

Developing Resources for the Culinary Domain

Cvetana Krstev
University of Belgrade
Faculty of Philology



<SYNSET>
<ID>ENG30-07734017-n</ID>
<SYNONYM>
<LITERAL>paradajz</LITERAL>
<DEF>paradajz</DEF>
<DEF>blago kiselo crveno ili žuto mekano voće koje se jede kao povrće</DEF>
<POS>n</POS>
<ILR>n</ILR>
<TYPE>ENG30-07710007-n</TYPE>
<ILR>ENG30-12905817-n</ILR>
<TYPE>holo_part</TYPE></ILR>
</SYNSET>

<SYNSET>
<ID>ENG30-07718472-n</ID>
<SYNONYM>
<LITERAL>krastavac</LITERAL>
<DEF>voće cilindričnog oblika sa korom zelene boje; unutrašnjost je bele boje a u ishrani se koristi kao povrće; pripada porodici tikvi</DEF>
<POS>n</POS>
<ILR>ENG30-07707451-n</ILR>
<TYPE>hypemym</TYPE>
<ILR>ENG30-12165384-n</ILR>
<TYPE>holo_part</TYPE></ILR>
</SYNSET>

Outline

- Introduction and motivation
- Resources at disposal
 - A corpus
 - Serbian Wordnet
 - Serbian e-dictionaries
- Enrichment of resources
- Problems to solve

Introduction and Motivation

- The culinary domain is one of the rare domains in which the general public and the scientific community are equally interested today.
 - A number of web sites offer a huge number of recipes, in many languages, searchable by different criteria, and often populated by users.
 - Many TV shows, popular magazines, and culinary books and manuals worldwide are devoted to the art of cooking.
 - Various scientific institutions, many scientific publications, and applications from the domain exist.

Introduction and Motivation

- Various aspects of the culinary domain continuously attract the research community. The existence of various scientific institutions and many scientific publications from the domain can serve as evidence.
 - IEHCA – Institut européen d’histoire et des cultures de l’alimentation, in Tours, France.
 - The new application from IBM “Chef Watson with Bob Apetit” uses Watson’s capabilities to explore big data to create new recipes.
 - A course at Stanford University held by Dan Jurafsky „The Language of Food“
 - A book by Dan Jurafsky „The Language of Food – A Linguist Reads the Menu“ (W.W. Norton & Company, September 2014)

=> development of the culinary linguistics

Initiating an informal project – a support for the culinary domain

- It was essential
 - to create the corpus of culinary content in Serbian that can be used for research;
 - to enrich the Serbian lexical resources with the appropriate terms from the domain:
 - Serbian WordNet (SWN) and
 - Serbian morphological electronic dictionaries

to provide a basis for the development for the culinary domain:

- An ontology and
- more complex natural language processing applications.

Outline

- Introduction and motivation
- Resources at disposal
 - A corpus
 - Serbian Wordnet
 - Serbian e-dictionaries
- Enrichment of resources
- Problems to solve

Details of the Culinary Text Corpus

- Corpus of Serbian written culinary recipes in the Latin script was formed from web texts.
- Existing programs were adjusted to particular web pages, their content and also the meta-data that can be useful for ongoing work.
- The created text corpus contains approximately 14,000 recipes (approximately 1,600.000 simple word forms).

Serbian WordNet

- The production of the SWN was initiated by the BalkaNet project (2001 – 2004).
- The structure of all SWN is linked to the Princeton WordNet (PWN), through the so-called Interlingual Index.
- Before the beginning of the (informal) culinary project, concepts belonging to the culinary domain were not given special attention.

Serbian WordNet

- However,
 - 393 such concepts were already present in the SWN,
 - 99 of which belong to basic concept sets and
 - 91 to Balkan- or Serbian-specific concepts.
- After enhancement, Serbian WordNet has
 - 1544 sysets from the culinary domain, of which
 - 364 belong to Balkan- or Serbian-specific concepts

Electronic Dictionaries for Serbian

- Follow the methodology and format known as DELA, covering both simple words and multi-word units (MWU).
- The system of Serbian e-dictionaries (SED) covers both general lexica and proper names (approximately 28.5%). All inflected forms are generated from **133,500** simple forms and **13,500** MWU lemmas.

Domain Specific Semantic Markers for Serbian Electronic Dictionaries

- Before starting the enrichment process, there were:
 - **218** simple word entries with the semantic marker **+Food**, and
 - **217** multi-word entries.
- All entries with the **+Food** marker should also have been assigned the **+Conc** marker (for concrete object, as a more general category), but it was not the case.

