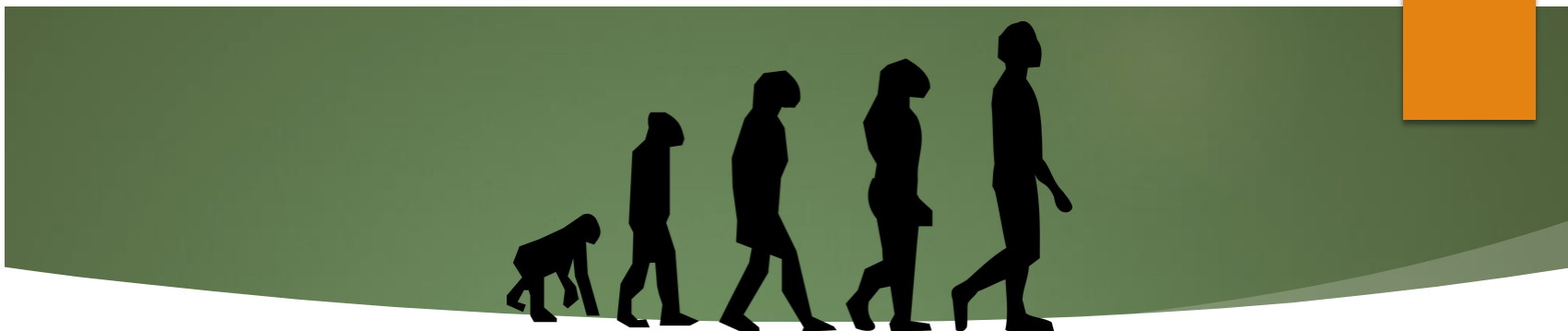


# Трансформација електронских речника

ОД DELA ДАТОТЕКА ДО  
ЛЕКСИЧКЕ БАЗЕ ПОДАТАКА

БИЉАНА ЛАЗИЋ, УНИВЕРЗИТЕТ У БЕОГРАДУ, РУДАРСКО-ГЕОЛОШКИ ФАКУЛТЕТ





- ▶ Matematički model morfologije srpskohrvatskog jezika : (imenska fleksija) : doktorska disertacija / Duško Vitas. - Beograd : [D. Vitas], 1993. - 284 lista
- ▶ Vitas, D., Pavlović-Lažetić, G., and Krstev, C. (1993). Electronic dictionary and text processing in Serbo-Croatian. Sprache–Kommunikation–Informatik, 1:225.
- ▶ Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije : doktorska disertacija / Cvetana Krstev. - Beograd : [C. Krstev], 1997. - 287 lista
- ▶ ...

# DELA речници

**DELA** (фр. Dictionnaires Electroniques du LADL)  
**формат**

- речник лема
  - simple-word lemmas (DELAS)
  - multi-word lemmas (DELAC)
- речник облика
  - simple word forms (DELAF)
  - multi-word forms (DELACF)

# Запис у речнику лема

recynik,N9+Ek+FLX=N9

lemma, POS#fst [**+Marker**]\*

semantic/ontology e.g. **+Hum**

pronunciation e.g. **+Ek**

derivation e.g. **+GM**

domain e.g. **+DOM=Mining**

syntactic e.g. **+Ref**

variation e.g. **+VAR=SatiRati**

information e.g. **+SI=d1**

# Запис у речнику облика

`form, lemma[:categories]*`



облик леме



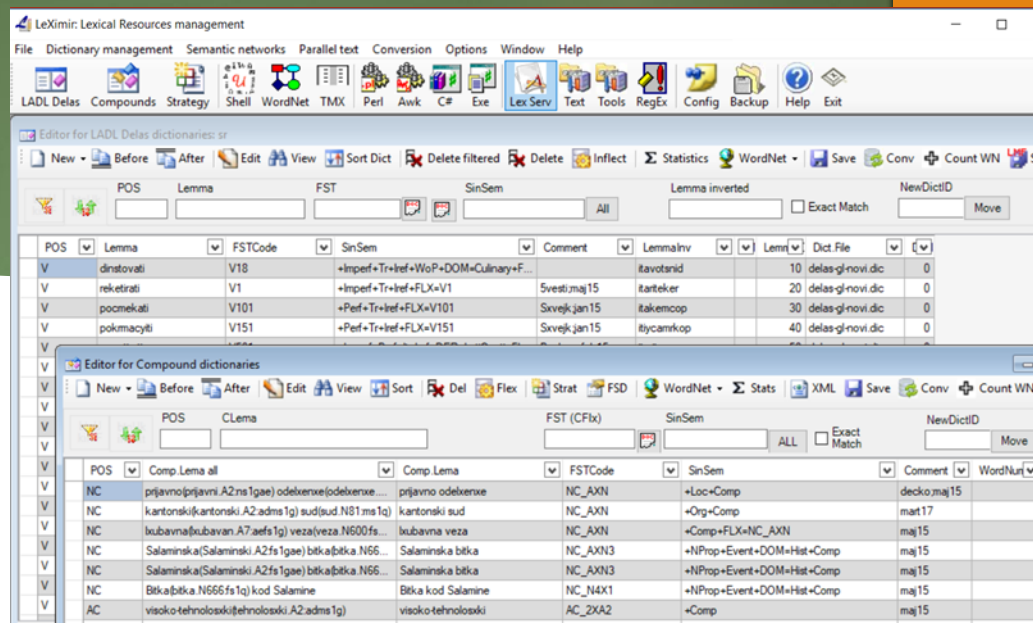
могуће граматичке категорије  
придružене облику речи,  
представљене једнокарактерским  
КОДОМ

recynici,recynik.N:mp1q:mp5q

# Мотивација

WS4LR\*

LeXimir



Нова  
апликација

# Модели за представљање речничких података

- ▶ Иницијатива за обележавање текста - TEI (енг. Text Encoding Initiative)
- ▶ Оквир за лексичко обележавање - LMF (енг. Lexical Markup Framework)
- ▶ *lemon* - Модел лексикона за онтологије (енг. Lexicon Model for Ontologies)

