

i COST

Distant Reading WG2

?

i COST

Distant Reading WG2

?

Ranka Stanković

SlaviCorp 2018



- Prague, Czech Republic,
24 –26th September 2018

- Gröbe Villa →

- Michal Křen (chair)
Charles University



<https://slavicorp.ff.cuni.cz/programme/presentations/>

SlaviCorp 2018:

Panel on language variation



Douglas Biber (Northern Arizona University):

Exploring cross-linguistic universals of register variation through multi-dimensional analysis

- Dimenzije: identifikovanje i interpretacija glavnih parametara lingvističkih varijacija
- Svaka dimenzija je skup lingvističkih svojstava (fičera) koja teže da se zajedno pojavljuju (co-occur) u tekstu
- Registri se mogu opisati i porediti
 - Prema lingvističkim karakteristikama
 - Funkcionalnim vezama (interpretacije dimenzija)
 - Kvantitativnim sličnostima i razlikama

Faktorska analiza za identifikaciju 'dimenzija' (1988)



67 lingvističkih fičera se koristi u analizi

- 1) 'Involved vs Informational Production'
- 2) 'Narrative Discourse'
- 3) 'Situation-dependent vs Elaborated Reference'
- 4) 'Overt Expression of Argumentation'
- 5) 'Abstract Style'

Dimension 1: Involved vs informational production [‘oral’ versus ‘literate’]



Positive features:

Verbs: present tense verbs, mental verbs, do as pro-verb, be as main verb, possibility modals

Pronouns: 1st person pronouns, 2nd person pronouns, *it*, demonstrative pronouns, indefinite pronouns

Adverbs: general emphatics, hedges, amplifiers

Dependent clauses: *that* complement clauses (with *that* deletion), causative adverbial clauses, *WH* clauses

Other: contractions, analytic negation, discourse particles, sentence relatives, *WH* questions, clause coordination

=====

Negative features:

Nouns, long words, prepositional phrases, attributive adjectives, lexical diversity

SlaviCorp 2018:

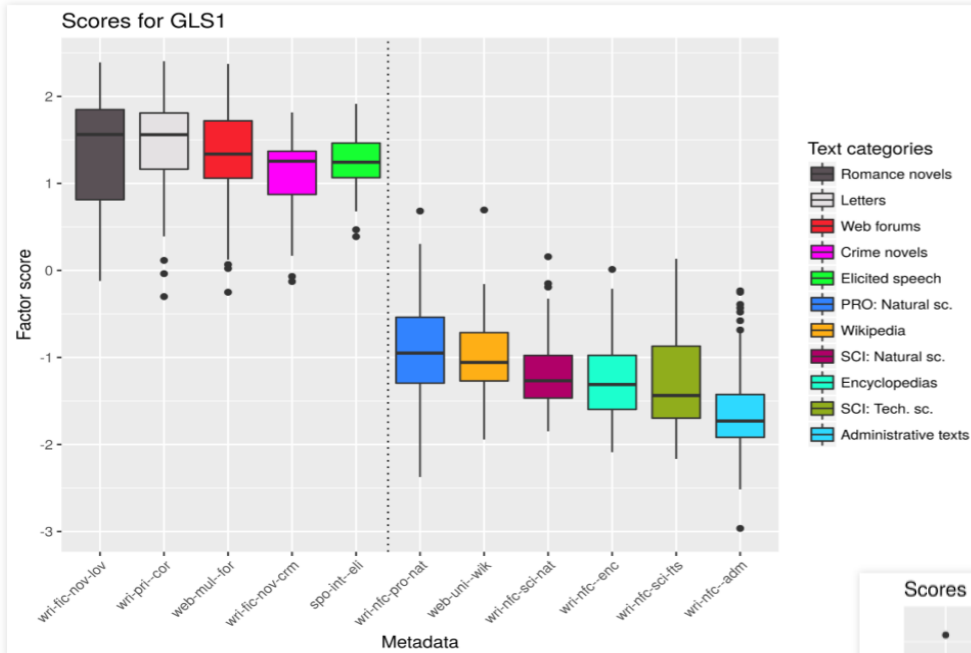
Panel on language variation



Václav Cvrček i drugi (Charles University, Prague): Multi-dimensional analysis of Czech