Conclusion: markers in SED were inconsistently assigned.

Domain Specific Semantic Markers for Serbian Electronic Dictionaries

Semantic marker	Description
+Culinary	culinary domain
+Food	food (e.g. senf 'mustard')
+Alim	aliment (e.g. mleko 'milk')
+Prod	product (e.g. sirće 'vinegar')
+Meal	meal (e.g. doručak 'breakfast')
+Course	course (e.g. puding 'pudding')
+Uten	utensil (e.g. šolja 'cup')
+Ing	ingredient (e.g. so 'salt')
+MesApp	approximate measures (e.g. kašičica 'spoonful')
+Taste	taste (e.g. slatkokiseo 'sweet-sour')
+WoP	way of preparation (e.g. dinstati 'to stew'; dinstanje 'stewing')
+Cond	condition (e.g. bajat 'stale')

Overview of the newly proposed semantic markers, that could be used individually or in combination.

Serbian Electronic Dictionaries after the first phase of the informal culinary project

- Entries with the **+Culinary** marker **3194**
 - 1617 simple word entries, and
 - 1577 multi-word entries.
- Among them
 - **+Food** – 2800 (1296 simple; 1504 MWUs);
 - **+Uten** – 173 (111 simple; 62 MWUs);
 - **+WoP** – 137 (137 simple)
 - **+MesApp** – 105 (96 simple; 9 MWUs)

Outline

- Introduction and motivation
- Resources at disposal
 - A corpus
 - Serbian Wordnet
 - Serbian e-dictionaries
- **Enrichment of resources**
- Problems to solve

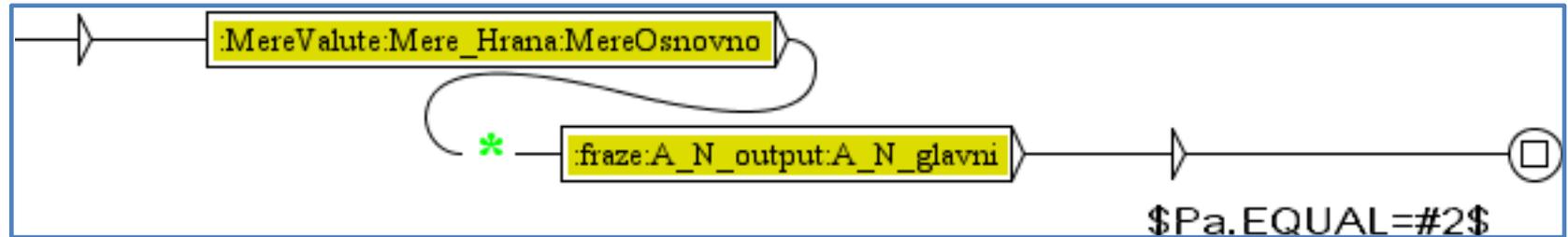
Further enhancement of resources

- Detaction of verbs that belong to culinary domain
- Their description:
 - How specific they are?
 - What syntactic properties they have (in the culinary domain)?
 - How can they be classified and marked?
 - Can we retrieve which verbs are used with which food?

Detection of verbs used in the culinary domain (1)

- If an **adjective** derived from a **verb past participle** is preceded by **numeric expressions with units of measure** (standard and approximate) and followed by a **noun in the genitive case that refers to food** (and possibly preceded by an **adjective** in the corresponding case, gender, number and animacy) then it can be **an adjective referring to a way of preparation of food**.

A Unifex graph that performs it



xeg belog mesa 5 kasxika
 kiseke pavlake 5 kasxika
 og belog mesa 3 sxolxice
 raju ili mleku 1/2 cyasxe
 endane jabu 3/4
 2/4
 2 sa
 2 sxake
 xixiku
 onzerva
 a banana
 log sira 1 m
 log sira 1 manxi
 ala bundevica 1 sxolxica
 i malo vode, 1 sxolxicu
 uvim zacyinima, 1 sxolxu
 eg origana ili 1 kasxika
 i koricu limuna i 2 rebra
 r mlevenih oraha, 3 rebra
 g mlevenih oraha 3 rebra
 g mlevenih oraha 3 rebra
 g mlevenih oraha 3 rebra
 drugi dodajte 2 sxtangle
 r jedan dodati 2 sxtangle
 og grozdxia i 3 sxtangle
 oraha, ruma i 4 sxtangle
 lan deo dodati 5 sxtangli
 uski baget, 1/2 sxolxice
 lijskog oraha 1-2 kasxike
 asxicyica soli 3 kasxike

