

Hungarian verbal valence dictionary in NooJ

Kata Gábor
Linguistics Institute, HAS

9th INTEX/NooJ conference



The Hungarian module: syntactic analysis

- aims:
 - 1) to parse Hungarian sentences
 - 2) further goal: semantic interpretation
- in NooJ:
 - recognize phrases with a strict word order
 - annotate them according to their *syntactic / semantic relation* to the predicate



State of the art

- we already have:
 - 1) morphological analyser
 - 2) sentence splitting, clause boundary detection
 - 3) NP, AdjP, AdvP grammars etc...
- next step:
 - verbal valence frames & adjuncts

Annotation of the argument structure I.

- *What the output should be like? What does "relation to the predicate" mean?*
- 1. COMPLEMENT vs ADJUNCT dichotomy
 - in some languages it can be captured by purely syntactic tests
 - semantic difference: semantic arguments of the predicate are realized syntactically as complements, while adjuncts express circumstances

Semantic basis of syntactic phenomena



semantic arguments = syntactic complements?

eat sg with sg

vs

poke sg with sg

syntactic tests:

This idea is reflected upon by a lot of scholars.

**This office is worked at by a lot of people.*

**This stick is poked with by John.*

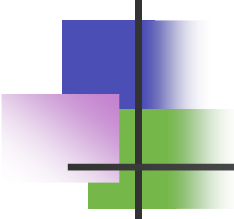
→ **semantic arguments are not always realized as complements**

Complements and adjuncts in the Hungarian sentence

- syntactic tests for complementness fail to account for the Hungarian data:
 - complements are not always obligatory
 - word order does not provide clues either:

A fiúk tegnap megnézték a moziban a filmet.

[The boys][yesterday][watched] [in the movie] [the film]



Annotation of the argument structure II.

- complement-adjunct dichotomy is not sufficient:
 - semantic argumentness is not encoded
 - doesn't reveal fine-grained semantic relations
 - transformation possibilities are not predictable from this information

→ for further semantic processing we need more information



Semantic verb classes

- verb classes defined by semantic metapredicates are characterized by:
 - similar complement structures
 - similar transformational properties
 - *possible expansion by the same adjunct types*



Resources: verbal valence frame tables

- developed at the LI, 2001- 2003
- its vocabulary is constantly being expanded
- 19,000 frames for 9,000+ verbs
- format: Excel tables → converted to XML
- goal: conversion into NooJ dictionary



Vocabulary

- base: 2,800 most frequent verbs in the HNC
- 6,200 entries from other Hungarian corpora (short business news, Szeged treebank)
- automatic retrieval of patterns from sub-corpora of the HNC (B. Sass)
 - manually checked
 - manually written ex. sentences for each frame



Structure of valence frames

- dependency grammar framework
- 3 complement relations: SUBJ, OBJ, COMPL
- POS of complements: N, A, ADV, Partic, INF, CLAUSE
- except for ADV and INF, every complement bears a morphological case
- first step: recognize nominal complements

Difficulties I.:

The quantity of frame types

- 2092 N.NOM N.ACC
- 1843 N.NOM
- 1093 N.NOM+Human
- 1046 N.NOM+Human N.ACC
- 367 N.NOM N+Dir
- 288 N.NOM N.ACC+Abstract
- 276 N.NOM+Human N.ACC+Human
- 267 N.NOM+Anim
- 253 N.NOM N.ACC N+Dir
- 253 N.NOM N.ACC+Human
- 231 N.NOM N+Dir N+Source
- 204 N.NOM CLAUSE.ACC

Difficulties I.:

The quantity of frame types

- 2,903 frame types
- frames present different degrees of specification from general patterns to frozen expressions
- bottom of the list: structures which occur only once, with the lemma of one or more complements given

Difficulties II.:

Optionality

- SUBJ and OBJ are always optional (encoded in verbs' morphological features)
 - many other complements are also optional
- increases the number of possible sentence structures in the case of verbs with several complements



Experiments with INTEX

1) use the Excel tables and generate different graph-patrons for any possible word order, taking into consideration the (theoretically) unlimited quantity of adjuncts at any position →

verb + 4 possible complements,

MOT* between them =

120+24+6+2+1

compilation failed



In NooJ

- lexicon-grammars → dictionary entries
- cascaded analysis is facilitated by the annotation of the text: annotation needs to be considered only when needed
- basic question: how to represent complement structures in the dictionary?

Complement structures in the dictionary



- one complement structure type → one lexical feature

problems: 1) quantity of different types
2) doesn't allow generalizations

- one complement → one lexical feature

problem: difficult to account for the interdependence of complements

Possible solution: semantic classes



- supposition: verbs' distributional and transformational properties depend on their semantic content
- verb classes that share a semantic metapredicate can be described by the same grammars
 - semantic metapredicates are coded in the lexicon
 - graphs will refer to these semantic features



The workflow

- 1. basic vocabulary: 2,000 most frequent verbs
- 2. choosing a syntactic property that characterizes a group of verbs
- 3. manual checking and creation of sub-classes on the basis of other morphosyntactic properties
- 4. finding the semantic metapredicates for the V classes and coding it in the dictionary
- 5. collecting simple sentences from the HNC that contain a finite form of a V from given class
- 6. writing grammars to cover all of their occurrences



Example

- 1. syntactic property: complement in ablative case
- 2. list of all the verbs that have this type of complement in the tables
- 3. manual checking of the list: creation of sub-classes, e.g.:
 - movement verbs (source) : *gyalogol* /walk/
 - verbs with a patient subject (direct cause) : *elalszik* /fall asleep/
 - distinction: on the basis of other complements or adjuncts



State of the art

- 3 syntactic properties with the corresponding verb lists have been checked
- we established some basic classes (movement verbs, change_in_state, cause_change...) on the basis of their syntactic properties
- they have been marked in the dictionary



Further work

- verification of the existing classes:
 - representation of syntactic properties as NooJ grammars
 - testing the grammars on sentences from the HNC
- creation of new classes



Thank you for your attention!

gkata@nytud.hu