From DELA Dictionaries to Leximirka Lexical Database

Biljana Lazić, Mihailo Škorić
University of Belgrade, Serbia

Serbian Unitex Day

11.3.2019

- *Managing dictionaries?*

BL2
BL3
BL4

- Serbian morphological dictionaries – SMD
25 years from textual DELA
to relational database format

Slide 2

BL2 Treba reći da se ideja o SMD javila pre 25 godina i da se razvijaju...
Biljana Lazić, 3/9/2019

BL3 potkrepljeno:
Biljana Lazić, 3/9/2019

BL4 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
- .

Biljana Lazić, 3/9/2019

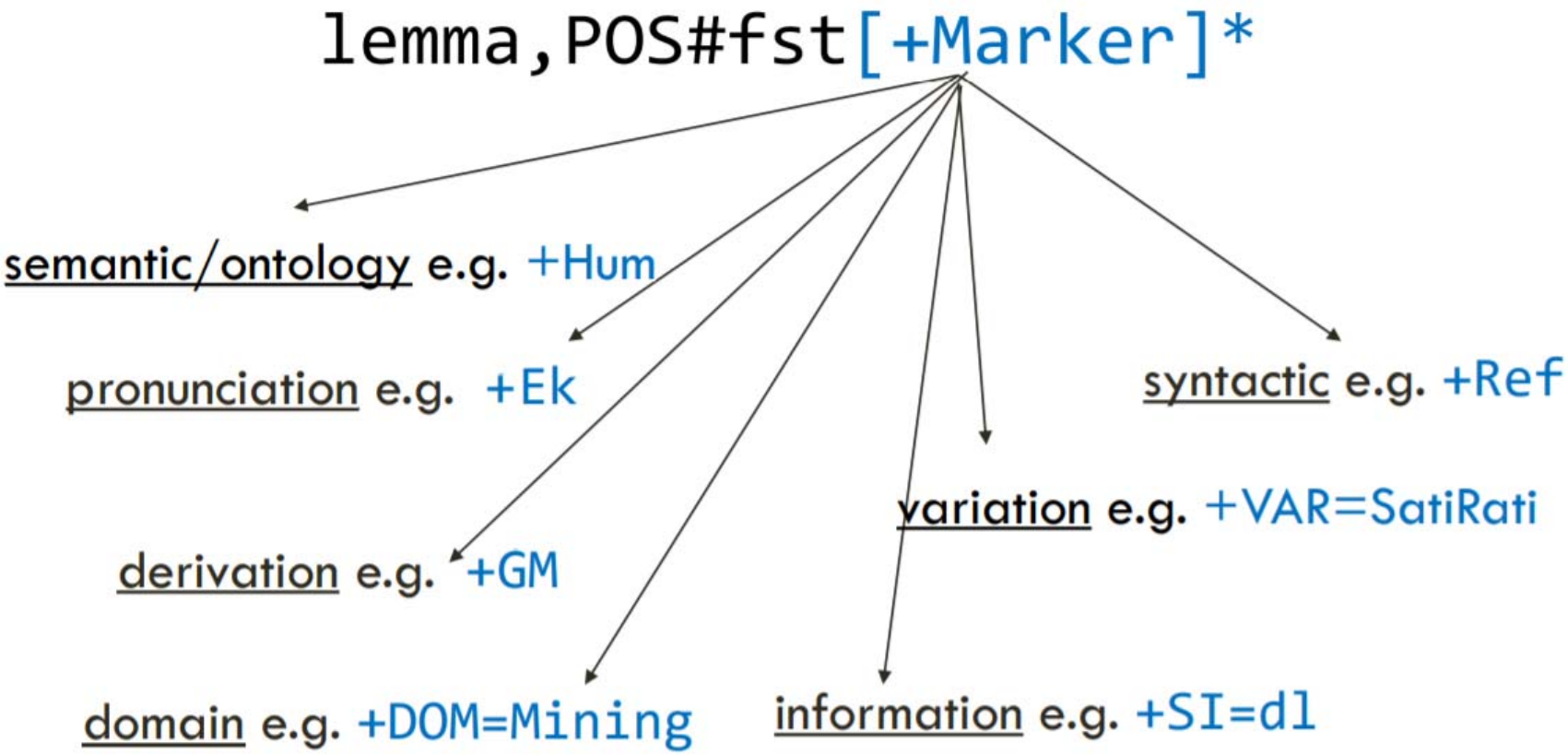
DELA dictionaries

- simple word lemmas (DELAS)
- multi-word lemmas (DELAC)

- simple word forms (DELAF)
- multi-word forms (DELACF)

