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.

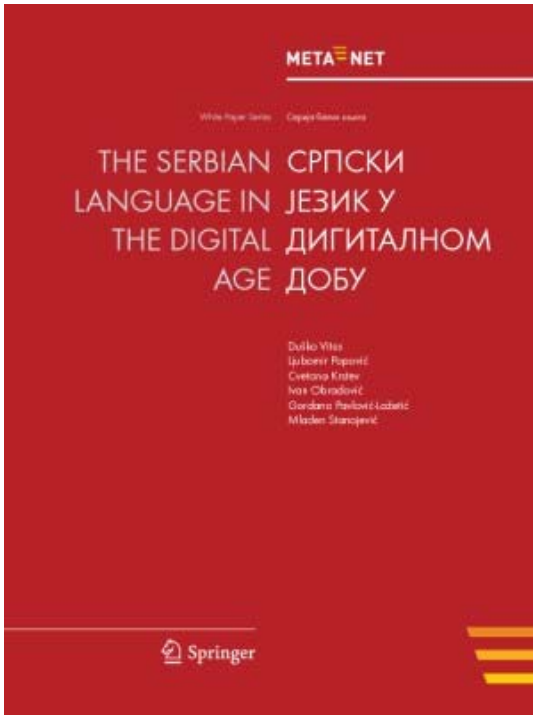
Co-funded by the 7th Framework Programme of the European Commission through the contract T4ME, grant agreement no.: 249119.

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

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
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 prevođenje**
- **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
Генерисање 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: Стање језичких технологија за српски језик

