



Kako je projekat CESAR doprineo obradi srpskog jezika?

Duško Vitas

Matematički fakultet, Univerzitet u Beogradu
vitas@matf.bg.ac.rs

Dan jezičkih tehnologija
Beograd, Hotel Hyatt, 29. oktobar 2012.



Co-funded by the 7th Framework Programme of the European Commission through the contract T4ME, grant agreement no.: 249119.



Co-funded by the ICT PSP Programme of the European Commission through the contract CESAR, grant agreement no.: 271022.



Pregled

- 
- Promocija JT
 - Bela knjiga
 - Meta-share
 - Resursi za srpski

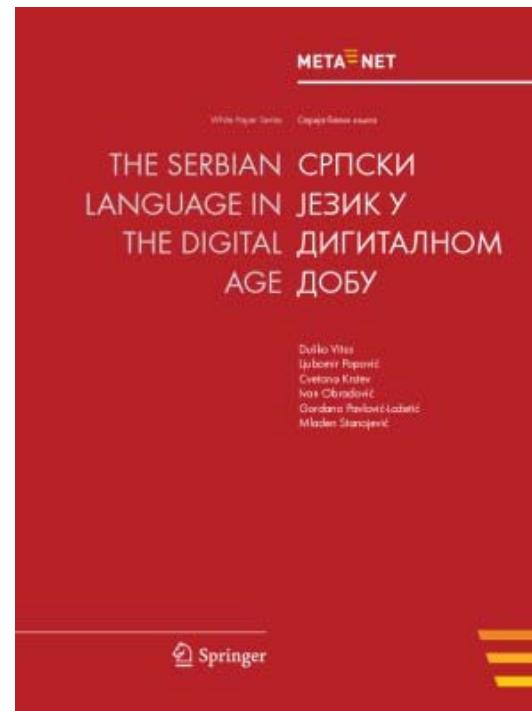


Bela knjiga *Srpski jezik u digitalnom dobu*

Springer, 2012.

<http://www.meta-net.eu/whitepapers/e-book-serbian.pdf>

ISBN: 978-3-642-30754-6



Sadržaj

- Opasnost po naše jezike i izazovi pred jezičkim tehnologijama
- Srpski jezik u evropskom informacionom društvu
- Jezičke tehnologije za srpski jezik

Analiza resursa

| | Q u a n t i t y | A v a b i l i t y | Q u a l i t y | C o v e r a g e | M a t u r i t y | S u s t a i n a b i l i t y | A d a p t a b i l i t y |
|---|--------------------------------------|---|---------------------------------|--------------------------------------|--------------------------------------|--|--|
| Language Technology (Tools, Technologies, Applications) | | | | | | | |
| Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation) | 4 | 3 | 5 | 5 | 5 | 4 | 4 |
| Parsing (shallow or deep syntactic analysis) | 1 | 2 | 5 | 3 | 2 | 2 | 2 |
| Sentence Semantics (WSD, argument structure, semantic roles) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Text Semantics (coreference resolution, context, pragmatics, inference) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Advanced Discourse Processing (text structure, coherence, rhetorical structure / RST, argumentative zoning, argumentation, text patterns, text types etc.) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Information Retrieval (text indexing, multimedia IR, crosslingual IR) | 3 | 1 | 3 | 3 | 2 | 2 | 3 |
| Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics) | 1 | 2 | 2 | 2 | 3 | 2 | 3 |
| Language Generation (sentence generation, report generation, text generation) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Summarization, Question Answering, advanced Information Access Technologies | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| Machine Translation | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| Speech Recognition | 2 | 2 | 1 | 1 | 1 | 1 | 0 |
| Speech Synthesis | 2 | 2 | 4 | 4 | 5 | 5 | 1 |
| Dialogue Management (dialogue capabilities and user modelling) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Language Resources (Resources, Data, Knowledge Bases) | | | | | | | |
| Reference Corpora | 2 | 4 | 2 | 4 | 4 | 4 | 4 |
| Syntax-Corpora (treebanks, dependency banks) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Semantics-Corpora | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Discourse-Corpora | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Parallel Corpora, Translation Memories | 3 | 3 | 3 | 2 | 2 | 2 | 3 |
| Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data) | 1 | 2 | 4 | 4 | 3 | 3 | 3 |
| Multimedia and multimodal data (text data combined with audio/video) | 1 | 2 | 2 | 1 | 2 | 1 | 2 |
| Language Models | 1 | 3 | 2 | 3 | 2 | 2 | 3 |
| Lexicons, Terminologies | 2 | 3 | 4 | 4 | 3 | 3 | 3 |
| Grammars | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| Thesauri, WordNets | 2 | 4 | 3 | 2 | 4 | 2 | 4 |
| Ontological Resources for World Knowledge (e.g. upper models, Linked Data) | 1 | 1 | 0 | 1 | 0 | 1 | 1 |