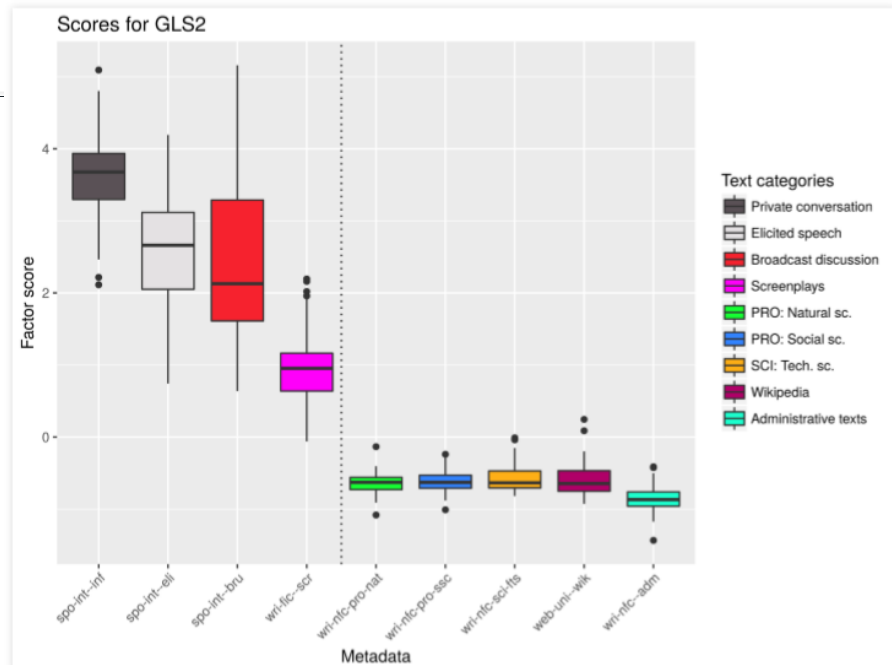
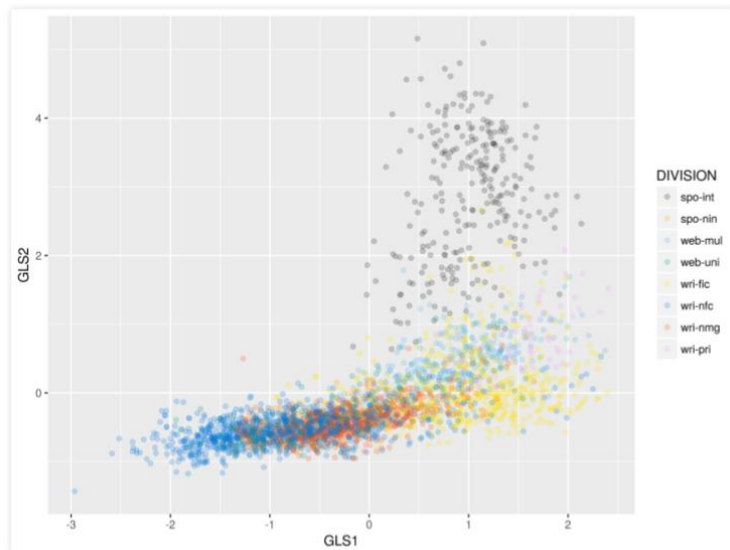
- Dimensions of variability
 1. dynamic (+) vs. static (-)
 2. spontaneous (+) vs. prepared (-)
 3. higher (+) vs. lower (-) level of cohesion
 4. polythematic (+) vs. monothematic (-)
 5. higher (+) vs. lower (-) amount of addressee coding
 6. general/intension (+) vs. particular/extension (-)
 7. prospective (+) vs. retrospective (-)
 8. attitudinal (+) vs. factual (-)

Dim 1: dynamic (+) × static (-)



Dim 2: spontaneous (+) × prepared (-)

2D-plot: dim 1 and dim 2



Plenarna predavanja



- Björn Hansen & Edyta Jurkiewicz-Rohrbacher:
Detecting constraints on clitic climbing with
the help of Web corpora

