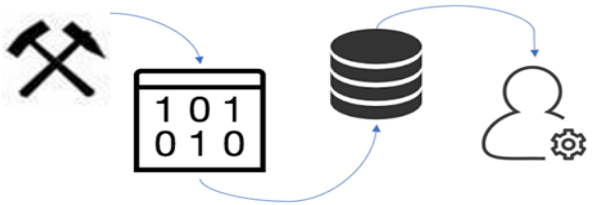
Development and application of a domain specific corpus for mining engineering

Ranka Stanković, Miloš Utvić, Aleksandra Tomašević, Ivan Obradović, Biljana Lazić
University of Belgrade, Serbia

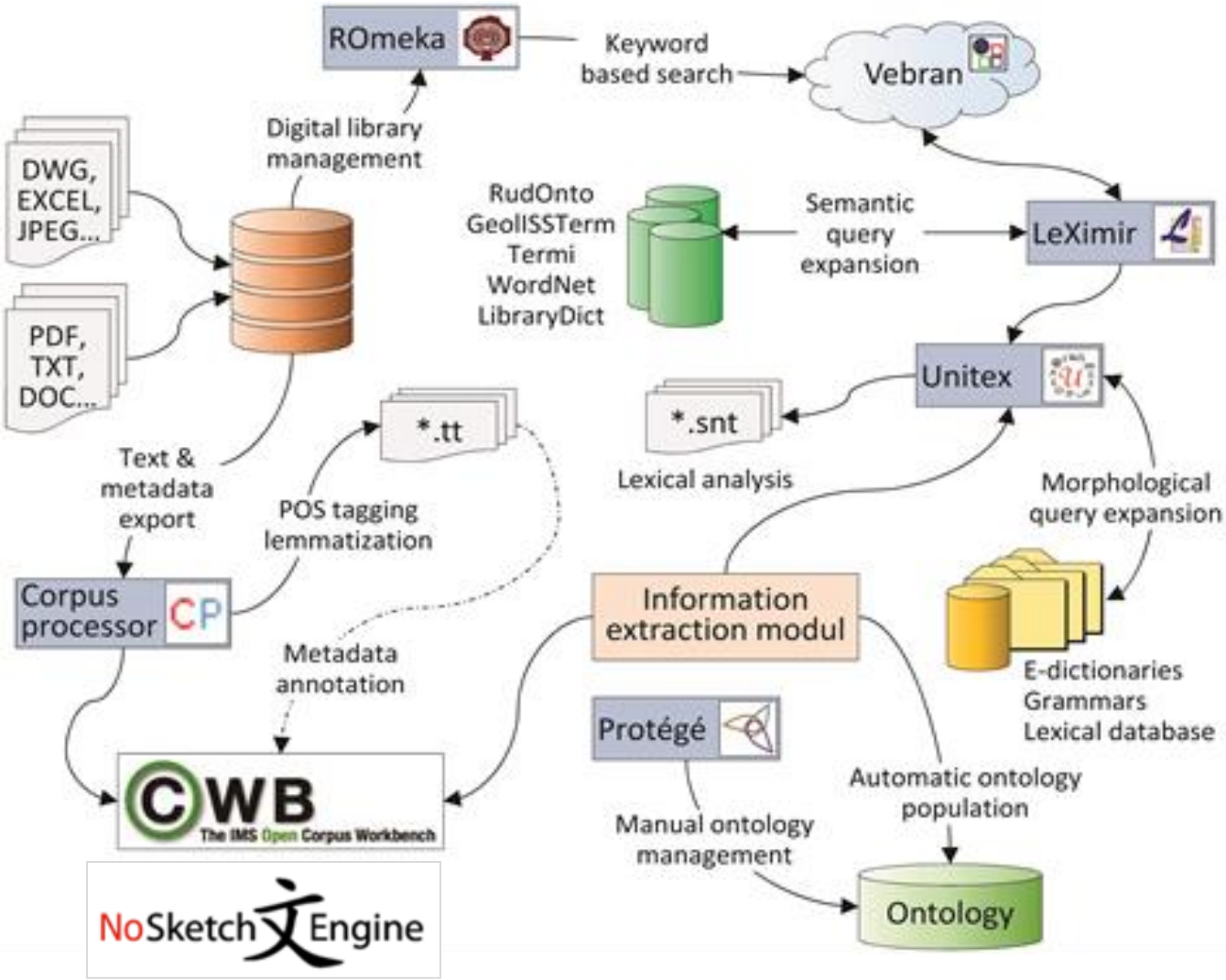


SlaviCorp 2018, 24-26. September, Prague

Introduction

- Lexical resources play an important role in management of project documentation in a specific domain
  - general lexical resources for Serbian have reached a considerable size (SMD: >180,000 lemmas, SWN: >25,000 synsets)
  - resources covering domain specific terminology still require further development for many fields, including mining engineering
- We present RudKorp - corpus of engineering documentation in the mining domain and how it is used to
  - enrich Serbian lexical resources by adding terminology specific for the mining domain
  - support corpus querying combined with lexical resources

HLT based mining documentation management system overview



RudKor

Developed at the University of Belgrade from ROmeka@RGF digital library

- as a means of improving the search of the digital library based on linguistic annotation, and
- as a resource for various linguistic and terminological research, including extraction.