Grupisanje resursa i alata

Na sastanku u Berlinu je, polazeći od prethodnih tabela, izvršeno poređenje opisa resursa za 30 evropskih jezika. Prepoznaju se četiri osnovne linije razvoja jezičkih tehnologija:

- **Obrada govora**
- **Automatsko prevodenje**
- **Gramatička analiza**
- **Resursi za obradu govora i teksta**

Analiza posle poređenja

| | Квантитет | Доступност | Квалитет | Покривеност | Зрелост | Одрживост | Прилагодљивост |
|---|-----------|------------|----------|-------------|---------|-----------|----------------|
| Језичке технологије (алати, технологије, апликације) | | | | | | | |
| Препознавање говора | 2 | 2 | 1 | 1 | 1 | 1 | 0 |
| Синтеза говора | 2 | 2 | 4 | 4 | 5 | 5 | 1 |
| Граматичка анализа | 1 | 1 | 2,5 | 2 | 2 | 1,5 | 1,5 |
| Семантичка анализа | 1 | 1 | 1 | 1,5 | 1 | 1 | 1,5 |
| Генерирање текста | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Машинско превођење | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| Језички ресурси (ресурси, подаци, базе знања) | | | | | | | |
| Текстуални корпуси | 0,5 | 1 | 0,5 | 1 | 1 | 1 | 0,5 |
| Говорни корпуси | 1 | 2 | 4 | 4 | 3 | 3 | 3 |
| Паралелни корпуси | 3 | 3 | 3 | 2 | 2 | 2 | 3 |
| Лексички ресурси | 1 | 2 | 2 | 2 | 2 | 2 | 2,5 |
| Граматике | 1 | 1 | 0 | 1 | 0 | 1 | 1 |

Табела 5: Стане језичких технологија за српски језик

Resursi

| | | | |
|----------|---|--|---|
| енглески | немачки француски холандски шведски чешки мађарски пољски италијански шпански | баскијски бугарски дански естонски фински галицијски грчки кatalонски хрватски норвешки португалски румунски српски словачки словеначки | ирски исландски летонски литвански малтешки |
|----------|---|--|---|

