Semantic annotation and expansion for keyword queries part2, hands-on

Ranka Stanković
ranka@rgf.rs

University of Belgrade, Serbia

http://rgf.rs/ranka2.pptx

COST Action IC1302

2nd KEYSTONE Training School

Keyword search in Big Linked Data

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Univesiy of Santiago de Compostela, Spain, 18-22 July 2016

keystone

Main topics

- Tools and guidelines
- Tool for enhanced search of bilingual digital libraries
- Examples of application in GIS
- Linguistic analysis and annotation
  - Math
  - Unitex
  - WebCQP
- Where To Go From Here?

Ontotex - GraphDB
Free RDF triplestore

- Semantically Rich NoSQL Graph Database
- data management for very large sets of structured, semi-structured or unstructured data.
- NoSQL (‘not only SQL’) graph databases serve organizations to access, integrate and analyze both unstructured data and data stored in the cloud, thus helping them with their big data and social media analytics.
- More on http://ontotext.com/knowledge-hub/
- Any experiance?
- Discussion?
- Beter option?

keystone

W3C Multilingual Linked Open Data Community Group

- Best Practices for Multilingual Linked Open Data Community Group (December 2015) https://www.w3.org/community/bpmlod/
  - Guidelines for Linguistic Linked Data Generation: Multilingual Dictionaries (BabelNet)
  - Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries
  - Guidelines for Linguistic Linked Data Generation: Multilingual Terminologies (TBX, TermBase eXchange)
  - Guidelines for developing NIF-based NLP Web Services
  - Guidelines for LLD Exploitation

keystone

NLP Interchange Format (NIF)

- NIF is an RDF/OWL-based format that aims to achieve interoperability between NLP tools language resources and annotations
- best practices to follow for the implementation of RESTful NLP web services that rely on NIF
- They implemented NIF wrappers for the *Stanford POS tagger* and *Stanford parser*
- http://www.w3.org/2015/09/bpmlod-reports/nif-based-nlp-webservices/

keystone

Parallel Corpora

- United Nations Parallel Corpus
  http://conferences.unite.un.org/UNCorpus
- OPUS open parallel corpus search & browse
- http://opus.lingfil.uu.se/
  - OPUS multilingual search interface
  - Europarl v7 search interface
  - Europarl v3 search interface
  - OpenSubtitles search interface
  - EUconst search interface
  - Word Alignment Database

keystone

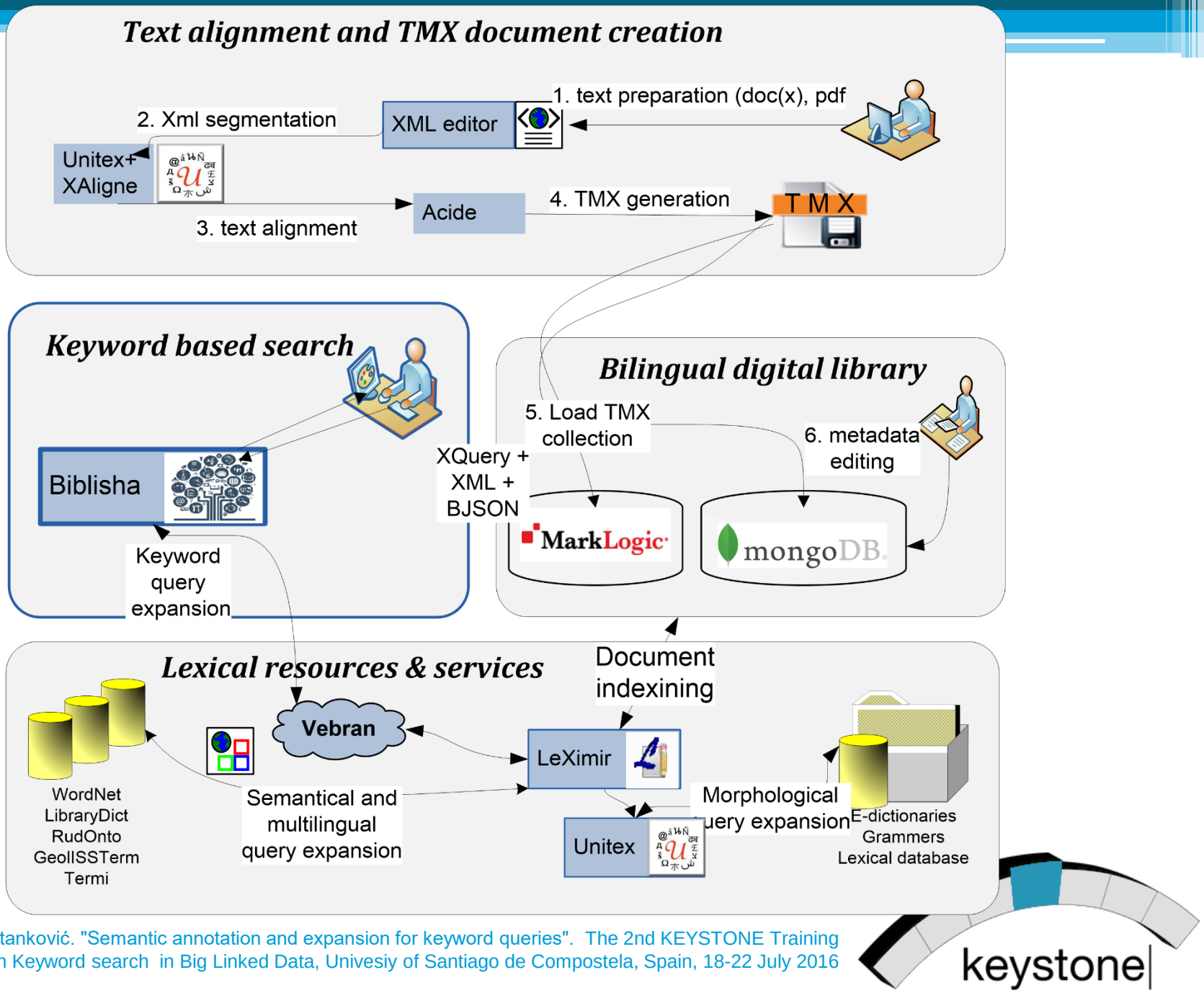
Task 1

- Find and select parallel corpus on ofered links
- Search & browse
- Find another (not listed on slides) in your country / region /language
- Compare: size, user interface, content