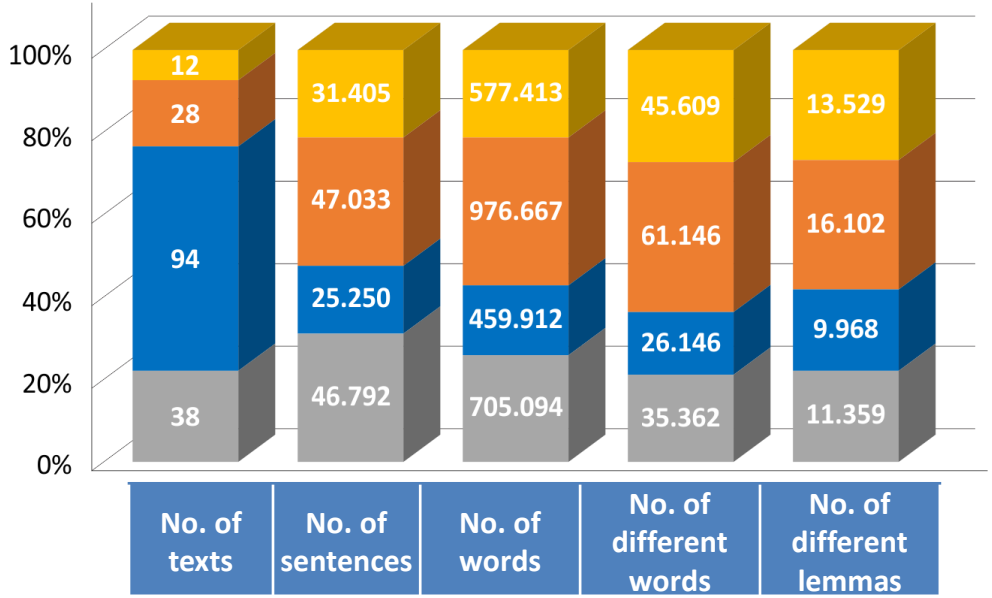
Three different systems for diverse types of usage scenarios

- IMS Open Corpus Workbench (CWB) and an adaptation of CQPweb, a web-based graphical user interface
- Unitex, used to create a second corpus from the same texts for custom information extraction tasks and
- NoSketch Engine

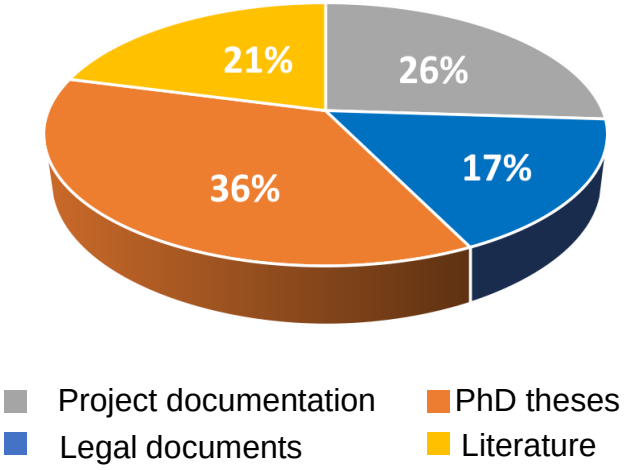
RudKor in numbers

Text processing results

Text type	No. of texts	No. of sentences	No. of words	No. of different words	No. of different lemmas	Text size (MB)	LR index (%)
Project documentation	38	46.792	705.094	35.362	11.359	10,1	1,61
Legal documents	94	25.250	459.912	26.146	9.968	6,6	2,17
PhD theses	28	47.033	976.667	61.146	16.102	14,2	1,65
Literature	12	31.405	577.413	45.609	13.529	8,2	2,34
Corpus	**172**	**150.365**	**2.719.086**	**100.414**	**22.875**	**39,1**	



Percent of words per text type



Markers

Domain markers

Marker	Description	Examples
+Mining	Mining	mine, miner, ore, exploitation,...
+Mach	Machine engineering	bucket-weel excavator, dragline...
+Safety	Occupational safety	ventilation, protective equipment, noise, vibrations, ...
+Transport	Transport	conveyor belt, truck, tipper,...
+RockMech	Rock mechanics	pressure, shear, slope stability,...
+Surveying	Mining surveys, geodesy	mine network, pit polygon,...
+EnvProt	Environment protection	air pollution, noise, pollution, monitoring,...

Sub-domain markers

Маркер	Опис	Примери
+Surface	Open pit exploitation	surface exploitation, floor,...
+Underground	Underground exploitation	underground exploitation, shaft,...
+MinProcess	Mineral resources processing	movement of mass, sample, shredding,...
+Petroleum	Petroleum and gas exploitation	oil, gas, well, pipeline,...

Semantic markers

Маркер	Опис	Примери
+MinStatus	Status of the mine	active mine, closed mine,...
+Ore	Mineral resources	coal, lignite, brown coal, gravel,...
+Activity	Mining activity	mine design, drilling,...
+Object	Mining facilities	mining window, hall, pit,...
+Prof+Hum	Professions	mining engineer, miner, geologist,...
+Org	Organisations	Kolubara, Kostolac, RTB Bor, EPS,...
+Instrum	Mining equipment	excavator, truck, conveyor belt,...
+Exploration	Mineral resource research	exploratory drilling,...
+Conc+Fashion	Personal protection equip.	lamp, helmet,...
+Text	Mining documentation	mining project documentation,...