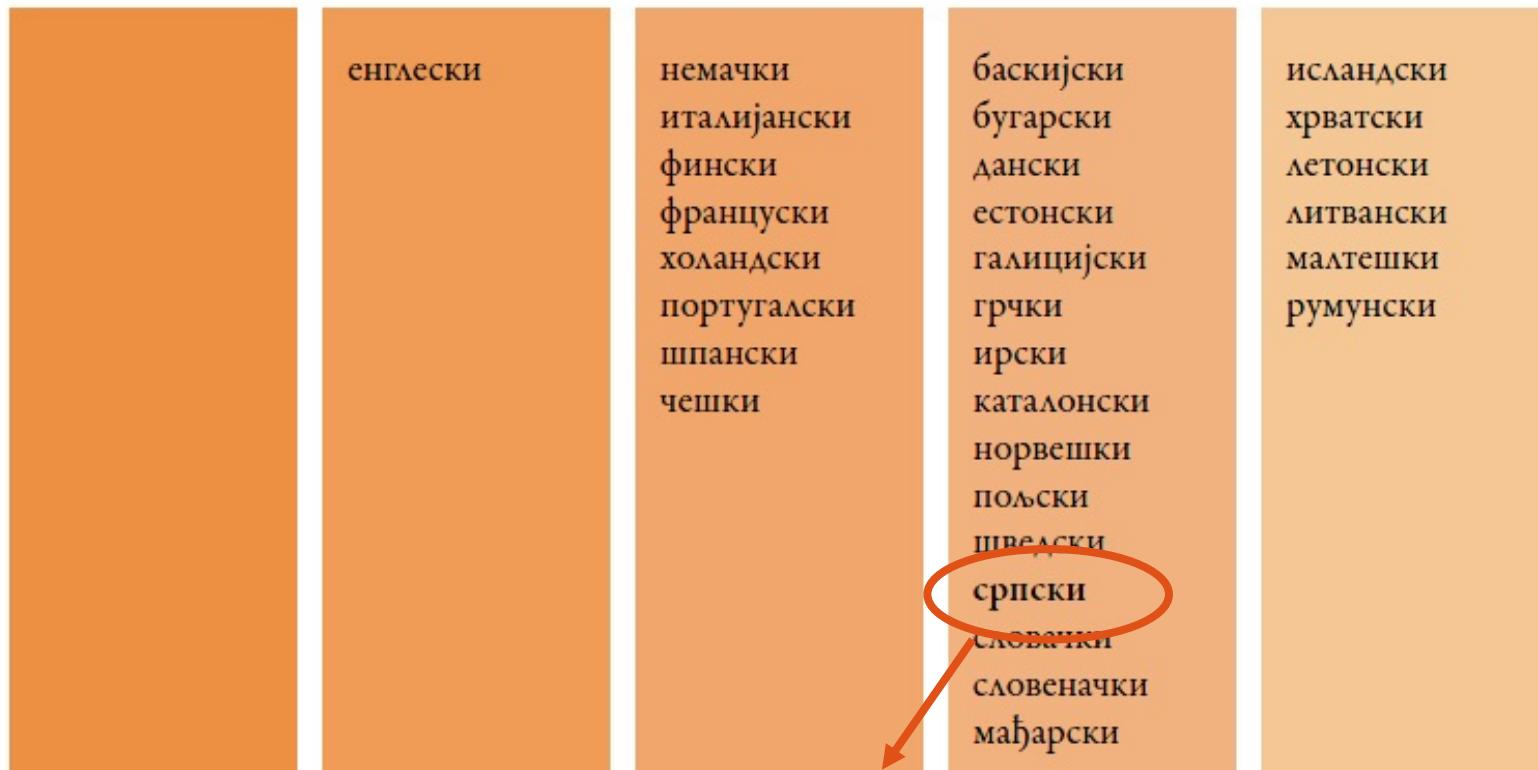
odlična, dobra, umerena, fragmentarna, slaba podrška jeziku

Gramatička analiza



odlična, dobra, umerena, fragmentarna, slaba podrška jeziku

Prepoznavanje i analiza govora



odlična, dobra, umerena, fragmentarna, slaba podrška jeziku

Automatsko prevodenje

| | | | |
|----------|----------------------|---|---|
| енглески | француски шпански | немачки италијански каталонски холандски пољски румунски мађарски | баскијски бугарски дански естонски фински галицијски грчки ирски исландски хрватски летонски литвански малтешки норвешки португалски шведски српски словатски словеначки чешки |
|----------|----------------------|---|---|

odlična, dobra, umerena, fragmentarna, slaba podrška jeziku

Meta-razmena



META-SHARE Welcome to META-SHARE!

META-SHARE is developed within the META-NET Network of Excellence

About the project

META-NET is designing and implementing META-SHARE, a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources. Data and tools can be both open and with restricted access rights, free and for-a-fee. META-SHARE targets existing but also new and emerging language data, tools and systems required for building and evaluating new technologies, products and services.



About the partners

META-SHARE will start by integrating nodes and centres represented by the partners of the META-NET consortium. It will gradually be extended to encompass additional nodes/centres and provide more functionality with the goal of turning into an as largely distributed infrastructure as possible.

Select network node

Please select one of the following META-SHARE network nodes to proceed:

META-SHARE Managing Nodes



CNR — National Research Council of Italy



DFKI — Deutsches Forschungszentrum für Künstliche Intelligenz



ELDA — Evaluations and Language resources Distribution Agency



FBK — Fondazione Bruno Kessler



ILSP — Institute for Language and Speech Processing

Other META-SHARE Nodes



CESAR — Central and South-East European Resources



META-NORD — Baltic and Nordic Parts of the European Open Linguistic Infrastructure

This is the first prototype version of META-SHARE. © META-NET 2010, some rights reserved.
Except where stated otherwise, this website is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License](#).
Co-funded by the 7th Framework Programme of the European Commission through the grant agreement no. 249119.



