Treatment of Multiword Expressions and Compounds in Bulgarian

Petya Osenova and Kiril Simov
Linguistic Modelling Deparment, IICT-BAS
Acad. G. Bonchev 25A, 1113 Sofia, Bulgaria
petya@bultreebank.org and kivs@bultreebank.org

Abstract

The paper shows that catena representation together with valence information can provide a good way of encoding Multiword Expressions (beyond idioms). It also discusses a strategy for mapping noun/verb compounds with their counterpart syntactic phrases. The data on Multiword Expression comes from BulTreeBank, while the data on compounds comes from a morphological dictionary of Bulgarian.

1 Introduction

Our work is based on the annotation of Multiword Expressions (MWE) in the Bulgarian treebank — BulTreeBank (Simov et al., 2004). We use this representation for parsing and analysis of compounds. BulTreeBank exists in two formats: HPSG-based (original - constituent-based with head annotation and grammatical relations) and Dependency-based (converted from the HPSG-based format). In both of them the representations of the various kinds of Multiword Expressions is an important problem. We need a mechanism for connecting the MWE in the lexicon with their actual usages within the sentences. As an interesting case of MWE at the interface of morphology and syntax we consider Compounds. They are usually derived from several lexical units, have an internal structure with a specified derivation model and semantics. In the paper we are especially interested in the mapping among deverbal compounds and their counterpart syntactic phrases.

Since there is no broadly accepted standard for Multiword Expressions (see about the various classifications in (Villavicencio and Kordoni, 2012)), we will adopt the Multiword Expressions classification, presented in (Sag et al., 2001). They divide them into two groups: lexicalized phrases and institutionalized phrases. 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 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 of MWE when necessary. The 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 - catena - for their representation and analysis. Catena is a path in the syntactic or morphemic analysis that is continuous in the vertical dimension. Its potential is discussed further in the text. The paper is structured as follows: In the next section a brief review of previous works on catena is presented. In Section 3 a typology of the Miltiword Expressions in BulTreeBank is outlined. Section 4 considers possible approaches for

consistent analyses of MWE. Section 5 introduces the relation of syntax with compound morphology. Section 6 concludes the paper.

2 Related works on catena

The notion of catena (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 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. In recent years the notion of catena revived again and it was applied also to dependency representations. Catena is used successfully for modelling of problematic language phenomena.

(Gross, 2010) presents the problems in syntax and morphology that have led to the introduction of the subconstituent catena level. Constituency-based analysis faces non-constituent structures in ellipsis, idioms, verb 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 advocated his approach on providing a surface-based account of the non-constituent phenomena via the contribution of catena. Here the author 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) presents 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 catena in NLP applications is additional motivation for us to use it in the modeling of an interface between the valence lexicon, treebank and syntax.

In this paper we consider catena as a unit of syntax and morphology[1]. 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. 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 catena can be extended with additional syntactic or morphological relations to elements outside of the catena. The relations within the catena 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 lexicons each MWE represented as a catena might specify the potential extension of each element of the 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 analysis.

3 Multiword Expressions in BulTreeBank

In its inception and development phase, the HPSG-based Treebank adopted the following principles: When the MWE is fixed, which is inseparable, with fixed order and can be viewed as a part-of-speech, it receives lexical treatment. This group concerns the multiword closed class parts-of-speech: multiword prepositions, conjunctions, pronouns, adverbs. There are 1081 occurrences of such multiword closed class parts-of-speech 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 shows variation in its ending part: какви-то и да е/са/било (whatever). The varying

[1] http://en.wikipedia.org/wiki/Catena_(linguistics)

part is a 3-person-singular-present-tense-auxiliary, 3-person-plural-present-tense-auxiliary or its 3-person-neuter-singular-past-participle. The semi-fixed expressions (mainly proper names) have been interpreted as Multiword Expressions. However, all the idioms, light verb constructions, etc. have been treated syntactically. 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 coreferencing 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 its specific semantics from the semantics of the verbal phrases with semantically non-vacuous verbs. The same holds for the idioms.

4 Possible Approaches for Encoding Multiword Expressions in Treebanks

There are a number of possible approaches for handling idioms, light verb constructions and collocations. The approaches are not necessarily conflicting with each other. However, we also seek for an approach that would give us the mapping between compounds and their syntactic paraphrases.

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. For example, a verb subcategorizes for only one lexical item or a very constrained set of lexical items. When combined with nouns, such as време (time), форма (shape), надежда (hope), the verb forms idioms - губя време '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 latter cases, verbs like - обръщам 'to turn' pay - take only noun - внимание 'attention' 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 expressions with desemantisized 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 dependants. In this case we mark the information only on the head of these Multiword Expressions. In this approach the assumption is that the verb posits its requirements on its dependants. However, a very detailed valency lexicon is required. One problem with this approach is when the dependant elements allow for modifications.