What are we upgrading?

Existing web interfaces for searching corpora

- IMS CWB
- NoSketch Engine

Serbian corpora

- RudKor: professional texts from the mining area
- SrpKor: contemporary Serbian texts

How do we upgrade?

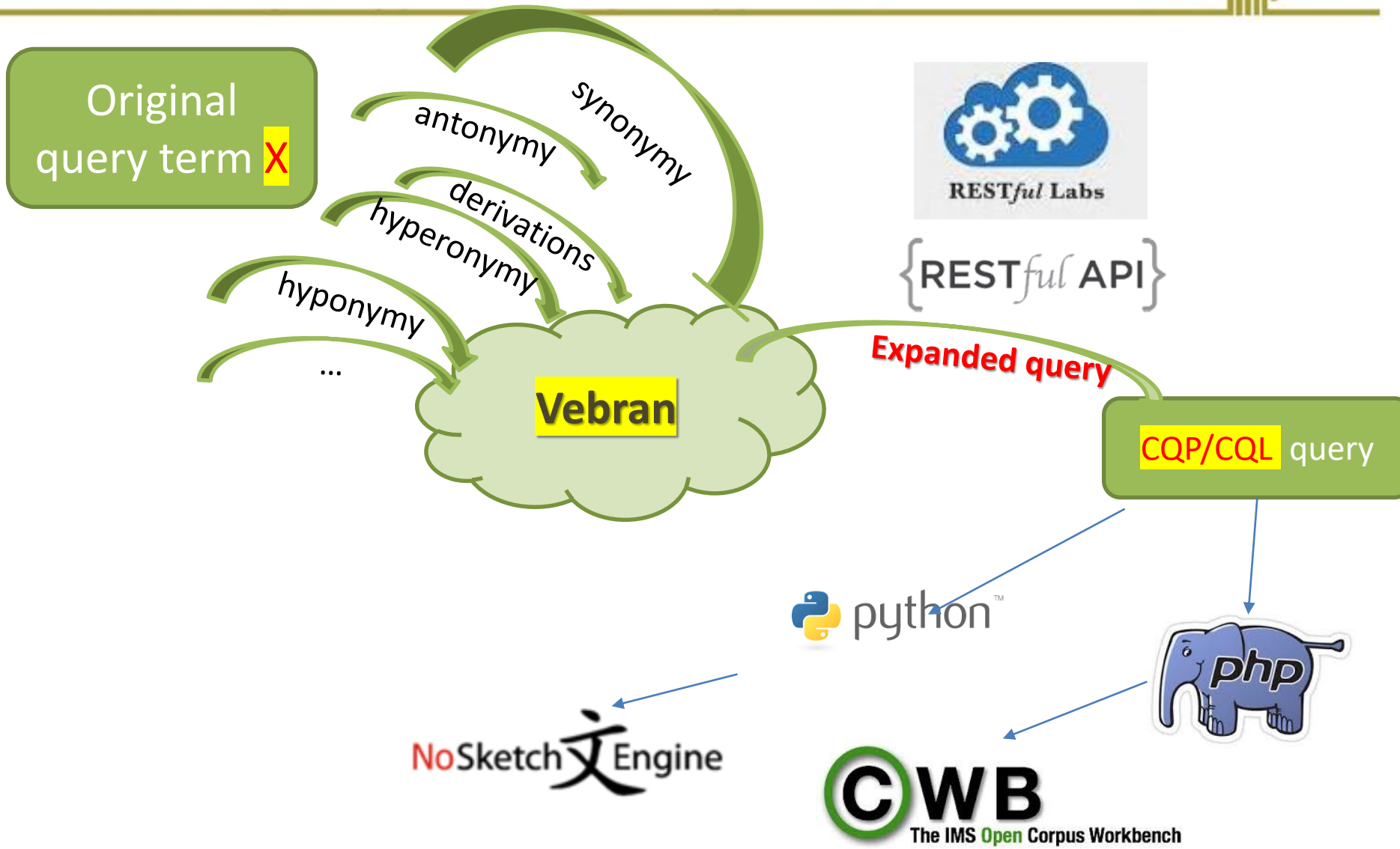
Query expansion is based on external lexical resources for Serbian

- Morphological electronic dictionaries
- Semantic network Wordnet
- Terminological databases Termi, RudOnto, GeolISS, RBI

Implementation

- Vebran, an adapted set of web services (RESTfull, MVC .Net)
- Modification of the web-based corpus search interface (open source code in PHP, Python)

How do we expand?



Expansion using synonyms

- [synlemma="bandvagen"]
  - synonyms (Wordnet and terminological databases)

- ([lemma="bandvagen|odlagač"] | ([lemma="samohodni"][lemma="transporter"]))

Results: IMS CWB / CQPweb

- **bandvagen** is not a recognized word (although it exists in the corpus) but its **synonyms** odlagač and samohodni transporter are

Your query "([lemma="bandvagen|odlagač"] | ([lemma="samohodni"][lemma="transporter"]))" returned 309 matches in 28 different texts (in 3,542,016 words [172 texts]; frequency: 87.24 instances per million words), ordered randomly [0.002 seconds]

|< << >> >1 Show Page: 1 Line View Show in corpus order New query Go!