Resursi

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

odlična, dobra, umerena, fragmentarna, slaba podrška jeziku

Gramatička analiza

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

odlična, dobra, umerena, fragmentarna, slaba podrška jeziku

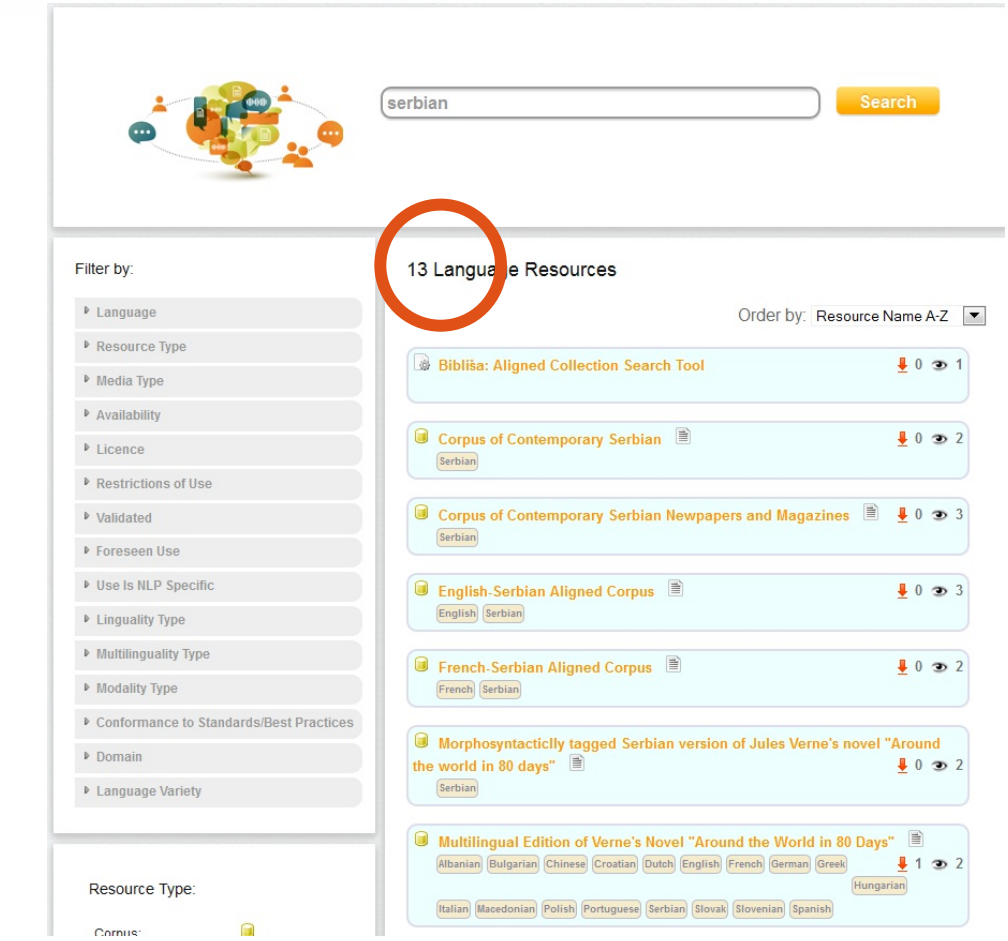
Prepoznavanje i analiza govora

Automatsko prevođenje

Meta-razmena

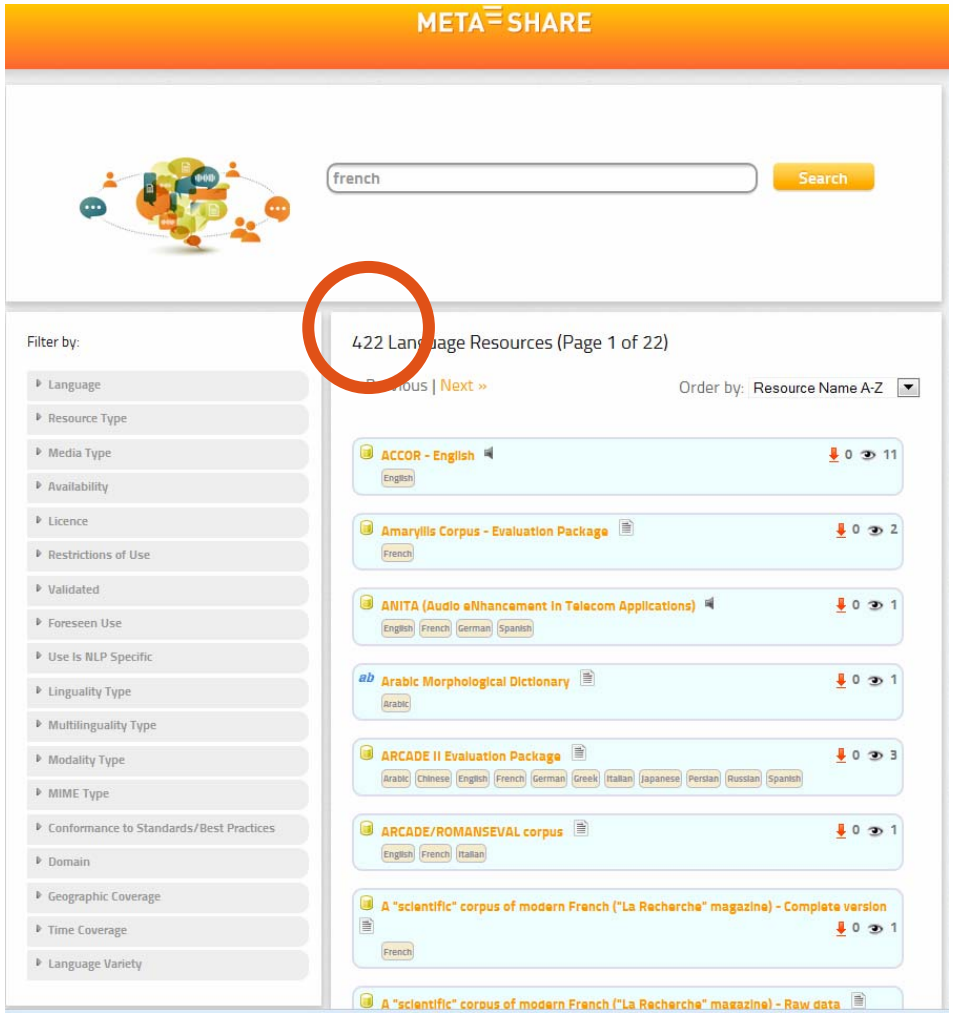
Meta-razmena na CESAR-serveru

Meta-razmena



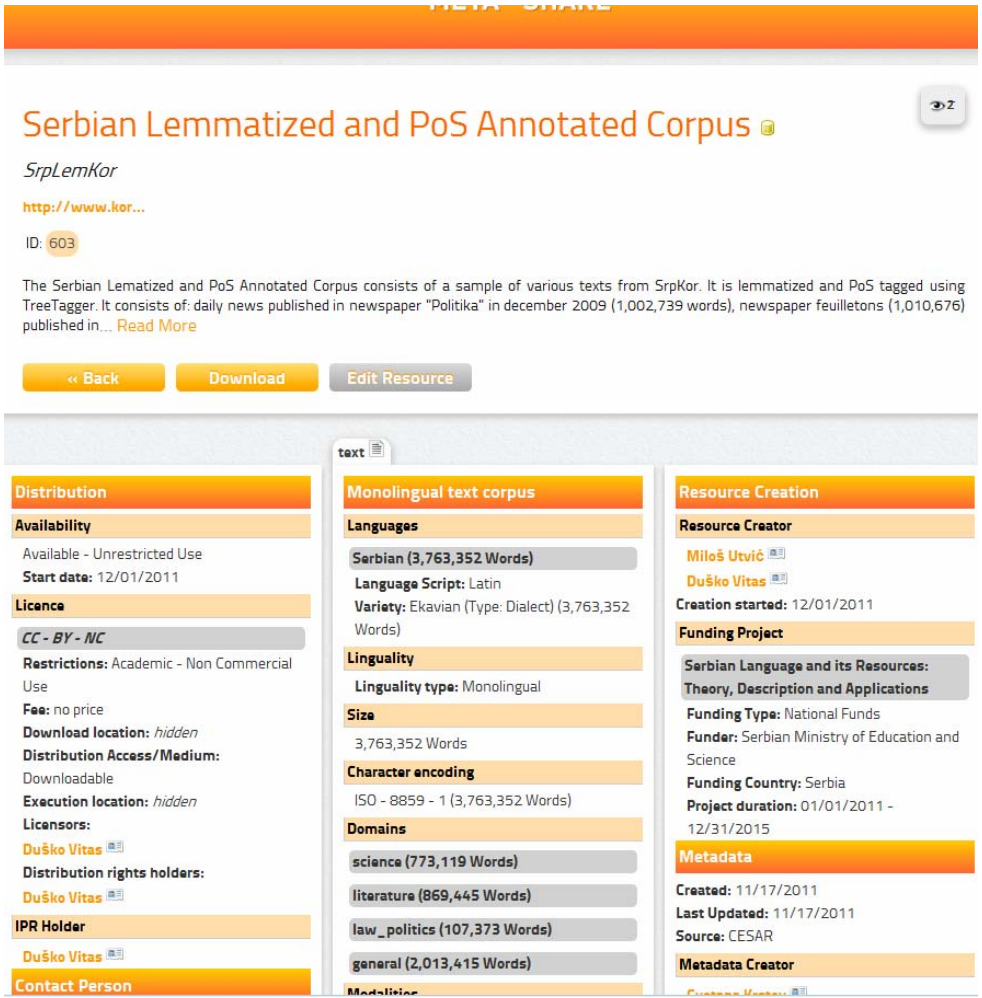
Devet resursa dolazi od Grupe za jezičke tehnologije na Matematičkom fakultetu

Poređenje sa bolje predstavljenim jezicima



ELDA

Meta-podaci



Serbian Lemmatized and PoS Annotated Corpus

SrpLemKor

http://www.kor...

ID: 603

The Serbian Lematized 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

text

Distribution

Availability

Available - Unrestricted Use

Start date: 12/01/2011

Licence

CC - BY - NC

Restrictions: Academic - Non Commercial Use

Fee: no price

Download location: *hidden*

Distribution Access/Medium: Downloadable

Execution location: *hidden*

Licensors:

Duško Vitas

Distribution rights holders:

Duško Vitas

IPR Holder

Duško Vitas

Contact Person

Monolingual text corpus

Languages

Serbian (3,763,352 Words)

Language Script: Latin

Variety: Ekavian (Type: Dialect) (3,763,352 Words)