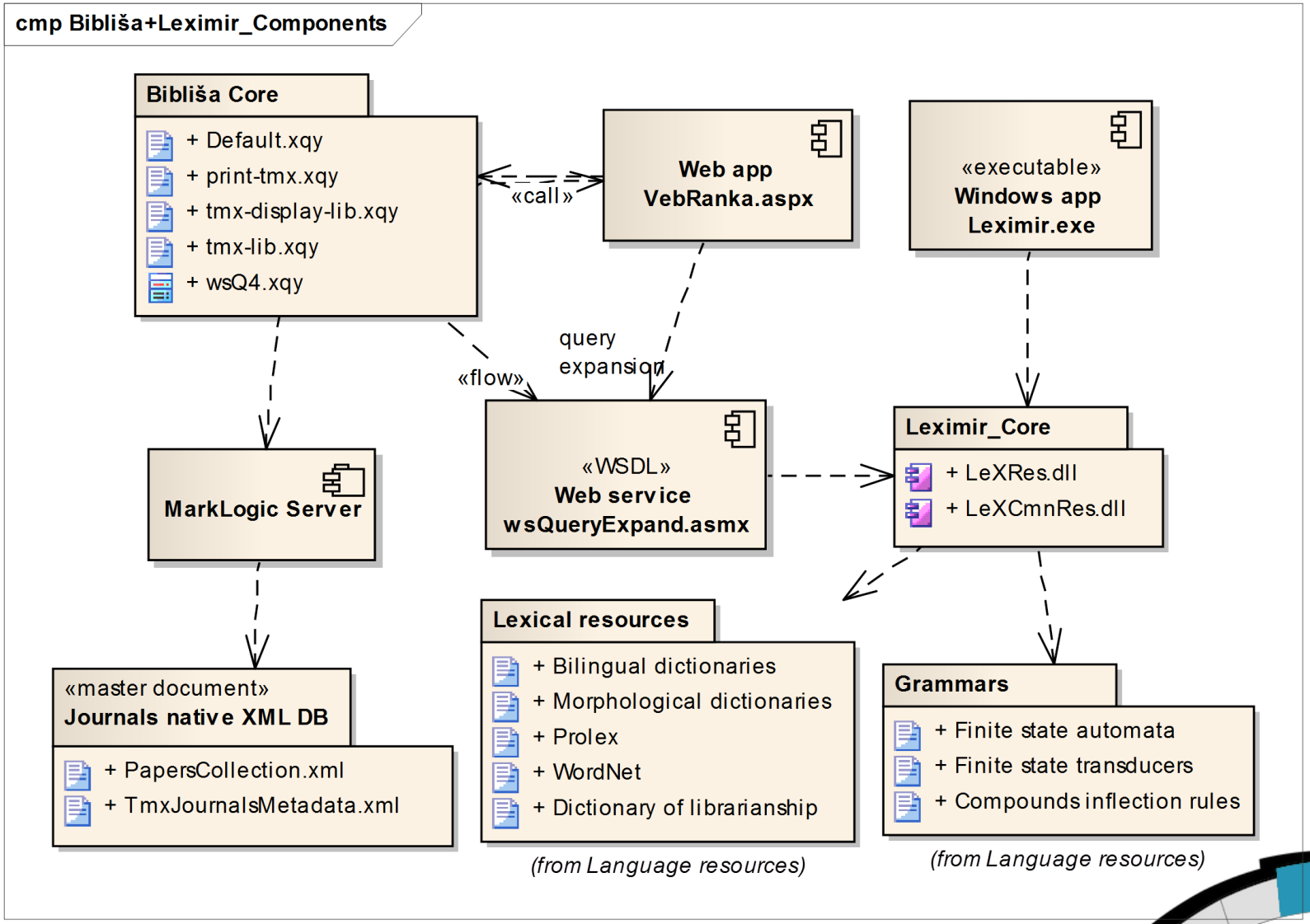
Query expansion for aligned text

- Bibliša (http://jerteh.rs/Biblisha) is a tool for enhancement of search possibilities in multilingual digital libraries of e-journals
  - Cross-lingual search functions for large collections of aligned texts
  - Queries with simple and multiword keywords in Serbian and English
  - Queries expanded semantically and morphologically
- Previously enhanced cross-lingual search implemented in LeXimir with XPath queries works only for one document
- Extension of cross-lingual search for collections relies on XML databases and XQuery language

keystone


Text alignment and TMX document creation
1. text preparation (doc(x), pdf
2. Xml segmentation
XML editor
Unitex+
XAligne
3. text alignment
Acide
4. TMX generation
TMX
Keyword based search
Biblisha
Keyword
query
expansion
Bilingual digital library
5. Load TMX
collection
6. metadata
editing
XQuery +
XML +
BJSON
MarkLogic
mongoDB.
Lexical resources & services
Document
indexing
Vebran
LeXimir
WordNet
LibraryDict
RudOnto
GeoIISSTerm
Termi
Semantical and
multilingual
query expansion
Morphological
query expansion
Unitex
E-dictionaries
Grammers
Lexical database

keystone


cmp Bibliša+Leximir_Components
Bibliša Core
+ Default.xqy
+ print-tmx.xqy
+ tmx-display-lib.xqy
+ tmx-lib.xqy
+ wsQ4.xqy
«call»
Web app
VebRanka.aspx
«executable»
Windows app
Leximir.exe
query
expansion
«flow»
MarkLogic Server
«WSDL»
Web service
wsQueryExpand.asmx
Leximir_Core
+ LeXRes.dll
+ LeXCmnRes.dll
«master document»
Journals native XML DB
+ PapersCollection.xml
+ TmxJournalsMetadata.xml
Lexical resources
+ Bilingual dictionaries
+ Morphological dictionaries
+ Prolex
+ WordNet
+ Dictionary of librarianship
(from Language resources)
Grammars
+ Finite state automata
+ Finite state transducers
+ Compounds inflection rules
(from Language resources)

keystone

Resources

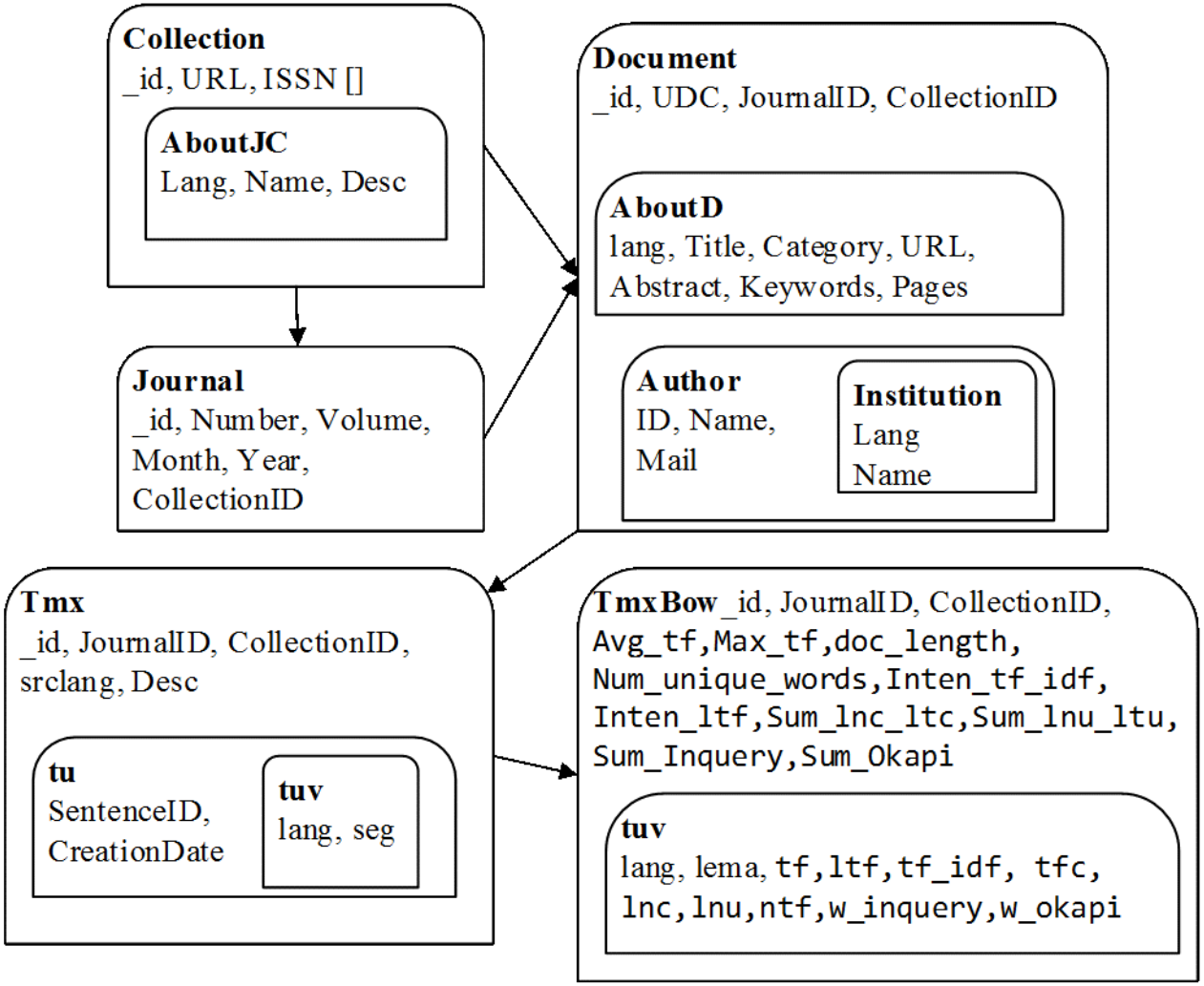
```
<tu>
  <prop type="Domain">Tomašević et al., 2012, vol. XX:20, ID: 2.2012.20.4
  </prop>
  <tuv xml:lang="en" creationid="n5 " creationdate="20140525T090842Z">
    <seg>It is the authors' wish to bring closer to the mining public at
    least a small part of the possibilities offered by GIS, both in mining
    activities and in environment protection. </seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n5 " creationdate="20140525T090842Z">
    <seg>Želja autora je da rudarskoj javnosti približi samo mali deo
    ogromnih mogućnosti koje GIS pruža, kako pri rudarskim aktivnostima,
    tako i u zaštiti životne sredine. </seg>
  </tuv>
</tu>
```

```
"Desc": "Tomašević et al., 2012, vol. XX:20, ID: 2.2012.20.4",
"SentenceID": "n5",
"tuv": [{
    "lang": "en",
    "seg": "It is the authors' wish to bring closer to the mining
    public at least a small part of the possibilities offered by
    GIS, both in mining activities and in environment protection. "
},
{
    "lang": "none",
    "seg": "Želja autora je da rudarskoj javnosti približi samo mali
    deo ogromnih mogućnosti koje GIS pruža, kako pri rudarskim
    aktivnostima, tako i u zaštiti životne sredine. "
}]
```