an
 adjective
 derived
 from the
 past
 participle

A measurement unit -
 Traditional or
 approximate

A noun expressions
 – using nouns
 representing food

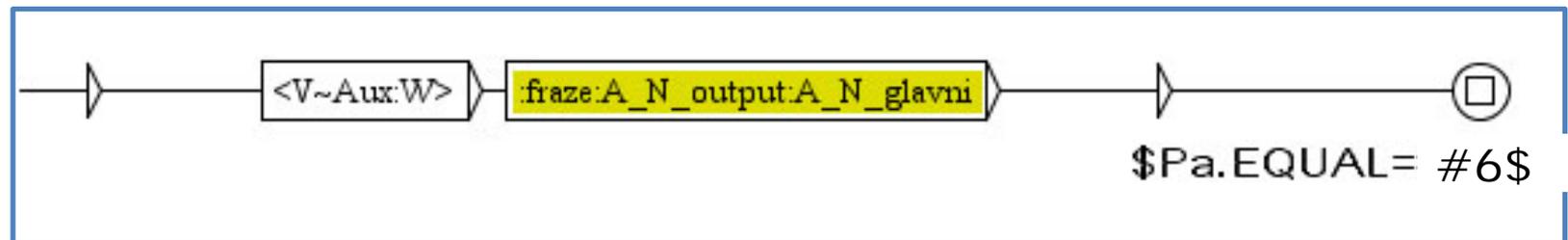
incyia 2 jaja i praziluk pola kasxicy.
 nacxa 1 praxak za pecivo 200 g kacyl
 incyia 1 glavica crnog luka malo soli
 ulxa u slanoj vodi 1/2 cyasxe socyiva
 sxolxe meda 1/4 sxolxe ulxa 2 k
 anxa 1 kasxika kakao 100 g lomlxene
 sxanxa 1 kasxika kakao 100 g lomlxene
 g krema. {S} Ovim nadevom premazati svaku
 olxusxtenog paradajza 1 cyasxa belog suvog vina
 omeksxanog butera, 1 solka sxecxera, 2 jajeta, 1
 Oprane tikvice, zajed
 sa korom izrendati u dubi
 opranog pirincyia
 opranog pirincyia
 opranog pirincyia P
 onxa berog luka, olxusxte
 ovim filom puniti jabuke
 cyokolade. {S} Ovim
 cyokolade, 3 kasxike brasxna i pola pra
 cyokolade 3 kasxike brasxna 1/2 praxka
 cyokolade 3 kasxike brasxna 1/2 praxka
 cyokolade 3 kasxike brasxna 1/2 praxka
 cyokolade. {S}U pomasxcxen i brasxnom p
 cyokolade, a drugi ostaviti da bide sve
 cyokolade. {S}Filovati: kora, krem sa k
 cyokolade napraviti nadev, pa nxime do
 cyokolade. {S}Tamni deo razvucxi u odgo
 putera ili margarina, 1/4 sxolxice sve
 Sxampinxone ocystiti sa mokrir
 maslaca 140 g brasxna pomesxanih 1 pra

An example

- The verb *otopiti* 'to melt'
- The past participle *otopljen*
 - Used with:
 - (bela, crna) čokolada '(white, black) chocolate'
 - (maslac, puter) 'butter'

Detection of verbs used in the culinary domain (1)

- If a **verb** in the **infinitive** or in the **imperative** (the second person plural) is followed by a **noun phrase** (e.g. Adj_N) from the culinary domain (a noun is marked by +Food) than the verb is the „**way of preparation**“ **verb** from the culinary domain
- Resulting concordances



