbody>
</table>

Here the wordform is a concatenation of morphs. The part-of-speech of the whole form is specified for the root morph. Also the grammatical features of the form are given for the root morph.

The CoNLL format of catenae is used in the rest of the paper to represent the elements of lexical entries.

6. Structure of Lexical Entry and Representation of MWEs

The current valency lexicon of Bulgarian, constructed on the basis of syntactic analyses in BulTreeBank, includes information about the main form (lemma) of the word, the valency frame with all the elements, their forms, grammatical features and semantics. The lexical entry for each lexical item also includes the semantics of the main form and information on how this semantics incorporates the semantics of each frame element.
Here we present the structure of the lexical entry for the verb ‘бежам’ (run) in the sense “ignore the facts”. The verb takes an indirect object in the form of prepositional phrase starting with preposition ‘от’ (from).

In this model we use catena for representation of a single word and MWE, because by definition single words are also catenae. Using the formal definition of catena from above, we can specify all grammatical features of the lexical item. The semantics in the lexical entry attaches to each node in the lexicon-catena. In this example, there is just one node of the lexicon-catena. In the example, the verb introduces three elementary predicates: ignore_rel(e, x₀, x₁), fact(x₁), [1](x₁). The predicate ignore_rel(e, x₀, x₁) represents the event and its main participants: x₀, x₁. The predicate fact(x₁) is part of the meaning of the verb in the sense that the agent represented by x₀ ignores some fact. There is also one underspecified predicate [1](x₁) which has to be compatible with the predicate fact(x₁).

Compatibility of two predicates is defined with respect to a hierarchy of predicates: (P, ⊆), where ⊆ is a partial order over P. Two predicates p and q are compatible if and only if there exists a predicate r such that r ⊆ p and r ⊆ q. This predicate is used for incorporating the meaning of the indirect object. The valency frame is given as a set of valency elements. They are defined as a catena and semantic description. The catena describes the basic structure of the valency element including the necessary lexical information, grammatical features, the syntactic relation to the main lexical item. The semantic description determines the main semantic contribution of the frame element and via structural sharing it is incorporated in the semantics of the whole lexical item. In the example there is only one frame element. It is introduced via the preposition ‘от’ (from). The semantics comes from the dependent noun which has to be compatible with fact(x) predicate and via the underspecified predicate [1](x₁) which
could specify a more concrete predicate. Via the structure sharing index [1] this specific predicate is copied to the semantics of the main lexical item.

The lexical entry of a MWE uses the same format: a **lexicon-catena**, **semantics** and **valency**. The lexicon-catena for the MWEs is stored in the CoNLL format as described above. The semantics part of a lexical entry specifies the list of elementary predicates for the MRS analysis. When the MWE allows for some modification (also adjunction) of its elements, i.e. modifiers of a noun, the lexical entry in the lexicon needs to specify the role of these modifiers. For instance, the MWE from the example ‘затварям си очите’ (close one’s eyes) which is synonymous to the verb ‘бягам’ (run) in the sense “ignore the facts” presented above, is encoded as follows:

```
[ 
  lexicon-catena:

    No Wf Le POS ExPOS GramFeat   Head   Rel
  CNo1 _  затварям V Vpi _       CNo0   CRoot
  CNo2 си _  си P Pp _          CNo1   elitic
  CNo3 очите _  око N Nc _       CNo1   obj

  semantics:
  CNo1:  { ignore_rel(e, x0, x1), fact(x1), [1](x1) }

  valency:
  { < }

    No Wf Le POS ExPOS GramFeat   Head   Rel
    No1 пред _  пред R R _         CNo1   indobj
    No2 _ _ _ N Nc _              No1   pobj

  semantics: No2:  { fact(x), [1](x) } > ,

  <

    No Wf Le POS ExPOS GramFeat   Head   Rel
    No1 за _  за R R _           CNo1   indobj
    No2 _ _ _ N Nc _             No1   pobj

  semantics: No2:  { fact(x), [1](x) } >

  ]
```

The lexical entry is similar to the one shown earlier. The main differences are: the lexicon-catena is for a MWE instead of a single word. The semantics is the same, because the verb and the MWE are synonyms.
The valency frame contains two alternative elements for indirect object introduced by two different prepositions. The situation that the two descriptions are alternatives follows from the fact that the verb has no more than one indirect object. If there is also a direct object then the valency set will contain elements for it as well.

