The Catena Approach
From Syntax to Compound Morphology

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

1. Overview

- We need a mechanism for connecting the MWEs in the lexicon with their usages in text
- Compounds are viewed as a phenomenon at the interface of Morphology and Syntax
- 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.
- We use catena as "path in the syntactic or morphemic analysis that is continuous in the vertical dimension"

2. 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**.

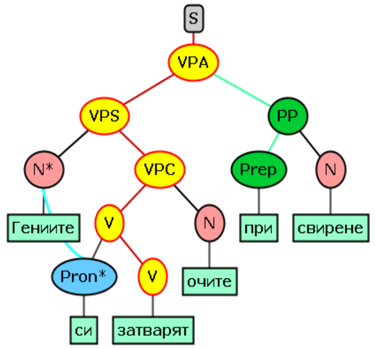
3. MWE Annotation: Perspectives

- **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

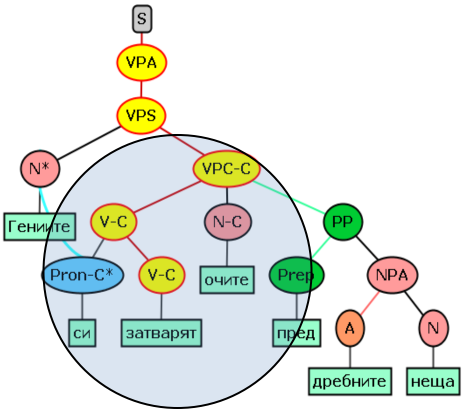
(**VPS** *Той* (**VPC-C** (**V-C** *ритна*) (**N-C** *камбаната*)))
= He kicked the bucket.

Selection-based + Catena-based = powerful analysis instrument

4. Catena Record in Lexicon



'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.

[
form: < затварям си очите >
catena: (VPC-C (V-C (V-C затварям)(Pron-C си))(N-C очите))
semantics: not-pay-attention-to-facts_rel(e,[1]fact)
valency: < indobj; (PP (P x) (N [1]y)) : ∊ { пред, за } >
]

5. Compound Morphology

Deverbal nouns inherit the syntactic structure from the source syntactic phrase

билколечение ('herbcuring', curing by herbs)
* билколекувам (*'herbcure.1PERS.SG', to cure with herbs)
лекувам с билки ('cure.1PERS.SG with herbs', to cure with herbs)

ръкомахане ('handwaving', gesticulating)
ръкомахам ('handwave.1PERS.SG', gesticulate)
махам с ръка ('wave with hand', gesticulate)

A previously done survey in (Osenova, 2012):
- Performed over an extracted data from a morphological dictionary
- Shows that in Bulgarian head-dependant compounds are more typical for the nominal domain (with a head final structure)
- The free syntactic phrasing is predominant in the verbal domain

[
form: < билколечение >
catena: (MorphVIObj-C (MorphIObj-C [1]билк-)(MorphV-C [2]леч-))
derivational catena: (VPC-C (V-C [2]лекувам (PP-C (P-C с) (N-C [1]билки))))
semantics: cure_rel(e,x,y,[4]билки) & nominal_rel(e)
valency: <mod; (PP (P с) [4](NP ModB* (N билки) ModA*)):ModB* or ModA* is not empty>
]

6. Conclusion & Future Work

Summary
The annotation in the treebank and the creation of lexicon with Multiword Expressions and Compounds encoded as catena is in process

Future work
Next step is to incorporate the catena lexicon in Bulgarian processing pipeline

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