Meta-razmena na CESAR-serveru

The screenshot shows the META-SHARE homepage. At the top, the META-SHARE logo is displayed. Below it, a search bar contains the word "serbian". To the right of the search bar is a yellow "Search" button. Above the search bar, a message reads "127 language resources at your disposal". A red oval highlights the "serbian" search term in the bar. Another red oval highlights the message above the search bar. On the left side of the page, there is a decorative graphic of overlapping colored shapes (orange, green, blue) resembling a brain or a network. On the right side, there is a section titled "What is it? - About the project" which contains text about the META-SHARE project's purpose and implementation.

127 language resources at your disposal

serbian

Search

What is it? - About the project

META-SHARE, the open language resource exchange facility, is devoted to the sustainable sharing and dissemination of language resources (LRs) and aims at increasing access to such resources in a global scale.

META-SHARE is an open, integrated, secure and interoperable sharing and exchange facility for LRs (datasets and tools) for the Human Language Technologies domain and other applicative domains where language plays a critical role.

META-SHARE is implemented in the framework of the [META-NET Network of Excellence](#). It is designed as a network of distributed repositories of LRs, including language data and basic language processing tools (e.g., morphological analysers, PoS taggers, speech recognisers, etc.).

Meta-razmena

The screenshot shows a search interface with an orange header bar. Below it is a search bar containing 'serbian' and a 'Search' button. To the left is a sidebar titled 'Filter by:' with various options like Language, Resource Type, Media Type, etc. A red circle highlights the title '13 Language Resources'. The main area displays a list of resources, each with a thumbnail, title, and download/eye count. One resource is highlighted with a yellow box: 'Multilingual Edition of Verne's Novel "Around the World in 80 Days"' which includes links to various languages including Albanian, Bulgarian, Chinese, Croatian, Dutch, English, French, German, Greek, Hungarian, Italian, Macedonian, Polish, Portuguese, Serbian, Slovak, Slovenian, and Spanish.

| Resource Title | Language Options |
|---|--|
| Bibliša: Aligned Collection Search Tool | English, Serbian |
| Corpus of Contemporary Serbian | Serbian |
| Corpus of Contemporary Serbian Newspapers and Magazines | Serbian |
| English-Serbian Aligned Corpus | English, Serbian |
| French-Serbian Aligned Corpus | French, Serbian |
| Morphosyntactically tagged Serbian version of Jules Verne's novel "Around the world in 80 days" | Serbian |
| Multilingual Edition of Verne's Novel "Around the World in 80 Days" | Albanian, Bulgarian, Chinese, Croatian, Dutch, English, French, German, Greek, Hungarian, Italian, Macedonian, Polish, Portuguese, Serbian, Slovak, Slovenian, Spanish |

Devet resursa
dolazi od Grupe za
jezičke tehnologije
na Matematičkom
fakultetu

Poređenje sa bolje predstavljenim jezicima

META-SHARE

The screenshot shows the META-SHARE search interface. At the top, there is a search bar with the word "french" and a "Search" button. Below the search bar is a decorative graphic of people connected by lines and bubbles containing various symbols. To the left, a sidebar titled "Filter by:" lists numerous categories such as Language, Resource Type, Media Type, Availability, Licence, Restrictions of Use, Validated, Foreseen Use, Use Is NLP Specific, Linguality Type, Multilinguality Type, Modality Type, MIME Type, Conformance to Standards/Best Practices, Domain, Geographic Coverage, Time Coverage, and Language Variety. A red circle highlights the "422 Language Resources (Page 1 of 22)" heading. The main content area displays a list of resources, each with a thumbnail, title, language(s), download count, and view count. The first few items are:

- ACCOR - English (English) - 0 downloads, 11 views
- Amaryllis Corpus - Evaluation Package (French) - 0 downloads, 2 views
- ANITA (Audio eNhanement in Telecom Applications) (English, French, German, Spanish) - 0 downloads, 1 view
- Arabic Morphological Dictionary (Arabic) - 0 downloads, 1 view
- ARCADE II Evaluation Package (Arabic, Chinese, English, French, German, Greek, Italian, Japanese, Persian, Russian, Spanish) - 0 downloads, 3 views
- ARCADE/ROMANEVAL corpus (English, French, Italian) - 0 downloads, 1 view
- A "scientific" corpus of modern French ("La Recherche" magazine) - Complete version (French) - 0 downloads, 1 view
- A "scientific" corpus of modern French ("La Recherche" magazine) - Raw data (French) - 0 downloads, 1 view