DFG funded Project HA 2659/6-1

'Microvariation of the Pronominal and Auxiliary Clitics in Bosnian, Croatian and Serbian. Empirical Studies of Spoken Languages, Dialects und Heritage Languages' (09/2015 bis 02/2019)

Team

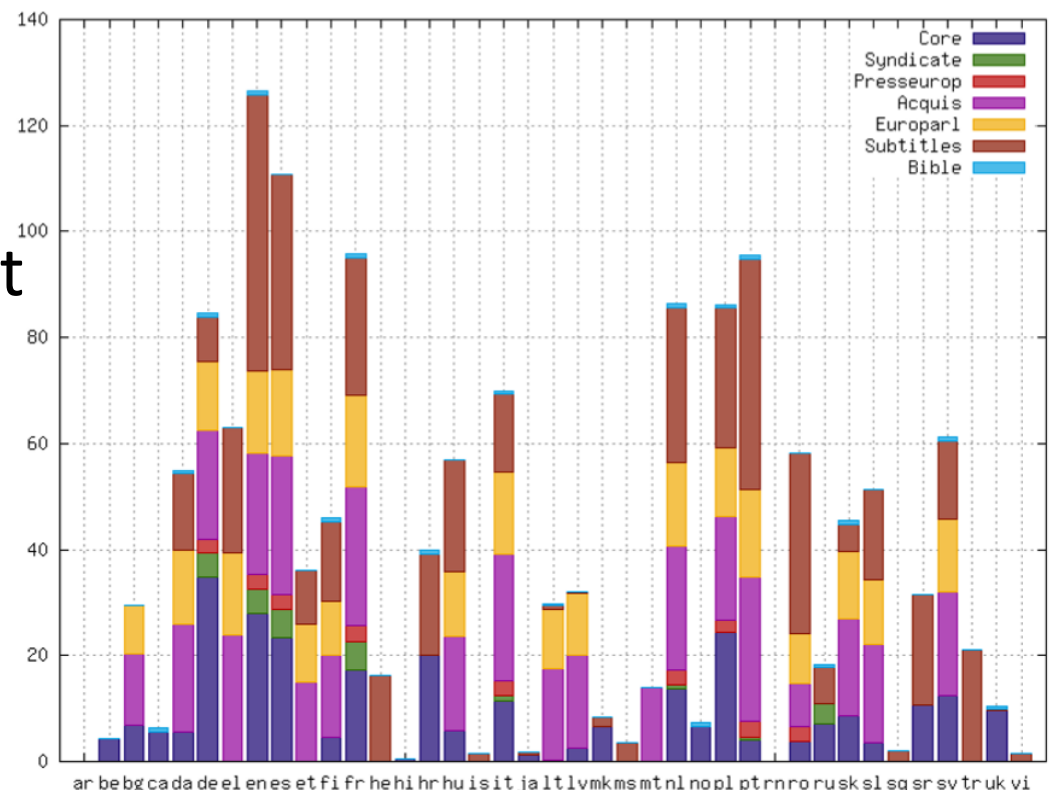
- ✓ Björn Hansen, (PI)
- ✓ **Zrinka Kolaković**, U Regensburg (main research assistant)
- ✓ **Edyta Jurkiewicz-Rohrbacher** U Regensburg /U Helsinki (research assistant)
- ✓ Nataša Dimitrijević (research assistant)
- ✓ **Dušica Filipović Đurđević**, U Novi Sad (consultant)
- ✓ Petar Vuković, U Zagreb (consultant)
- ✓ Theodora Tiha Loos (student helper)

Plenarna predavanja



- Ruprecht von Waldenfels: Variation on many levels: why and how comparing corpora and (Slavic) languages makes sense
- Alexandr Rosen: The merits of a parallel corpus and how to get the most out of it (InterCorp)

