CLARA

Building the HINA system for processing newswires

Language technologies view

Marko Tadić

University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics

marko.tadic@ffzg.hr

CLARA Career Course
Dubrovnik
2011-09-22

Overview

- on Croatian language
- state of the art of LT for Croatian (in 2008)
  - available language resources
  - available language tools
- how to shape the project
  - analysis of user needs
  - which existing LT could meet user requests
- adaptation of language resources and tools to the project
- possible future developments

CLARA

On Croatian Language

On language name

- Croatian (ISO 639: hr/hrv) = South Slavic language
  - Slovenian, Bosnian, Serbian, Macedonian, Bulgarian,…
- term “Serbo-Croatian”: Kopitar suggested (1836)
  - in Šafárik’s classification of Slavic languages (1842)
  - influenced by Serbian nationalist philologist Vuk Karadžić, supported by Austrian empire
- compound name used as one of means of political oppression in all Yugoslav states
  - Kingdom of Yugoslavia: Serbo-Croato-Slovenian (!?)
  - communist Yugoslavia: “srpskohrvatski”, “srpsko-hrvatski”, “hrvatsko-srpski”, “hrvatski ili srpski”;
    in fact: none of them representing Croatian only
- Croatian had its own standardisation development until 20th ct
- http://www.ethnologue.com/show_lang_family.asp?code=hrv

Croatian language structure

- morphology: inflectional, fusion language
  - 10 PoS: Nouns, Verbs, Adjectives, Numerals, Pronouns, Adverbs, Conjunctions, Prepositions, Interjections, Particles
  - inflectional morphosyntactic descriptions (MSD)
    - N: 7 cases, 2 numbers, 3 genders (non predictable, lexical info)
    - A: 7 cases, 2 numbers, 3 genders, 2 definite forms, 3 grades
    - V: 3 persons, 2 numbers, 3 simple tenses, 3 complex tenses, 2 simple moods, 2 complex moods, + passive, 2 participles with 3 genders/2 numbers, 2 verbal adverbs
    - …
  - N: 14 word-forms (WF); A: 227 WFs; V: 30 simple WFs; Num: like N, A or Adv; Pro: like A; Adv have comparison…
  - productive derivative system (incl. compounding)

Croatian language structure 2

- syntax: dominantly SVO language
  - relatively free word-order
  - scrambling of higher level constituents/chunks
    - to a considerable extent describable internally with regular grammars (Abney 1996: “islands of certainty”)
    - CF grammars needed for their combinations
  - !but: position of clitics within the 1st phonetic word
    - long-distance dependencies, branch-crossing…
  - complex verbal aspectual system
    - inperfective verbs, perfective verbs, iterative verbs

  no problem for derivative morphology, but for syntax/semantics
- sentence semantics
  - verbal valencies / semantic roles (verbs with 4 slots)
    - e.g. *Ona mu otvara vrata ključem.*
- illustration of the initial linguistic complexity

CLARA

State of the art of LT for Croatian

(in 2008)

Existing resources: corpora

- tradition: Institute of Linguistics, University of Zagreb
  - 1967: 1st hr computer corpus: Bujas
  - 1968-1973: 1st en-hr parallel corpus: Filipović
  - ’70: corpora of Croatian old authors (typical LLC)
  - 1976-1996: 1M Corpus of Croatian Literary Language
    - 1Mw in size, time-span: 1938-1976, 5 genres
    - Moguš-Bratanić-Tadić (1999) *Croatian Frequency Dictionary*
  - 1998-2003: Croatian National Corpus v 1.0: Tadić (LREC2002)
- Croatian National Corpus (HNK) v 2.0, 2004-
  - currently 101.2 million tokens
  - original texts written in standard Croatian produced 1990-
  - 74% faction, 23% fiction, 3% mixed; only prose
  - XML XCES encoding, stored on Manatee server
  - freely accessible using Bonito client
  - corpus web-page: http://hnk.ffzg.hr ➲