# Иницијатива за обележавање текста - TEI

- ▶ Широко прихваћени и заступњени стандард за обележавање најразличитијих врста текстуелних извора
- ▶ Поглавље 9 TEI 5 смерница посвећено искључиво речницима
- ▶ Али „само“ стандард за ОБЕЛЕЖАВАЊЕ текста
- ▶ Нешто више се користи за обележавање традиционалних дигитализованих речника али... постоји тенденција да се ово промени



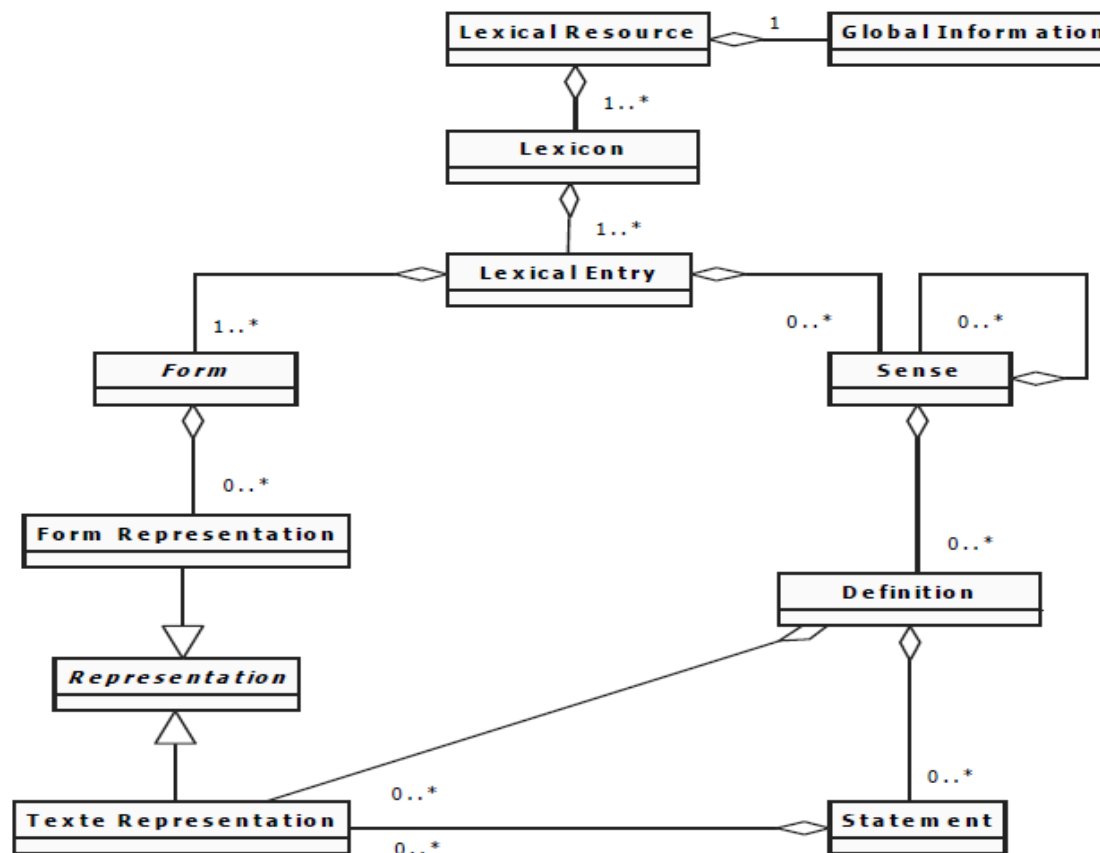
# Пример описа једног облика из речника DELAF

```
<form type="inflected">  
<orth>hidrometeorolozima</orth>  
<gramGrp>  
<gram type="rod">m</gram>  
<gram type="broj">p</gram>  
<gram type="padez">7</gram>  
<gram type="animatnost">v</gram>  
</gramGrp>  
</form>
```

# Оквир за лексичко обележавање - LMF

- ▶ Прописује стандардизован оквир за бележење лингвистичких информација приликом креирања рачунарских лексикона
- ▶ Заснован на ISO 24613:2008
- ▶ Циљ - коришћење истих речника за различите апликације, изградња модела за употребу лексичких ресурса и спајања индивидуалних ресурса у један гоблани
- ▶ Пружа могућност описа лексичких објеката, њиховог морфолошког, семантичког или другог аспекта коришћењем одговарајућих пакета