Premazati kore filom i umucxenom cyokoladom. cyokolada pita **
toplu belu cyokoladu i umucxenom slatkom pavlakom i sxarenim m
sxati. {S} Vocxnu salatu umucxenom slatkom pavlakom i brusnicama
prebaciti na tacnu. {S} umucxenom slatkom pavlakom. torta pavla
to secyene komade umucxenom slatkom pavlakom i umucxenom slatkom pavlakom.
{S} Sipajte u cyinijice umucxenom slatkom pavlakom. ***** Glavna j
Servirajte u cyasxe i umucxenom slatkom pavlakom. umucxenom slatkom pavlakom. rum sxlag kolacy
. cyasxi. {S} Po zxelxi, umucxenom slatkom pavlakom i umucxenom slatkom pavlakom.
te cyokoladnim sosom i umucxenom slatkom pavlakom. banane keks
ne semenke lana mogu se umucxenom slatkom pavlakom.
ti ih i kada se ohlade umucxenom slatkom pavlakom.
pomorandye Zxuti rolat umucxenom slatkom pavlakom.
ponovo staviti list i umucxenom slatkom pavlakom.
pavlakom Zatim krusxke umucxenom slatkom pavlakom.
le. {S} Listove rasxtana umucxenom slatkom pavlakom.
na pola vodoravno, pa umucxenom slatkom pavlakom.
prohladiti, presecki i umucxenom slatkom pavlakom.
te. {S} Svaku palacyinku umucxenom slatkom pavlakom, pa ih pored
svati. {S} Jednu oblandu umucxenom slatkom pavlakom, a drugu staviti preko nxe
ugu staviti preko nxe i umucxenom slatkom pavlakom, te ih zajedno uviti u rolat
: fil - kora Celu tortu umucxenom slatkom pavlakom, i ukrasiti po zxelxi. torta
pice ohladiti, a zatim umucxenom slatkom pavlakom i iseckanim vocxem po
vi ili serbetom i odmah umucxenom slatkom pavlakom, dok se fil ne stegne. {S} O
stavite da se ohladi pa umucxenom slatkom pavlakom. {S} Za fil: {S} Paradajz (s
niju za posluzxivanxe, umucxenom slatkom pavlakom nane. ***** Sos i
niju za posluzxivanxe, umucxenom slatkom pavlakom nane. sir pavlaka li:
vati u duboke tanxire i umucxenom slatkom pavlakom ***** Glavn
preko svega celu tortu umucxenom slatkom pavlakom. ananas visx
umucxenom slatkom pavlakom. II, fil III. {S} Kolacy umucxenom slatkom pavlakom. plazma keks malin

A verb in
the
infinitive or
imperative

A noun expressions
– using nouns
representing food

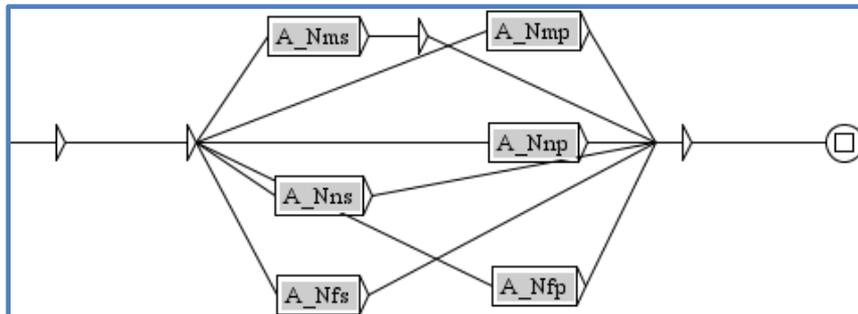
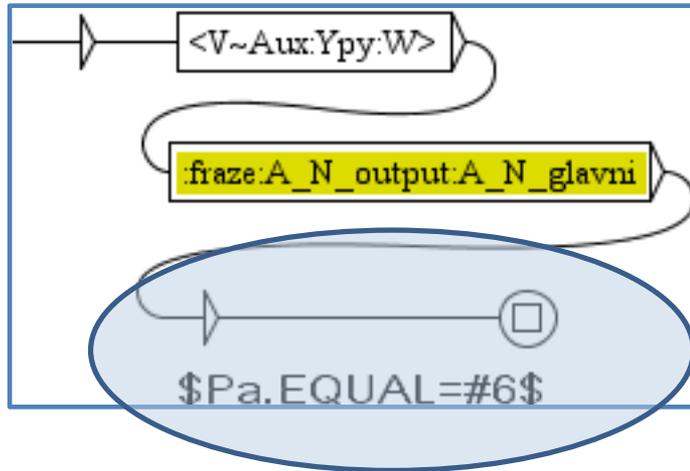
An example

- The verb *filovati* 'to fill; to stuff'
- To fill with what?
 - Used with:
 - (*braon, crni, svetli, tamni, čokoladni, vruć*) *fil*
'(brown, black, light, dark, chocolate, hot) fill'
 - *mleveno meso* 'minced meat'
 - (*umućena*) *pavlaka* '(whisked) cream'

Outline

- Introduction and motivation
- Resources at disposal
 - A corpus
 - Serbian Wordnet
 - Serbian e-dictionaries
- Enrichment of resources
- Problems to solve