Existing resources: corpora 2

- Croatian-English Parallel Corpus
  - Tadić (LREC2000)
- single direction parallel corpus
  - source language: Croatian
  - target language: English
- newspaper corpus
  - *Croatia Weekly* (113 issues)
  - from 1998-01 until 2000-04
- corpus size

	hr	en
articles	4.748	4.748
sentences	74.638	82.898
tokens	1.636.246	1.968.874

Existing resources: corpora 3

- Slovenian-Croatian Parallel Corpus (150 Kw)
- French-Croatian Parallel Corpus (130 Kw)
- Bulgarian-Croatian comparable corpus (3.5 Mw)
  - Bekavac et al. (LREC2004)
- South-East European (Parallel) Corpus
  - based on SETimes portal
    - Albanian, Bosnian, Bulgarian, Croatian, Greek, English, Macedonian, Romanian, Serbian, Turkish
  - crawled since 2007, ca 3 Mw collected per language
- today large mono- or multilingual corpora being built within the EC-funded projects (ACCURAT, LetsMT!, CESAR) or national initiatives (hrWaC, 1.3 billion tokens)
- Croatian translations of Acquis Communautaire
  - obtained from Ministry of foreign affairs, ca 60 Mw

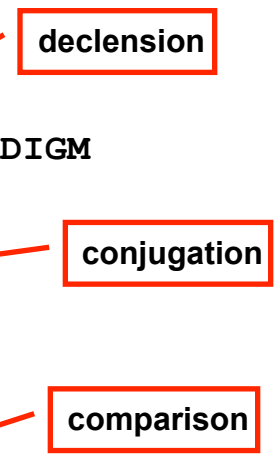
Existing resources: corpora 4

- Croatian Dependency Treebank
  - following Prague Dependency Treebank, adapted for Croatian
  - Tadić (2007)
  - http://hobs.ffzg.hr
- Institute of Croatian language and linguistics: text collection
  - Croatian Language Repository
  - http://riznica.ihjj.hr
- MulText East: set of recommendations
  - for encoding
    - corpora (example: translations of Orwell's *1984*)
    - lexica (example: inflectional lexicons)
    - tagsets: MT(E) compliant tagsets (following EAGLES 1996)
  - Erjavec (LREC2010): v 4.0, 16 languages (incl. hr since 1998)
  - http://nl.ijs.si/MTE/V4

Existing resources: lexica

- Croatian Morphological Lexicon (HML)
  - generated with Croatian Inflectional Generator (Tadić 1992, 1994)
    - model of the Croatian inflection
    - classification based: 614 inflectional paradigms
    - flat model, respecting linguistic units
    - not computationally optimized
    - covers all phenomena in Croatian inflection

LEMMA	STEM	INFLECTIONAL PARADIGM
bacati	bac	0/501/0
baciti	bac	0/511/0
bagatelizirati	bagatelizir	0/501,502/0
bageriranti	bagerir	0/501,502/0
bajati	baj	0/501/0
baktati	bakt	0/501/0
balansirati	balansir	0/501/0
balegati	baleg	0/501/0
baliti	bal	0/509,510/0
balzamirati	balzamir	0/501,502/0



