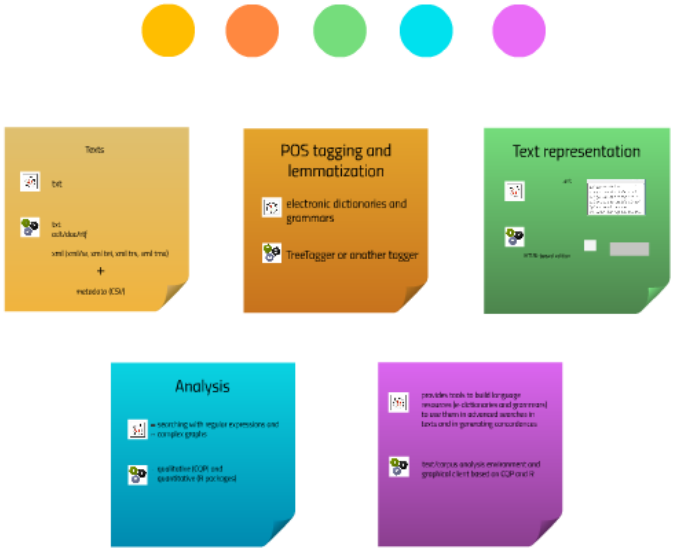
Serbian Unitex Day

UNITEX vs. TXM
What's the Difference?



open-source, cross-platform, multilingual
corpus processing environments

Jelena Jaćimović

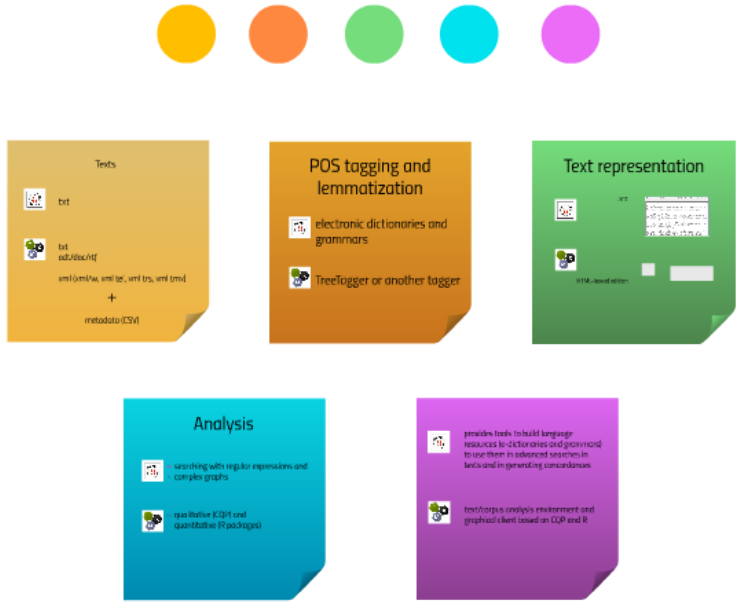
University of Belgrade, Faculty of Mining and Geology
March 11, 2019

Thank you!

Acknowledgements
COST Action 16204 – Distant Reading for European Literary History
STSM-CA16204-42562
Host institution: Institut d'Histoire des Représentations et des Idées dans les Modernités UMR 5317, École Normale Supérieure de Lyon, Lyon, France

Serbian Unitex Day

UNITEX vs. TXM
What's the Difference?



open-source, cross-platform, multilingual corpus processing environments

Jelena Jaćimović

University of Belgrade, Faculty of Mining and Geology
March 11, 2019

Texts

txt

txt
odt/doc/rtf

xml (xml/w, xml tei, xml trs, xml tmx)

+

metadata (CSV)