ELDA

Meta-podaci

Serbian Lemmatized and PoS Annotated Corpus 

SrpLemKor

<http://www.kor...>

ID: 603

The Serbian Lemmatized and PoS Annotated Corpus consists of a sample of various texts from SrpKor. It is lemmatized and PoS tagged using TreeTagger. It consists of: daily news published in newspaper "Politika" in december 2009 (1,002,739 words), newspaper feuilletons (1,010,676) published in... [Read More](#)

[« Back](#) [Download](#) [Edit Resource](#)

| | | |
|---|--|--|
| Distribution | Monolingual text corpus | Resource Creation |
| Availability | Languages | Resource Creator |
| Available - Unrestricted Use | Serbian (3,763,352 Words) | Miloš Utvić  |
| Start date: 12/01/2011 | Language Script: Latin | Duško Vitas  |
| License | Variety: Ekavian (Type: Dialect) (3,763,352 Words) | Creation started: 12/01/2011 |
| <i>CC - BY - NC</i> | Linguality | Funding Project |
| Restrictions: Academic - Non Commercial | Linguality type: Monolingual | Serbian Language and its Resources: Theory, Description and Applications |
| Use | Size | Funding Type: National Funds |
| Fee: no price | 3,763,352 Words | Funder: Serbian Ministry of Education and Science |
| Download location: <i>hidden</i> | Character encoding | Funding Country: Serbia |
| Distribution Access/Medium: | ISO - 8859 - 1 (3,763,352 Words) | Project duration: 01/01/2011 - 12/31/2015 |
| Downloadable | Domains | Metadata |
| Execution location: <i>hidden</i> | science (773,119 Words) | Created: 11/17/2011 |
| Licensors: | literature (869,445 Words) | Last Updated: 11/17/2011 |
| Duško Vitas  | law_politics (107,373 Words) | Source: CESAR |
| Distribution rights holders: | general (2,013,415 Words) | Metadata Creator |
| Duško Vitas  | Moduline  | Pavle Mihaljević  |
| IPR Holder | | |
| Duško Vitas  | | |
| Contact Person | | |

Pristup



Firefox Multilingual Edition of Verne'... SrpLemKor
www.korpus.matf.bg.ac.rs/SrpLemKor/

META CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

Resource Name:
Serbian Lemmatized and PoS Annotated Corpus

Resource Short Name:
SrpLemKor

Resource Description:
[SrpLemKor_2011_11.pdf](#)

Tagset:
[tagset.html](#)

Contact Person:
Miloš Utvić

Contact Email:
misko@matf.bg.ac.rs

Contact Person:
Duško Vitas

Contact Email:
vitas@matf.bg.ac.rs

CONDITIONS OF USE

Serbian Lemmatized and PoS Annotated Corpus (SrpLemKor) is distributed under the terms of the [CC BY-NC licence](#).

In your publications presenting the results obtained by using Serbian Lemmatized and PoS Annotated Corpus you should make attribution to

1. Utvić, M. ANNOTATING THE CORPUS OF CONTEMPORARY SERBIAN. *INFOtheica* 12, 2 (December 2011), 36a-47a;
2. Popović, Z. TAGGERS APPLIED ON TEXTS IN SERBIAN. *INFOtheica* 11, 2 (December 2010), 21a-38a.

You must accept the terms of [this license](#) to download and use this resource. Contact person(s) will send you the username and password required to access the resource.

DO YOU ACCEPT THE TERMS OF THIS LICENCE?

Yes, I accept the terms of this license No, I don't accept the terms of this license

Username
Password

Naredne isporuke

- Proširenja korpusa
- Geološka terminološka baza
- Proširenje Vernovog korpusa (svi balkanski, svi slovenski, svi romanski jezici)
- Moduli za imenovane entitete sa evaluacionim korpusom
- Klasifikacija stranica sa weba prema senzibilitetu
- Web-alat za poravnote tekstove
- ...



I još nekoliko primera!

Primer poravnatog teksta *(Put oko sveta za 80 dana)*

FR n569 : Vous savez que cette formalité du visa est inutile, et que nous n'exigeons plus la présentation du passeport?

ES n569 : Ya sabéis que la formalidad del visado no es necesaria, y que ya no exigimos la presentación del pasaporte.

PT n569 : Sabe que esta formalidade do visto é inútil, e que já não exigimos a apresentação do passaporte?