No	Filename	Solution 1 to 50		Page 1 / 7
1	projek0183	1800 (4 x 630 kW) , pet transportera , -	odlagač	A 2 RsB - 7200 . 95 (O 1) ,
2	projek0146	BTO sistem : • bager SRs 400 . 14 / 1 •	samohodni transporter	BRs - 2400 . 59 • bager ERs 710 . 17 ,
3	projek0154	BTO sistem : - bager : SRs 2000 . 28 : -	odlagač	: A 2 RsB 7200 : • IV BTO sistem : -
4	projek0144	sistemom BTO (rotorni bager ili vedričar , transporteri sa trakom ,	odlagač	) . Na otkopavanju otkrivke previđen je rad 6 BTO sistema sa
5	doktor0199	odnosno BTO sistema (rotorni bager - transporteri sa trakom -	odlagač	) [66] , [70] , [71]
6	projek0154	13 - 16 : 420 / h , 1000000 / god *	samohodni transporter	BRs 1400 . 17 , 5 / 32 , 5 : 595
7	projek0162	vršnu kotu + 105 mnv odlagališta . Duž ovog transportera kretaće se	odlagač	ARs 3000 , koji vršiti odlaganje površinskog sloja zemljišta f
8	projek0162	B - 1400 (4 x 315 kW) , pet transportera	odlagač	ARsB - 3000 . 50 Slika 5 . 2 . Dispozicija transportnog
9	projek0144	B - 1800 , pet transportera , L = 3835 m •	odlagač	A 2 RsB - 7200 . 95 * IV BTO sistem :
10	studij0196	transportni sistem od 5 transportera širine B = 1800 mm i	odlagač	A 2 RsB 7200 . Visina otkopne etaže na otkrivci ide od
11	projek0144	1800 (4 x 630 kW) , pet transportera , *	odlagač	A 2 RsB - 7200 Osnovne tehničke karakteristike opreme na I

No	Filename	Solution 1 to 1		Page 1 / 1
1	studij0133	BTO sistema , a sa II BTD sistema preko jednog od raspoloživih	bandvagena	interslojna jalovina se direktno odlaže u otkopani prostor unutrašnjeg odlagališta . Otkopan

Results: NoSketch Engine

NoSketch Engine

testsaska

Home
Search
Word list
Corpus info
My jobs
User guide

Save
Make subcorpus
View options
KWIC
Sentence
Sort
Left
Right
Node

Query bandvagen|odlagač, samohodni, transporter 309 (87.24 per million)

First | Previous Page 2 of 16 Go Next | Last

doc#30	odlagač sa kliznom strelom , b) Shevron metoda -	odlagač	sa fiksnom
doc#30	odlagač sa fiksnom strelom , c) Windrow metoda -	odlagač	sa kliznom
doc#30	u koji je uključen specijalno projektovan	odlagač	ili reklajme
doc#30	promeniti položaj pretovarnog uređaja , bio to	odlagač	ili " F traka
doc#35	. 500 m i dva odlagališna ukupne dužine 3 . 250 m) i	odlagač	klase A 2 R
doc#35	dva odlagališna sa ukupnom dužinom od 3 . 100 m) i	odlagač	A 2 RsB 72(
doc#35	kopa i dva odlagališna ukupne dužine 3 . 000 m) i	odlagač	A 2 RsB 72(
doc#39	bager ili vedričar , transporteri sa trakom ,	odlagač	) . Na otko
doc#39	je sledeća : * I BTO sistem : • bager SRs 400 . 14 / 1 •	samohodni transporter	BRs - 2400
doc#39	16 • transporteri širine B - 1400 mm , tri komada •	odlagač	ARsB - 350(
doc#39	. 50 * II BTO sistem : • bager SchRs 800 . 15 / 1 . 5 •	samohodni transporter	BRs - 2400
doc#39	transporter BRs - 2400 . 59 • bager SRs 470 . 20 / 3 •	samohodni transporter	BRs - 2400
doc#39	B - 1600 , pet transportera , L = 5140 m •	odlagač	A 2 RsB 55(
doc#39	B - 1800 , pet transportera , L = 3835 m •	odlagač	A 2 RsB - 7

