Annotation of Multiwords in BulTreeBank (current status and perspectives)

Petya Osenova and Kiril Simov (IICT, Bulgarian Academy of Sciences)
WG 4: Annotating MWEs in Treebanks (related also to the other WPs)

1. Overview

Task:
• In BulTreeBank only part of the Multi Word Expressions are annotated explicitly
• Current annotation is problematic for the correct linguistic generalization
• Main problems are related to training automatic parsers

Current Approaches:
Fixed-expressions (complex POS) - 1081 tokens within 214 000 tokens
• Semi-fixed expressions
- Idioms - treated syntactically (no difference between: kick the door and kick the bucket)
- Proper names – treated as constructs

• Syntactically-flexible expressions – treated syntactically

2. MWE Annotation: Perspectives

• **Selection-based** – depends on the lexical meaning of the elements, selected by the head (‘*lose time*’ = idiom, but ‘*lose wallet*’ = phrase)
• **Construction-based** – ‘*from needle to thread*’ (from the beginning to the end)
• **Catena-based** – esp. for idiosyncratic cases
(**VPS** *Той* (**VPC-C** (**V-C** *ритна*) (**N-C** *камбаната*)))
= He kicked the bucket.

3. Catena Definition

Here we consider catena as a unit of syntax. In a syntactic tree (constituent or dependency) ***Catena*** is:

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

It is applied to the syntax of 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.).

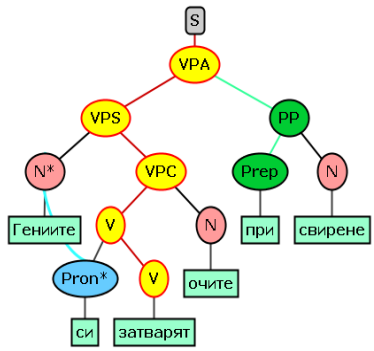
It provides a mechanism for a (partial) set of interconnected syntactic relations. **A good choice for Multiword expressions**.

4. Valency Representation

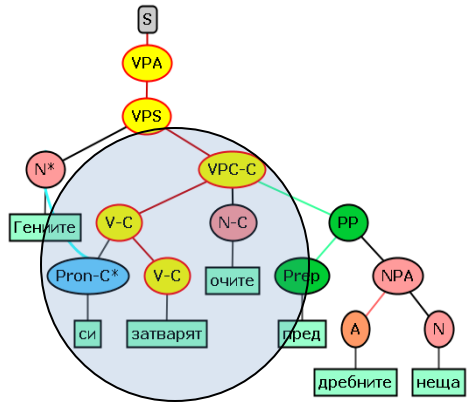
Catena represents a (partial) set of syntactic relations.
The set is partial in cases when the elements of the catena can be extended with additional syntactic relations to elements outside of the catena.
The syntactic relations within the catena cannot be changed.
These characteristics of catena make it a good candidate to represent MWE in lexicons and treebanks.
In the lexicons each MWE represented as a catena could specify the potential extension of each element of the catena.
In case the MWE does not allow any modifications, then for each element of the catena it is specified that the element does not allow any modifications.

In this way it is easy to see that catena is able to subsume the other two approaches:
- Selection-based
- Construction-based

5. Application to BulTreeBank



'Geniuses REFL.POSS.SHORT close eyes at playing'
Geniuses close their eyes when playing some instrument.



'Geniuses REFL.POSS.SHORT close eyes before minor things'
Geniuses run away from the minor things.

Белият дом обяви намаление на данъците.
'White-DEF house announced decreasing of taxes.'
The White House announced reduction of taxation.

Белият дом се намираше на хълма.
'White-DEF house REFL.PERS.ACC was at hill-DEF'
The white house was on the hill.

6. Conclusion & Future Work

Summary
Here we present an approach to annotation of Multiword Expressions in the current version of Bulgarian treebank – BulTreeBank. We have selected catena constructs as a mechanism for annotation of MWE in the treebank. During the annotation we determine only the boundaries of the MWE, but not their type. This will be done in the lexicon.

Future work
Completion of the annotation in the treebank.
Creation of lexicon with Multiword Expressions encoded as catena.

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