RO n569 : Știți că formalitatea vizei e inutilă și că noi nu mai cerem prezentarea pașaportului.

EN n569 : You know that a visa is useless, and that no passport is required?"

DE n569 : Sie wissen, daß diese Förmlichkeit des Visa unnütz ist, und wir verlangen die Ueberreichung des Passes nicht mehr?

RU n569 : Вам известно, что формальность с визой необязательна и мы не требуем больше предъявления паспорта?

PL n569 : Wie pan zapewne, że formalności wizowe nie są już konieczne i że nie wymagamy okazywania paszportu?

SR n569 : Vi znate da je ova formalnost viziranja izlišna i da se više ne traži pokazivanje isprava?

BG n569 : Знаете ли, че тази формалност с паспортите е безполезна и че ние вече не изискваме да представяте паспортите си?

GR n569 : Ξέρετε ότι αυτή η τυπική διαδικασία της βίζας δεν είναι αναγκαία και δεν απαιτείται πλέον η εμφάνιση του διαβατηρίου;

Primer eksplotacije poravnatog teksta

<A+Col>

2554 Puis, des frères quêteurs, des pèlerins en longues robes, de simples civils, chevelure lisse et d'un noir d'ébène, tête grosse, buste long, jambes grêles, taille peu élevée, teint coloré depuis les sombres nuances du cuivre jusqu'au blanc mat, mais jamais jaune comme celui des Chinois, dont les Japonais diffèrent essentiellement.

2662 Un autre, avec la fumée odorante de sa pipe, traçait rapidement dans l'air une série de mots bleuâtres, qui formaient un compliment à l'adresse de l'assemblée.

2827 Plus de sombreros, plus de chemises rouges à la mode des coureurs de placers, plus d'Indiens emplumés, mais des chapeaux de soie et des habits noirs, que portaient un grand nombre de gentlemen doués d'une activité dévorante.

2835 Mr. Fogg et Mrs. Aouda s'installèrent devant une table et furent abondamment servis dans des plats lilliputiens par des Nègres du plus beau noir.

2874 Le nombre des chapeaux noirs diminuait à vue d'œil, et la plupart semblaient avoir perdu de leur hauteur normale.

2901 Un énorme gaillard à barbiche rouge, au teint coloré, large d'épaules, qui paraissait être le chef de la bande, leva son formidable poing sur Mr.

Zatim kaludxeri koji mole za milostinxu, hodocynici u dugim halxinama, obicyni gradxani sa zalizanom kosom crnom kao abonos, velikom glavom, dugulxastim licem, tankih nogu, malog rasta, boje kozxe od bakarnosmedje do svetlosmedje, ali nikad zxute kao u Kineza, od kojih se Japanci bitno razlikuju.

2565 Drugi jedan, mirisavim dimom svoje lule brzo je pisao u vazduhu niz, blavicyastih slova koja su izrazzavala pohvalu publici.

2673 Nema viske sombrera, ni crvenih kosxulka koje su nosili kopacvi zlata, ni perjem, a i crnih uzxurbani Gospodin sedosxe z najlepsx posluzzxis tanxirime Crnih sxe a vecxine da je iz

crn - noir
bakarnosmeđ –
sombres nuances de cui
svetlosmeđ – blanc mat
žut - jaune

All sentences/Plain text

Matched sentences

All sentences/HTML

Aligned with target concordance

Aligned with source concordance

Locate... Clear alignment Align Save alignment Save alignment as... Locate...

upit
Iubav

semantičko proširenje koristeći relacije u Wordnetu

morfološko proširenje - e-rečnik srpskog

proširenje preko Wordneta za engleski

Lema

Search: Iubav is lemma

Use "+" for "OR" (example job+luck)

With inflection Unitex

Semantical extension WordNet

+Aurora +Latin +Cyrilic

+Another language extension

List of forms to search (Regular and XPath expression)

WordNet I WordNet II Prolex I Prolex II

Table Text XML View Literals

Include SS in relation: From (ILR) Relation Depth: 1

near_antonym To (RILR)