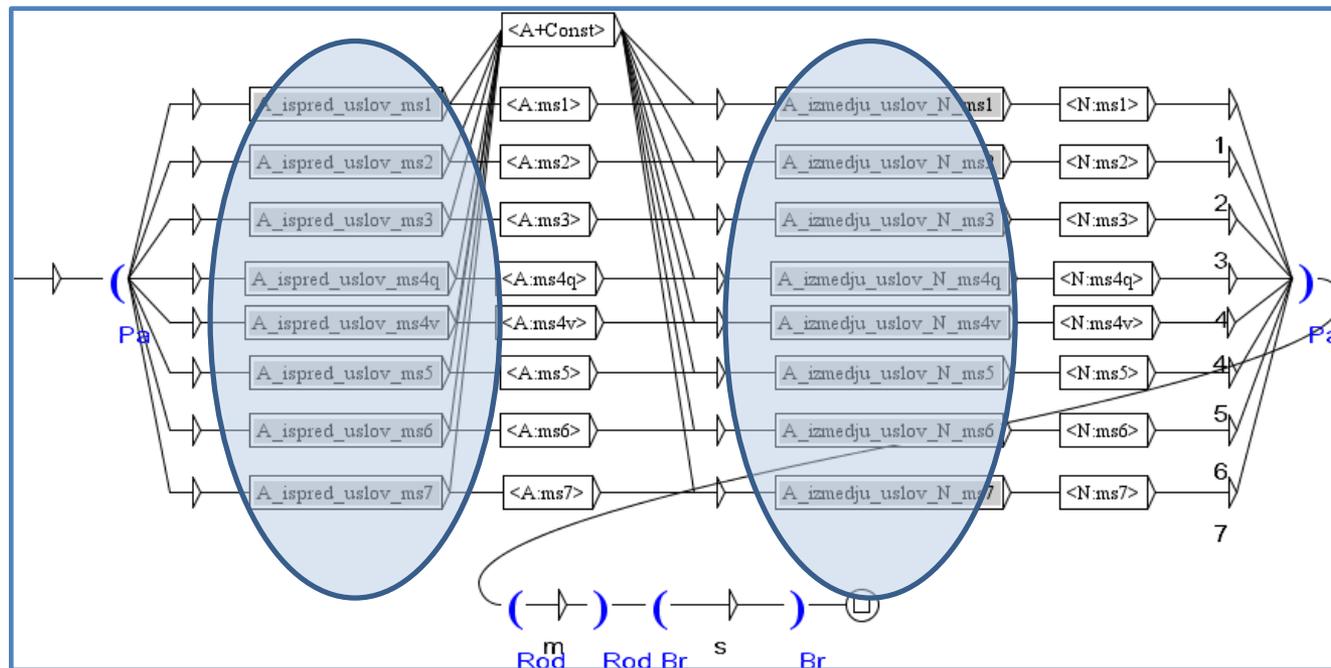
How do this extraction graphs work?



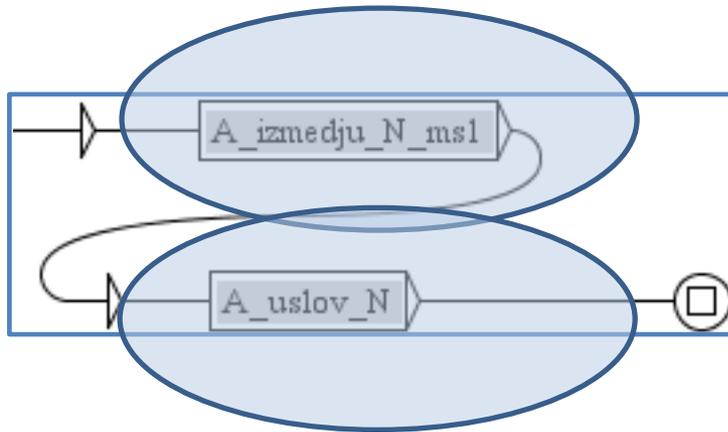
- It invokes a general graph for the Adj_Noun phrases that is placed in the repository
- This graph invokes six subgraphs, each for a pair of values for the number and the gender.
- We are looking only for Adj_Noun phrases that are in the instrumental case.

One of six subgraphs

- It takes care about the agreement in the case
- Subgraphs take care about
 - Insertions;
 - Conditions on adjectives or nouns.
- They remember grammatical categories as output variables