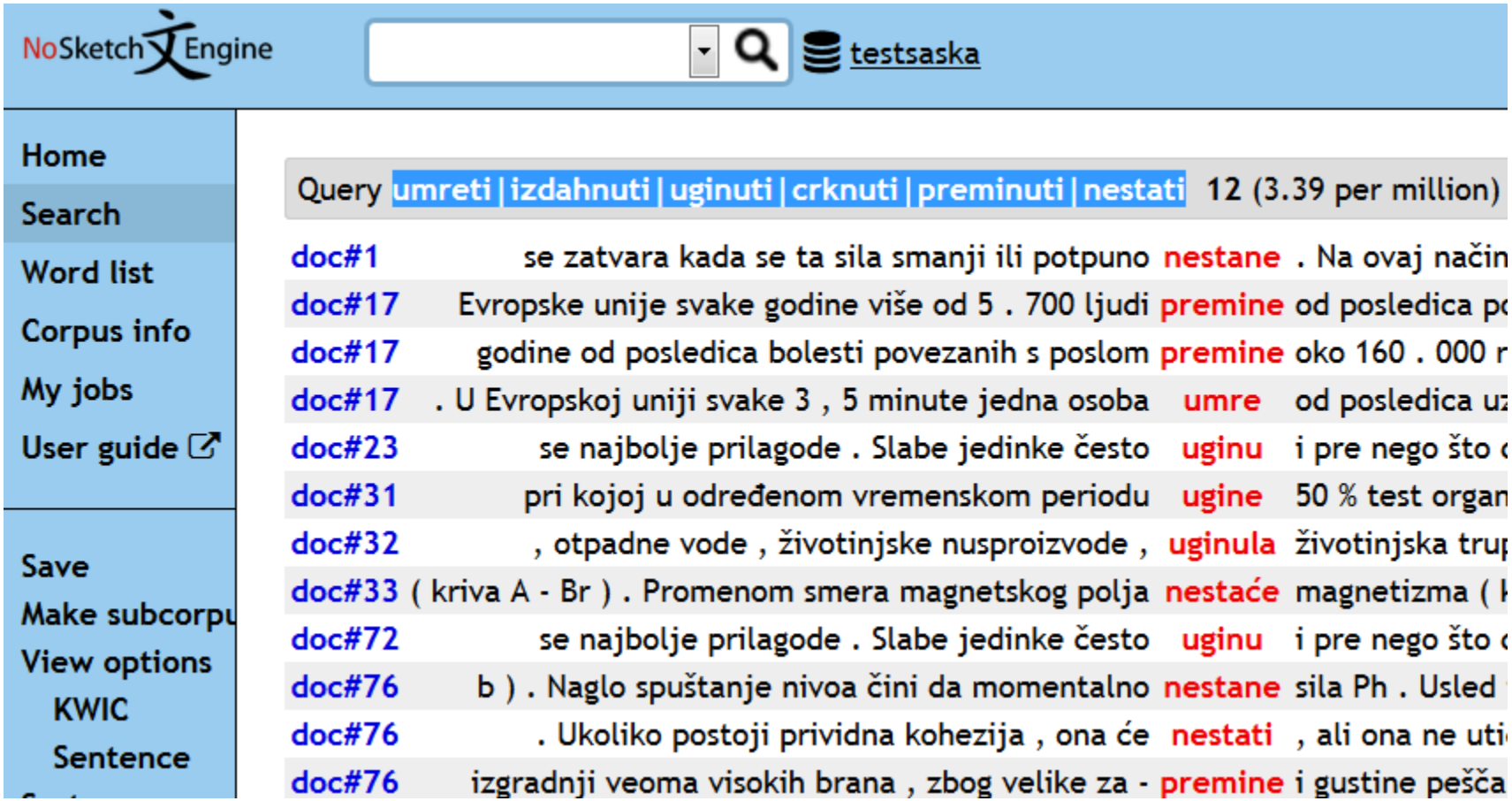
Semantic expansion

- Idea: expand the original query, in which term X is given, with other terms that are in a semantic relation with the term X (synonymy, antonymy, hyperonymy, etc.).
- Lexical relations
  - Wordnet: hypernym, hyponym, holo_member, holo_part, eng_derivative, near_antonym,…

[antlemma=“nastati”]

- *nastati* (to originate, to arise,…)
- antonyms:
  - *umreti* (to die)
  - *izdahnuti* (to expire)
  - *uginuti* (to die, used for a death of an animal)
  - *crknuti* (to drop dead)
  - *preminuti* (to pass away)
  - *nestati* (to disappear, to perish)

[antlemma="nastati"]