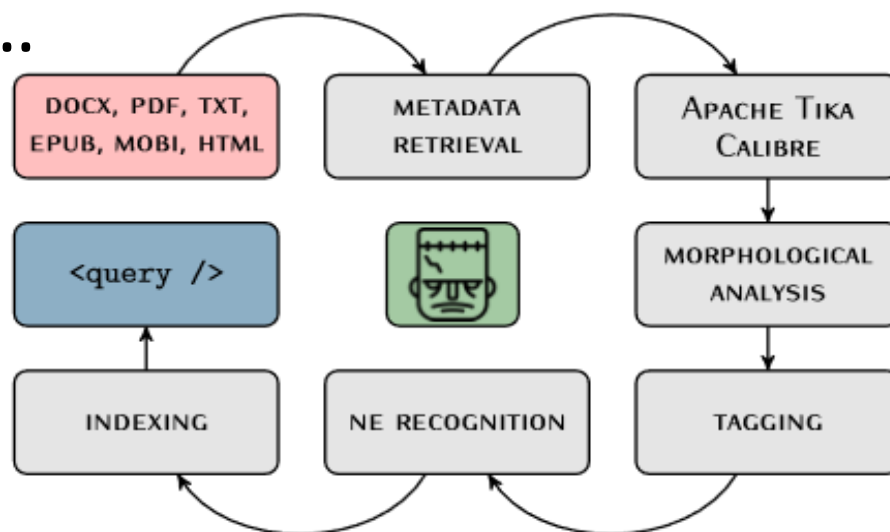
<https://kontext.korpus.cz/>



Još neke zanimljive teme



- From the old edition to the corpus
- Universal Dependencies ...
- Academic Dictionary ...
- Korpusomat