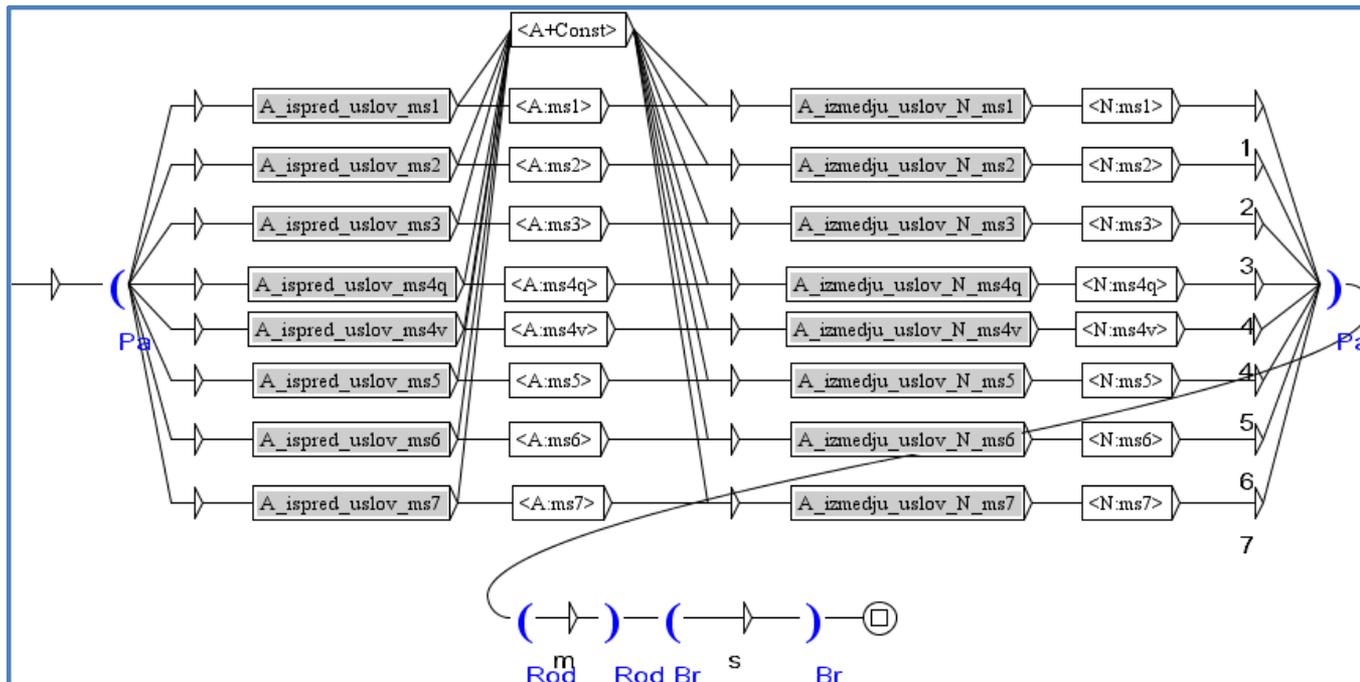
One insertion subgraph



- It invokes two subgraphs:
 - The first one is for possible insertion between the adjective and a noun.
 - In this case this subgraph is Empty.
 - The seconde one is for the condition the noun has to satisfy.
 - In this case a noun has to represent a food.

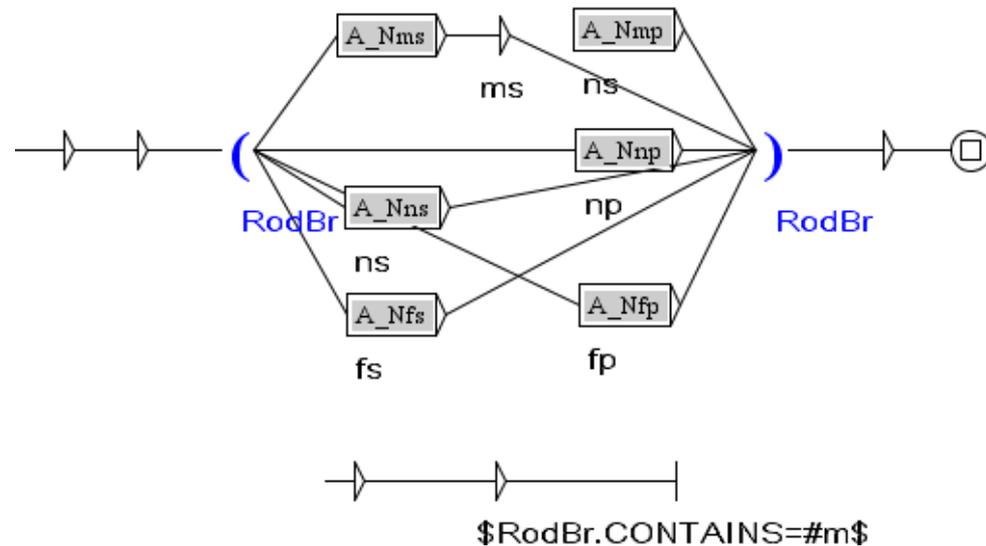
The first problem

- The values of grammatical categories are remembered at the lower level (in subgraphs) – repetition of the same nodes.
- It would be better to do it at the higher level.



