Formalizing MultiWords as Catenae in a Treebank and in a Lexicon

Kiril Simov and Petya Osenova (IICT, Bulgarian Academy of Sciences)
WG4: Annotating MWEs in Treebanks (related also to WG1)

1. Overview

Task:

- Definition of catena supporting representation of MWEs in syntactic parses in a treebank and in lexical entries in a lexicon
- Both representations have to be related
- Operations over catenae for realization in parse trees

Classification of MWEs:

[Sag et. al 2002] - Multiword Expressions: A Pain in the Neck for NLP:

- ***Lexicalized phrases***
  - Fixed expressions
  - Semi-fixed expressions
  - Syntactically-flexible expressions
- ***Institutionalized phrases***

2. MWE Types to Model

We define a formalization of MWE to cover the following three types:

- Noun phrases of type Adjective – Noun

  снежен човек 'snow man' (snowman)
- Noun phrases of type Noun – Prepositional Phrase

  среща на върха 'meeting-the at peak-the' (summit)
- Verb phrases of type Verb – Complement

  затварям си очите 'close own eys-the'

  (run away from the facts)

3. Tagged Dependency Tree

Tagged Dependency Tree:

Let LA be a set of POS tags, LE be a set of lemmas, WF be a set of word forms and D be a set of dependency tags (ROOT ∈ D). Let x = w1, …, wn be a sentence. A tagged dependency tree is a directed tree T = (V, A, π, λ, ω, δ) where:

1. V = {0,1,…, n} is an ordered set of nodes
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
4. $\lambda : V - \{0\} \rightarrow LE$ is a total labeling function from nodes to lemmas
5. $\omega : V - \{0\} \rightarrow WF$ is a total labeling function from nodes to word forms
6. $\delta : A \rightarrow D$ is a total labeling function for arcs
7. 0 is the root of the tree

Catena :

Any element (word) or any combination of elements that are continuous in the vertical dimension (y-axis)

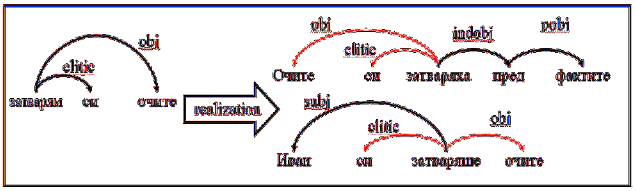
We model catena as a subtree of a tagged dependency tree

4. Catena Definition

A directed tree $G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G)$ (CatR $\in V_G$) is dependency catena of $T = (V, A, \pi, \lambda, \omega, \delta)$ iff:

1. $\psi : V_G \rightarrow V - \{0\}$
2. $A_G \subseteq A$
3. $\pi_G \subseteq \pi$
4. $\lambda_G \subseteq \lambda$
5. $\omega_G \subseteq \omega$
6. $\delta_G \subseteq \delta$

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



5. Treebank Representation

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	_

6. Representation in Lexicon

[**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=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 ∈ { пред, за } >
]

[**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	връх	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 research has received support by the EC's FP7 under grant agreement number 610516:
"QTLeap: Quality Translation by Deep Language Engineering Approaches"