# Оквир за лексичко обележавање - LMF



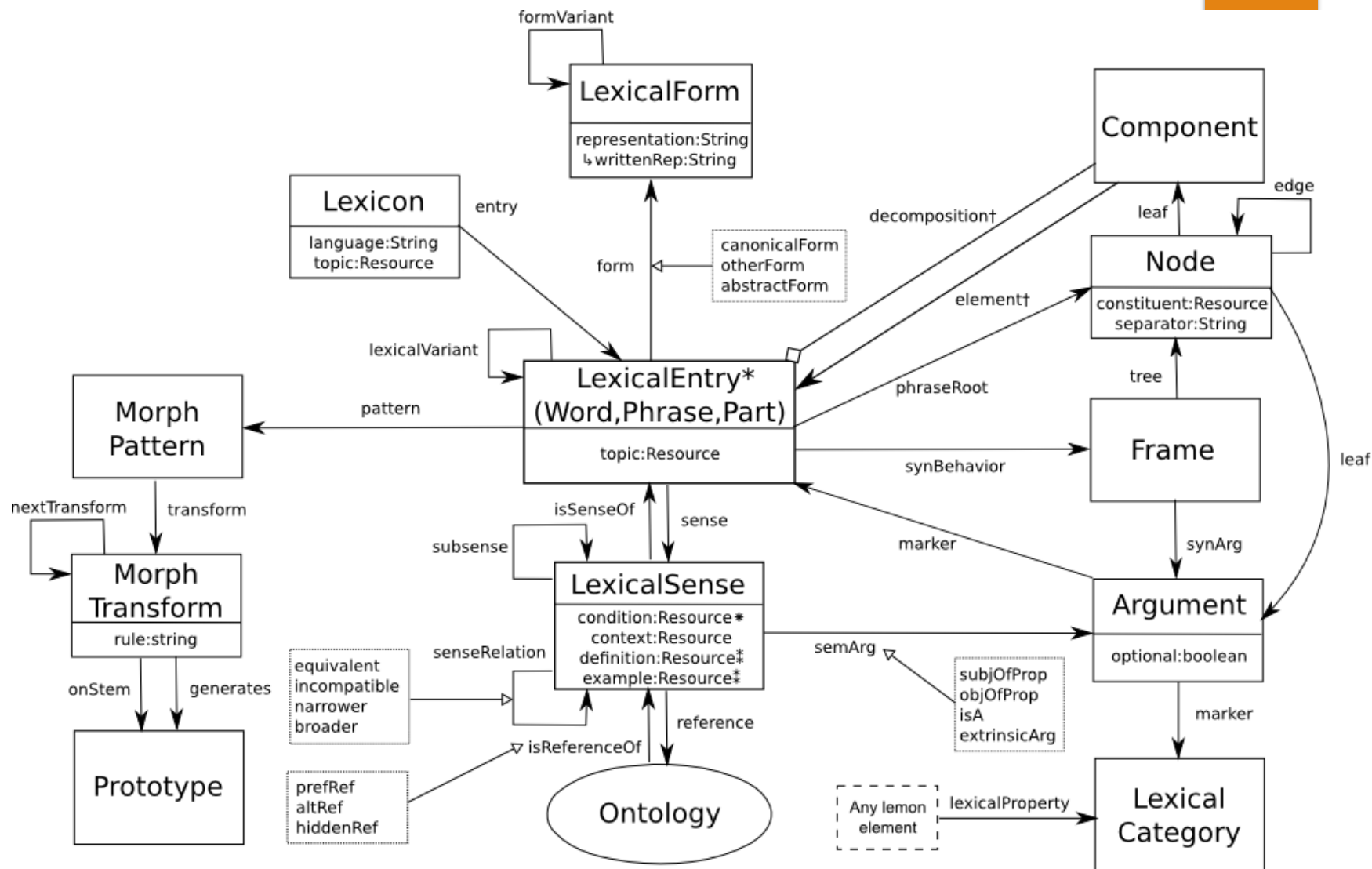
Francopoulo, Gil, и Monte George. 2013. „Model Description“. У *LMF Lexical Markup Framework*, 19–40. John Wiley & Sons, Inc.

# *Iemon* - Модел лексикона за онтологије

- ▶ Повезује некада неспојиве светове, веб онтологије и, с друге стране, лексичке и лингвистичке ресурсе
- ▶ Сврха модела је подршка језичкој заснованости онтологије – додавање информација о томе како су елементи онтологије лексикализовани у природном језику
- ▶ Заснован на RDF (енг. Resource description framework)

# *leton* - Модел лексикона за онтологије

- ▶ Заснован на 5 основних модула
  - Повезивање онтологије и лексикона (енгл. Ontology-lexicon interface - ontolex)
  - Синтакса и семантика (енгл. Syntax and semantics - synsem)
  - Декомпозиција (енгл. Decomposition - decomp)
  - Варијације и превод (енгл. Variation and Translation - vartrans)
  - Лингвистички метаподаци (енгл. Linguistic Metadata - lime)



\* LexicalEntry has three subclasses: Word, Phrase, Part

\* definition and example are stated as nodes with a value

\* condition has subproperties propertyDomain and propertyRange

† decomposition and element may also be used with Frames and Arguments resp.

<https://www.w3.org/2016/05/ontolex/>

lemon

# Поређење

- ▶ *Lemon* тежи коришћењу спољашњих лексикона за различите потребе (нпр. граматичке категорије)
- ▶ Има мање елемената од LMF-а, називи елемената су скраћени
- ▶ *Lemon* тежи онтолошком приступу дефинисања значења јер он нуди више могућности за изражавање нијансирања значења и специфичности значења у појединим језицима у односу на LMF
- ▶ TEI више служи за обележавање и не пружа толико могућности у производњи различитих релација
- ▶ АЛИ... Дешавају се промене тако TEI тежи повезивању са Повезаним отвореним подацима (енг. Linked Open Data, LOD)...

# Лексичка база

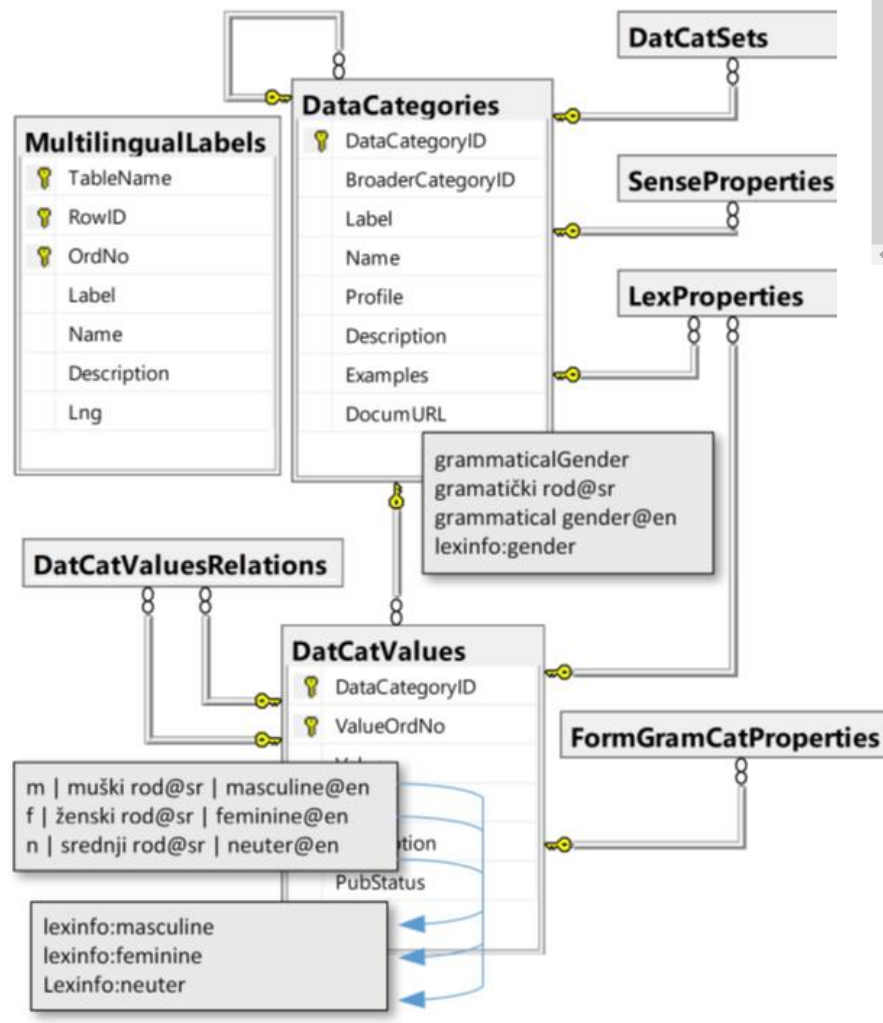
- ▶ Централни лексички репозиторијум
- ▶ Вишекорисничко управљање лексичким подацима
- ▶ Извоз у различитим форматима
- ▶ Аутоматска производња речничких издања (пуни речници, јавна верзија, уређени према неком критеријуму (екавски или ијекавски))
- ▶ ...