Existing resources: lexica 2

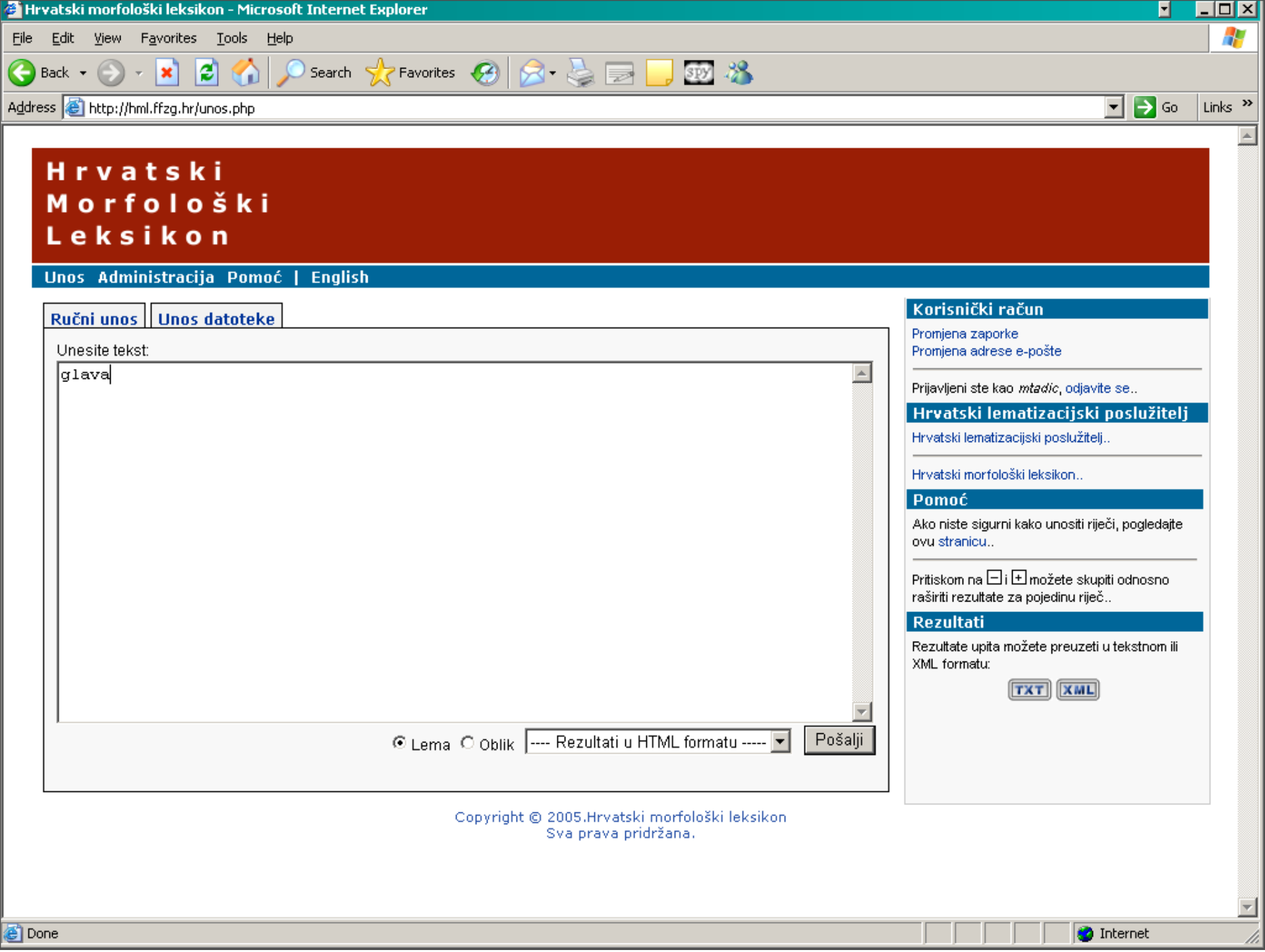
```
= abdikacija Ncfpg
abdikacija abdikacija Ncfsn
abdikacijama abdikacija Ncfpd
abdikacijama abdikacija Ncfpi
abdikacijama abdikacija Ncfpl
abdikacije abdikacija Ncfpa
abdikacije abdikacija Ncfpn
abdikacije abdikacija Ncfpv
abdikacije abdikacija Ncfsg
abdikaciji abdikacija Ncfsd
abdikaciji abdikacija Ncfsl
abdikacijo abdikacija Ncfsv
abdikacijom abdikacija Ncfsi
abdikaciju abdikacija Ncfsa
= abeceda Ncfsn
abecede abeceda Ncfsg
abecedi abeceda Ncfsd
abecedu abeceda Ncfsa
abecedo abeceda Ncfsv
abecedi abeceda Ncfsl
abecedom abeceda Ncfsi
abecede abeceda Ncfpn
abeceda abeceda Ncfpg
abecedama abeceda Ncfpd
abecede abeceda Ncfpa
abecede abeceda Ncfpv
abecedama abeceda Ncfpl
abecedama abeceda Ncfpi
```