In many languages the elements represented in the valency are not necessarily realized, i.e. they are optional. This is the case in Bulgarian — the objects and indirect objects of a verb might remain unexpressed. In such cases the semantics from the unrealized valency element represented by the underspecified predicate is assumed to be the most general one-like everything(x). In these cases the semantics is defined by the default predicate for the valency element. In the above example the default predicate is fact(x). In this way, the predicate assigned to the structure-sharing identifier [1] above will ensure a correct interpretation of the semantics expressed in the lexical entry for the multiword expression.

Since in the catena representation cell 9 is empty, it is not given in the lexicon table. The semantics and the valency information are attached to the corresponding nodes in the catena representation. In the example above only the information for the root node of the catena is given (node number 1 — CNo1). In cases when other parts of the catena allow modification, the information for the corresponding nodes is given. Here we provide examples of such cases. For example, the Multiword Expression ‘среща на върха’ (summit) allows for modification not only of the whole catena, but also of the noun within the prepositional phrase. The lexical entry from the lexicon is given as follows:

```
[  
lexicon-catena:

<table>
<thead>
<tr>
<th>No</th>
<th>Wf</th>
<th>Le</th>
<th>POS</th>
<th>ExPOS</th>
<th>GramFeat</th>
<th>Head</th>
<th>Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNo1</td>
<td>среща</td>
<td>N</td>
<td>Nc</td>
<td></td>
<td></td>
<td>0</td>
<td>CRoot</td>
</tr>
<tr>
<td>CNo2</td>
<td>на</td>
<td>R</td>
<td>R</td>
<td></td>
<td></td>
<td>CNo1</td>
<td>mod</td>
</tr>
<tr>
<td>CNo3</td>
<td>върха</td>
<td>връх</td>
<td>N</td>
<td>Nc</td>
<td>number=sg</td>
<td>definiteness=y</td>
<td>CNo2</td>
</tr>
</tbody>
</table>

semantics:

CNo1: {meeting_rel(e,x),member(y,x),head-of-a-country(y,z),country(z), [1](z)}

valency:
```
This lexical entry allows modifications like ‘европейски’ (European) — среща на европейския връх (meeting of the European top). This catena allows also modification of the head word, such as ‘успешна среща на върха’ (successful summit). This example also demonstrates that the semantics could include several elementary predicates that describe a complex event.

The last example presented here is for the multiword ‘снежен човек’ meaning “a man-like sculpture from snow” (snowman). It does not allow any modification of the dependent node ‘снежен’ (snowy), but it allows for modifications of the root like “large snow man” etc. The lexical entry from the lexicon is given as follows:

```
 lexicon-catena:

<table>
<thead>
<tr>
<th>No</th>
<th>Wf</th>
<th>Le</th>
<th>POS</th>
<th>ExPOS</th>
<th>GramFeat</th>
<th>Head</th>
<th>Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNo1</td>
<td>_</td>
<td>снежен</td>
<td>A</td>
<td>A</td>
<td>_</td>
<td>CNo2</td>
<td>mod</td>
</tr>
<tr>
<td>CNo2</td>
<td>_</td>
<td>човек</td>
<td>N</td>
<td>Ne</td>
<td>definiteness=n</td>
<td>0</td>
<td>CRoot</td>
</tr>
</tbody>
</table>

semantics:
CNo2: { snowman_rel(x) }

valency: {} 
```

The grammatical features for the head noun (definiteness=n) restricts its possible form. In this way, singular and plural forms are allowed. The empty valency ensures that the dependent adjective cannot be modified except for morphological variants like singular and plural forms, but also definite or indefinite forms depending on the usage of the phrase. The possible modifiers of the MWE are determined by the represented semantics. The relation `snowman_rel(x)` is taken from an appropriate ontology (appropriate
ontology could be WordNet, SUMO, etc.) where its conceptual definition is given.