Linguality

Linguality type: Monolingual

Size

3,763,352 Words

Character encoding

ISO - 8859 - 1 (3,763,352 Words)

Domains

science (773,119 Words)

literature (869,445 Words)

law_politics (107,373 Words)

general (2,013,415 Words)

Resource Creation

Resource Creator

Miloš Utvić

Duško Vitas

Creation started: 12/01/2011

Funding Project

Serbian Language and its Resources: Theory, Description and Applications

Funding Type: National Funds

Funder: Serbian Ministry of Education and Science

Funding Country: Serbia

Project duration: 01/01/2011 - 12/31/2015

Metadata

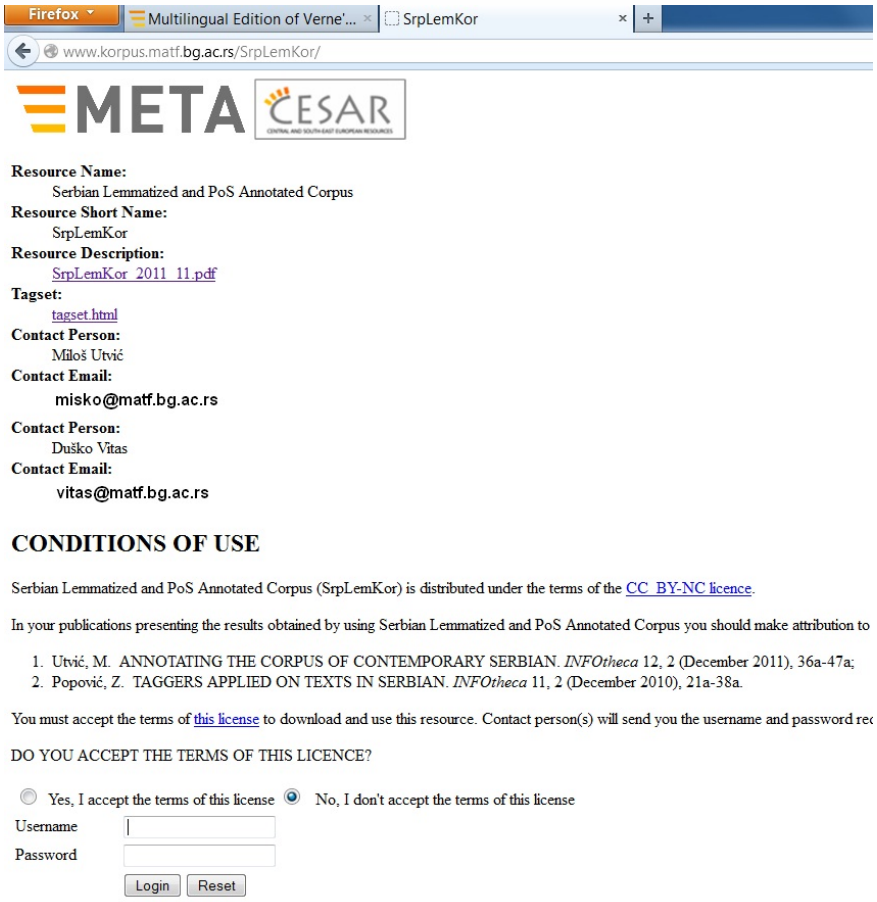
Created: 11/17/2011

Last Updated: 11/17/2011

Source: CESAR

Metadata Creator

Pristup


Firefox
Multilingual Edition of Verne'...
SrpLemKor
www.korpus.matf.bg.ac.rs/SrpLemKor/
META
CESAR
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. INFOtheca 12, 2 (December 2011), 36a-47a;
2. Popović, Z. TAGGERS APPLIED ON TEXTS IN SERBIAN. INFOtheca 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 rec
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
Login
Reset