DELA lexical entry structure

BL1



recynik,N9+Ek+FLX=N9

Slide 4

BL1 Možda napomenuti da su ovi markeri varijacija nastali nakon formiranja baze podataka, ovo +VAR=SatiRati
Biljana Lazić, 3/9/2019

DELA dictionaries

- simple word lemmas (DELAS)
- multi-word lemmas (DELAC)

- simple word forms (DELAF)
- multi-word forms (DELACF)

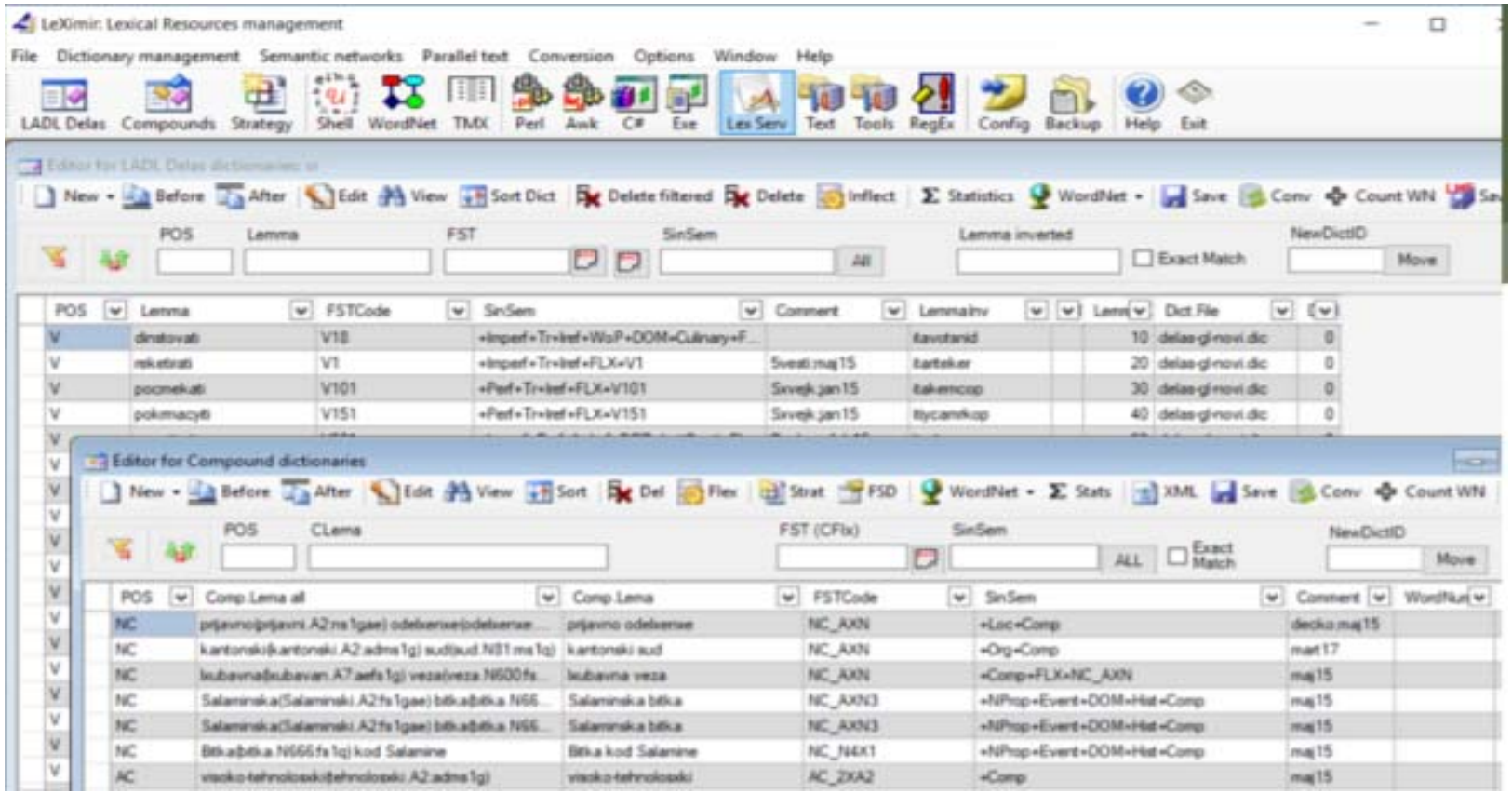
DELAF entries