- MarkLogic (http://www.marklogic.com) server for management of document collections consisting of aligned parallel texts converted in TMX (Translation Memory eXchange) format and MongoDB (recently)
- Collection of TMX documents generated by ACIDE with export routines for automatic generation of TMX versions of parallel texts aligned by Xalign

keystone

Resources -metadata



keystone

Software implementation

- Excerpt from XQuery code: variable $x creates a collection of “tu” XML nodes (elements), which can subsequently be queried, yielding the results obtained by $query tagged with :

```
for $x in cts:search(fn:doc()//tu,$query)
return
cts:highlight($x, $query,<em>{$cts:text}</em>)
```

- The results, a set of aligned concordances, are formatted and presented to the user
- The concordances are preceded by information identifying the document they originate from, and a link to summary metadata for this document in both languages

keystone

Supporting resources

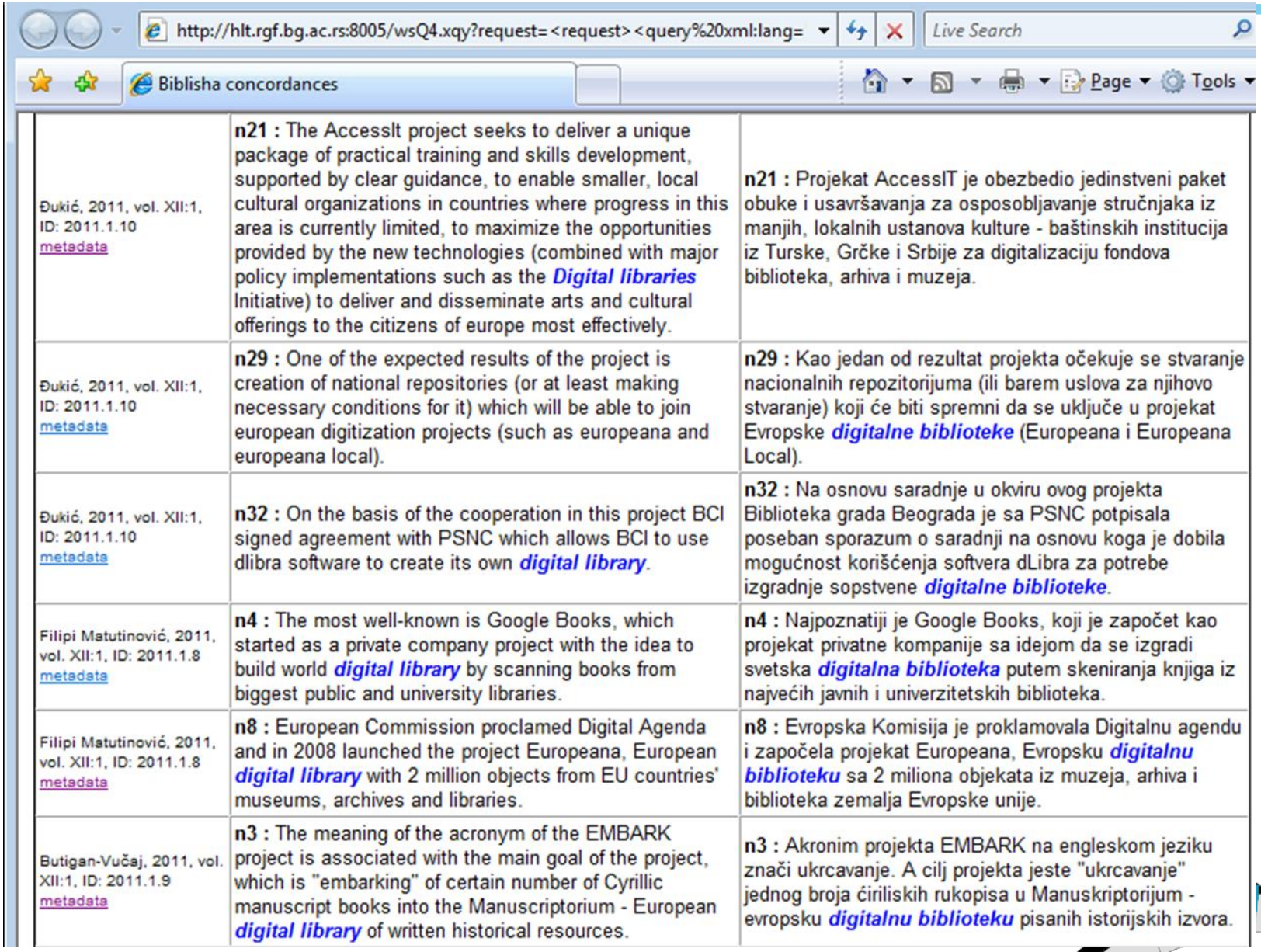
- Four types of resources are used for the expansion of queries submitted to our collection of documents.
- Serbian morphological dictionaries of simple words and multi-word units developed within: Unitex and Nooj. are used to generate all inflective forms of query keywords, thus improving the system recall without negative effects on precision: 135 000 lemmas of simple words and 20 000 compounds.
- Grammars in the form of finite state automata and transducers, amended by compounds inflection rules
- Princeton English Wordnet (PWN), version 2.0 and Serbian Wordnet (SrpWN), initially developed in the scope of the BalkaNet project and subsequently enhanced and upgraded, 17 500 synsets with 30 000 literals
- Dictionary of librarianship: English-Serbian and Serbian-English 23 400 terms – 11 300 in English and 12 100 in Serbia

keystone

```
<request>
    <query xml:lang='sr'>digitalna biblioteka</query>
    <query xml:lang='sr'>digitalnih biblioteka</query>
    <query xml:lang='sr'>digitalne biblioteke</query>
    <query xml:lang='sr'>digitalnim bibliotekama</query>
    ....
    <query xml:lang='en'>digital library</query>
</request>
```

```
cts:or-query ((cts:word-query ("digitalna biblioteka",      ("stemmed","lang= sr") ).
cts:or-query ((cts:word-query ("digitalnih biblioteka",     ("stemmed","lang= sr") ).
cts:or-query ((cts:word-query ("digitalne biblioteke",      ("stemmed","lang= sr") ),
cts:or-query ((cts:word-query ("digitalnim bibliotekama", ("stemmed","lang= sr") ),
.....
                cts:word-query (digital library,            ("stemmed", "lang= sr") ) ))
```