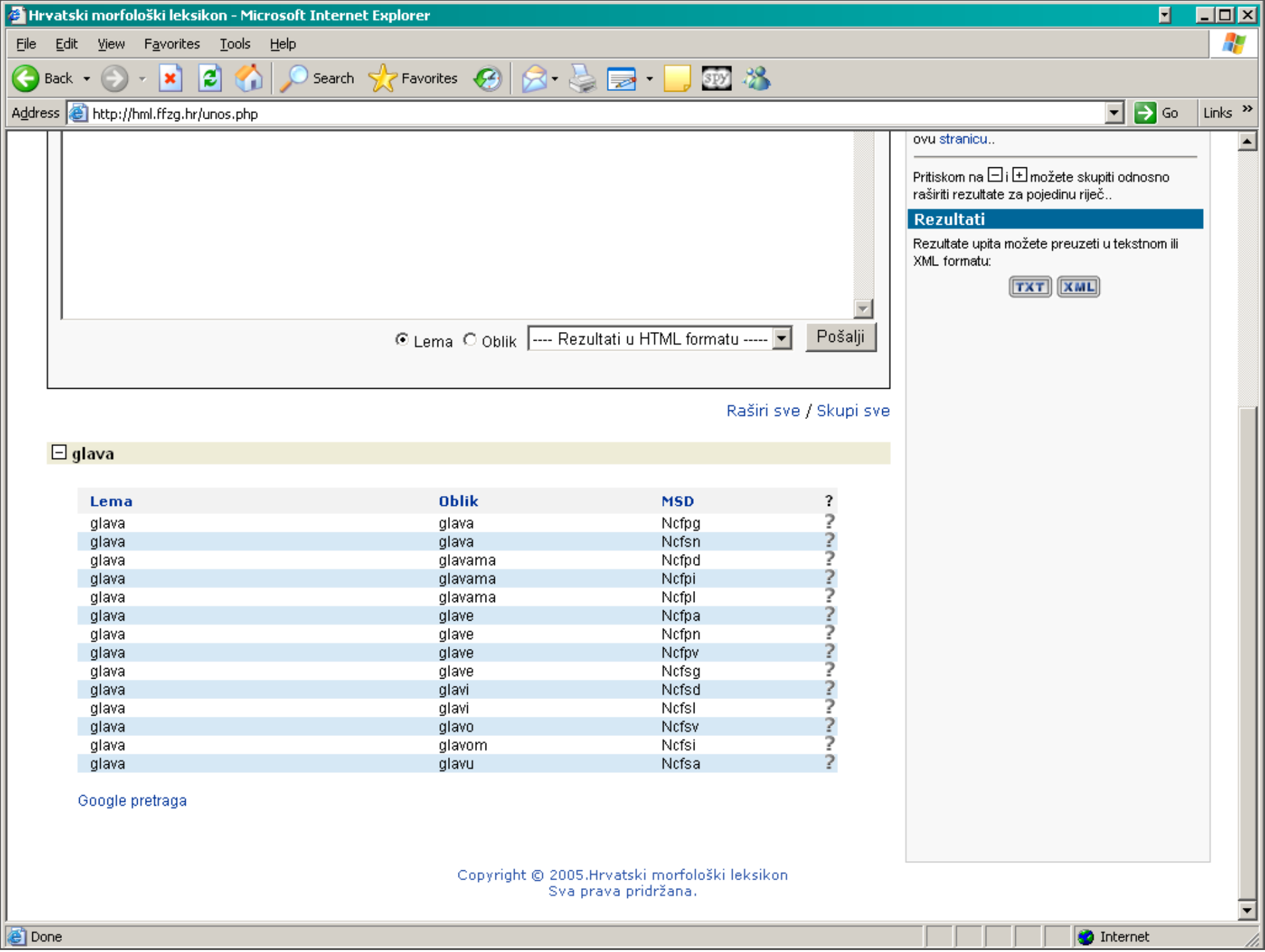
```
= abolicija Ncfsn
abolicije abolicija Ncfsg
aboliciji abolicija Ncfsd
aboliciju abolicija Ncfsa
abolicijo abolicija Ncfsv
aboliciji abolicija Ncfsl
abolicijom abolicija Ncfsi
abolicije abolicija Ncfpn
abolicija abolicija Ncfpg
abolicijama abolicija Ncfpd
abolicije abolicija Ncfpa
abolicije abolicija Ncfpv
abolicijama abolicija Ncfpl
abolicijama abolicija Ncfpi
= abrazija Ncfsn
abrazija abrazija Ncfpg
abrazijama abrazija Ncfpd
abrazijama abrazija Ncfpi
abrazijama abrazija Ncfpl
abrazije abrazija Ncfpa
abrazije abrazija Ncfpn
abrazije abrazija Ncfpv
abrazije abrazija Ncfsg
abraziji abrazija Ncfsd
abraziji abrazija Ncfsl
abrazijo abrazija Ncfsv
abrazijom abrazija Ncfsi
abraziju abrazija Ncfsa
```