The second approach is construction-based. In this case there is no head. 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 problematic for syntactically flexible Multiword Expressions.

The third approach marks all the parts of the Multiword Expressions. It is based on the notion of catena as introduced above. Here is an example of this annotation:

(VPS Той (VPC-C (V-C ритна) (N-C камбаната)))
(VPS He (VPC-C (V-C kicked) (N-C bell.DEF)))

He kicked the bucket

where the suffix "-C" marked the catena. This approach maybe adds some spurious compositionality to the idioms, but it would be indispensable for handling idiosyncratic cases, such as separable MWE. However, in order to model the various MWE and to ensure mappings among compounds and related syntactic phrases, the combination of catena with selection-based approach is needed. 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 selection-based approach is a powerful means for challenging analyses. 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 затварям 'to close' is used in its literal meaning and as a part of idiomatic expression. The catena is highlighted.

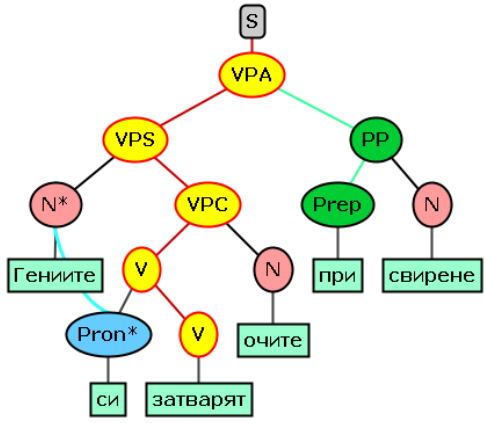

Figure 1: HPSG-based tree for the sentence "Гениите си затварят очите при свирене" ('Geniuses REFL.POSS.SHORT close eyes at playing' *Geniuses close their eyes when playing some instrument.*).

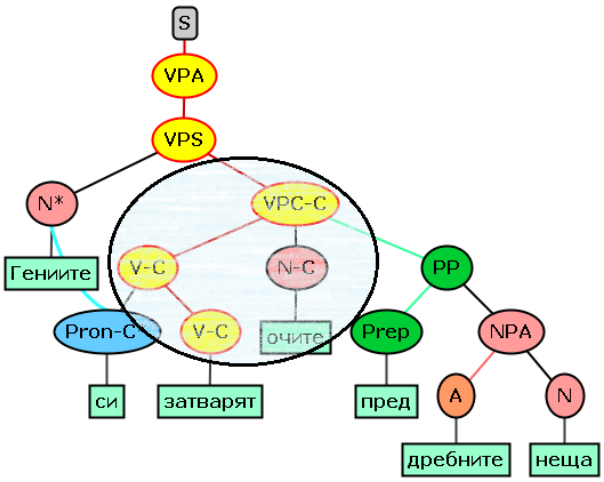

Figure 2: HPSG-based tree for the sentence "Гениите си затварят очите пред дребните неща" ('Geniuses REFL.POSS.SHORT close eyes before minor things' *Geniuses run away from the minor things.*).

In the lexicon each MWE is represented in its canonical (lemmatized) form. The catena is stored in the lexical unit. Additionally, the valency of MWE is expressed for the whole catena or for it 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 затварям си очите. Its representation in the lexicon could be as follows:

[**form:** < затварям си очите >
catena:
(VPC-C
(V-C (V-C затварям) (Pron-C си))
(N-C очите)
)
semantics:
не-обръщам-внимание-на-фактите_rel(e,[1]факт)
valency:
< indobj; (PP (P x) (N [1]y)) : x ∈ { пред, за } >
]

The specification above shows that the catena includes the elements 'shut my eyes' in the sense of 'run away from 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 'shut my eyes' reproduces some fact from the world. Also, if more than one preposition is possible, they are presented as a set of x-values.

5 From Syntax to Compound Morphology

The catena approach is also very appropriate for modeling the connection among compounds and their syntactic counterparts in Bulgarian. In (Gross, 2011) the notion of 'morph catena' has been explicitly introduced. By granting a node to each morpheme[2], the author makes the problematic morpheme a dominant element over the other depending morphemes. Thus, all these morphemes are under its scope. The catena set includes also the intended meaning.

Here we have in mind examples like the following: a) compound deverbal noun whose counterpart can be expressed only through a free syntactic 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 (ръкомахане ('handwaving', gesticulating), ръкомахам

[2]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]

('handwave.1PERS.SG', gesticulate) and махам 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-dependant 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 dependants 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 'snowrain', 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 dependants in the syntactic phrases relies on the valency information of the head verb only, while the realization of the dependants 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. Moreover, 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 verb compounds, presented via morphological catena. These words are then connected to the heads of the corresponding syntactic phrases (again in the lexicon), 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. However, in the syntactic catena a preposition is inserted according to the valence frame of the verb лекувам (cure): лекувам с билки ('cure.1PERS.SG with herbs', to cure with herbs). Using catena, we can safely connect the non-constituent phrase лекувам с (cure with) with the root morpheme of the head in the compound - леч. Also, all the possible modifiers of билки (herbs) in the syntactic phrase would be connected to the head morpheme билк.