These examples demonstrate the power of the combination of catenae (as syntactic units), MRS structures (as semantic units) and valency representation (as subcategorization units) to model MWEs in the lexicon. The catena is appropriate for representation of syntactic structure and variation on morphological level; the semantic part represents the idiosyncratic semantics of the MWE and determines the possible semantic modification, and the valency part determines the syntactic behavior of MWEs. One missing element of the lexical entry is a representation of constraints over the word order of the catena nodes. We envisage addition of such constraints as future work. The information from the lexical entries is combined by different grammar operations on them. For example, the position of the clitic in the above example is determined by the second clitic position in Bulgarian. The main operation on catenae is the realization in dependency trees. The two other operations are extension and composition of catenae. The extension is used when an MWE catena needs to be realized together with an auxiliary catena as in the case of sentence MWEs where the subject catena is detached from the verbal catena and realized as a subject of the auxiliary catena (see Fig. 5). The composition is used when the valency catena is realized with the main lexical catena (see Fig. 6).

7. From Syntactic Relations to Compound Morphology

The catena approach is also very appropriate for modeling the connection among compounds and their syntactic counterparts in Bulgarian. The compounds in question are, as mentioned before, one-word expressions with two roots and additional affixes/inflections. Most of the examples have the following root sequence: Noun + Verb11.

In Gross (2011a) the notion of ‘morph catena’ has been explicitly

---

11 Compare with differing compound types in English (Gross 2011b: p. 51): Verb + Noun; Verb + Particle; Particle + Particle and Auxiliary + Verb.
12 Such as, histor-ic-al novel-ist where the morpheme ’ist' dominates the rest of the morphemes, thus resolving the bracketing paradoxes of the type [historical [novel-ist]] and [[historical novel]-ist].
introduced. By granting a node to each morph\(^1\), the author makes the problematic morph a dominant element over the other depending morphs. Thus, all these morphs are under its scope and the catena set includes also the intended meaning.

In this paper two situations are considered: a) compound deverbal noun whose counterpart can be expressed only through a free syntactic phrase (N (compound) and VP (phrase)): " билколечение " ‘herbcuring’ curing by herbs, *" билколекувам *’herbcure.1PERS.SG’ (to cure with herbs) and " лекувам с билки " ‘cure.1PERS.SG with herbs’ (to cure with herbs); and b) compound deverbal noun whose verbal counterpart can be either a compound too, but verbal, or a free syntactic phrase (N (compound) and V (compound) or VP (phrase)): " ръкомахане " ‘handwaving’ (gesticulating); " ръкомахам " ‘handwave.1PERS.SG’ (gesticulate); and " максам с ръка " ‘wave with hand’, (gesticulate).

A previously done survey in Osenova (2012) performed over an extracted data from a morphological dictionary (Popov et al. 2003) shows that in Bulgarian head-dependent compounds are more typical for the nominal domain (with a head-final structure), while the free syntactic phrasing is predominant in the verbal domain. Also, regarding the occurrence of dependents in the compounds, subject is rarely present in the verbal domain, while complements and adjuncts are frequent – "гласоподавам " ‘votegive.1PERS.SG’ (vote) – where ‘vote’ is a complement of ‘give’. On the contrary, in the nominal domain also subjects are frequently present, since they are transformed into oblique arguments - "снеговалеж " ‘snowfall’ (snowing).