form,lemma[:categories]*

recynici,recynik.N:mp1q:mp5q

SMD maintaining

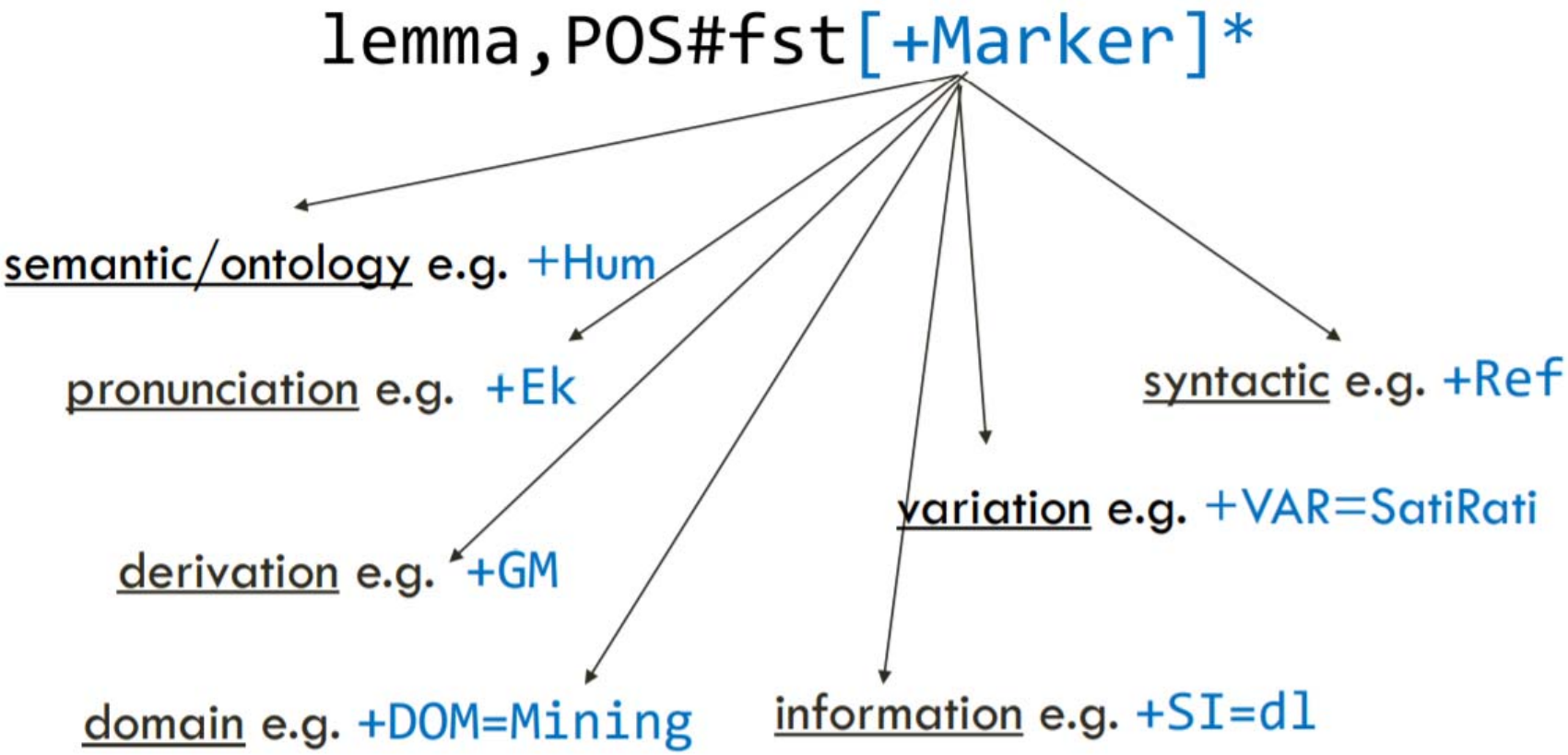
WS4LR* >> LeXimir >> Leximirka



* Workstation for Lexical Resources

DELA lexical entry structure

BL1



recynik,N9+Ek+FLX=N9

Slide 8

BL1 Možda napomenuti da su ovi markeri varijacija nastali nakon formiranja baze podataka, ovo +VAR=SatiRati
Biljana Lazić, 3/9/2019

