Syntactic Annotation of Slovene Corpora (SDT, JOS)

Nina Ledinek
ISJ ZRC SAZU
nledinek@zrc-sazu.si

Corpus Annotation

The practice of adding interpretative, linguistic information to a corpus of spoken/written language data

→ Human language technologies
→ Linguistic research

The added notations → transcriptions, part-of-speech tagging, semantic tagging, syntactic analysis, named entity recognition, anaphora resolution, etc.

Syntactically annotated corpus → treebank

MPŠ, 10. 12. 2008

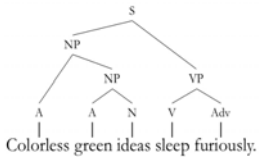
<w
lemma="primer" msd="Sometxxn"
lemmas="primer" msds="Sometxxn"
lemmass="primer primera" msdss="Somei,Sometxxn Sozmr,Sozdr">primer</w>

Syntax

→ Way in which linguistic elements (as words) are put together to form larger units, constituents (as phrases, clauses, sentences)

(morpheme) →
→ word → phrase → clause → sentence
→ (text)

→ Principles and rules for constructing (grammatical) sentences (with a certain meaning)



Dependency Grammar (Syntax)

Roots → Panini's grammar (Sanskrit), traditional grammar, medieval theories, Slavic linguistics, etc. Culmination: work of L. Tesnière (1959) → modern dependency grammar

→ Large and fairly diverse family of grammatical theories and formalisms that share certain basic assumptions about syntax

Syntactic structure
↓
→ Lexical elements linked by binary asymmetrical relations (dependencies, connexions)
→ Head/governor – dependent/subordinate
→ Valency

Problems

FGD

Functional Generative Description
↓
Prague Dependency Treebank

Multi-stratal framework
→ Analytical layer – surface syntactic annotation (subject, object, attribute, adverbial, coordination, etc.)
→ Tectogrammatical layer – deep syntactic/shallow semantic annotation → thematic roles, co-reference, topic-focus articulation (agent, patient, predicate, antecedent, etc.)

Dependency Parsing

→ Each node is assigned one head at most (single-head constraint)
→ All nodes have to be connected (connectedness)
→ Chains of dependency links do not contain cycles (acyclicity constraint)
↓
Syntactic tree structures

→ Dependency links are close to the semantic relationships (→ deep syntactic annotation, shallow semantic annotation)
→ Parsing is efficient (computationally)
→ Complexity of parsing – expressivity of syntactic representations (→ good compromise)

Treebank

A linguistically annotated corpus that includes some grammatical analysis beyond the part-of-speech

Empirical syntactic analysis of language patterns in large quantity of naturally occurring texts

Syntactic Annotation (Models)

Complexity of the annotation system
→ Chunking, skeletal, shallow parsing
→ Full parsing

Human vs. no human rule creation
→ Rule-based parsing (obsolete?)
→ Stochastic, data-driven parsing
↓
Robustness

Syntactic Annotation (Types)

Grammatical theories and formalisms/types of syntactic information

→ Dependency models
Asymmetric binary relations (connexions)
Governor – dependent(s)
Functional analysis
Inflectionally rich languages with free word order
→ Phrase structure/constituent models
Hierarchically embedded subparts (constituents)
Part – whole relations
Structural analysis
Languages with fixed word order, clear constituency structures

→ Hybrid models

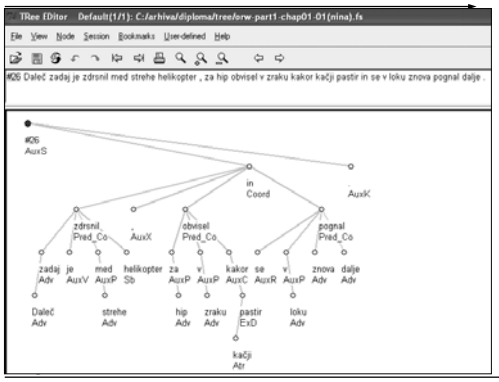
Slovene Dependency Treebank

SDT
http://nl.ijs.si/sdt/

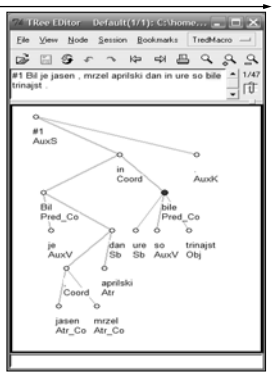
→ Dependency treebank of Slovene written texts
→Modeled after the Prague Dependency Treebank
→ Surface syntactic annotation
→ Two subcorpora (1984, SVEZ-IJS)
→ 2800 sentences, 45000 words
→ Experiments in inductive parsing
→ Freely available for research use

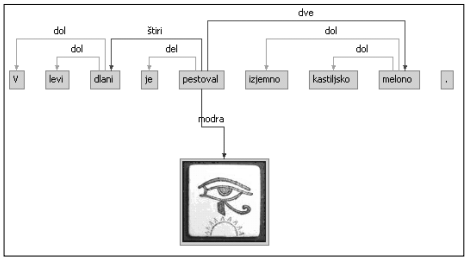
Problem → complexity of the theoretical framework

SDT: Syntactic Tree Structure I



SDT: Syntactic Tree Structure II





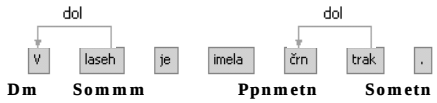
Linearity
Three types of connexions → green, red, blue
Connexions → intuitive names
Arrows
Connectedness
Root
Sentence → the maximal unit of parsing

JOS: Syntactic Tagset

Automatic annotation → robust linguistic units with clearly defined boundaries

Manageable tagset →
(SDT: >100), JOS: 10

Combining of the data: MSD + syntactic tags + etc.



First Level Tags

"Phrase structure connexions"
(Green)

Dol "attr"
Del "part"
Prir "coord"
Vez "conj"
Skup "together"

Second Level Tags

"Functional connexions"
(Red)

Ena "one"
Dve "two"
Tri "three"
Štiri "four"

Third Level Tags

"Residual"
(Blue)

Modra "blue"

Thank you!

nledinek@zrc-sazu.si