Formalizing MultiWords as Catenae in a Treebank and in a Lexicon

Kiril Simov and Petya Osenova

Linguistic Modeling Department, IICT, BAS
{kivs|petya}@bultreebank.org

February 6, 2015

Abstract is related to Work group 4 and Work group 1.

1 Encoding of Multiword Valency in a Treebank

We represent BulTreeBank 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. The following is an example for the sentence: Те си затварят очите пред истината (they run away from the truth), where the three elements of the MWE are marked up. In this way, its idiomatic reading is represented. Thus, each MWE in a dependency tree is represented via its realization.

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

2 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 catena in lexicon represents an underspecified dependency subtree which is a generalization over the catena realizations in the various sentences. 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=possesive	1	clitic
3	очите	око	N	Nc	number=pl|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 ∈ { 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.