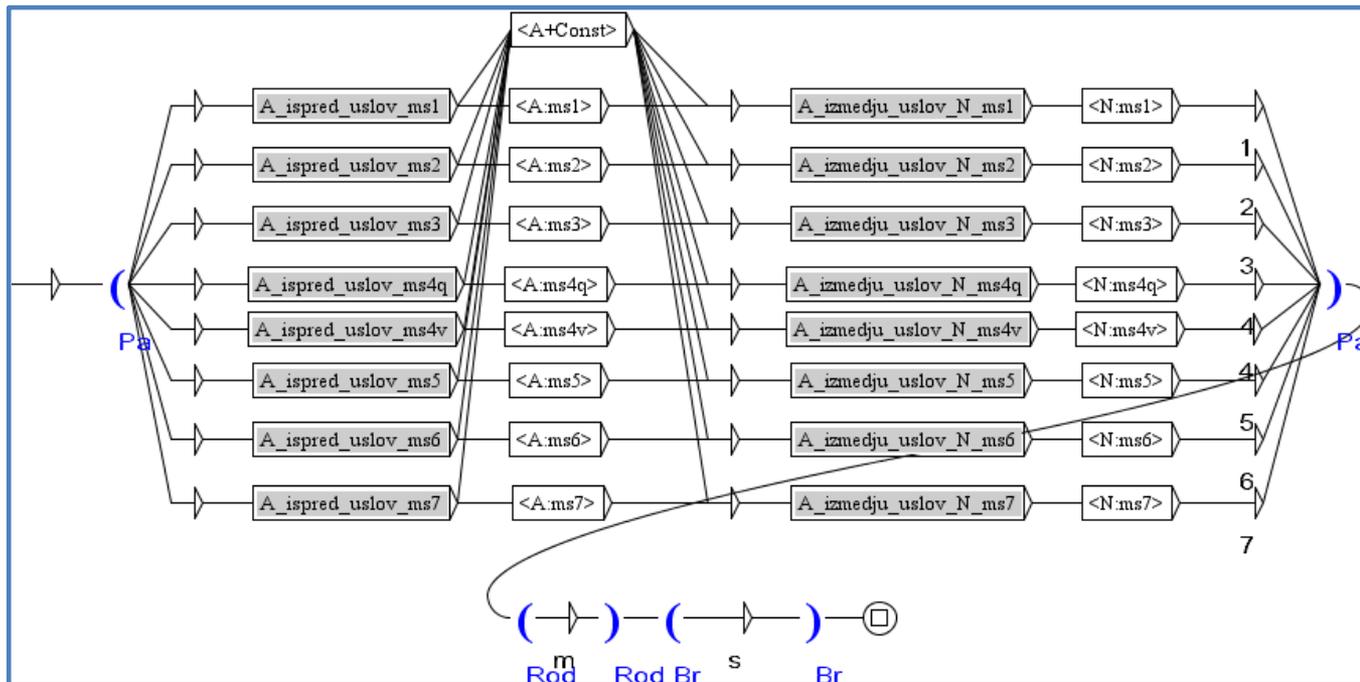
The solution to the first problem

- The values of grammatical categories can be remembered in the higher graph in a „compound variable“.
- The test on this value could be done with a new test like: **\$RodBr.CONTAINS=#m\$** (if it would exist).
- I learned two days ago in MLV that such a test actually exists but is not yet documented as **\$RodBr.SUBSTRING=#m\$**



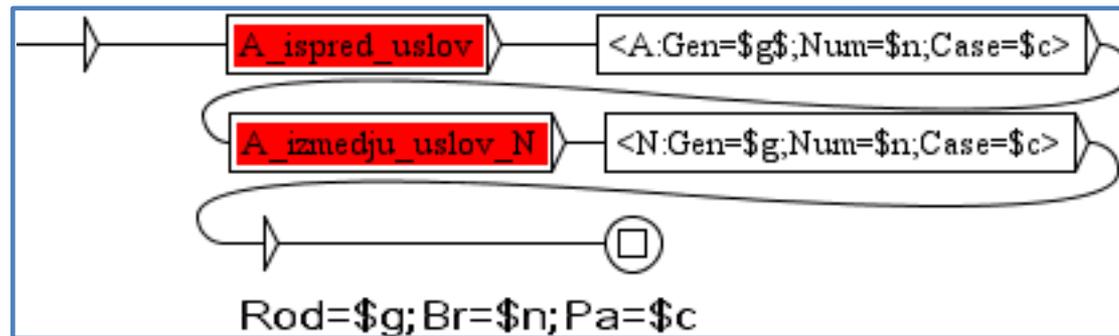
The second problem

- The graphs that test agreement are very complex.
- Every correction or enhancement in one path and one subgraph has to be redone for all paths (and all subgraphs).