keystone



Đukić, 2011, vol. XII:1, ID: 2011.1.10 metadata	**n21 :** The AccessIt project seeks to deliver a unique package of practical training and skills development, supported by clear guidance, to enable smaller, local cultural organizations in countries where progress in this area is currently limited, to maximize the opportunities provided by the new technologies (combined with major policy implementations such as the ***Digital libraries*** Initiative) to deliver and disseminate arts and cultural offerings to the citizens of europe most effectively.	**n21 :** Projekat AccessIT je obezbedio jedinstveni paket obuke i usavršavanja za osposobljavanje stručnjaka iz manjih, lokalnih ustanova kulture - baštinskih institucija iz Turske, Grčke i Srbije za digitalizaciju fondova biblioteka, arhiva i muzeja.
Đukić, 2011, vol. XII:1, ID: 2011.1.10 metadata	**n29 :** One of the expected results of the project is creation of national repositories (or at least making necessary conditions for it) which will be able to join european digitization projects (such as europeana and europeana local).	**n29 :** Kao jedan od rezultat projekta očekuje se stvaranje nacionalnih repozitorijuma (ili barem uslova za njihovo stvaranje) koji će biti spremni da se uključe u projekat Evropske ***digitalne biblioteke*** (Europeana i Europeana Local).
Đukić, 2011, vol. XII:1, ID: 2011.1.10 metadata	**n32 :** On the basis of the cooperation in this project BCI signed agreement with PSNC which allows BCI to use dlibra software to create its own ***digital library***.	**n32 :** Na osnovu saradnje u okviru ovog projekta Biblioteka grada Beograda je sa PSNC potpisala poseban sporazum o saradnji na osnovu koga je dobila mogućnost korišćenja softvera dLibra za potrebe izgradnje sopstvene ***digitalne biblioteke***.
Filipi Matutinović, 2011, vol. XII:1, ID: 2011.1.8 metadata	**n4 :** The most well-known is Google Books, which started as a private company project with the idea to build world ***digital library*** by scanning books from biggest public and university libraries.	**n4 :** Najpoznatiji je Google Books, koji je započet kao projekat privatne kompanije sa idejom da se izgradi svetska ***digitalna biblioteka*** putem skeniranja knjiga iz najvećih javnih i univerzitetskih biblioteka.
Filipi Matutinović, 2011, vol. XII:1, ID: 2011.1.8 metadata	**n8 :** European Commission proclamed Digital Agenda and in 2008 launched the project Europeana, European ***digital library*** with 2 million objects from EU countries' museums, archives and libraries.	**n8 :** Evropska Komisija je proklamovala Digitalnu agendu i započela projekat Europeana, Evropsku ***digitalnu biblioteku*** sa 2 miliona objekata iz muzeja, arhiva i biblioteka zemalja Evropske unije.
Butigan-Vučaj, 2011, vol. XII:1, ID: 2011.1.9 metadata	**n3 :** The meaning of the acronym of the EMBARK project is associated with the main goal of the project, which is "embarking" of certain number of Cyrillic manuscript books into the Manuscriptorium - European ***digital library*** of written historical resources.	**n3 :** Akronim projekta EMBARK na engleskom jeziku znači ukrcavanje. A cilj projekta jeste "ukrcavanje" jednog broja ćiriliskih rukopisa u Manuskriptorijum - evropsku ***digitalnu biblioteku*** pisanih istorijskih izvora.

Main features of *Bibliša*

- A complex system composed of several modules
- Targeted at textual resources in the form of collections of TMX documents and the corresponding metadata
- Able to expand search queries:
  - morphologically, using morphological e-dictionaries, the system of rules for compound inflection, and finite automata and transducers
  - semantically and bilingually, using Serbian and English wordnets and the bilingual Dictionary of Librarianship

E-dictionaries	Serbian wordnet	Dictionary of Librarianship
128,000 simple word lemmas	17,500 synsets	11,300 English terms
10,000 compound word lemmas	30,000 literals	12,100 Serbian terms

keystone

Query realization in *Bibliša*

- The user formulates the initial query as one or more keywords (simple or multiword)
- If the user so specifies, Bibliša forwards this query for further morphological and semantic expansion
- The expansion is handled by a web service that is part of the LeXimir software package, a multipurpose tool also developed by the HLT Group
- The web service invokes LeXimir's function library LeXimirCore, whose functions expand the query, using available lexical resources and Unitex routines
- The expanded query is transformed into an XQuery and used for searching the TMX document collection obtained from journal articles
- As a result a set of aligned concordances is obtained and presented to the user

What we can learn?

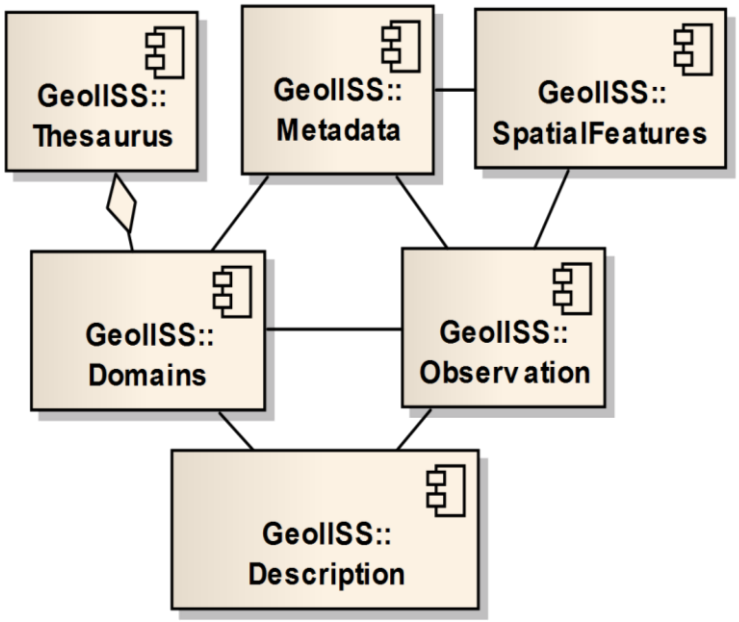
- Check context of terms
- Check other options for translation (partial match)
- Expand dictionary (with partial match)
- Use TMX for word alignement
- Aligned word for query expansion
- Machine translation

Task 2 Query expand using Biblisha

- http://jerteh.rs/biblisha/
- Metadata search
- Fultext search
- Expand query (have in mind domains of resources)
- Search in Mongo and Mark Logic
- Diferences?
- MongoDB multilingual search experiance?
- Best DBMS solution for multilingual paralel corpora?

Query expansion in GIS

Stanković Ranka, Obradović Ivan, Kitanović Olivera, "GIS Application Improvement with Multilingual Lexical and Terminological Resources", Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2010, Valetta, Malta, May 2010 , pp.2283-2287. European Language Resources Association Valetta, Malta 2010 ISBN:2-9517408-6-7 [link]