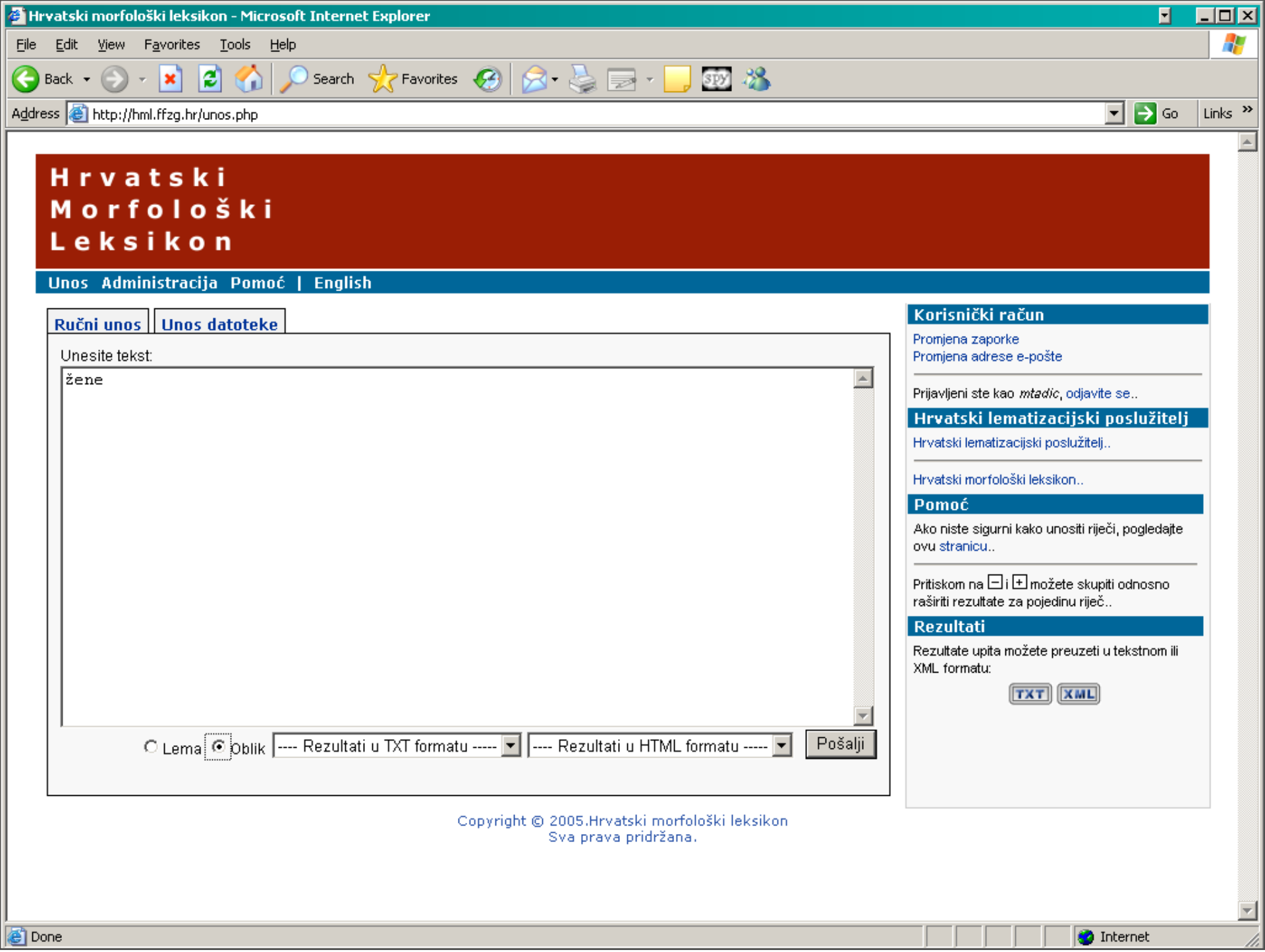
- tagset and lexicon format = MulTextEast compliant according to the specifications for hr