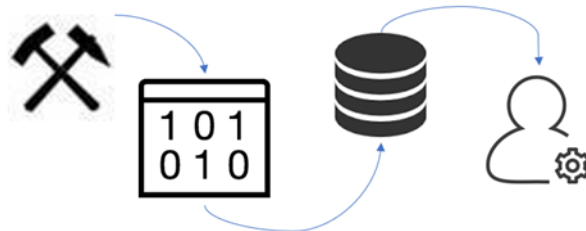


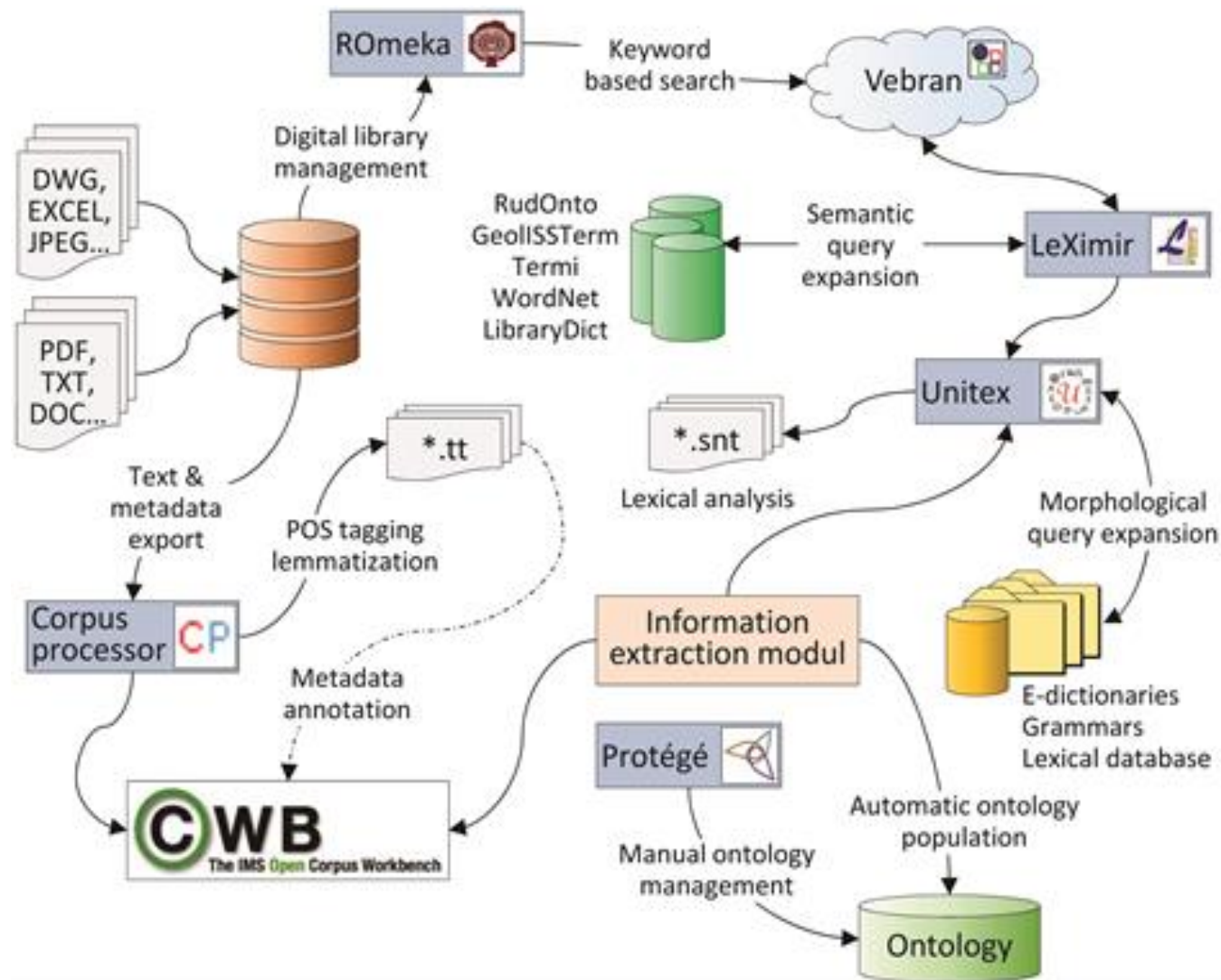
- I naravno.... mi



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





How do we upgrade?



Query expansion 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 open code corpus interfaces (PHP, Python)



Expansion using synonyms



- **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]

[|<](#)
[<<](#)
[>>](#)
[>|](#)
 Show Page:
[Line View](#)
[Show in corpus order](#)

[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 Γ

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

LeXimir: Lexical Resources management - [Editor for LADL Delas dictionaries: sr]

File Dictionary management Semantic networks Parallel text Conversion Options Window Help

LADL Delas Compounds Strategy Shell WordNet TMX Perl Awk C# Exe Lex Serv Text Tools RegEx Config Backup Help Exit

New Before After Edit View Sort Dict Delete filtered Delete InflectM Inflect Statistics WordNet Save Conv Count WN

POS Lemma FST SinSem Lemma inverted NewDictID

bandva All Exact Match Move

POS	Lemma	FSTCode	SinSem	Comment	LemmaInv	WN_Sinset	Lemn	Dict.File	
N	bandvagen	N1	+Instrum+DOM=Minin...	FST predict 29.0...	negavdnab		3400	delas-mining.dic	25

leximirka.jerteh.rs/LexicalEntry/View?uid=140039

Leximirka

bandvagen

Senses (1):

1. +Instrum+DOM=Mining

Domains: rudarstvo
Properties: oprema i delovi opreme

rudonto.rgf.bg.ac.rs/StabloRO.aspx

УНИВЕРЗИТЕТ У БЕОГРАДУ
РУДАРСКО - ГЕОЛОШКИ ФАКУЛТЕТ

Pretraga rečnika Rudarski rečnici OER Taksonomija O Projektu Terminološki resursi Korisnici

RUDARSKA TERMINOLOGIJA I NOMENKLATURA

SAMOHODNI TRANSPORTER
Taksonomija rudarskih pojmova
L rudarska oprema
L klasifikacija rudarske opreme
L transportna oprema

Definicija: -

Podređeni koncepti: [konstrukcija samohodnog transportera](#); [tehnologija samohodnog transportera](#);

Sinonimi: bandvagen

Potrebna nova verzija tagera će da uključi reči unete u morfološki rečnik od poslednje proizvodnje tagera

Document segment extraction

```

<item type="Naslov" n="A."> STUDIJA OPRAVDANOSTI SA IDEJNIM PROJEKTOM O
</list>
<item type="Deo" n="1."> OGRANIČENJE EKSPLOATACIONOG POLJA
</list>
<item type="Glava" n="1.1."> Geološka granica kopa
</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>
</list>
<head type="Glava" n="2"><seg>1.2. Geološki model ležišta
kopa</seg></head>
<p><seg>Proračun eksploatacionih rezervi kao i geometrija
(određivanje granica kopa i podela na etaže), urađena je na o
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 radovima
fizičkih informacija koje se koriste u procesu
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>