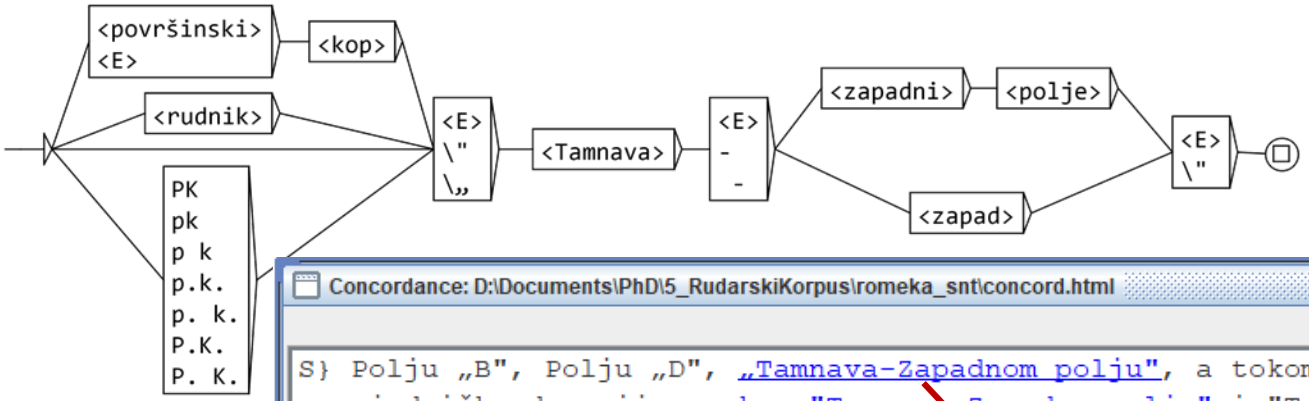
Identifying patterns

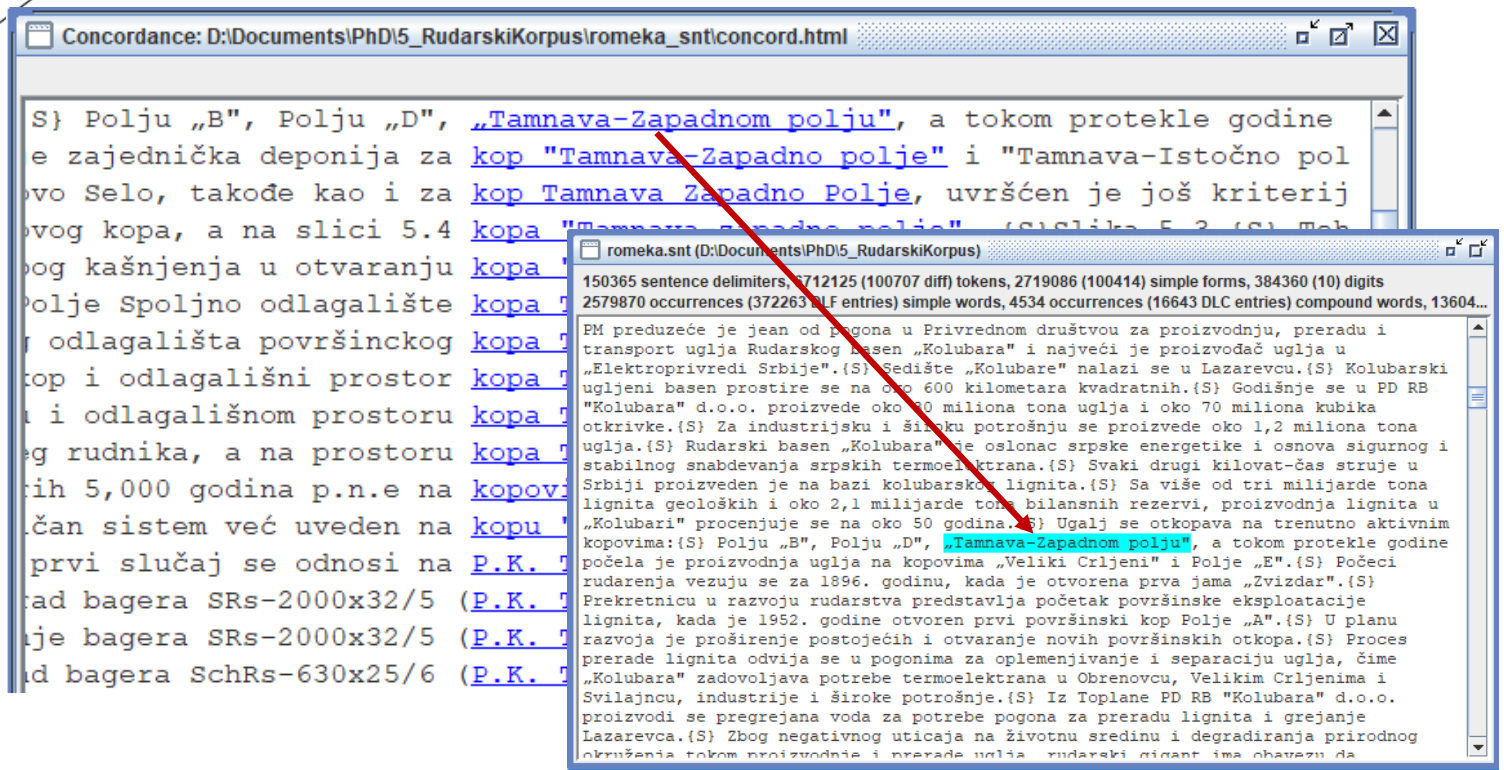
37 different ways of writing a concept

Recognized expression	Recognized expression
"Tamnava – Zapad"	Površinski kop "Tamnava-zapadno polje"
"Tamnava – Zapadno polje"	Površinski kop Tamnava Zapadno Polje
"Tamnava-Zapad"	površinski kop Tamnava Zapadno Polje
"Tamnava-zapadno polje"	Površinski kop Tamnava-Zapadno polje
"Tamnava-Zapadno polje"	površinskog kopa "Tamnava-zapadno polje"
"Tamnava-Zapadnom polju"	površinskog kopa Tamnava Zapadno Polje
kop "Tamnava-Zapadno polje"	površinskog kopa Tamnava Zapadno polje
kop Tamnava Zapadno Polje	površinskog kopa Tamnava-Zapadno Polje
kopa "Tamnava-Zapadno polje"	površinskom kopu "Tamnava-zapadno polje"
kopa "Tamnava-zapadno polje"	površinskom kopu "Tamnava-Zapadno polje"
kopa Tamnava Zapadno Polje	površinskom kopu TAMNAVA ZAPAD
kopa Tamnava-Zapad	površinskom kopu Tamnava Zapadno Polje
kopovima Tamnava Zapad	rudniku Tamnava - Zapad
kopu "Tamnava – Zapad"	Tamnava – Zapad
kopu "Tamnava-Zapadno polje"	Tamnava – Zapadno polje
kopu Tamnava – zapadno polje	Tamnava Zapadno Polje
P.K. TAMNAVA ZAPAD	Tamnava Zapadno polje
P.K. TAMNAVA-ZAPAD	Tamnava-Zapadno polje
	Tamnave Zapadnog polja