# Модел лексичке базе за Српске морфолошке речнике

- ▶ Заснована на *lemon* моделу
- ▶ Како би се на неко време омогућио паралелни развој, у LeXimir су интегрисане развијене процедуре за трансформацију података из постојећих речника у *lemon* модел
- ▶ Главне класе за лексичке записе, морфолошке, синтаксичке и семантичке особине су контролисане интерним тезаурусом категорија података
- ▶ База података садржи маркере који имају хијерархијску структуру која служи као контролор за домене или нека поља у бази

# LexInfo



## Data Categories Board

Root category

- ⊕ (Kategorija ne postoji)
- ⊕ semantički markeri (semantic/ontology)
- ⊕ sintaksički markeri (SintMark)
- ⊕ derivacioni markeri (DerivMark)
- ⊕ izgovor (pronunciation)
- ⊕ varijacije (variation)
- ⊖ domenski (domain)
  - ⊖ domeni (DOM)
    - ⊖ (domenski markeri za rudarstvo)
    - ⊖ (domenski markeri za bibliotekarstvo i informatiku)
- ⊖ informacije (Information)
  - ⊖ country code (CC2)
  - ⊖ anglosaksonske mere (AS)
  - ⊖ jedinice (SI)
  - ⊖ kodovi država (ISO)
- ⊖ gramatičke kategorije ( )
  - ⊕ vrste reči (POS)
  - ⊖ gramatički rod (grammaticalGender)
  - ⊖ gramatički broj (grammaticalNumber)
  - ⊖ padež (grammaticalCase)
  - ⊖ živo (grammaticalAnimateness)
  - ⊖ komparacija (grammaticalDegree)
  - ⊖ pridevski vid (grammaticalDefinitness)
  - ⊖ lice (grammaticalPerson)
  - ⊖ oblik (grammaticalForm)
  - ⊖ enklitika (grammaticalClitic)
  - ⊖ negacija (grammaticalNegation)

[+ Add Sibling Category](#)
[+ Add Child Category](#)

## Manage Category Details

|                                 |                   |
|---------------------------------|-------------------|
| ID                              | 10002             |
| Parent ID                       | 1015              |
| Label                           | grammaticalNumber |
| Name                            | gramatički broj   |
| Profile                         | gramcats          |
| Description                     | ***               |
| Examples                        |                   |
| DocumURL                        |                   |
| <a href="#">Update database</a> |                   |

## Manage Category Values

[Add New / Cancel](#)

| Value | Label   | Description | Status                                      |
|-------|---------|-------------|---------------------------------------------|
| s     | jednina |             | <a href="#">Edit</a> <a href="#">Delete</a> |
| p     | množina |             | <a href="#">Edit</a> <a href="#">Delete</a> |
| w     | paukal  |             | <a href="#">Edit</a> <a href="#">Delete</a> |