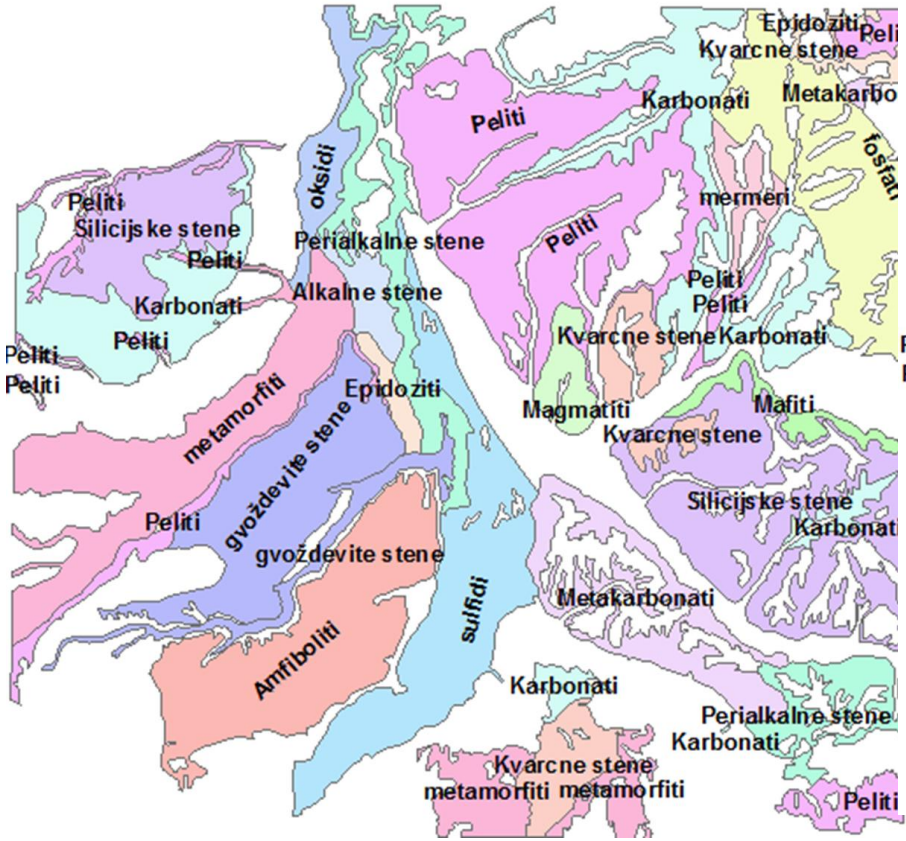


- GeolISS
  - Digital archiving, query, retrieving, analysis and visualization of geological data
  - Integration of existing geologic archives, data from published maps, newly acquired field data
  - Domains controlled by the geologic thesaurus
  - Implemented using .NET + ESRI ArcGIS

