

Formalizing MultiWords as Catenae in a Treebank and in a Lexicon

Kiril Simov and Petya Osenova

Linguistic Modeling Department
Institute of Information and Communication Technologies, BAS
{kivs|petya}@bultreebank.org

Abstract

The paper presents formalization of multiwords as catenae in a treebank and in a lexicon. We view catenae as a dependency subtree, which reflects non-constituents and non-standard dependencies. Since the multiword classifications vary to great extent, starting from very narrow ones and proliferating to extended ones which include also valences, the focus in the paper is not on the multiword typology per se, but on the general formalization of multiwords.

1 Introduction

Multiwords (or Multiword Expressions (MWEs)) have been approached from various perspectives. It seems that most efforts go into introducing various classifications with respect to various NLP tasks, such as annotation, parsing, etc. Since there is no broadly accepted standard for Multiwords (see about the various classifications in [4]), we adopt the Multiword classification, presented in the seminal work of [8]. The authors divide multiwords 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). The latter group handles semantically and syntactically compositional, but statistically idiosyncratic phrases (such as, traffic lights).

We follow the understanding of [6] that multiwords 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 multiwords (when existing together with the idiomatic one). Thus, in this paper we focus on the formal representation of multiwords as catenae in the treebank as well as in the lexicon. Also, examples of formalization are provided for the most frequent multiword types in BulTreeBank.

The paper is structured in the following way: in Section 2 some previous works on catena are presented; Section 3 discusses the most frequent specific multiword types in the treebank; Section 4 outlines the formal definition of catena; Section 5 demonstrates the encoding of catenae in a dependency treebank; Section 6 shows the encoding of the catena in the lexicon; Section 7 concludes the paper.

2 Previous Work on Catenae

The notion of catena (chain) was introduced in [6] 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. [2] 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 [3] 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.

Apart from the linguistic modeling of language phenomena, catena was used in a number of NLP applications. [5], for example, 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 treebank and the lexicon.

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 as shown in [7]. In this paper we focus on the formal representation of multiwords in the treebank and lexicon from the perspective of the syntactic relations among their elements. Thus, irrespectively of the multiword classifications, the challenging issue remains the representation of multiwords with syntactic variability in the syntactic resource and the lexicon.

3 Multiwords from the Perspective of Syntactic Relations

Here we outline some frequent multiword types 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 in the syntax-lexical interface.

The adjunction is expressed in the following multiword types:

1. Noun phrases of type Adjective - Noun

вътрешен министър, 'interior minister' (Minister for Internal Affairs) снежен човек, 'snow 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 snow man', etc. Semantically, the first phrase is a metonymical synthetic form of the phrase 'Minister for Internal Affairs'. The second phrase conveys its literal meaning of: (1) a man-like sculpture from snow or (2) hypothetical man leaving in Himalayas or some other regions.

2. Noun phrases of type Noun - Prepositional Phrase

срещата на върха, 'meeting-the at peak-the' (summit)

Here 'meeting' can inflect in all its forms and allows for some modifications: 'past meetings', etc.

The complementation is expressed in the following multiword type:

1. Verb phrases of type Verb-Complement

знае си работата, 'knows-he his business-the' (one knows one's business); затварям си очите, 'close own eys-the' (to hide from the facts);

Here 'business' allows for only various possessive forms (one *knows* their business), but the nominal phrase always has to be definite, singular. The verb 'know' can vary in all its word forms and it allows for modification: one knows his business *well*.

4 Formal Definition of Catena

Here we follow the definition of catena (originally called chain, but later changed to catena, because of the ambiguity of the term chain) provided by [6] and [2]: a **catena** is a word or a combination of words directly connected in the dominance dimension. In reality this definition of catena for dependency trees is equivalent to

a subtree definition. We prefer to use the notion of catena to that of dependency subtree, because its high usage in modeling MultiWord Expressions. However, we have to utilize the notion of catena for two purposes: for annotation of MultiWord Expressions in the actual trees expressing the analysis of sentences as well as for representation of MultiWord Expressions in the lexicon.

Let us have the sets: LA — a set of POS tags, LE — a set of lemmas, WF — a set of word forms and a set D of dependency tags ($ROOT \in D$). Let us have a sentence $x = w_1, \dots, w_n$. A **tagged dependency tree** is a directed tree $T = (V, A, \pi, \lambda, \omega, \delta)$ where:

1. $V = \{0, 1, \dots, n\}$ is an ordered set of nodes, that corresponds to an enumeration of the words in the sentence (the root of the tree has index 0);
2. $A \subseteq V \times V$ is a set of arcs;
3. $\pi : V - \{0\} \rightarrow LA$ is a total labeling function from nodes to POS tags. π is not defined for the root;
4. $\lambda : V - \{0\} \rightarrow LE$ is a total labeling function from nodes to lemmas. λ is not defined for the root;
5. $\omega : V - \{0\} \rightarrow WF$ is a total labeling function from nodes to word forms. ω is not defined for the root;
6. $\delta : A \rightarrow D$ is a total labeling function for arcs;
7. 0 is the root of the tree.