# Маркери

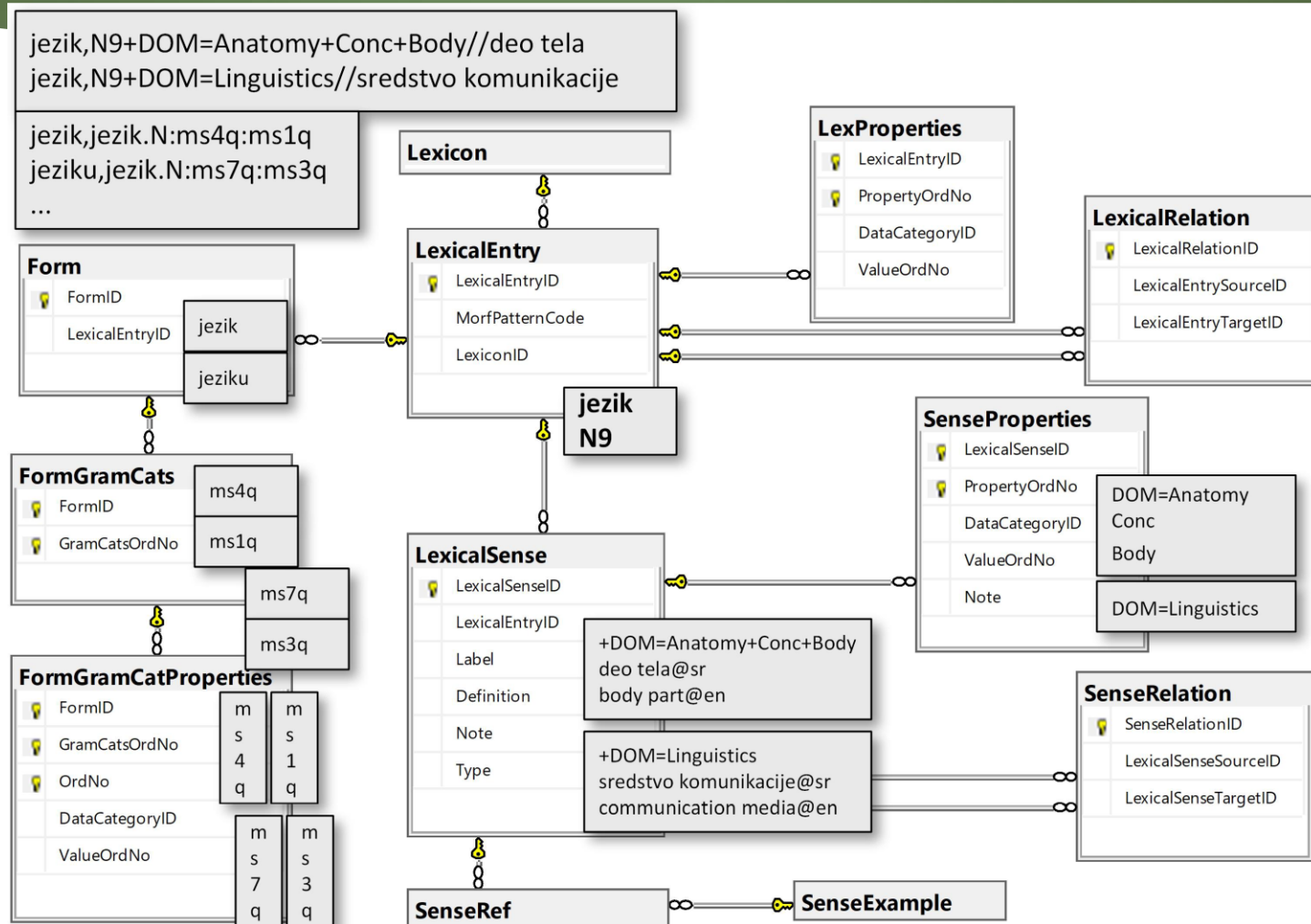
- ▶ Током транзиције у лексичку базу која подржава контролу домена открили смо неке недоследности у коришћењу маркера, као нпр. коришћење истих маркера у различите сврхе, непостојање маркера, маркери придружени неодговарајућим категоријама података итд.
- ▶ 23 семантичка маркера (нпр. +Hum)
- ▶ 17 синтаксичких (нпр. +Ref)
- ▶ 24 граматичка (нпр. V)

# lemon klase

jezik,N9+DOM=Ling

jezik,N9+Conc+Body+DOM=Anatomy

jezik,N9+Conc+Food+Prod+DOM=Culinary





N

Lemma: jezik  
 Canonical form: jezik [See all forms](#)

Properties:

Add property:

Morf pattern code description: N9

sr S Lexicon : delas-im.dic

Note:

Relations:

- To [jezicyki](#) using k\_cyki
- To [jezicyni](#) using k\_cyni

## Lexical Senses (3):

[+Add Sense](#)

1 . 14107 +DOM=Ling+FLX=N9 (sredstvo komunikacije)

Label: 1

Definition: +DOM=Ling+FLX=N9

Domains:

**DOM=Ling**

Properties:

Add property:

Add domain:

Note:

sredstvo komunikacije

Is a component of:

- [gestovni jezik](#)
- [govorni jezik](#)
- [materni jezik](#)
- [razgovoran jezik](#)
- [romanski jezik](#)
- [ruskoslavenski jezik](#)
- [strani jezik](#)
- [zvanicyni jezik](#)
- [sxatrovacyki jezik](#)
- [jezik mrzxnxe](#)
- [aglutinativni jezik](#)
- [arnautski jezik](#)

# Побољшања речника

- ▶ Корекција неисправних маркера (најчешће насталих грешкама при куцању), дуплих маркера (који означавају исти концепт)
- ▶ Нпр. облици `hilxadarka` и `ilxadarka` су означени истим маркером `DER=H0` али би облик `ilxadarka` требало означити маркером `DER=0H`
- ▶ Конверзија свих маркера који наводе везу између лема у праве везе међу лексичким записима

# Успостављање веза међу лексичким записима

- ▶ Везе између екавског и ијекавски облика Ek-ljk  
devojka, N617+Hum+**Ek**  
d**j**evojka, N617+Hum+**ljk**
- ▶ Деривационе везе  
grebat**i**, V19+Imperf+Tr+lref+Ref+**VAR=TiNje**  
greban**je**, N300+VN+**VAR=NjeTi**
- ▶ Везе међу језичким варијантама  
afirmir**ati**, V1+Imperf+Perf+Tr+lref+Ref+**VAR=RatiSati**  
afirmi**sati**, V21+Imperf+Perf+Tr+lref+Ref+**VAR=SatiRati**  
**h**leb, N81+**VAR=HO**+Ek+Conc+Course+Food+DOM=Culinary  
leb, N81+**VAR=OH**+Ek+Conc+Food+Prod+DOM=Culinary



# Повезивање лексичких записа

- ▶ Развијен је сет Unitex графова, SQL процедура и C# алатки за аутоматизацију задатка повезивања лексичких записа
- ▶ Систем заснован на правилима интегрисан у базу и апликацију
- ▶ Иако је главни циљ био повезивање постојећих записа, ове алатке су омогућиле нове могућности за анотацију и проширење речника