Irrespectively of the blocking on some compound verbs, there is a need to establish a mapping between the nominal compound and its free syntactic phrase counterpart. Both expressions are governed by the selection-based rules. Thus, the realization of the dependents in the syntactic phrases relies on the valency information of the head verb only, while the realization of the dependents in the nominal or verb compounds respects also the compound-building constraints.

A mechanism is needed which relates the external syntactic representations with the internal syntax of the counterpart morphological compounds. Some ideas on the catena presentation of intra-word and inter-
word dependencies are presented in Gross (2011b) and Gross (2011c). Thus, the intra-word dependency is defined in the following way: A morph M1 is an intra-word dependent of another morph M2, if the morph combination M1-M2 distributes more like an M2-type than like an M1-type unit (Gross 2011c: p. 59).

For our purposes, we will use a morph-based catena analysis inside the complex word and a syntactic-based catena analysis in the syntactic phrases. However, we simplify the approach in discussing relations among the roots only rather than taking into account also the affixes. We assume that the morph-based analysis with affixes applies straightforwardly to the discussed compounds.

When moving from compounds to free syntactic phrases, some external arguments which are missing in the compound structures may well appear in the free syntactic phrases, such as: ръкомахам с лявата ръка ‘handwave.1PERS.SG with left.DEF hand’, (I am gesticulating with my left hand), where the complement ръка (hand) is further specified and for that reason is explicitly present. Thus, we can imagine that in the lexicon we have the deverbal noun compounds as well as verbal compounds, presented via morph catena and with specified head-dependent relations. The corresponding parts of the words are then connected to the heads of the corresponding syntactic phrases (in the valency lexicon) and their dependents, but this time the relations are presented via a syntactic catena tree. We can think of the morphological catena as a rather fixed one, while of the syntactic catena as a rather flexible one, since it would allow also additional arguments or modifiers in specific contexts.

Let us see in more detail how this mapping will be established. The first case is the one where the deverbal nominal compound connects directly to a syntactic phrase (with no grammatical verb compound counterpart). The morph catena will straightforwardly present the tree of: билк-о-леч-ен-ие ‘herb-linkingMorph-cure-affix-inflection’ where each morph from left to right dominates the previous one.

However, in the syntactic catena a preposition is inserted according to the valence frame of the verb лечувам (cure): лечувам с билки ‘cure.1PERS.
SG with herbs’, to cure with herbs). Using a catena, we can safely connect
the non-constituent phrase лекувам с (cure with) to the root morpheme
of the head in the compound – леч (cure-). Also, all the possible modifiers
of билки (herbs) in the syntactic phrase would be connected to the head
morpheme билк (herb-).

The second case is the one where the nominal compound has mappings
to both counterparts – a verb compound and a syntactic phrase. The
connection between the noun and verb compounds is rather trivial, since
only the inflections differ:

- ръкомахае ‘handwaving’ (gesticulating);
- ръкомахам ‘handwave.1PERS.SG’ (gesticulate); and
- махам с ръка ‘wave with hand’ (gesticulate).

Here is the representation of the lexical unit for the compound noun:
билколечение ‘herbcuring’ (curing by herbs):