The second case is the one where the nominal compound has mappings to both - verb compound and syntactic phrase. The connection among the nominal and verb compounds is rather trivial, since only the inflections differ. (ръкомахане ('handwaving', gesticulating), ръкомахам ('handwave.1PERS.SG', gesticulate) and махам с ръка ('wave with hand', gesticulate): рък-о-мах-а-не vs. рък-о-мах-а-м. The connection with the syntactic phrase follows the same rules as in the previous case.

Here is the representation of the lexical unit for compound nouns: (билколечение ('herbcuring', curing by herbs):

[**form:** < билколечение >
catena:
(MorphVerbObj-C
(MorphVerb-C [1]билк-) (MorphoObj-C [2]леч-)
)
derivational catena:
(VPC-C
(V-C [2]лекувам (PP-C (P-C с) (N-C [1]билки)))
)
semantics:
лекувам_rel(e,x,y,[4]билки) & номинал_rel(e)
valency:
< mod; (PP (P с) [4](NP ModP* (N билки) ModS*)) :
ModP* or ModS* is not empty >
]

In this example we present two relations. First, the morph catena is presented with its roots (the role of affixes omitted for simplicity). Then, the catena reflecting the derivational syntactic phrase is shown. The correspondences are marked with tags [1] and [2]. The second relation is at the semantic level, where the semantics of the syntactic phrase (лекувам_rel(e,x,y,[3]билки)) is represented fully, and additionally the event is nominalized by the second predicate номинал_rel(e). In the valency list we might 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 български билки ('herbcuring with Bulgarian herbs', curing with Bulgarian herbs) and билколечение с билки, които са събрани през нощта ('herbcuring with Bulgarian herbs that are collected during the night', curing with Bulgarian herbs that were collected at night) are allowed. But phrases with dublicate 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. Among the omitted phenomena are the representation of the subject and patient information as well as the inflection of the compounds.

As a result, we propose a richer valence lexicon, extended with information on mappings between compounds and their counterpart syntactic phrases. The morph catena remains steady, while the syntactic one is flexible in the sense that it encodes the predictive power of adding new material. 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 nontrivial and difficult for realization.

6 Conclusions and Future Work

The paper confirms the conclusions from previous works that catena is indispensable means for encoding idioms. Especially for cases where the literal meaning also remained a possible interpretation in addition to the figurative meaning in the respective contexts. We also extend this observation to other types of MWE.

Apart from that, we show that catena is a tool that together with the selection-based approach can ensure mappings between the expressions which have paraphrases on the level of morphology as well as syntax. While at the morphological level the catena is stable, in syntax domain it handles also additional material on prediction from valence lexicons and beyond them.

7 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 Multi-word Expressions. Towards linguistic precision and computational efficiency in natural language processing."

We are grateful to the two anonymous reviewers, whose remarks, comments, suggestions and encouragement helped us to improve the initial variant of the paper. All errors remain our own responsibility.

References

Thomas Gross. 2010. Chains in syntax and morphology. In Otoguro, Ishikawa, Umemoto, Yoshimoto, and Harada, editors, *PACLIC*, pages 143–152. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Thomas Gross. 2011. Transformational grammarians and other paradoxes. In Igor Boguslavsky and Leo Wanner, editors, *5th International Conference on Meaning-Text Theory*, pages 88–97.

K. Tamsin Maxwell, Jon Oberlander, and W. Bruce Croft. 2013. Feature-based selection of dependency paths in ad hoc information retrieval. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 507–516, Sofia, Bulgaria.

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

Petya Osenova. 2012. The syntax of words (in Bulgarian). In Diana Blagoeva and Sia Kolkovska, editors, *"The Magic of Words", Linguistic Surveys in honour of prof. Lilia Krumova-Tsvetkova*, Sofia, Bulgaria.

Dimitar Popov, Kiril Simov, Svetlomira Vidinska, and Petya Osenova. 2003. *Spelling Dictionary of Bulgarian (in Bulgarian)*. Nauka i izkustvo, Sofia, Bulgaria.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for nlp. In *In Proc. of the CICLing-2002*, pages 1–15.

Kiril Simov, Petya Osenova, Alexander Simov, and Milen Kouylekov. 2004. Design and implementation of the Bulgarian HPSG-based treebank. In *Journal of Research on Language and Computation*, pages 495–522, Kluwer Academic Publishers.

Aline Villavicencio and Valia Kordoni. 2012. There's light at the end of the tunnel: Multiword Expressions in Theory and Practice, course materials. Technical report, Erasmus Mundus European Masters Program in Language and Communication Technologies (LCT).