# Успостављање деривационих веза

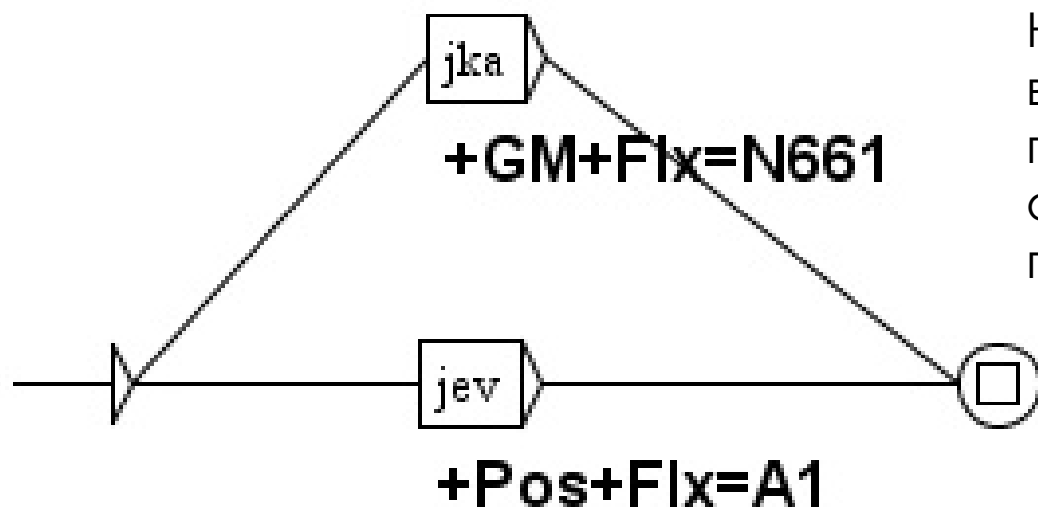
- ▶ У морфолошким речницима се налази око 18 000 презимена пореклом из српског језика и око 7 500 презимена страног порекла транскрибованих према српској ортографији
- ▶ Само 850 присвојних придева и 25 женских презимена је записано у речницима
- ▶ Lazić - Lazićev - Lazićka
- ▶ Ešton (Ashton) - Eštonov – Eštonka
- ▶ Како би се ове леме произвеле, развијено је 16 трансдуктора

# Černi - Černijev - Černijka

Černi,  $N1064 + NProp + Hum + Last + SR$

Černijka,  $N661 + GM + NProp + Hum + Last + SR + BASE = Černi\_N1064$

Černijev,  $A1 + Pos + NProp + Hum + Last + SR + BASE = Černi\_N1064$



На овај начин је произведено више од 24 000 присвојних придева и око 20 000 женских облика презимена који су повезани са основним лемама

# Правила за лексичке варијанте

- ▶ У СМР је 2539 лексичких парова обележено са 86 VAR маркера
- ▶ Процедура се заснива на affix\_variations правилима (218 правила)

Проналази кандидате који имају исту поднису (једно или више слова, али и празну нису која указује на то да је могуће изостављање) било где у лексичком запису и одговарајући маркер

# VAR=0H

- ▶ Пример: VAR=0H – лексички запис где је слово h изостављено и може бити уметнуто у варијанти (нпр. leб - hleб)
- ▶ Инверзно правило VAR=H0 наводи да слово h може бити изостављено
- ▶ hleб,N81+VAR=H0+Ek+Conc+Course+Food +DOM=Culinary
- ▶ leб,N81+VAR=0H+Ek+Conc+Food+Prod +DOM=Culinary