```

```

<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='Odeljak' 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>

```

2. Geološki model ležišta i eksploatacione granice

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

2.1. Geološka baza podataka i model ležišta Drmno

2.1.1. Baza podataka istražnih radova

Baza podataka o istražnim geološkim radovima je osnovni izvor informacija koje se koriste u procesu modelovanja ležišta. Podaci su dobijeni iz konačnoj dubini bušenja, litologiji, startigrafskim članovima, hemijskim analizama, podacima o devijaciji bušotina itd. Validana baza podataka je osnov za pouzdanu procenu mineralnih resursa i rudnih rezervi.

COST-DR WG2

“Methods and tools”



- sharing, evaluating and improving methods and tools for DR research
- “scholarly hackers,” active in computational linguistics, text mining, computational stylistics, and digital literary studies
- fertile crossovers between different disciplinary traditions and coding schools
- stimulate the awareness of the many frameworks and libraries through the organization of hands-on training schools and through the development of dedicated online tutorials on specific topics.
- unify existing language and text technologies in Digital Humanities (e.g. Named Entity Recognition) and bring them together as much as possible into a single platform/library, that allows users to seamlessly switch between different languages or corpora
- strengths and weaknesses of established modelling approaches in the field, for instance for topic modelling and text classification, including the standards and metrics that are necessary for their empirical evaluation

<https://www.distant-reading.net/wg-2/>

Training school



three-day training school (TS) in Galway,
preconference WS at EADH: 5-7 Dec. 2018

- day 1: corpus management and analysis with TXM (**Serge Heiden**)
- day 2 (parallel program?): distributional semantics (**Steffen Pielström**) + network analysis (**Meliha Handzic**, Gephi, Palladio)
- day 3: stylometry (and machine learning) (**Joanna Byszuk**)

<https://www.distant-reading.net/calls/call-galway/>

Corpus pipeline



Annotation pipeline for the novels in the ELTeC corpus, for the texts (the so-called level 2), arranged in 5 subsequent levels of increasing complexity:

- 2.1: word and sentence tokenization
- 2.2: POS tagging, including morphological analysis
- 2.3: Lemmatization
- 2.4: Named Entity Recognition
- 2.5: direct speech detection
- 2.6: more high-level classification tasks, including sentiment analysis

Annotating European Novels (1840-1920) for Distant Reading: A Critical Assessment of Tools



- Set up different repository for the sample (Lou)
- 20+ novels in ELTeC level 0 (deadline: **15th November 2018**):
 - Certain / minimum: English (Lou), Hungarian (DONE), Slovenian (Marco), Italian (Fabio), French (Serge + Christof), German (Fotis), Portuguese (Diana)
 - Possible: Spanish (), Norwegian () Mi?
- Select sample (responsible: Lou; deadline: **22 November 2018**):
 - 5 random passages of 400 “tokens” from each text: 100 samples per language. Whitespace tokenizer: 2,000 whitespace-delimited tokens per text.
- Groups: apply (presumably) the best tool that they have and return the output (**29 November 2018**) (Ranka NER)
- Annotators correct the annotation (deadline: **11 January 2019**)
- Corrections go back to the groups and they do their thing: perform comparison of tools (for each language); error analysis; suggestions for improvement and finetuning; comparison to baseline **figures for “normal”, contemporary text (9 March 2019)**. Write-up in Google Doc
- Write up (**12 April 2019**)

Za NER (i ostale zadatke)



- Task definition
- State of the art
- Suitable evaluation metrics
- Uniform labeling system
- Common tools
- Baseline results (for non-literary modern newspaper-like text)
- Annotation standard proposal for test data
- Performance off-the-shelf on 19c novel data

РАДНА



АКЦИЈА

На Семинару у новембру више
речи о кориговању текстова и
како се сви заинтересовани могу
прикључити акцији

Дођите...





Zainteresovani za
istraživanje,
pridružite se...



cvetana@matf.bg.ac.rs