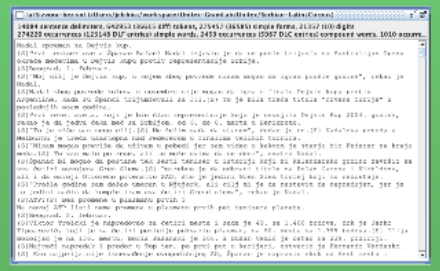
POS tagging and lemmatization

electronic dictionaries and grammars

TreeTagger or another tagger

Text representation

.snt



HTML-based edition

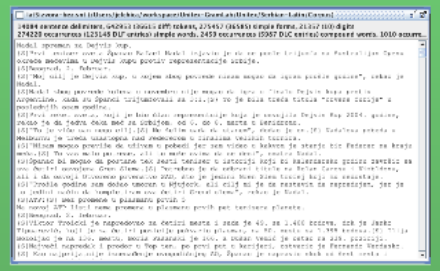
lat5izvora-bez.snt (/Users/jelchica/workspace/Unitex-GramLab/Unitex/Serbian-Latin/Corpus)

14084 sentence delimiters, 642953 (36615 diff) tokens, 275457 (36585) simple forms, 21357 (10) digits
274220 occurrences (123148 DLF entries) simple words, 2459 occurrences (5987 DLC entries) compound words, 1010 occurre...

Nadal spreman za Dejvis kup.
{S}Prvi teniser sveta Španac Rafael Nadal izjavio je da se posle trijumfa na Australijan Openu okreće mečevima u Dejvis kupu protiv reprezentacije Srbije.
{S}Beograd, 2. februar.
{S}"Moj cilj je Dejvis kup, u kojem zbog povrede nisam mogao da igram prošle godine", rekao je Nadal.
{S}Nadal zbog povrede kolena u novembru nije mogao da igra u finalu Dejvis kupa protiv Argentine, kada su Španci trijumfovali sa 3:1.{S} To je bila treća titula "crvene furije" u poslednjih osam godina.
{S}Prvi reket sveta, koji je bio član reprezentcije koja je osvojila Dejvis Kup 2004. godine, rekao je da jedva čeka meč sa Srbijom, od 6. do 8. marta u Benidromu.
{S}"To je više san nego cilj.{S} Ne želim sada da stanem", dodao je on.{S} Nadalova pobeda u Melburnu je treća uzastopna nad Federerom u finalima velikih turnira.
{S}"Nisam mogao previše da uživam u pobedi jer sam video u kakvom je stanju bio Federer na kraju meča.{S} To vas malo potrese, ali to može svima da se desi", smatra Nadal.
{S}Španac bi mogao da postane tek šesti teniser u istoriji koji bi kalendarsku godinu završio sa sva četiri osvojena Gren Slema.{S} Potrebno je da odbrani titule na Rolan Garosu i Vimbldonu, ali i da osvoji Otvoreno prvenstvo SAD, što je jedini Gren Slem trofej koji mu nedostaje.
{S}"Prošle godine sam došao umoran u Njujork, ali cilj mi je da nastavim da napredujem, jer je to jedini način da kompletiram sva četiri Grend slema", rekao je Nadal.
{S}ATP:{S} Bez promene u plasmanu prvih 5
Na novoj ATP listi nema promena u plasmanu prvih pet tenisera planete.
{S}Beograd, 2. februar.
{S}Viktor Troicki je napredovao za četiri mesta i sada je 49. sa 1.400 bodova, dok je Janko Tipsarević, koji je za četiri pozicije pokvario plasman, na 50. mestu sa 1.395 bodova.{S} Ilija Bozoljac je na 136. mestu, Boris Pašanski je 166. a Dušan Vemić je ostao na 239. poziciji.
{S}Najveći napredak i prodor u Top ten, po prvi put u karijeri, ostvario je Fernando Verdasko.
{S} Kao najprijatnije iznenađenje ovogodišnjeg AO, Španac je napravio skok od šest mesta i

Text representation

.snt



HTML-based edition

- 1 -

ДЕСЕТ ПАРА

(НЕШТО ИЗ ЖИВОТА У ВАРОШИ)

I

Проводадџија

«Метни дрво уз дрво да боље гори!» —

Проводад. пословица

Улицом од Беле Чешме ка Правоме Раскршћу, једнога дана после по дне, жураше се једна жена која на десну ногу малко храмаше. Уз пут се ником не јављаше, и нешто се толико бејаше замислила да се сама са собом гласно разговараше.