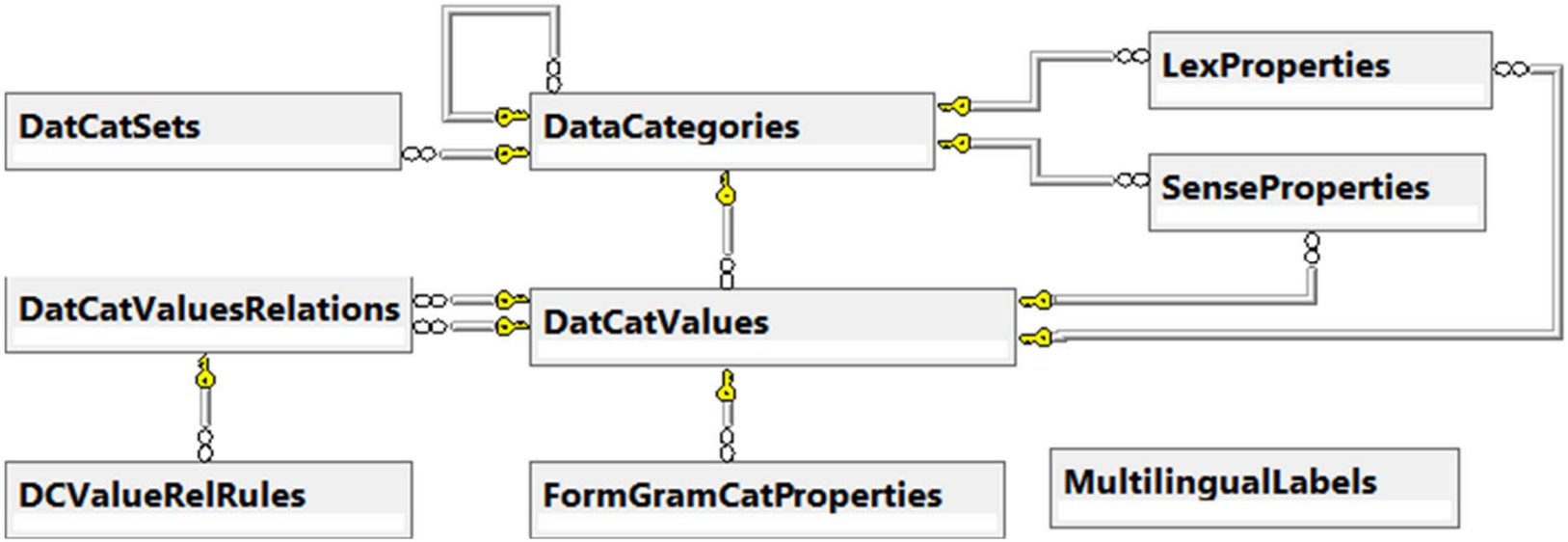
Serbian Unitex Day

Relational lexical database

- Lexical repository
- Based on lemon, LMF and DCR models

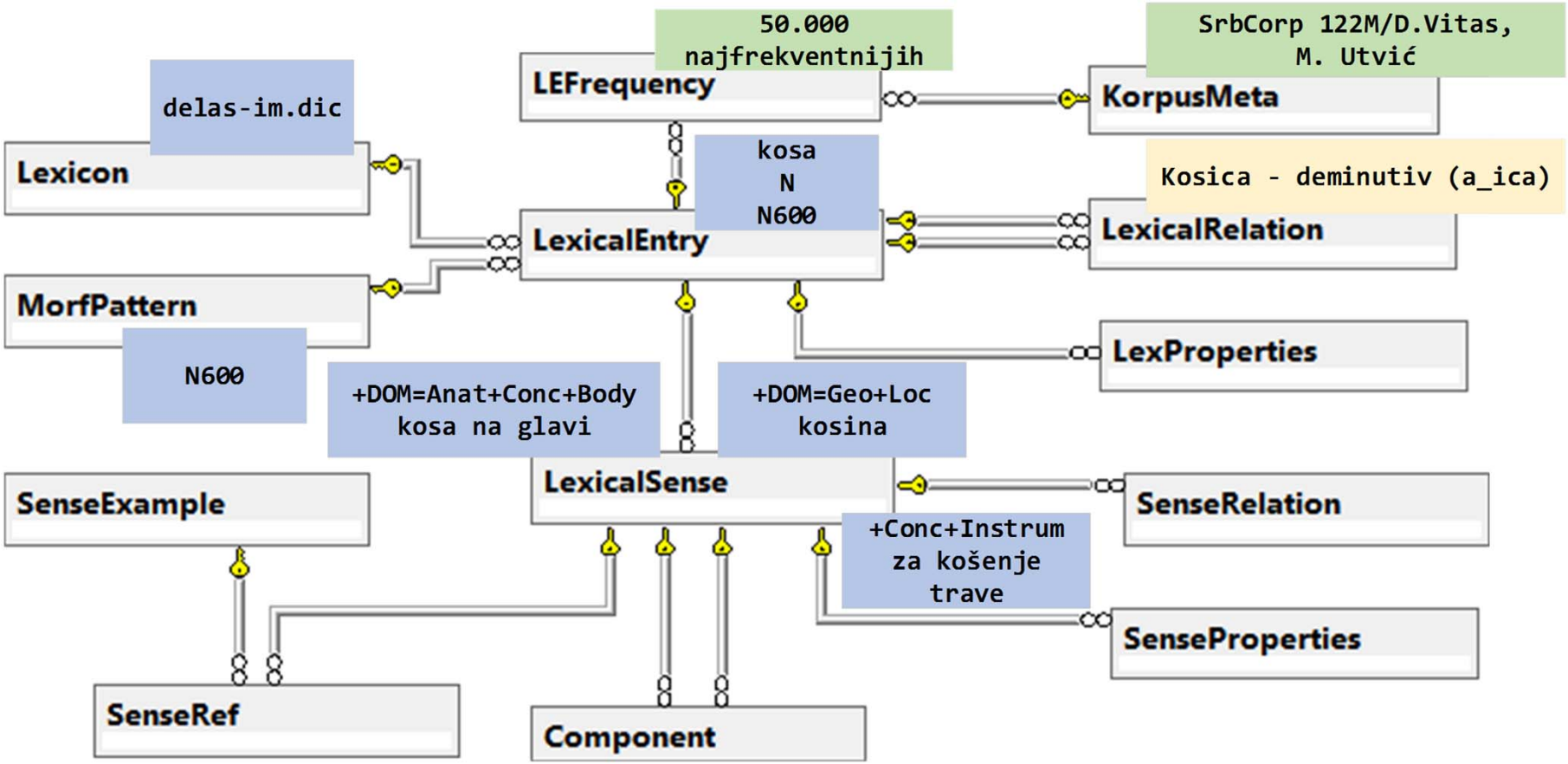
- Multiuser access
- In real time
- Lexical categories and relations
- Lexical data import and management
- Lexical data export (full dictionaries…)