List literals of the selected and related synsets Inflect listed literals

Literal Inflected forms

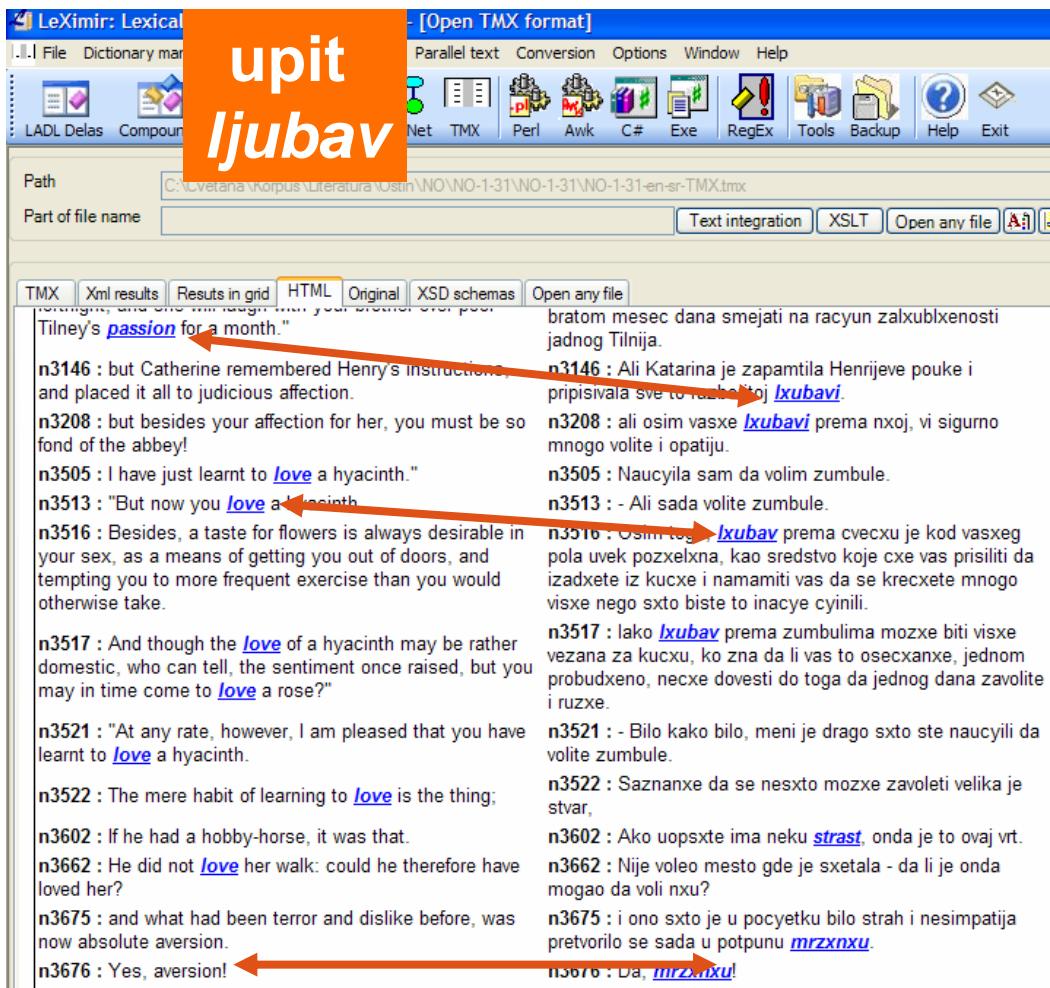
- beloved
- dear
- dearest
- loved one

volkeni|volkenaldragan|draganaldraganu|dragane|draganom|draganildragan|draganol|draganama|xubavima|xubavbu|xubavi|xubav|cakan|cakanamalcakan|cakanol|cakanul|cakanilca|kanel|cakanajstrastima|strastxci|strasti|strast|mrzxna|mrxnxel|mrxnxim|rznxu|mrxnxu|mrxnxo|mrxnxamalhepriateljstvo|voljeni|voljenal|jubavimal|jubavljul|jubavil|jubav|strašću|mrnja|mrnje

beloved|dear|dearest|loved one|honey|love|passion|hate|hated

Rezultat

upit
Ijubav



The screenshot shows the LeXimir Lexical tool interface. The main window displays a list of search results for the term 'Ijubav'. Several instances of the word are highlighted in blue, indicating they are matches for the search query. Red arrows point from the search term 'Ijubav' at the top left to specific matches in the results list.

Pronađen je ključ (*Ijubav*), ali i sinonim (*strast*) i antonim (*mržnja*)

Za kraj