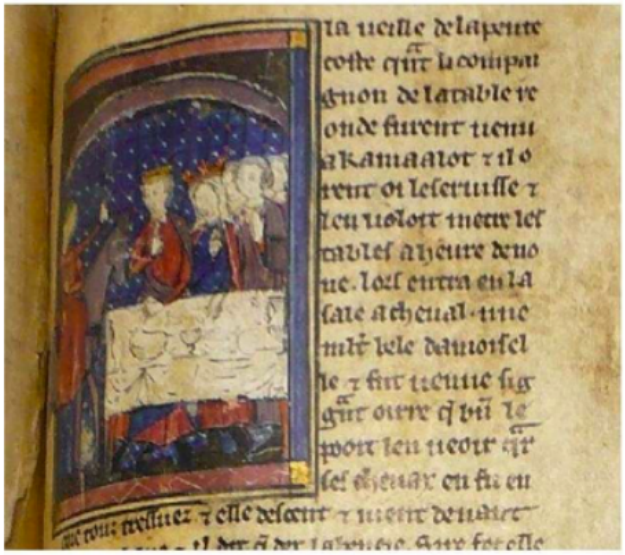
<160a>

Ici commence la version de la Queste del saint Graal donnée par le manuscrit K (Bibliothèque Municipale de Lyon, Palais des Arts n° 77), folios 160 recto à 224 verso. Tout le début du texte a été mutilé : la première grande lettre a été découpée, comme on le voit sur la reproduction du manuscrit, et quelques lignes du texte manquent, que nous donnons ici entre crochets, en bleu, d'après le manuscrit Z (Paris, BNF n. acq. fr. 1119, folio 138 recto, colonne a) qui est un manuscrit proche de celui que nous éditons ici.

(p. 1)

[§ 1]

[A la veille de la Pente-
coste quant li compai-
gnon de la table re-
onde furent venu
a Kamaalot et il o-
rent oï le servisse et
l'en voloit metre les
tables a heure de]
nonne.´ lors en[tra] [a cheval en la][1] sale une mout bele
damoisele, et fu venue si grant oirre que bien le pot
l'en veoir, car ses chevaux en fu encore toz suanz, et ele
descent et vient devant le roi si le salue, et il dist que
Diex la beneïe. « Sire, fet ele, por Dieu dites moi se Lancelot



Retour au manuscrit K (Lyon, BM, P.A. 77, col. 160a)