Corpus search using patterns

- Query: **površinski kop Tamnava Zapadno polje**





Search results

Recognized expression	Frequency
"Tamnava – Zapad"	1
"Tamnava – Zapadno polje"	2
"Tamnava-Zapad"	1
"Tamnava-zapadno polje"	3
"Tamnava-Zapadno polje"	2
"Tamnava-Zapadnom polju"	1
kop "Tamnava-Zapadno polje"	1
kop Tamnava Zapadno Polje	1
kopa "Tamnava-Zapadno polje"	1
kopa "Tamnava-zapadno polje"	1
kopa Tamnava Zapadno Polje	4
kopa Tamnava-Zapad	1
kopovima Tamnava Zapad	1
kopu "Tamnava – Zapad"	1
kopu "Tamnava-Zapadno polje"	1
kopu Tamnava – zapadno polje	1
P.K. TAMNAVA ZAPAD	1
P.K. TAMNAVA-ZAPAD	3

Recognized expression	Frequency
Površinski kop "Tamnava-zapadno polje"	1
Površinski kop Tamnava Zapadno Polje	4
površinski kop Tamnava Zapadno Polje	1
Površinski kop Tamnava-Zapadno polje	1
površinskog kopa "Tamnava-zapadno polje"	1
površinskog kopa Tamnava Zapadno Polje	5
površinskog kopa Tamnava Zapadno polje	2
površinskog kopa Tamnava-Zapadno Polje	2
površinskom kopu "Tamnava-zapadno polje"	2
površinskom kopu "Tamnava-Zapadno polje"	1
površinskom kopu TAMNAVA ZAPAD	2
površinskom kopu Tamnava Zapadno Polje	1
rudniku Tamnava - Zapad	1
Tamnava – Zapad	1
Tamnava – Zapadno polje	1
Tamnava Zapadno Polje	8
Tamnava Zapadno polje	1
Tamnava-Zapadno polje	1
Tamnave Zapadnog polja	1

Advance search templates

(({površinski})? {kop})| {rudnik} | pk | (p (.)? k (.)?))? ([",”,“,„])?

{Tamnava} ((-|–|—)? ({zapadni} {polje})|{zapad})) ([",”,“,„])?

Your query "((({površinski})? {kop})| {rudnik} | pk | (p (.)? k (.)?))? ([",”,“,„])? {Tamnava} ((-|–|—)? ({zapadni} {polje})|{zapad}))([",”,“,„])?" returned 64 matches in 10 different texts (in 3,542,016 words [172 texts]; frequency: 18.07 instances per million words), ordered randomly [0.001 seconds - retrieved from cache]

1 ≤	<<	>>	>|	Show Page:	1	KWIC View	Show in corpus order	New query ▾	Go!

No	Filename	Solution 11 to 20 Page 2 / 7
11	monogr0121	Tamnava - istočno polje " Slika 5 . 4 . Tehnološki profil **površinskog kopa " Tamnava - zapadno polje "** Površinski kop " Tamnava - zapadno polje " prvobitno je bio namenjen
	doktor0099	Očigledno je na ovom primeru Bogutovog sela , kao i u slučaju **Tamnave Zapadnog polj** pošumljavanja imala apsolutno najniža cena od 1
13	doktor0099	6 . - Principijelna tehnološka šema otkopavanja i odlaganja otkrivke na **površinskom kopu Tamnava Zapadno Polje** Spoljno dlagalište kopa Tamnava Zapadno Polje nalazi ce u cactavu unutrašnjeg odlagališta

Link to metadata

Link to document segment

Metadata for text *monogr0121*

Text identification code	monogr0121
Naslov dokumenta	Upravljanje kvalitetom uglja
Godina nastanka dokumenta	2007
Autor(i) dokumenta	Ignjatović Dragan|Knežević Dinko|Kolonja Božo|Lilić Nikola|Stanković Ranka
Poreklo (tip) dokumenta	monografije
Lista ključnih reči dokumenta	ugalj,kvalitet,homogenizacija,površinski kop,Tamnava
Jezik dokumenta	srpski
No. words in text	49673

Displaying extended context for query match in text *monogr0121*

File info for text monogr0121 ▾ | Go! | Show tags

Tehnološki profil **površinskog kopa " Tamnava - zapadno polje "** Površinski kop " Tamnava - zapadno polje " prvobitno je bio namenjen za potrebe snabdevanja ugljem buduće termoelektrane " Kolubara B " , instalisane snage u prvoj fazi 2 350 MW .

Kako ova elektrana još uvek nije izgrađena , kop se trenutno koristi kao dopuna kapacitetu površinskog kopa " Tamnava - istočno polje " .

Eksploatacija uglja na ovom kopu počela je 1995 .

godine , a projektovani razvoj će se odvijati kroz dve faze : - prva faza obuhvata period do kraja eksploatacionog veka površinskog kopa " Tamnava - istočno polje " .