We will hereafter refer to this structure as a parse tree for the sentence x . Let $T = (V, A, \pi, \lambda, \omega, \delta)$ be a tagged dependency tree.

A directed tree $G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G)$ is called **dependency catena of T** if and only if:

1. G is a connected directed tree with root $CatR$ ($CatR \in V_G$);
2. $\psi : V_G \rightarrow V$, there is a mapping from the nodes V_G into $V - \{0\}$. V_G is the set of nodes of G ;
3. $A_G \subseteq A$, the set of arcs of G ;
4. $\pi_G \subseteq \pi$ is a partial labeling function from nodes of G to POS tags;
5. $\lambda_G \subseteq \lambda$ is a partial labeling function from nodes to lemmas;
6. $\omega_G \subseteq \omega$ is a partial labeling function from nodes to word forms;
7. $\delta_G \subseteq \delta$ is a partial labeling function for arcs.

A directed tree $G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G)$ is a **dependency catena** if and only if there exists a dependency tree T such that G is a dependency catena of T .

Having partial functions for assigning POS tags, dependency labels, word form and lemmas allows us to construct arbitrary abstractions over the structure of catena. The mapping ψ parameterizes the catena with respect to different dependency trees. Using the mapping there is a possibility to use different word orders of the nodes of the catena, for example. Also catena could be underspecified for some of the node labels like grammatical features, lemmas and also some dependency labels.

The image (mapping) of a catena in a given dependency tree we will call **realization of the catena in the tree**. We consider the realization of the catena as fully specified subtree including all node and arc labels. For example, the catena for “to spill the beans” will allow for any realization of the verb form like in: “they spilled the beans” and “he spills the beans”. Thus, the catena in the lexicon will be underspecified with respect to the grammatical features and word form for the verb.

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

5 Encoding of Multiword Valency in a Treebank

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 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 as they were used in the CoNLL 2006 format. Here column 9 is used for annotation of the node as being part of a catena or not. 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, column 9 of the corresponding line contains an underscore symbol. Since a sentence might contain more than one catena, each one is numbered in different way. We do not allow any catena overlapping.

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

1. The nodes of $V - \{0\}$ are represented in the first cell of each row in the table;
2. For each arc $\langle d, h \rangle \in A$, the head node h is represented in cell 7 of the row for node d ;

3. 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 ;
4. For each node $n \in V - \{0\}$ the value $\lambda(n)$ is represented in cell 3 of the row for node n ;
5. For each node $n \in V - \{0\}$ the value $\omega(n)$ is represented in cell 2 of the row for node n ;
6. For each arc $\langle d, h \rangle \in A$ the label $\delta(\langle d, h \rangle)$ is represent in cell 8 of the row for node d .
7. the root 0 is not represented in the table.

The following is an example for the sentence: Те си затварят очите пред истината (they run away from the truth):

No	Wf	Le	POS	ExPOS	GramFeat	Head	Rel	Catena
1	Те	те	P	Pp	number=pll case=nom	3	subj	—
2	си	си	P	Pp	form=possesive	3	clitic	c_1
3	затварят	затварям	V	Vpi	number=pll person=3	0	Root	c_1
4	очите	око	N	Nc	number=pll definiteness=y	3	obj	c_1
5	пред	пред	R	R	—	3	indobj	—
6	истината	истина	N	Nc	number=sgl definiteness=y	5	prepobj	—

In the table it can be seen that three elements are part of the catena: си затварят очите 'their close eyes' (they close their eyes). 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. Here we will not discuss the role of catena in deep level dependency analysis (e.g. the tectogrammatical level in the Prague dependency treebank).

In order to model the behavior in a better way we need to add semantics to the dependency representation. We will not be able to do this in full in this paper. In order to represent the MWEs in the lexicon, we assume a semantic analysis based on Minimal Recursion Semantics (see [1]). For dependency analyzes the MRS structure are constructed in a way similar to the one presented in [9]. 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 will use the root of the corresponding catena. In the dependency tree for the corresponding sentence the catena root will provide the interpretation of the MWE and its dependent elements, if any. In the lexicon we will provide the corresponding structure to model the idiosyncratic semantic content of MWE.

6 Encoding of Multiword Valency in a Lexicon

The lexical entry of a MWE consists of a **form**, a **catena**, **semantics** and **valency**. The form is represented in its canonical form which corresponds to one of its realizations. The catena for the multiwords 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 example, the multiword from the above example затварям си очите is represented as follows:

[**form:** < затварям си очите >

catena:

No	Wf	Le	POS	ExPOS	GramFeat	Head	Rel
1	–	затварям	V	Vpi	–	0	CRoot
2	си	си	P	Pp	form=possessive	1	clitic
3	очите	око	N	Nc	number=pll definiteness=y	1	obj

semantics:

No1: { run-away-from_rel(e, x_0, x_1), fact(x_1), [1](x_1) }

valency:

No1: < :indobj: x/Prep :prepobj: y/N[1] || $x \in \{ \text{пред, за} \}$ >
]

The lexical entry 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 set of elementary relations. In this case we have the relation run-away-from_rel(e, x_0, x_1) which determines that the multiword expression is denoting an event with two main participants denoted by the subject (x_0) and the indirect object (x_1). In the lexical entry we represent the restriction on the indirect object which has to be a fact. The actual fact in this part is indicated via a structure-sharing mechanism with a valency part — [1]. This is necessary, because in the valency part of the lexical entry the noun within the subcategorized PP by the catena ‘shut my eyes’ reproduces some fact from the world.