Naredne isporuke

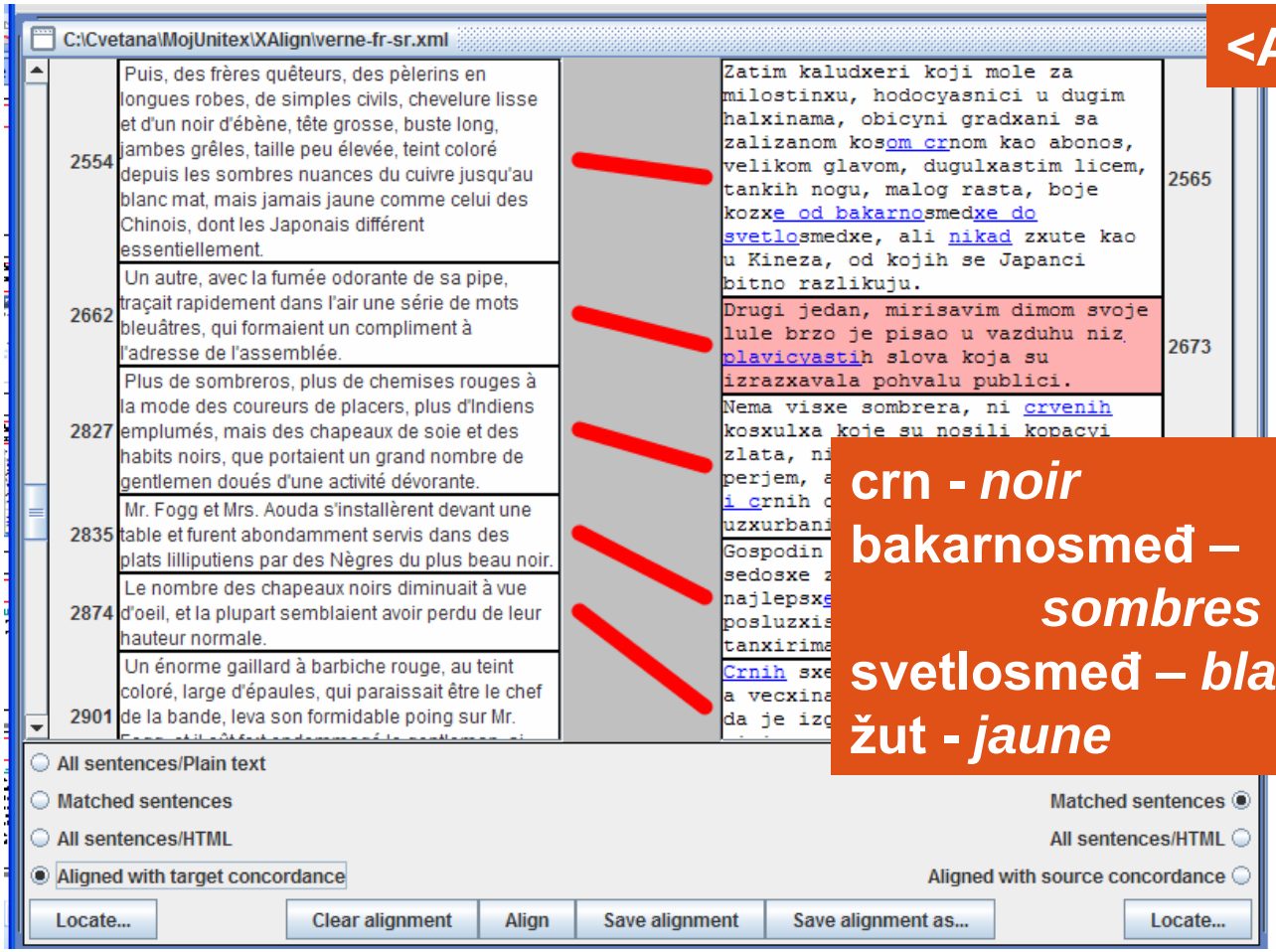
- Proširenja korpusa
- Geološka terminološka baza
- Proširenje Vernovog korpusa (svi balkanski, svi slovenski, svi romanski jezici)
- Moduli za imenovane entitete sa evaluacionim korpusom
- Klasifikacija stranica sa weba prema senzibilitetu
- Web-alat za poravnate tekstove
- ...

I još nekoliko primera!