Leximirka lexical database Categories

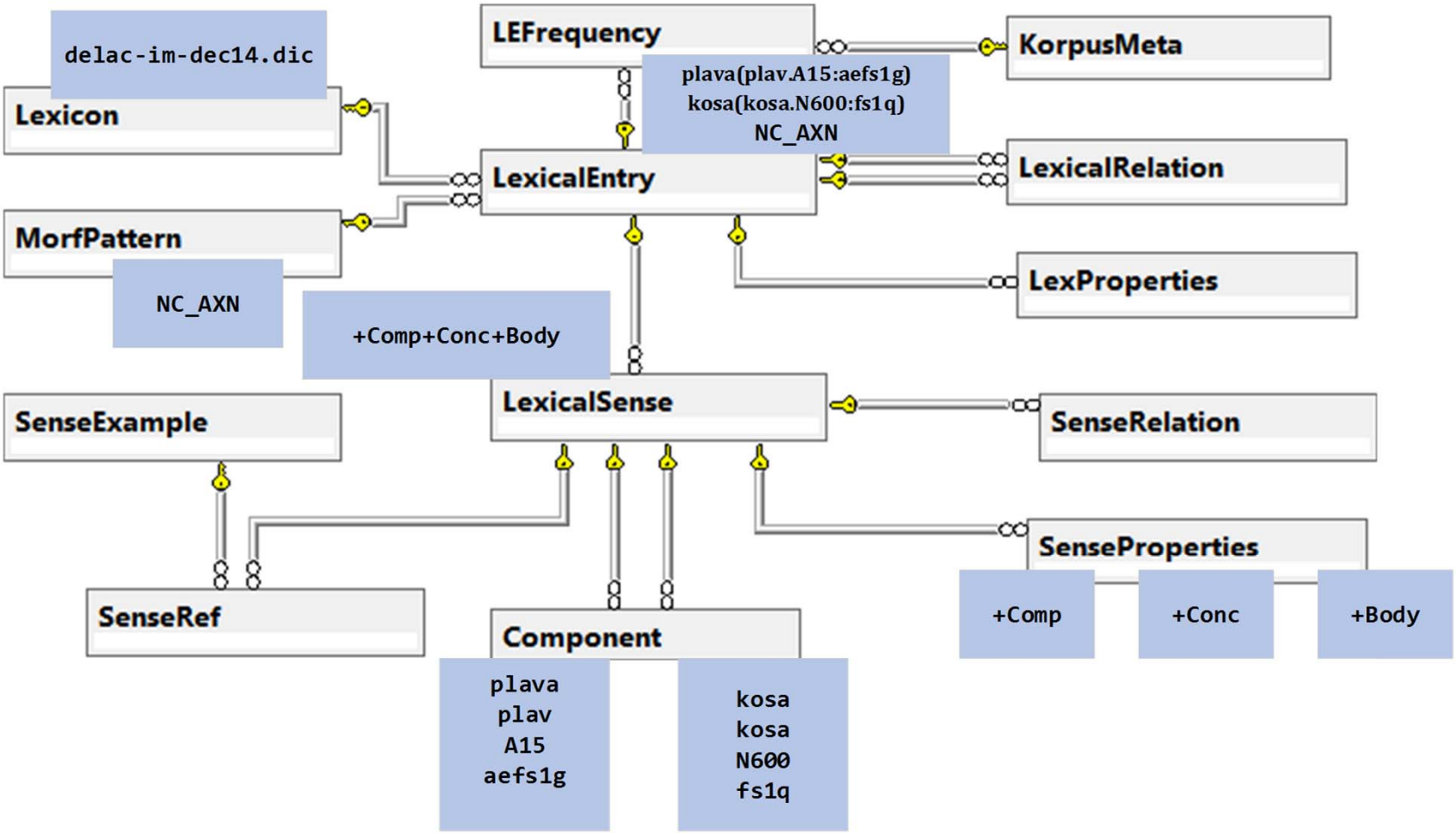


Serbian Unitex Day

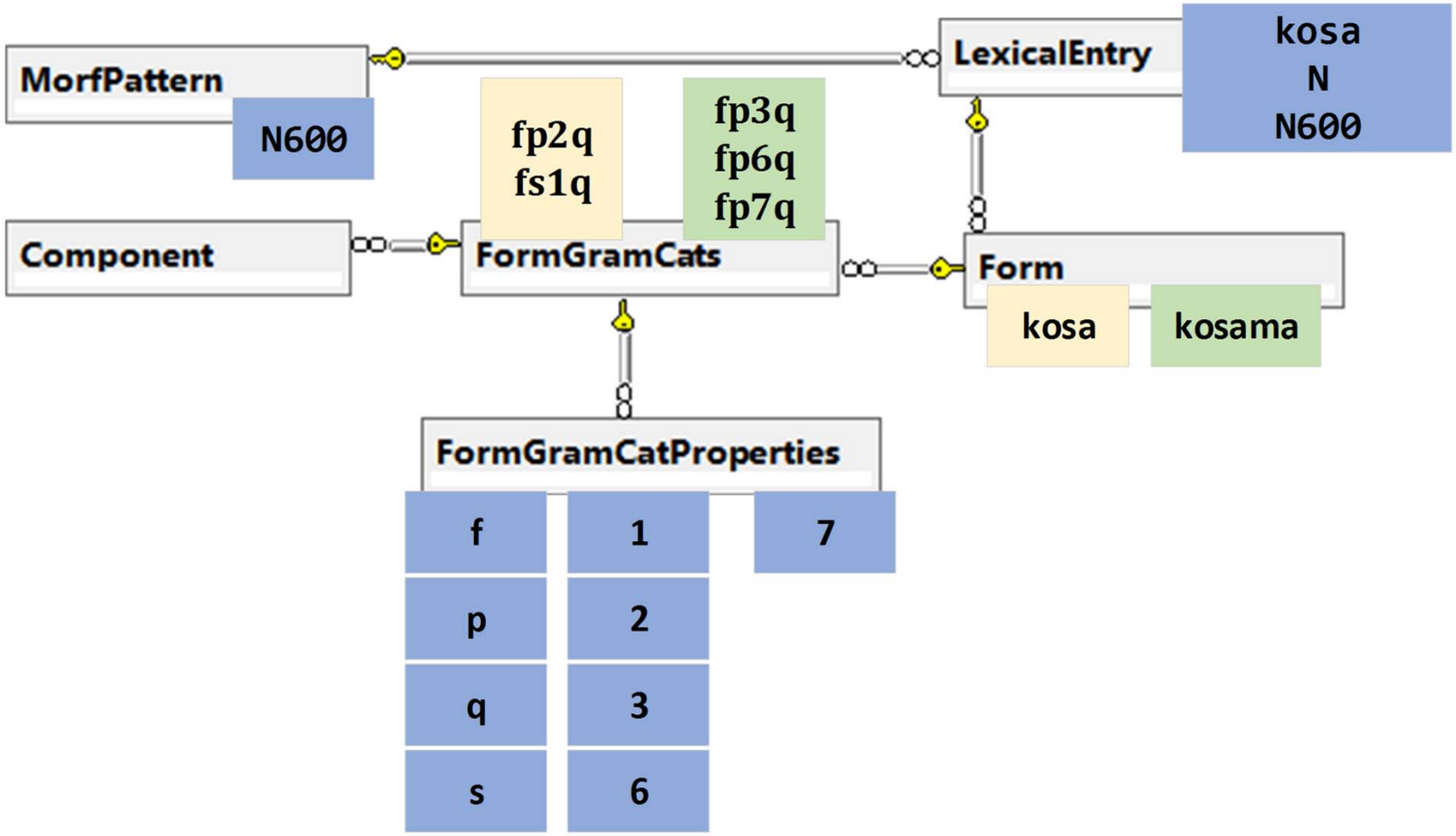
DELAS dictionary model



DELAC dictionary model



DELAF dictionary model



leximirka.jerteh.rs/?q=kosa

Leximirka Categories Lexicons Entries Corpora Evaluation Relations Bibliography Hello labil