Document segment extraction

```
<div2>
<head type='Deo' n='1.'>
<seg>
OGRANIČENjE            N       ograničenje
EKSPLOATACIONOG        A       EKSPLOATACIONI
POLjA                  N       polje
</seg></head>
<div3>
<head type='Glava' n='1.1.'>
<seg>
Geološka               A       geološki
granica                N       granica
kopa                   N       kop
</seg></head>
<div4><head type='Odelxak' n='1.1.1.'>
<seg>
Geološke               A       geološki
rezerve                N       rezerva
i                      CONJ    i
ocena                  N       ocena
istraženosti           N       istraženost
ležišta                N       ležište
sa                     PREP    sa
potrebnim              A       potreban
doistraživanjima       N
                       doistraživanjima
</seg></head></div4>
```

```
<item type="Naslov" n="A."> STUDIJA OPRAVDANOSTI SA IDEJNI
<list>
<item type="Deo" n="1."> OGRANIČENJE EKSPLOATACIONOG POLJA
<list>
  <item type="Glava" n="1.1."> Geološka granica kopa</item
  <list>
    <item type="Odeljak" n="1.1.1."> Geološke rezerve i oc
    <item type="Odeljak" n="1.1.2."> Inženjersko-geološke
    <item type="Odeljak" n="1.1.3."> Hidrogeološke karakte
  </list>
  <item type="Glava" n="1.2."> Eksploatacione granice povr
  <list>
    <item type="Odeljak" n="1.2.1."> Proračun eksploatacio
    <item type="Odeljak" n="1.2.2."> Proračun gubitka uglj
    <item type="Odeljak" n="1.2.3."> Proračun ukupne količ
    <item type="Odeljak" n="1.2.4."> Proračun koeficijenta
  </list>
</list>
```

A. СТУДИЈА ОПРАВДАНОСТИ СА ИДЕЈНИМ ПРОЈЕКТОМ О
КОЛИЧИНА УГЉА ЗА РАД ПОСТОЈЕЋИХ ТЕ У ТЕ-КО КОС
БЗ (350MW)

1. ОГРАНИЧЕЊЕ ЕКСПЛОАТАЦИОНОГ ПОЉА
1.1. Геолошка граница копа
1.1.1. Геолошке резерве и оцена истражености
истраживањима
1.1.2. Инжењерско-геолошке карактеристике л
1.1.3. Хидрогеолошке карактеристике лежишта
1.2.Експлоатационе границе површинског копа
1.2.1.Прорачун експлоатационих резерви угља
1.2.2.Прорачун губитка угља у јаловој зони
1.2.3.Прорачун укупне количине откривке и међ
1.2.4.Прорачун коефицијента откривке

2. ТЕХНИЧКО-ТЕХНОЛОШКИ ДЕО

2.1.Одређивање капацитета и век експлоатације копа
2.2. Вертикална подела копа на етаже
2.3.Геомеханичка провера стабилности
2.3.1
2.3.2
2.4.Димен
2.4.1
2.4.1
2.4.1

```
<head type="Glava" n="2"><seg>1.2. Geološki model ležišt
kopa</seg></head>
<p><seg>Proračun eksploatacionih rezervi kao i geometriz
(određivanje granica kopa i podela na etaže), urađena je
modela ležišta Drmno.</seg></p>
<div3>
    <head type="Odeljak" n="1"><seg>1.2.1. Geološka baza
    </seg></head>
    <div4>
        <head type="Pododeljak" n="1"><seg>1.2.1.1. Baza
        <p><seg>Baza podataka o istražnim geološkim rado
        fizičkih informacija koje se koriste u procesu r
        ustima bušotina, konačnoj dubini bušenja, litolo
        njihovim kodovima, hemijskim analizama, podacima o devijaciji bušotina itd.</seg><seg>
        Validana baza podataka je osnov za pouzdanu procenu mineralnih resursa i rudnih rezervi.
        </seg></p>
```

1.2. Geološki model ležišta i eksploatacione granice

Proračun eksploatacionih rezervi kao i geometrizacija pov
(određivanje granica kopa i podela na etaže), urađena je na o
geološkog modela ležišta Drmno.

1.2.1. Geološka baza podataka i model ležišta Drmno

1.2.1.1. Baza podataka istražnih radova

Baza podataka o istražnim geološkim radovima je osnovni izvo
informacija koje se koriste u procesu modelovanja ležišta. P
konačnoj dubini bušenja, litologiji, startigrafskim članovima
hemijskim analizama, podacima o devijaciji bušotina itd. Validan
za pouzdanu procenu mineralnih resursa i rudnih rezervi.

Next steps?

Resource enrichment

- Besides expert terminology, colloquial lexica needed for processing operational reports, plans, project documentation…

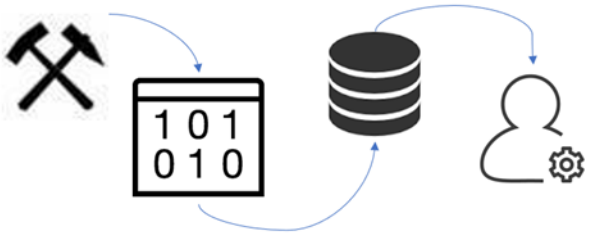
Examples

- ruč = landslide
- škarpa = sloping slopes
- rukanje = moving the conveyor
- grabuljar = self-propelled portable separator
- rolna = belt conveyor roller segment
- papuča = crawler system segment

Also planned…

- Query expansion based on generating the inflective paradigm of a given lemma, instead of search based on the lemma that has been automatically assigned during corpus annotation.
- New Serbian lexicon for TreeTagger
- New corpora
- A more user friendly search interface

Thank you for your attention



ranka@rgf.rs