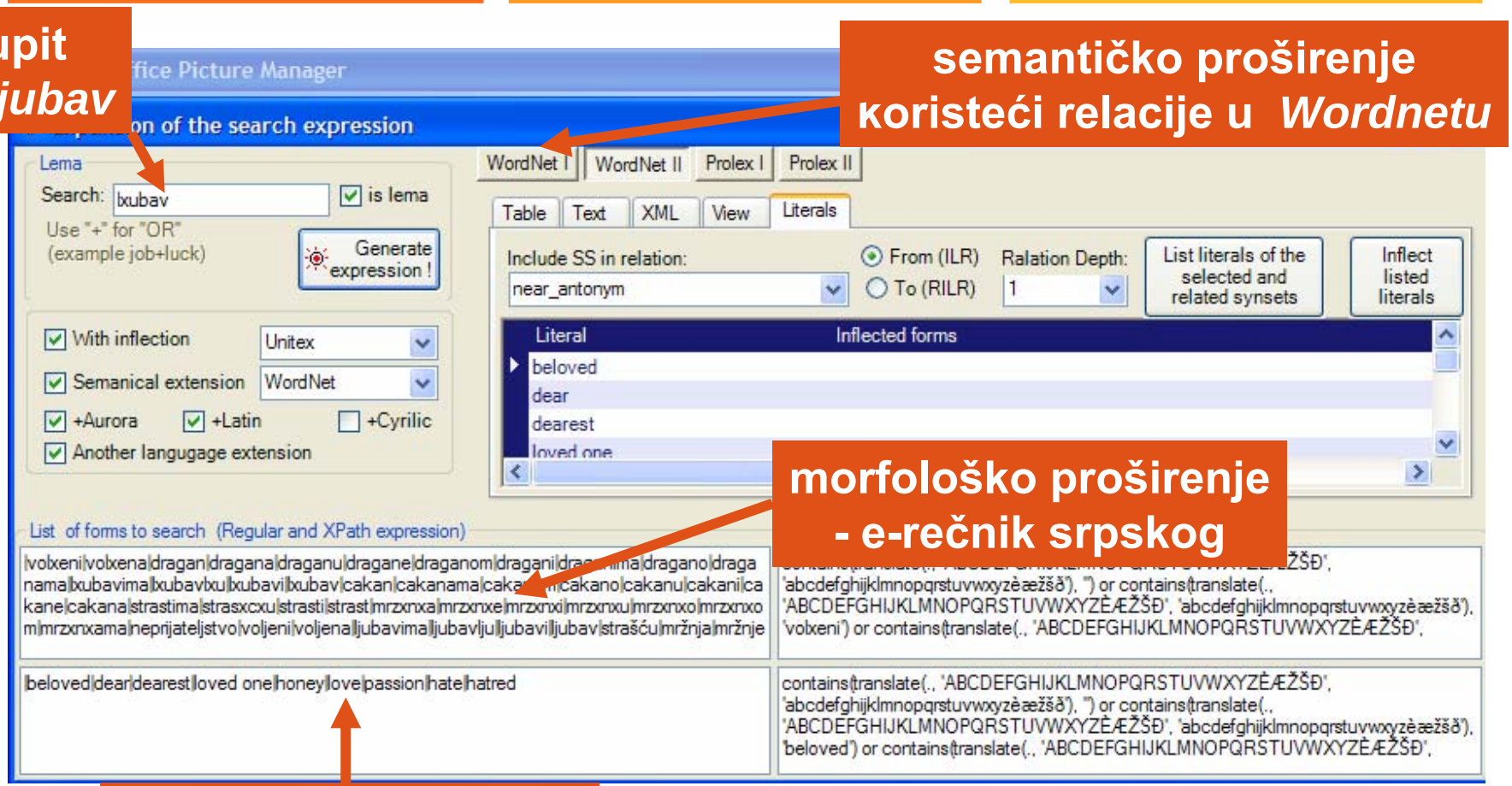
Primer poravnatog teksta (*Put oko sveta za 80 dana*)

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 : Ştiţi cã formalitatea vizei e inutilã şi cã noi nu mai cerem prezentarea paşaportului.
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;