Lexical Entry #30083

Edit

kosa

N600 delas-im.dic

Relations:

- To kosica using **deminutiv (a_ica)**

Frequencies:

- Top 50000 most frequent in SrbCorp122M Corpus by D.Vitas, M.Utvić

Senses (3):

1. +Conc+DOM=Anat+Body

Domains:	anatomija
Properties:	konkretna imenica, telo
Note:	kosa na glavi

Is a component of:

- bela kosa
- bijela kosa
- seda kosa
- sijeda kosa
- plava kosa
- Bežanijska kosa
- Banska kosa

2. +Loc+DOM=Geo

Domains:	geografija
Properties:	mesto
Note:	kosina

3. +Conc+Instrum

Domains:	
Properties:	konkretna imenica, oprema i delovi opreme
Note:	za kosenxe trave

About Contact

Lexical Entry #30083

Save all Changes

Lemma: kosa

Cannonical form: kosa

Language: sr

Entry Type: S

Part of Speech: N

Morf pattern code: N600

Lexicon: delas-im.dic

Note:

Properties: Add:

Relations:

- To kosica using **deminutiv (a_ica)**

Add Sense

Save all Changes

1. Sense 200064 +Conc+DOM=Anat+Body (kosa na glavi)

Label: 1

Sense Definition: +Conc+DOM=Anat+Body

Note: kosa na glavi

Properties: Add:

Conc x Body x

Domains: Add:

DOM=Anat x

Is a component of:

- bela kosa
- bijela kosa
- seda kosa
- sijeda kosa
- plava kosa
- Bezxanijska kosa
- Banska kosa

References:

- None

2. Sense 203077 +Loc+DOM=Geo (kosina)

Lexical Entry #214665

Save all Changes

Lemma: seda(sed.A6:aefs1g) kosa(kosa.N600:fs1q)

Cannonical form: seda kosa

Language: sr

Entry Type: C

Part of Speech: N

Morf pattern code: NC_AXN

Lexicon: Delac-im-new.dic

Note:

Properties: Add:

Relations:
- None

Add Sense

Save all Changes

1. Sense 377351 +Ek+Conc+Body+Comp (MWElista;jun18)

Label: 1

Sense Definition: +Ek+Conc+Body+Comp

Note: MWElista;jun18

Properties: Add:

Conc x Body x Comp x

Domains: Add:

Is composed of:

Form	Lemma	FST Code	Gram Cat	Separator
seda	sed	A6	aefs1g	
kosa	kosa	N600	fs1q	

References:
- None

Serbian Unitex Day

Data categories

Data Categories Board

- ekavski-ijekavski (izgovor)
- kroatizam, engl,... (jezik)
- varijacije (variation)
- domenski (domain)
- informacije (information)
- gramatičke kategorije ()
  - vrste reči (POS)
    - imenica (N)
    - broj (NUM)
    - pridev (A)
    - glagol (V)
    - predlog (PREP)
    - veznik (CONJ)
    - partikula (PAR)
    - uzvik (INT)
    - prilog (ADV)
    - zamenice (PRO)
    - skraćenica (ABB)
    - prefiks (PREF)

Add Sibling Category | Add Child Category | Save all Changes

Manage Category Details for 10057

Parent ID	1006
Label	jezik
Name	kroatizam, engl,...
Profile	semcats
Description	
Examples	
DocumURL	

Manage Category Values

Add Value

Value	Label	Description	Status
Cr	kroatizam	(e.g. nogomet)	
EN	anglicizam	iz engleskog jezika (e.g. Džon)	
Sr	srbizam	srbizam (e.g. fudbal)	

Serbian Unitex Day

Lexicons

Leximirka | Categories | Lexicons | Entries | Corpora | Evaluation | Relations | Bibliography | Hello labilja! | Log off

Lexicons

Add a Lexicon

ID	Name	Type	Language	Description
0	delas-gl-novi.dic	S	sr	
1	delas-im-nove.dic	S	sr	
2	delas-zm.dic	S	sr	
3	delas-br.dic	S	sr	
4	delas-im.dic	S	sr	
5	delas-gl.dic	S	sr	
6	delas-pr.dic	S	sr	
7	delas-abb.dic	S	sr	
8	delas-ad.dic	S	sr	
9	delas-con.dic	S	sr	