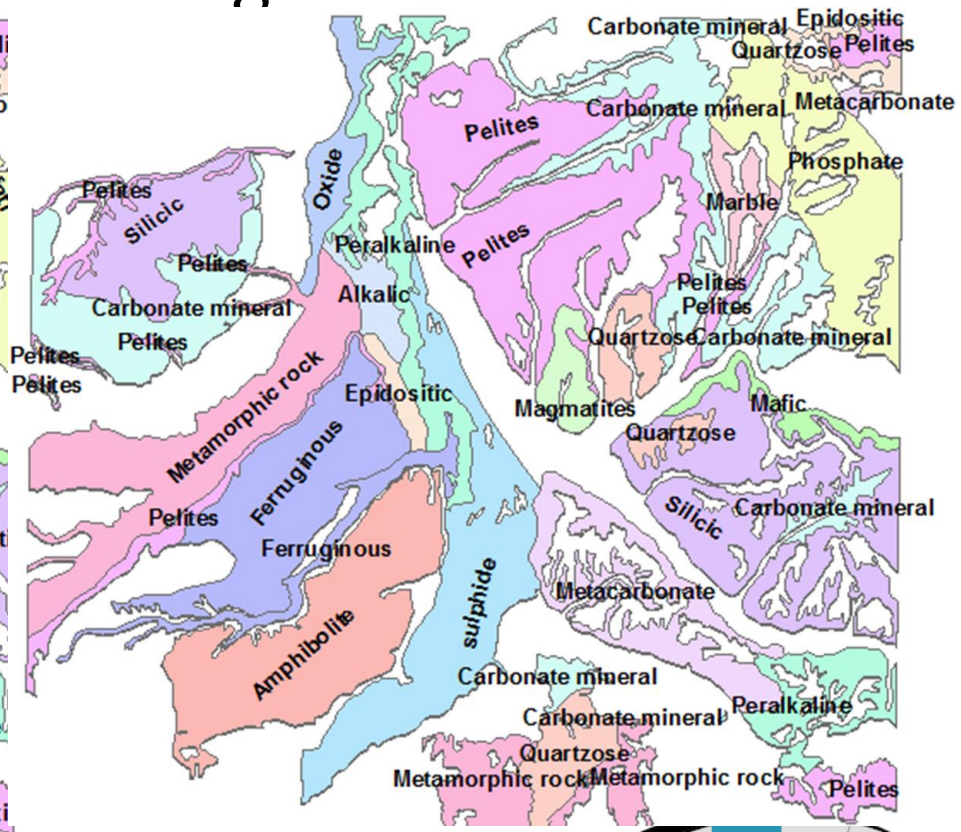
keystone

Multilingual search and annotation

Serbian



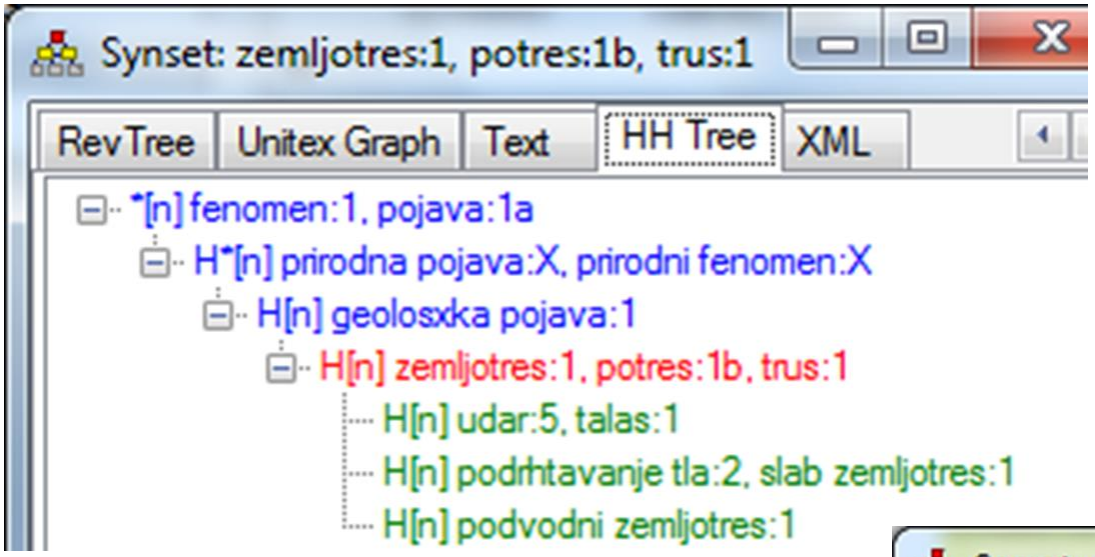
English

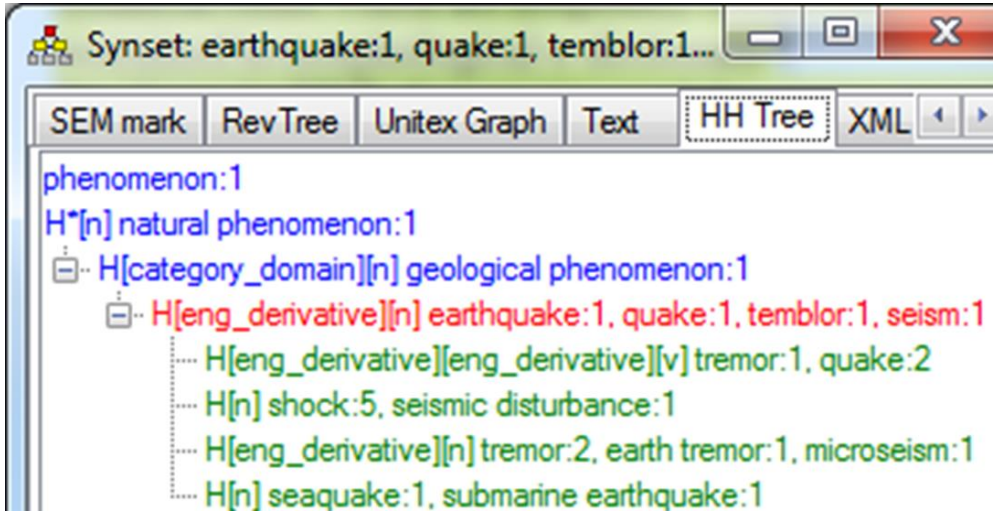


keystone

QE by WordNet

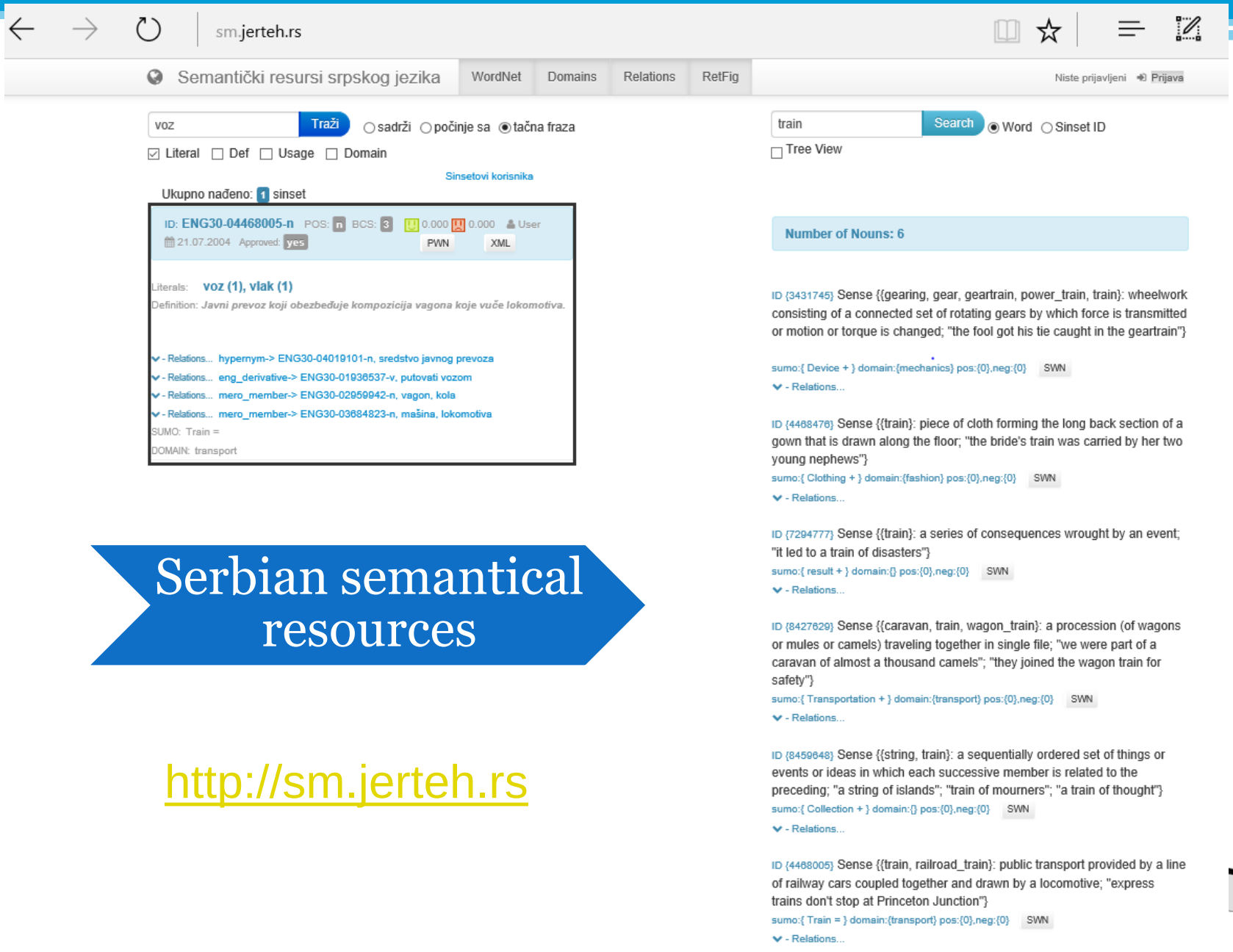
Serbian & English WordNet
HH tree for synsets from domain 'geology





- Other HLT Group resources
  - morphological e-dictionaries in LADL format
  - finite state transducers developed within the Unitex system
  - electronic thesauri
  - ontologies
- HLT Group applications
  - LeXimir (various language processing related tasks)
  - Web Service for Query Expansion

keystone



Serbian semantical resources

http://sm.jerteh.rs

keystone

Task 3. Search Wordnet

- Search Serbian WN (sebian left side, english right side)
- Search Prinston on their site
- Find WN in your (favorite) language
- Compare?
- Sugesstions?
- Best solution?

QE by Geological thesaurus

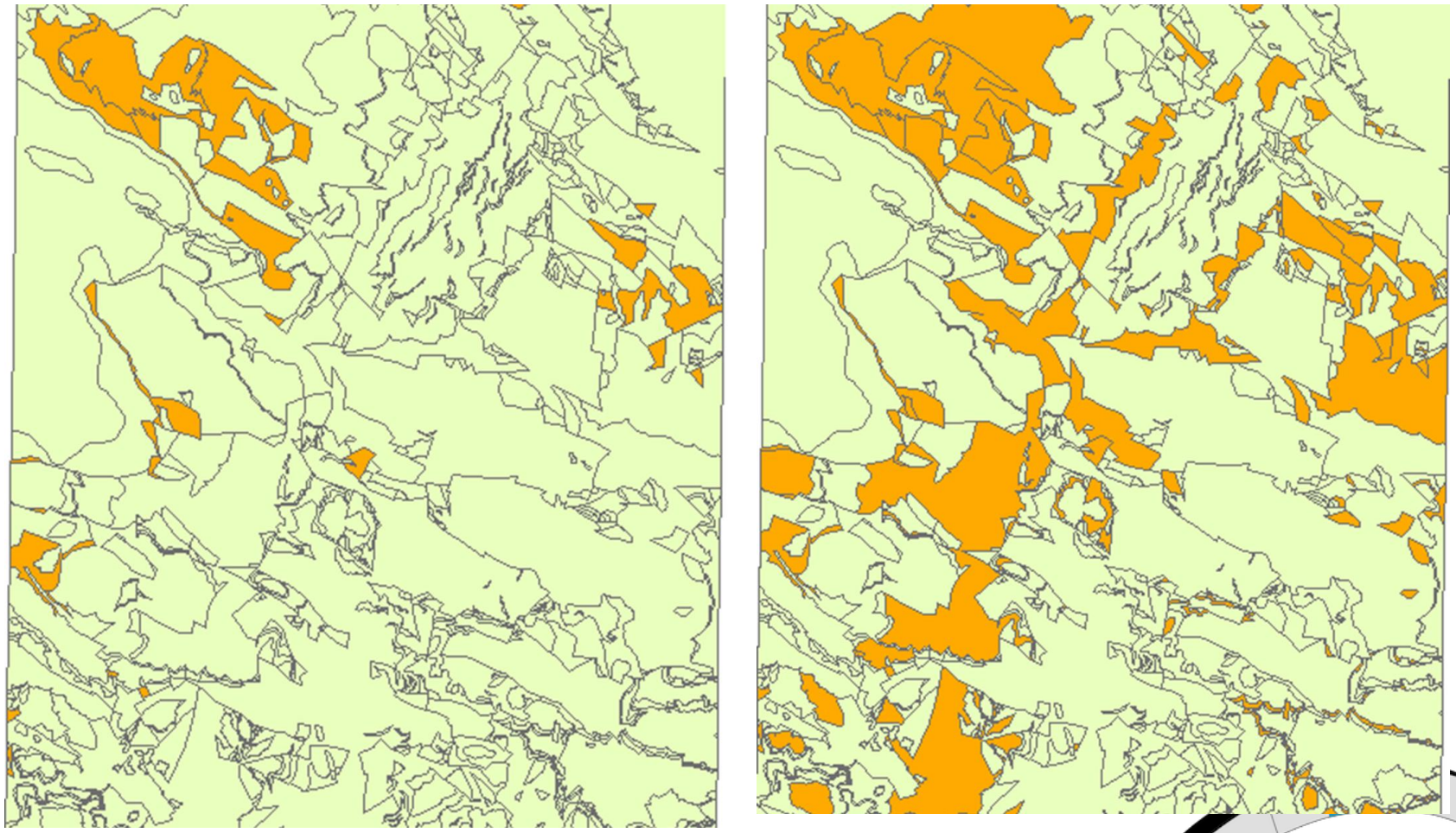
- GeolISSTerm, aggregation of geological vocabularies
  - {ID=2356, Name=Piroklastična stena, Def= Stena nastala depozicijom i litifikacijom piroklasticnih naslaga; fragmenti ove stene obrazovani su direktnom eksplozivnom fragmentacijom, Synonym= Piroklastit, Hyperonym= 2123}
- multilingual
  - {ID=12345, Name=Pyroclastic rock, Def=A volcanoclastic rock formed by direct explosive volcanic activity, OrigID=2356, Lng=EN)
  - {ID=12367, Name= Roches pyroclastiques, Def= Une pierre volcanoclastiques formées par l'activité volcanique explosive direct, OrigID=2356, Lng=FR)