<table>
<thead>
<tr>
<th>lexicon-catena:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>MNo1</td>
</tr>
<tr>
<td>MNo2</td>
</tr>
<tr>
<td>MNo3</td>
</tr>
<tr>
<td>MNo4</td>
</tr>
<tr>
<td>MNo5</td>
</tr>
<tr>
<td>MNo6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>derivational template:</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>CNo1</td>
</tr>
<tr>
<td>CNo2</td>
</tr>
<tr>
<td>CNo3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>semantics:</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNo1: { [3]cure_rel(e,x,y,z), [4]herbs_rel (z) }</td>
</tr>
<tr>
<td>&gt;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>semantics:</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNo6: { [3]cure_rel(e,x,y,z), [4]herbs_rel (z), <a href="z">5</a>, nominal_rel(e) }</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>valency:</th>
</tr>
</thead>
<tbody>
<tr>
<td>{ &lt;</td>
</tr>
</tbody>
</table>
In this example we present two relations. First, the morph catena (lexicon-catena) is presented with its morph structure. Then, the derivational template reflecting the derivational syntactic phrase (catena) and its semantics is shown. The correspondences on morph level are marked with tags [1] and [2]. The second relation is at the semantic level, where the semantics of the syntactic catena {{\text{cure_rel}(e,x,y,z), \text{herbs_rel}(z)}} is represented fully in the semantics of the compound noun via the indices [3] and [4]. Additionally, the event is nominalized by the predicate nominal_rel(e). In the valency list we would have a PP modifier (corresponding to the indirect object in the verb phrase) of the compound only if the preposition is ‘с’ (by), the head noun of the preposition complement is the same as the noun in the verbal phrase ‘билки’ (herbs) and there is at least one modifier of the noun. Thus, phrases like: ‘билколечение с пресни билки’ (‘herbcuring with fresh herbs’ curing with fresh herbs) and ‘билколечение с билки, които са събрани през нощта’ (‘herbcuring with herbs that are collected during the night’ curing with herbs that were collected at night) are allowed. The semantics of the noun modifiers are added to the compound noun semantics via the label [5]. But phrases with duplicate internal and external arguments like *‘билколечение с билки’ (‘herbcuring with herbs’ curing with herbs)
are not allowed. Many of the other details are left out here in order to put the focus on the important relations.

As a result from the examples, discussed above, we propose the combination of the catena approach with the selection-based one for encoding the information at the lexicon-syntax interface. This means that all the words, MWEs and compounds are stored in an extended valency lexicon using catenae with appropriate mappings among them. The morph catenae remain steady, while the syntactic ones are flexible in the sense that they also encode the predictive power of adding new material. These observations are in line with Osborne et al. (2012: 374) where it is stated that catenae are stored in the lexicon but might be interrupted in texts by various syntactic processes.

When connectors (such as prepositions) are added, the prediction is easy due to the information in the valence lexicon. However, when some modifiers come into play, the prediction might become non-trivial and difficult for realization.

8. Conclusions and Future Work

The paper shows on Bulgarian data that the modeling on the level of catena is appropriate for encoding language units (including multiword expressions and compounds) at the lexicon-syntax interface. Catena together with the selection-based approach ensures mappings between the expressions which have paraphrases. While at the morphological level the catena is stable due to the rigid structure of the word, in syntax domain it allows additional material to be inserted, based on the information from valence lexicons and contexts. Additionally, semantics component is added (in our case this is MRS) for ensuring the correct interpretation of the language units.

The paper confirms the conclusions from previous works that catena is a possible means for encoding idioms and idiosyncratic language material. With respect to idioms it is very useful for cases where in addition to the figurative meaning the literal meaning also remains a possible interpretation.

The novelties in our survey are as follows: formalization of the catena; lexicon and syntax representation of language units; establishing relations between compounds and their syntactic paraphrases. We suppressed the
morph-based analysis to the roots only, since the focus of the paper is on the connection between the valency lexicon and the syntactic behavior of the lexical units. However, a morph-based lexicon for Bulgarian, or any other language, would be really helpful for the purposes of analysis and generation.

Our future work envisages the implementation of the proposed representations in the valency lexicon and the syntactic treebank, and extending the number of related compound-to-phrases counterparts in the lexicon; also extending the semantic component beyond the participation of the elementary predicates only.

Acknowledgements

This research has received support by the EC’s FP7 (FP7/2007-2013) under grant agreement number 610516: “QTLeap: Quality Translation by Deep Language Engineering Approaches” and by European COST Action IC1207: “PARSEME: PARSing and Multiword Expressions. Towards linguistic precision and computational efficiency in natural language processing.”

References

Gross, Thomas. 2010. Chains in syntax and morphology. In Otooguro, Ishikawa,


Stoyanova, Ivelina. 2010. Factors influencing the performance of some methods for