1 2 3 ... 5 »

items per page

Serbian Unitex Day

Entries

Leximirka | Categories | Lexicons | Entries | Corpora | Evaluation | Relations | Bibliography | Hello labilja! | Log off

Lexical Entries

find entry

Lexicon	Lemma	Canon	POS	Type	Morf Pattern	Status	Lang	Note
0	dinstovati	dinstovati	V	S	V18	I	sr	
0	reketirati	reketirati	V	S	V1	I	sr	BL
0	pocmekati	pocmekati	V	S	V101	I	sr	
0	pokrmacyiti	pokrmacyiti	V	S	V151	I	sr	
0	raportirati	raportirati	V	S	V501	I	sr	
0	apretirati	apretirati	V	S	V501	I	sr	
0	odmeksxati	odmeksxati	V	S	V101	I	sr	
0	omeksxati	omeksxati	V	S	V101	I	sr	
0	napabircyiti	napabircyiti	V	S	V151	I	sr	
0	zavrecyati	zavrecyati	V	S	V641	I	sr	

1 | 2 | 3 | ... | 20683 | »

items per page

Data Category Values Relation

Save Changes

Label: ekavski - ijekavski

Relation type: Ekljk

Relation simetric: yes no

Source Value: izgovor/Ek

Destination Value: izgovor/Ijk

Rules (9):

Add New Rule

		POS	Fix	Afix	Marker	Example	Stem End
101/1	From	MOT		ije	Ijk	dijelove	
	To:	MOT		e	Ek	delove	
101/2	From	MOT		je	Ijk	bezbjednost	
	To:	MOT		e	Ek	bezbednost	
101/3	From	MOT		lxe	Ijk	slxedecxi	
	To:	MOT		le	Ek	sledecxi	
101/4	From	MOT		nxe	Ijk	snxesxko	
	To:	MOT		ne	Ek	snesxko	
101/5	From	MOT		ijel	Ijk	bijel	
	To:	MOT		eo	Ek	beo	
101/6	From	MOT		ilxe	Ijk	bilxezxnički	
	To:	MOT		ele	Ek	belxezxnicyki	
101/7	From	MOT		lxe	Ijk	polxevacy	
	To:	MOT		li	Ek	polivacy	

Serbian Unitex Day

leximirka.jerteh.rs/Relation/Rule?di

Not secure | leximirka.jerteh

A total of **60** entry pairs were e

on relation **68** using rule **befs2**

Connected Entries

Source Lemma
polagano
otvoreno
ubrzano
kicosxko
naucyno
docno
juzxno
severno
blagonaklono
mogucxno
studiozno
uputno
brzo

leximirka.jerteh.rs/Relation/Rule?did=50&n=2&r=0 - Google Chrome

Not secure | leximirka.jerteh.rs/Relation/Rule?did=50&n=2&r=0

A total of **1020** entry pairs were established,

on relation **50** using rule **_jka**.

Connected Entries

Source Lemma	Target Lemma	Status
Gvardi	Gvardijka	AU
Kaneti	Kanetijka	AU
Hajnrici	Hajnricijka	AU
Toskanini	Toskaninijka	AU
Klementi	Klementijka	AU
Halximi	Halximijka	AU
Hiseni	Hisenijka	AU
Kadri	Kadrijka	AU
Leka	Lekajka	AU
Petko	Petkojka	AU
Rami	Ramijka	AU
Remzi	Remzijka	AU
Ruzxdi	Ruzxdijka	AU

Serbian Unitex Day

Evaluation

Leximirka | Categories | Lexicons | Entries | Corpora | Evaluation | Relations | Bibliography | Hello labilja! | Log off

Evaluate

CorpusID	Graf	Form	Lemma	MWE?	Lemma?	Term?	General?	Sin Sem
5	grf01	plu	fiksni troškovi	True	True	True	False	
5	grf01	sin	ručna korekcija	True	True	False	True	
5	grf01	sin	tehnološko rešenje	True	True	False	False	
5	grf01	sin	temperaturno polje	True	True	True	True	
5	grf03	sin	analiza varijanse	True	True	False	False	
5	grf03	sin	brzina agregata	True	True	True	True	
5	grf06	sin	zrno kvalitet	False	False	True	False	
5	grf06	sin	traktor točkaš	True	True	True	True	
5	grf06	sin	obrtaj motor	True	False	True	True	
5	grf06	sin	povećanje produktivnost	True	False	True	True	

1 | 2 | 3 | ... | 931 | »

items per page

Future work

- WordNet connection
- Connection to other web resources
- Adding multimedia contents
- Dictionary export (Turtle, RDF/XML)
- API for NLP applications

Thank you!