The valency information is presented by a dependency path. The arc labels are given between column marks, the node information is given after the arc information and could include a variable for the word (we also plan to add lemma information) and grammatical features. The structure-sharing identifier [1] denotes the semantics of the noun phrase that is indirect object. Its main variable is made equal to the variable for indirect object in the semantic representation of MWE — x_1 . This ensures that the expected noun phrase has to denote a fact. Additionally, if one or more (but small amount of) words are possible for a node, they can be given as a set. In the example only two prepositions are possible for node x.

In many languages the elements represented in the valency are not realized. This is the case for Bulgarian — the objects and indirect objects of a verb could

be unexpressed. In such cases the semantics is assumed to be empty, expressed via the most general predicate like *everything*(*x*) which will agree with any other predicate. 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.

In the catena representation cell 9 is empty and this is why it is not given in the lexicon. The semantics and the valency information is 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 — No1). In cases when other parts of the catena allow modification, the information for the corresponding nodes will be given.

For example, the multiword *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:

[**form:** < среща на върха >

catena:

No	Wf	Le	POS	ExPOS	GramFeat	Head	Rel
1	—	среща	N	Nc	—	0	CRoot
2	на	на	R	R	—	1	mod
3	върха	върх	N	Nc	number=sg definiteness=y	2	prepobj

semantics:

No1: { meeting_rel(*e*, *x*), member(*y*,*x*), head-of-a-country(*y*,*z*), country(*z*), [1](*z*)) }

valency:

No3: < :mod: *x*/Adj[1] >

]

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.

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

[**form:** < снежен човек >

catena:

No	Wf	Le	POS	ExPOS	GramFeat	Head	Rel
1	—	снежен	A	A	—	2	mod
2	—	човек	N	Nc	definiteness=n	0	CRoot

semantics:

No2: { 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 can not 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 multiword expression are determined by the represented semantics. The relation `snowman_rel(x)` is taken from an appropriate ontology where its conceptual definition is given.

These three examples demonstrate the power of the combination of catenae, MRS structures and valency representation to model multiword expressions 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 MWE. One missing element of the lexical entry is a representation of constraints over the word order of the nodes of the catena. We envisage addition of such constraints as future work.

7 Conclusion

In this paper a formalization of the multiwords as catenae was presented. The focus was on their modeling in the treebank and in the lexicon. Although the catenae approach provided a good apparatus for this, there are specificities in the syntax and lexical representation that had to be reflected. The common perspective for the syntax-lexical interface of multiwords lies in the syntactic relations among their elements (adjunction and complementation).

Sag et. al (2002) [8] enumerated several problems for MWEs representation which we hope our representation of MWEs in the lexicon solves to a great extent. The **overgeneration problem** is solved by an appropriate combination of syntactic, morphological, semantic and valency constraints. They are enough to rule out the impossible realizations of the multiword expressions. The **idiomaticity problem** is also solved because any peculiarities on these levels can be expressed in the lexical entry. The **flexibility problem** is solved by the definition of catena which allows for different realizations in the actual dependency trees. The **lexical proliferation problem** is manageable by using the valency constraints. In this way we can incorporate semantic constraints on the dependents. In the case of light verb, for example, the semantic of the verb in most cases is very general, but the actual semantic is coming from the direct object.

In future we will develop a lexicon for the MWEs appearing in the Bulgarian treebank. Then we will develop a mechanism to use the created multiword lexicon in parsing and generation processing. In addition, the formalization represented above needs to be extended with word order constraints and statistical information for institutionalized phrases.

8 Acknowledgements

This research has received partial funding from the EC’s FP7 (FP7/2007-2013) under grant agreement number 610516: “QTLeap: Quality Translation by Deep Language Engineering Approaches”.

References

- [1] Copestake, A., Flickinger, D., Pollard, C., and Sag, I. (2005) Minimal Recursion Semantics: an Introduction. *Research on Language and Computation*, 3(4).
- [2] Thomas Gross (2010) *Chains in syntax and morphology*. In Ootoguro, Ishikawa, Umemoto, Yoshimoto, and Harada, editors, PACLIC, pages 143–152. Institute for Digital Enhancement of Cognitive Development, Waseda University.
- [3] 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.
- [4] 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).
- [5] 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.
- [6] William O’Grady (1998) The syntax of idioms. *Natural Language and Linguistic Theory*, 16:279–312.
- [7] Petya Osenova and Kiril Simov (2014) *Treatment of Multiword Expressions and Compounds in Bulgarian*. In Proceedings of the Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014), ESSLLI, Tuebingen, Germany, pages 41–46.
- [8] 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.
- [9] Simov, K., and Osenova, P. (2011) *Towards Minimal Recursion Semantics over Bulgarian Dependency Parsing*. In Proc. of the RANLP 2011.