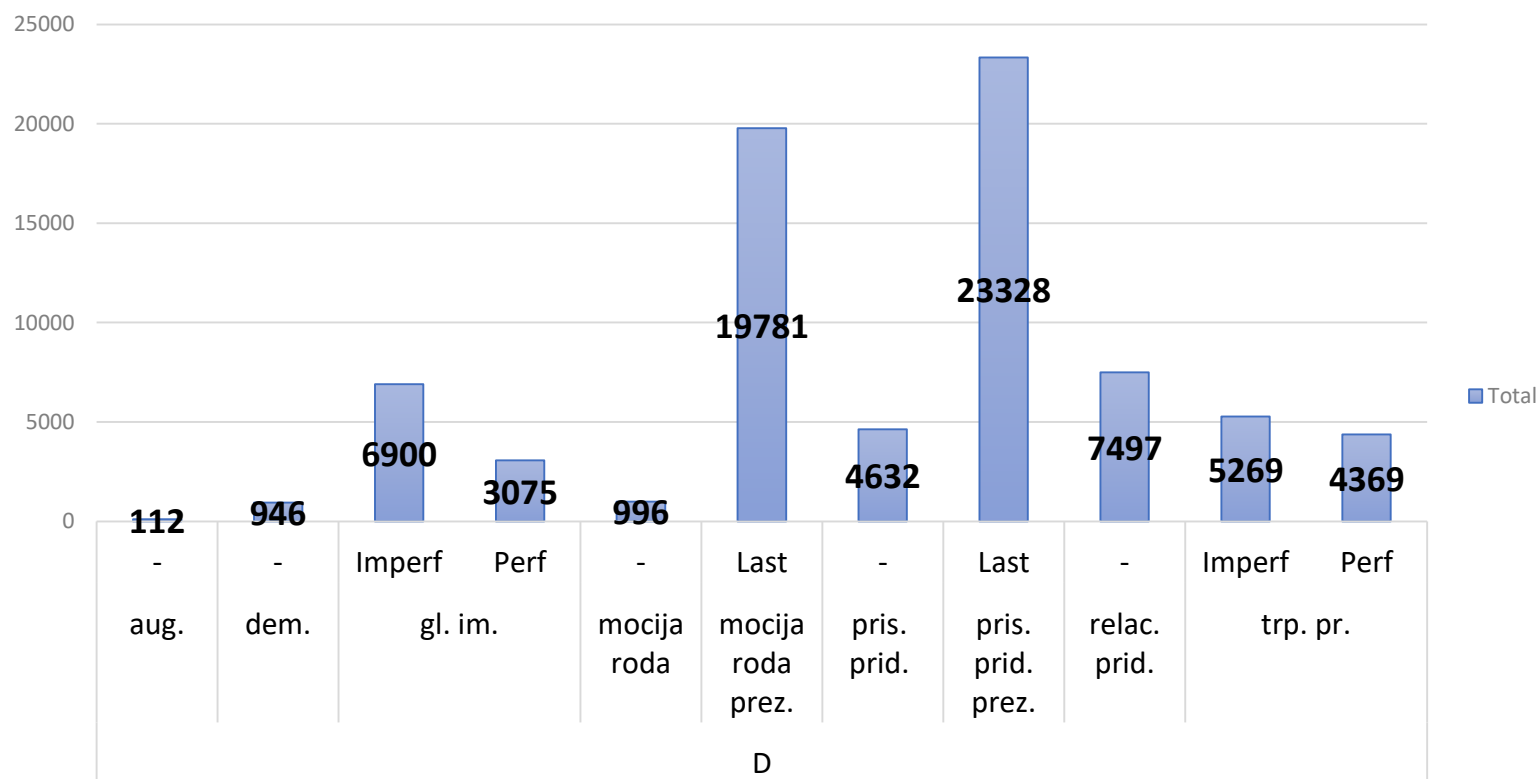
# VAR=CS

- ▶ Правило VAR=CS понаша се слично али уместо изостављања/уметања уводи замену
- ▶ Тако слово s може бити замењено словом c, нпр. sufinan**c**iranje од sufinan**s**iranje
- ▶ Ова правила нису превише успешна у проналажењу кандидата за проширење речника због великог броја кандидата које је могуће произвести услед непрецизираних позиција подниске
- ▶ Нпр. sufinansiranje са AffixFrom које је слово s и AffixTo који је слово c могу произвести: sufinan**c**iranje, **c**ufinansiranje и **c**ufinan**c**iranje, с тим што је само прво тачно

# Статистика на данашњи дан

| Тип везе            | Број повезаних парова |
|---------------------|-----------------------|
| Деривације          | 78 467                |
| Екавски - ијекавски | 4 863                 |
| Варијације          | 2 410                 |
|                     | 85 740                |

# Статистика - деривације





## Data Category Values Relation

Save

Source: markeri ličnih imena/Last

Destination: markeri (derivacije)/Pos

Change Source Value  
Change Destination Value

Relation simetric: ☐ yes ☒ no

Relation type: D Label: pris. prid. prez.

Rules:

+Add New Rule

|       | POS From | POS To | Fix From | Fix To | Afix From | Afix To | Marker From | Marker To | Example From | Example To  | Stem End |  |  |
|-------|----------|--------|----------|--------|-----------|---------|-------------|-----------|--------------|-------------|----------|--|--|
| 48/1  | N        | A      |          |        |           | ev      |             |           | Tanasicx     | Tanasicxev  |          |  |  |
| 48/2  | N        | A      |          |        |           | ov      |             |           | Abot         | Abotov      |          |  |  |
| 48/3  | N        | A      |          |        |           | jev     |             |           | Musolini     | Musolinijev |          |  |  |
| 48/4  | N        | A      |          |        |           | v       |             |           | Tintoreto    | Tintoretov  |          |  |  |
| 48/5  | N        | A      |          |        | a         | ov      |             |           | Sxenoa       | Sxenoov     |          |  |  |
| 48/6  | N        | A      |          |        | a         | in      |             |           | Bajra        | Bajrin      |          |  |  |
| 48/7  | N        | A      |          |        | ac        | cyev    |             |           | Mostarac     | Mostarcyev  |          |  |  |
| 48/8  | N        | A      |          |        | ac        | cyin    |             |           | Javorac      | Javorcyin   |          |  |  |
| 48/9  | N        | A      |          |        | ak        | kov     |             |           | Cyanak       | Cyankov     |          |  |  |
| 48/10 | N        | A      |          |        | ak        | kov     |             |           | Cyanak       | Cyankov     |          |  |  |

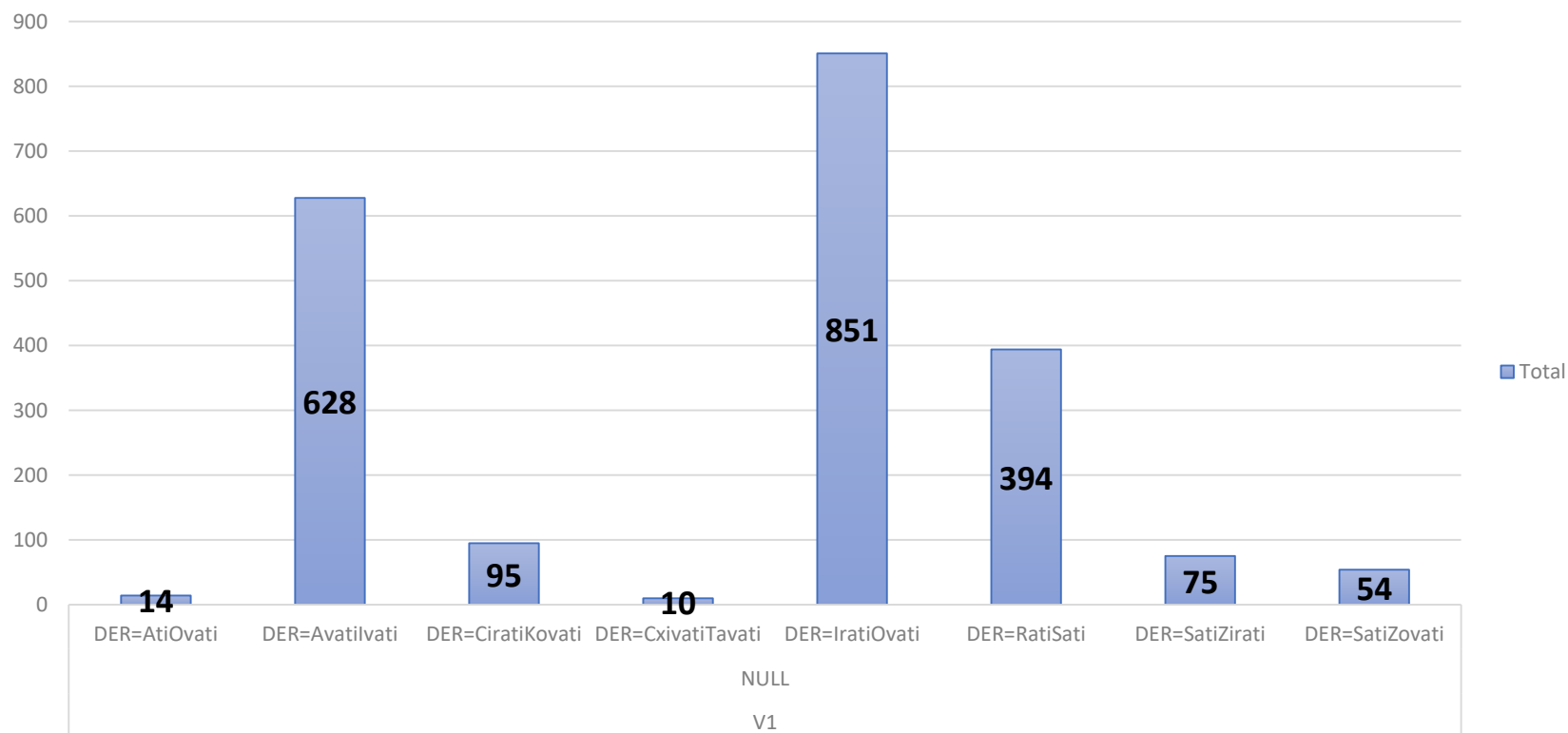
A total of 13015 entry pairs were established,  
and a total of 0 additional candidates are proposed.