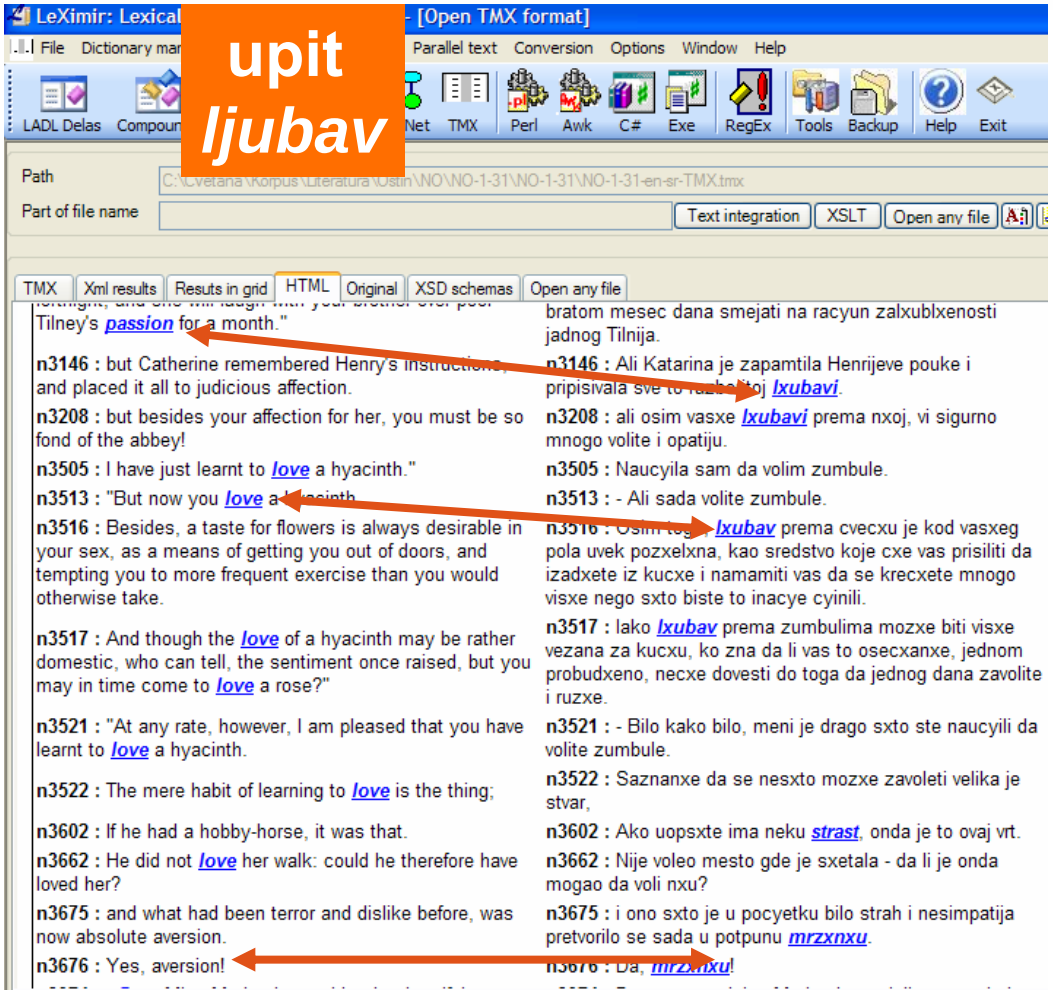
Primer eksploatacije poravnatog teksta



Leximir

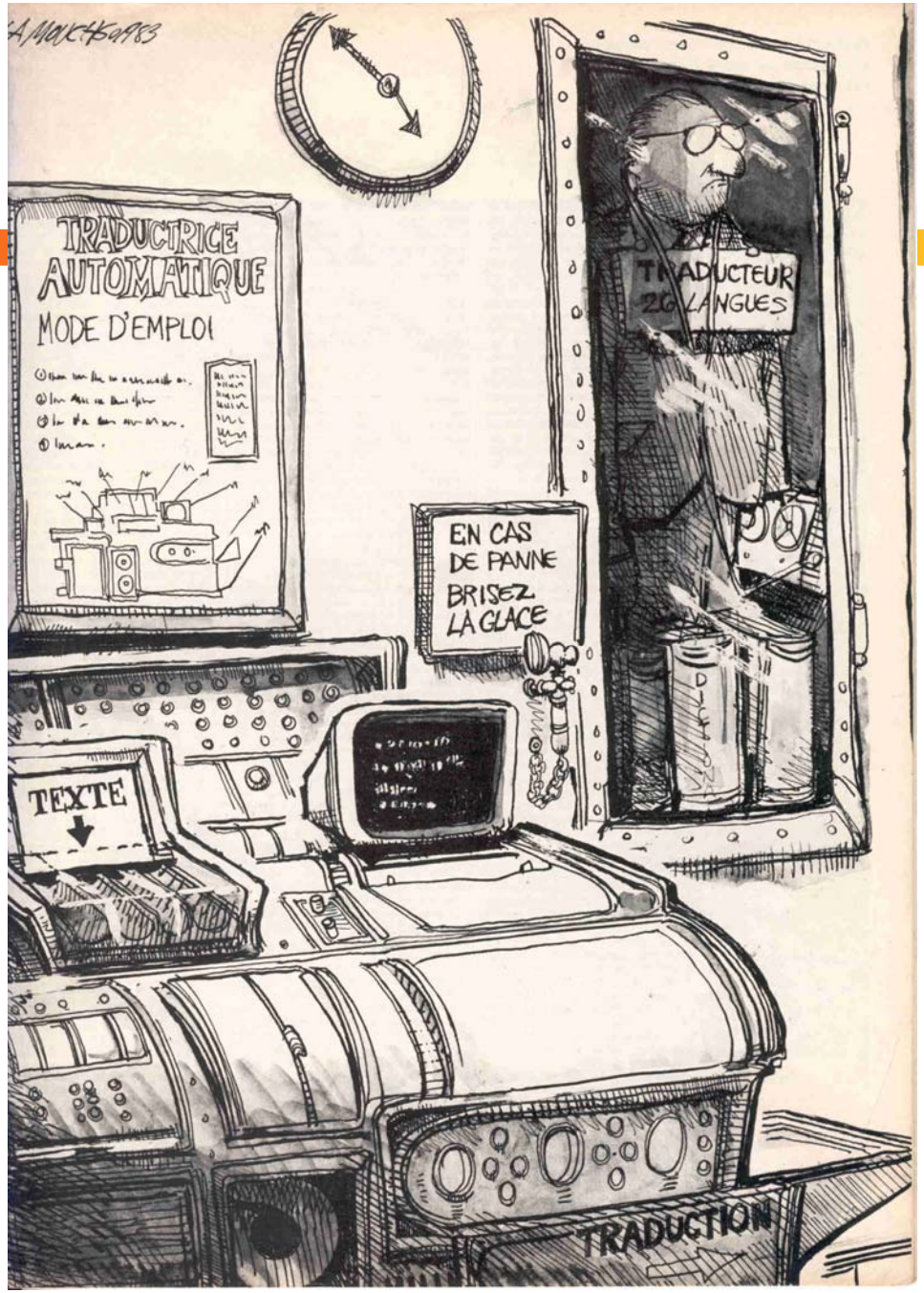

upit
ljubav
semantičko proširenje
koristeći relacije u Wordnetu
Lema
Search:
lxubav
is lema
Use "+" for "OR"
(example job+luck)
Generate expression !
With inflection
Unitex
Semanical extension
WordNet
+Aurora
+Latin
+Cyrilic
Another langugage extension
WordNet I
WordNet II
Prolex I
Prolex II
Table
Text
XML
View
Literals
Include SS in relation:
near_antonym
From (ILR)
To (RILR)
Ralation Depth:
1
List literals of the selected and related synsets
Inflect listed literals
Literal
Inflected forms
beloved
dear
dearest
loved one
morfološko proširenje
- e-rečnik srpskog
List of forms to search (Regular and XPath expression)
|beloved|dear|dearest|loved one|honey|love|passion|hate|hatred
proširenje preko
Wordneta za engleski

Rezultat



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

Za kraj


TRADUCTRICE AUTOMATIQUE
MODE D'EMPLOI
TRADUCTEUR 26 LANGUES
EN CAS DE PANNE BRISEZ LA GLACE
TEXTE
TRADUCTION