Semantic Annotations in COROLA, the corpus of Contemporary Romanian Language

Dan Cristea
COROLA

• Contemporary Romanian: 1945 => nowadays
• Registers: bellettristic, drama, essays, poetry, journalism, scientific
• Domains: as many as possible
• Representation: perhaps proportional with the reading density, more than with the writing density
• Heavily annotated on more levels:
  – compulsory: sub-syntactic, morpho-syntactic
  – optional: syntactic, semantic, discourse
COROLA

• Who will be the beneficiaries?
  – language researchers
  – professors & students
  – Romanian language users
  – writers
  – historians of language (imprint of the language at a certain moment in its evolution)
Goals of our endeavour

• Propose annotation conventions for the semantic level
• Develop a lexical-semantic knowledge base to be used by Natural Language Processing (NLP) tools to learn how to reproduce the same type of annotation on other texts
• Investigate the level of trust on these annotations
“Quo Vadis”

- Romanian translation of Henrik Sinkiewich’s Nobel price novel
  - translation by Remus Luca and Elena Lință, Tenzi Publishing House, 1991
# Development of the “QV” Corpus

<table>
<thead>
<tr>
<th></th>
<th>#sent</th>
<th>#entities</th>
<th>#ref.links</th>
<th>#AKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sept. 2013</td>
<td>1,500</td>
<td>3,663</td>
<td>2,045</td>
<td>75</td>
</tr>
<tr>
<td>May 2014</td>
<td>7,150</td>
<td>22,310</td>
<td>21,439</td>
<td>1,133</td>
</tr>
<tr>
<td>June 2014</td>
<td>7,281</td>
<td>24,643</td>
<td>21,605</td>
<td>759</td>
</tr>
</tbody>
</table>


Many false relations eliminated
The annotation process

Extremely complex:

– done originally by students in CL on chunks of text
– verified by a nucleus of experts
– sewed at chunks boundaries
– finally re-verified by an expert
– filters applied
Layers of initial annotation

- sentence borders
- tokens (words or word compounds, punctuation marks)
- part-of-speech tags
- lemmas
What do we annotate?

• Singular and group entities of type PERSON and GOD and body parts of them
  – Individual: [Ligia], [ea], [Marcus Vinicius], [împăratul], [al lui Petronius], [al lui], [un grup mare de credincioși], [care], [Christ]
  – Imbricated: [alte femei din [societatea înaltă]], [[his] hand], [[[his] mother’s] sister]
  – Realisation included: dar 1:[îl] și 2:[iubeau, REALISATION="INCLUDED"])) din tot sufletul
What do we annotate?

• Relations between entities
  – Referential:
    • coreferential: 1:[Ligia]... 2:[tânăra libertă]...
    • ...

What do we annotate?

• Relations between entities
Affective: love

Acteea, fosta amantă

Această societate

toți

propriul său copil

Subrius Flavius

Poppea

Nero

Rubelius Plautus

Octavia

același popor

Agrippina

Thraseas

Ligia, Callina

un ligian

soția lui Plautius

familia lui Plautius

Vinicius

Gulo

doi oameni

învățător divin

Petronius

poporul

Ursus

Aulus

Eunice