|              |                |    |
|--------------|----------------|----|
| Agicyicx     | Agicyicxev     | AU |
| Aleksicx     | Aleksicxev     | AU |
| Abrasxevicx  | Abrasxevicxev  | AU |
| Andxelkovicx | Andxelkovicxev | AU |
| Anticx       | Anticxev       | AU |
| Arsicx       | Arsicxev       | AU |
| Andricx      | Andricxev      | AU |
| Antosijevicx | Antosijevicxev | AU |

Relation Candidates

| Comment | RelType | POS | FSTCode | LemaID | LemaFrom | LemaGener | relName | LemaTo |
|---------|---------|-----|---------|--------|----------|-----------|---------|--------|
|---------|---------|-----|---------|--------|----------|-----------|---------|--------|

# Статистика - варијације



# Data Category Values Relation

Save

Source:

Destination:

[Change Source Value](#)  
[Change Destination Value](#)

Relation simetric: ☐ yes ☒ no

Relation type:  Label:

Rules: [+Add New Rule](#)

|      | POS From                         | POS To                           | Flx From             | Flx To               | Afix From                           | Afix To                             | Marker From          | Marker To            | Example From                               | Example To                                 | Stem End             |  |  |
|------|----------------------------------|----------------------------------|----------------------|----------------------|-------------------------------------|-------------------------------------|----------------------|----------------------|--------------------------------------------|--------------------------------------------|----------------------|--|--|
| 11/1 | <input type="text" value="A"/>   | <input type="text" value="A"/>   | <input type="text"/> | <input type="text"/> | <input type="text" value="iran"/>   | <input type="text" value="ovan"/>   | <input type="text"/> | <input type="text"/> | <input type="text" value="civiliziran"/>   | <input type="text" value="civilizovan"/>   | <input type="text"/> |  |  |
| 11/2 | <input type="text" value="ADV"/> | <input type="text" value="ADV"/> | <input type="text"/> | <input type="text"/> | <input type="text" value="irano"/>  | <input type="text" value="ovano"/>  | <input type="text"/> | <input type="text"/> | <input type="text" value="civilizirano"/>  | <input type="text" value="civilizovano"/>  | <input type="text"/> |  |  |
| 11/3 | <input type="text" value="N"/>   | <input type="text" value="N"/>   | <input type="text"/> | <input type="text"/> | <input type="text" value="iranxe"/> | <input type="text" value="ovanxe"/> | <input type="text"/> | <input type="text"/> | <input type="text" value="civiliziranxe"/> | <input type="text" value="civilizovanxe"/> | <input type="text"/> |  |  |
| 11/4 | <input type="text" value="V"/>   | <input type="text" value="V"/>   | <input type="text"/> | <input type="text"/> | <input type="text" value="irati"/>  | <input type="text" value="ovati"/>  | <input type="text"/> | <input type="text"/> | <input type="text" value="civilizirati"/>  | <input type="text" value="civilizovati"/>  | <input type="text"/> |  |  |



A total of 265 entry pairs were established,  
and a total of 0 additional candidates are proposed.

|                   |                   |    |
|-------------------|-------------------|----|
| antropomoriiziran | antropomoriizovan | AU |
| apsolutiziran     | apsolutizovan     | AU |
| apsorbiran        | apsorbovan        | AU |
| apstrahiran       | apstrahovan       | AU |
| arhaiziran        | arhaizovan        | AU |
| asimiliran        | asimilovan        | AU |
| atomiziran        | atomizovan        | AU |
| autoriziran       | autorizovan       | AU |

Ralation Candidates

| Comment | RelType | POS | FSTCode | LemaID | LemaFrom | LemaGener | relName | LemaTo |
|---------|---------|-----|---------|--------|----------|-----------|---------|--------|
|---------|---------|-----|---------|--------|----------|-----------|---------|--------|

# Уз помоћ развијених правила решавају се подзадаци:

- ▶ Повезивање лексичких записа који задовољавају неко од дефинисаних правила
- ▶ Проналажење лексичких записа којима недостаје маркер (нпр. *alxina* има маркер DER=0H, док запис *halxina* није означен маркером DER=H0)
- ▶ Проналажење нових кандидата за речник (нпр. у речнику постоји реч *prehistorija* обележена маркером +DER=H0, док реч *preistorija* не постоји у речнику али је сасвим валидан кандидат)

# ПЛАНОВИ...

- ▶ 1. фокус је план на рударском речнику, као и речник геохазарда
- ▶ Повезивање записа са WordNet-ом
- ▶ Интегрисање са терминолошком базом
- ▶ Употреба принципа за омогућавање отвореног објављивања и повезивања језичких ресурса на Вебу,
- ▶ Извоз речина у различитим форматима (нпр. Turtle, RDF/XML)
- ▶ Развој API-ја за подршку потреба различитих апликација које се баве обрадом природних језика



► Хвала на пажњи!