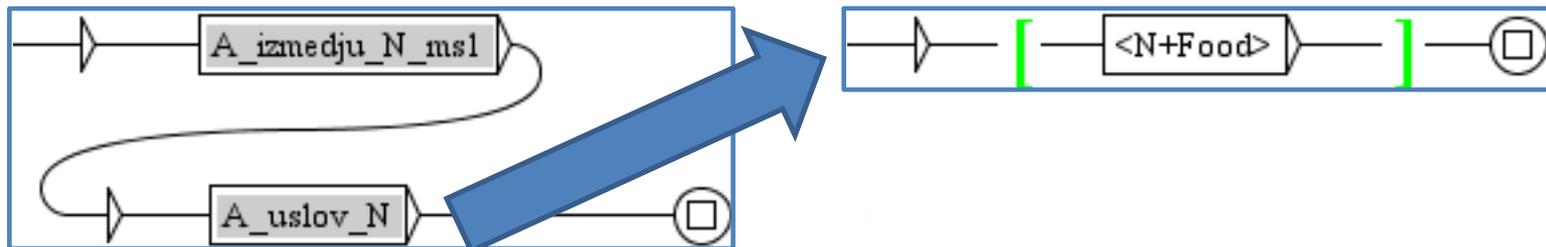
The wishful solution to the second problem

- They could be made much simpler if graphs could use unification variables (in a way similar to **Multiflex**)
- In that case used grammatical categories would have to be documented (like **Morphology** and **Equivalence** files).



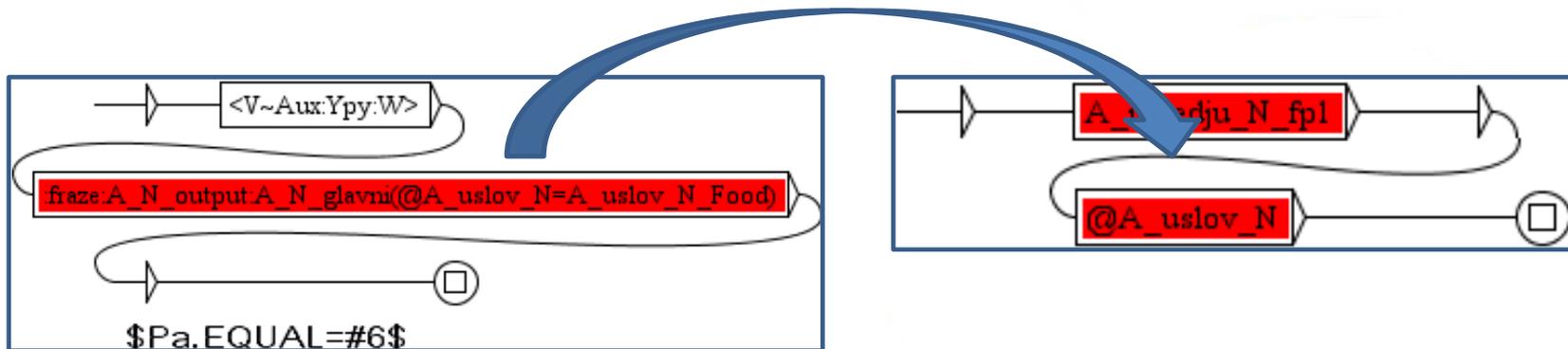
The third problem

- We wanted to make the **Adj_N** graph as general as possible so that it can be used in different context for different tasks.
- For that reason, the special conditions on nouns and adjectives, as well as possible insertions are put in separate graphs.
- However, the graph is not general enough because for different purposes we must have different copies of it (and all its subgraphs) with all inconveniences that follow from this.



The wishful solution to the third problem

- It would be easier if we could invoke subgraphs with parameters – subgraph names.
- These subgraph names would replace „formal graph names“.



The more realistic solution to the third problem

- Use the morphological mode to extract from the dictionary the semantic markers;
- Remember these markers in an output variable;
- Test for these values with the use of a **CONTAINS** or a **SUBSTRING** test.
- The cost of using the morphological mode!

**Thank you
for your
attention!**

cvetana@matf.bg.ac.rs