courante | ms-colonne

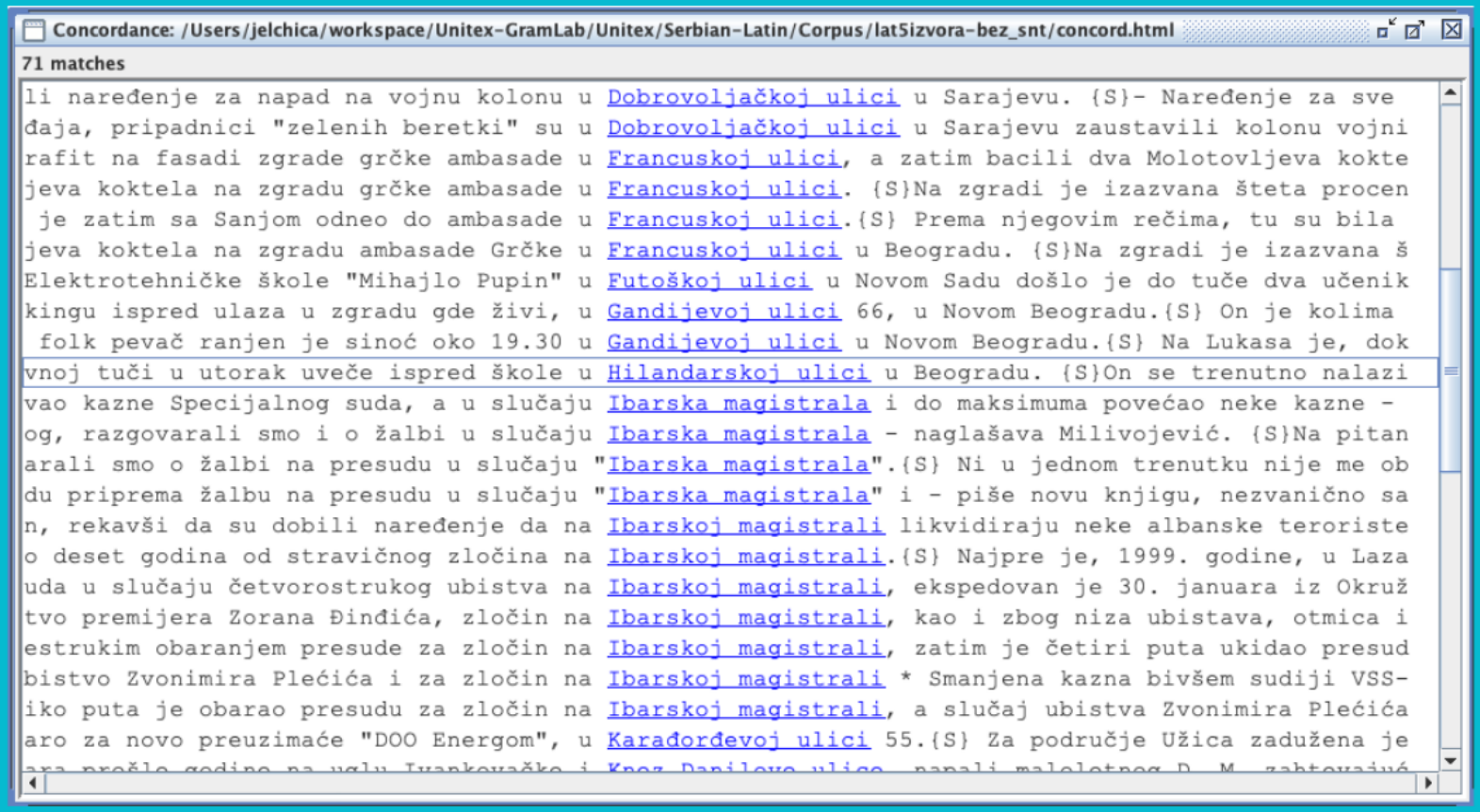
160a / 268

Analysis

- searching with regular expressions and complex graphs

- qualitative (CQP) and quantitative (R packages)

to search a text for simple patterns



using powerful regular expressions, which include all morphosyntactic categories available in dictionaries

applying complex graphs which can describe both morphological and syntax phenomena

Analysis

- searching with regular expressions and complex graphs

- qualitative (CQP) and quantitative (R packages)

- whole lexicon and frequencies display;

- index and frequencies of some patterns expressed by CQL expressions for the search engine;

- concordance display of CQL patterns;

- reading of text editions;

- progression analysis of some CQL patterns;

- sub-corpus building;

- specificity analysis;

- cooccurrence analysis.

provides tools to build language resources (e-dictionaries and grammars) to use them in advanced searches in texts and in generating concordances

text/corpus analysis environment and graphical client based on CQP and R

Thank you!

Acknowledgements
COST Action 16204 – Distant Reading for European Literary History
STSM-CA16204-42562
Host institution: Institut d'Histoire des Représentations et des Idées dans les Modernités UMR 5317, École Normale Supérieure de Lyon, Lyon, France