GeolISS Query Expansion

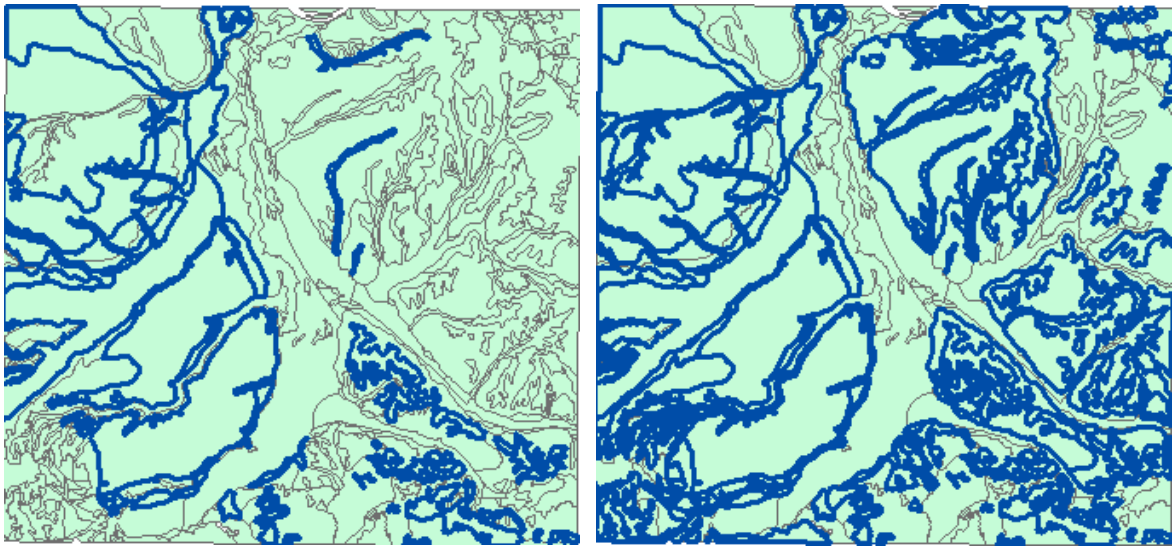
- Selection of keywords can be improved by using lexical resources, morphological dictionaries and transducers
- Morphological expansion
  - retrieval of geological units containing ‘limestone’ in their description field ‘ (‘krečnjak’ in Serbian)
    >>> 95
  - with morph.exp.: ‘krečnjak, krečnjaka, krečnjaku, krečnjakom, krečnjače, krečnjaci, krečnjacima, krečnjake’
    >>> 249
- GeolISSTerm Semantic expansion
  - ‘Middle Jurassic’ (srednja jura) >>> 53 + doger (dogger) >>> 134
  - Lithologic constituent (+ Lithologic Member + Lithologic 'Package') no improvement
  - WordNet Semantic expansion ‘izvor, vrelo’ (eng. ‘spring, outflow, outpouring, natural spring’)

keystone

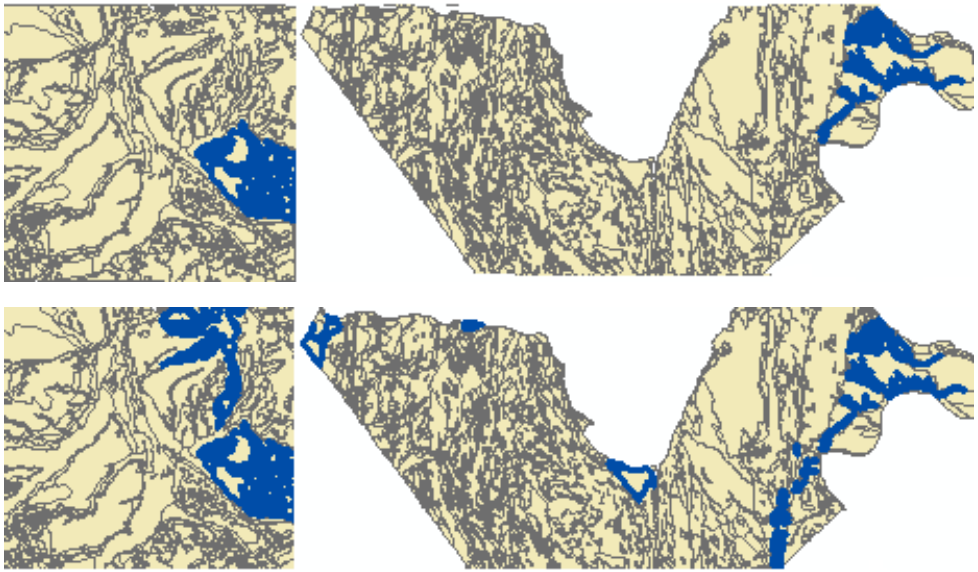
Selected map units with the original and morphologically expanded query



QE in GIS



<u>Morphological expansion</u>
glina (clay) 51
→
glina, glinama, glinom, glinu, glini, gline 118



<u>Semantic expansion</u>

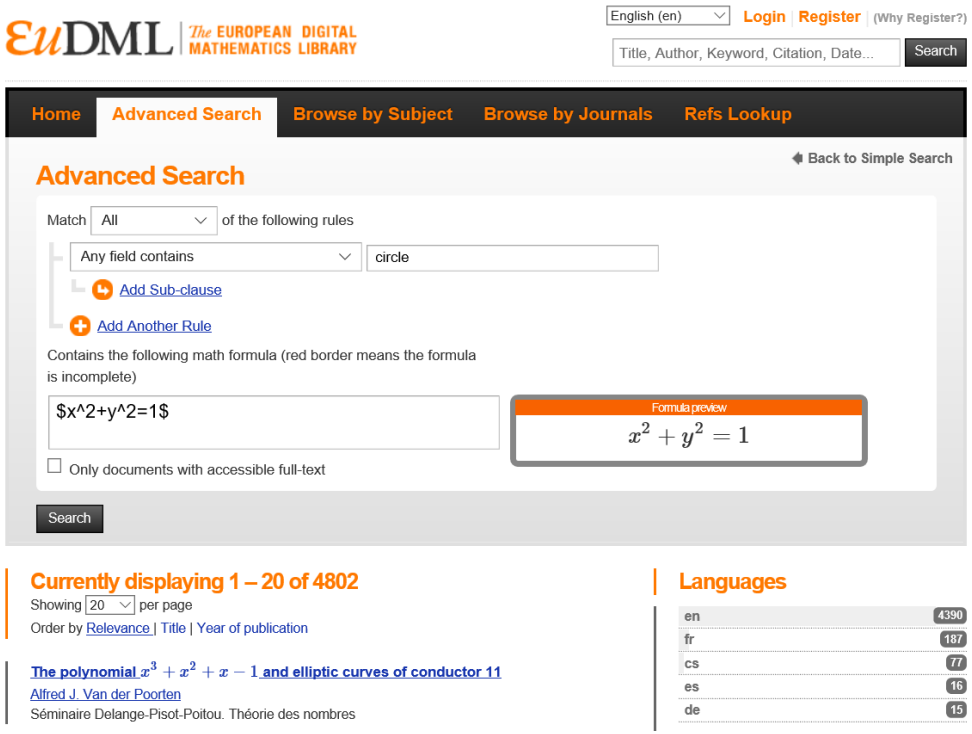
Lower Pliocene (donji pliocen) 13
→
donji pliocen, pont 33

keystone

What is EuDML?

https://eudml.org/search/

EuDML makes the mathematics literature available online in the form of an enduring digital collection, developed and maintained by a network of institutions.



keystone

Task 4. Search EuDML

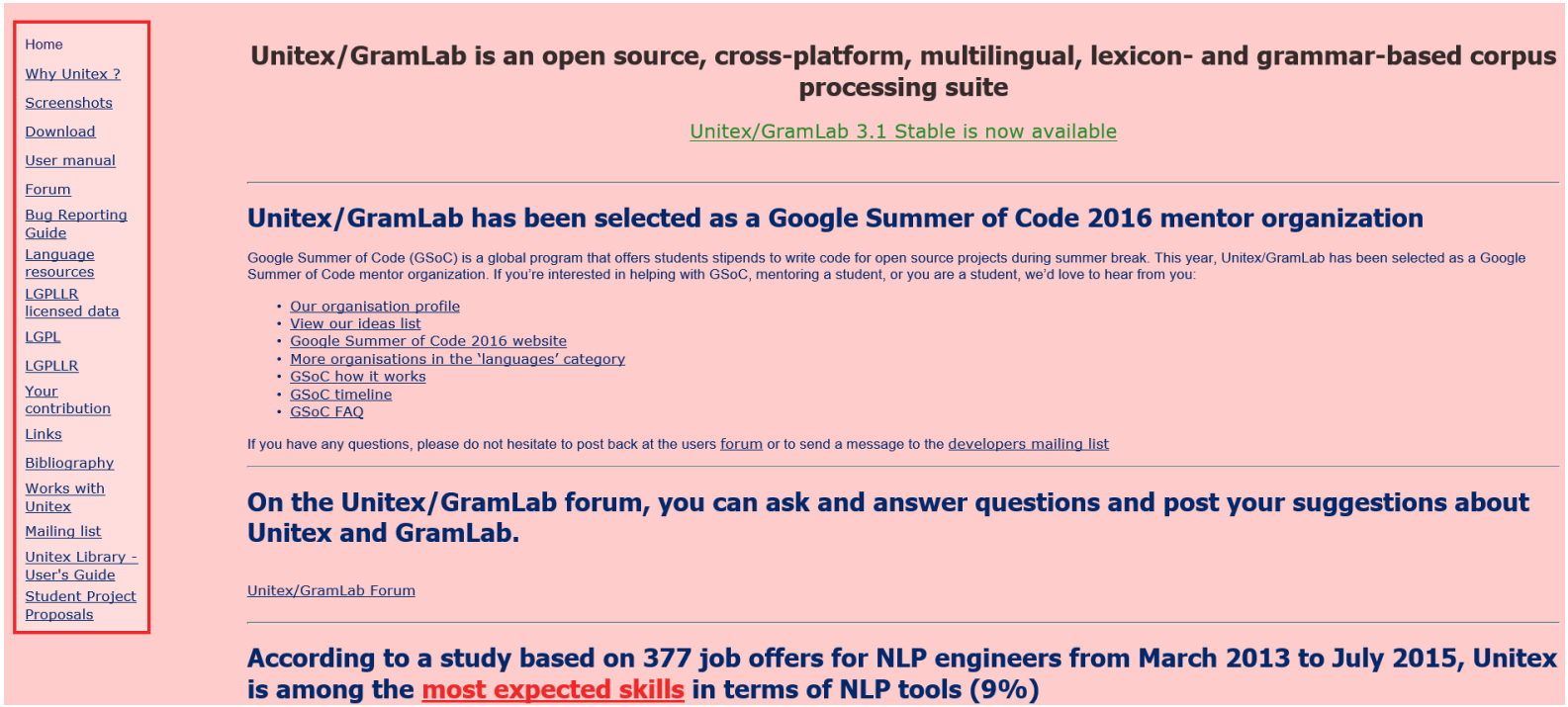
- Find documents for circle with or without formula
- Change names od variables. Compare results.
- Filter language
- Same on other library (math)
- Find new (not listed)
- Best solution?

keystone

Unitex

Demo for english, french, serbian,...

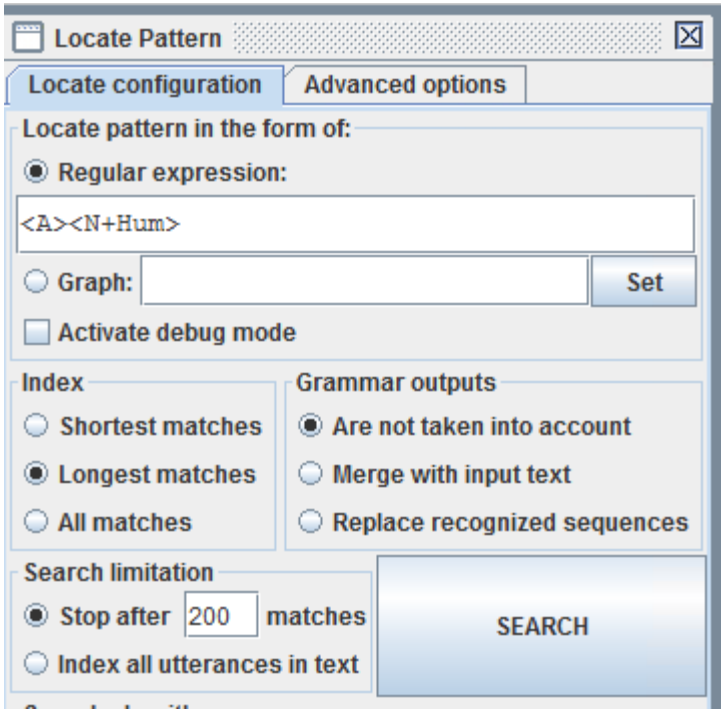
- http://www-igm.univ-mlv.fr/~unitex



keystone

Task 5a. Unitex

- Run
- Select language
- Open text document (Text->Open)
- Process (automatic, default)
- Text -> Locate patern (fig.)
- Build concordances



Paumier, S. "Unitex 3.1 User Manual (2016).„ http://www-igm.univ-mlv.fr/~unitex/

keystone

Task 5b. Query in Unitex

- **<to be>** - all word forms linked to lemma
- **<N>** - search for POS – all nouns
- **<DET><A><N+Hum>** - noun preceeded by determiner and adjective
- **<N+Hum>** - nouns with semantic tag Human
- **<N+NProp+Hum~Inh>** - ...not Inhabitans
- **(<A>+<PRO+ProA>) <love>**
- **(<this>+<that>) <A> <N>**
- **Try yours idea!**

keystone

Task 6 CQPweb

- Try
  - http://cwb.sourceforge.net/cqpweb.php
  - https://cqpweb.lancs.ac.uk
- CQP language
  - http://cwb.sourceforge.net/files/CQP_Tutorial/
- Find other
- Favorite?

keystone

Thank you for your attention

Hvala na pažnji

Хвала на пажњи

keystone