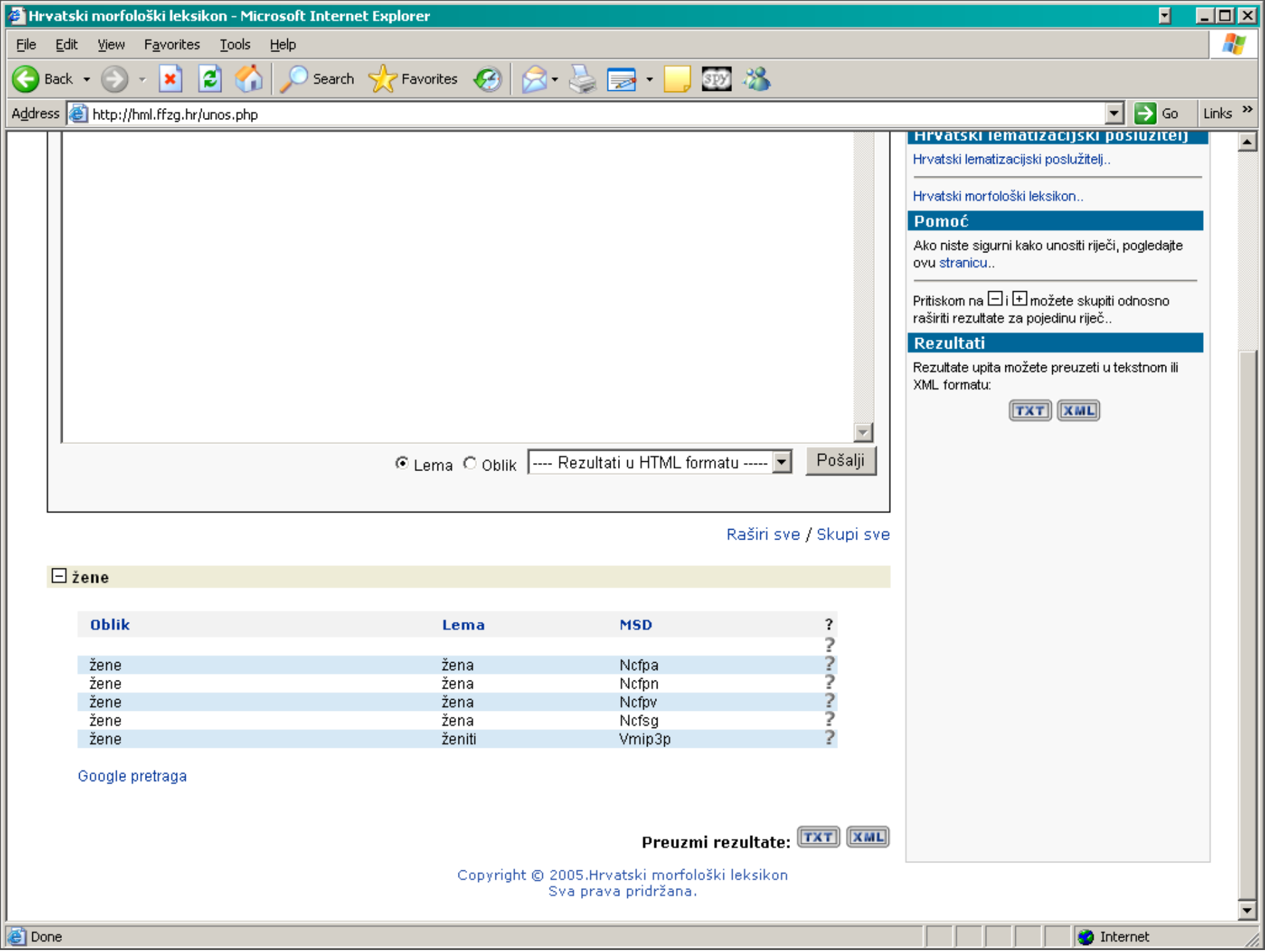
Existing resources: lexica 3

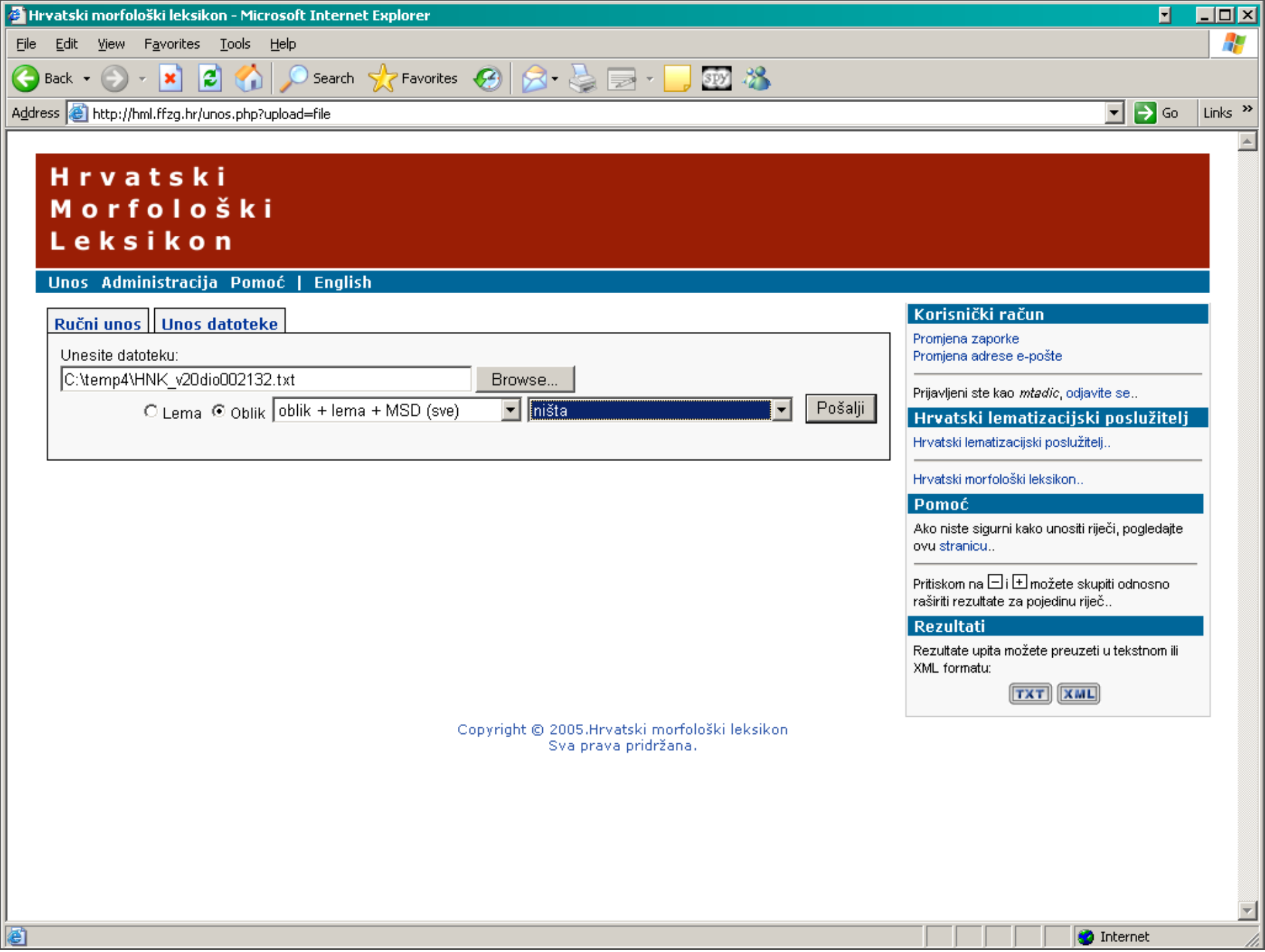
- HML v4.6 = lexicon of WFs stored in a database
  - 45,000+ lemmas of general language
  - 15,000+ lemmas of personal fe/male names (Boras-Mikelić 2003)
  - 50,000+ lemmas of surnames registered in Croatia (ibid.)
  - 4.0+ million of generated WFs
- freely accessible web service: http://hml.ffzg.hr
- input
  - lemma(s) or WF(s) in HTML web interface
  - tokenized (verticalized) text (e.g. XML document)
- output in different formats
  - HTML: web page with all WFs or lemmas of input
  - text file: tok-lem-MSD, tok-lem-PoS, tok-lem, lem
  - !not disambiguated: all possible interpretations of homographs

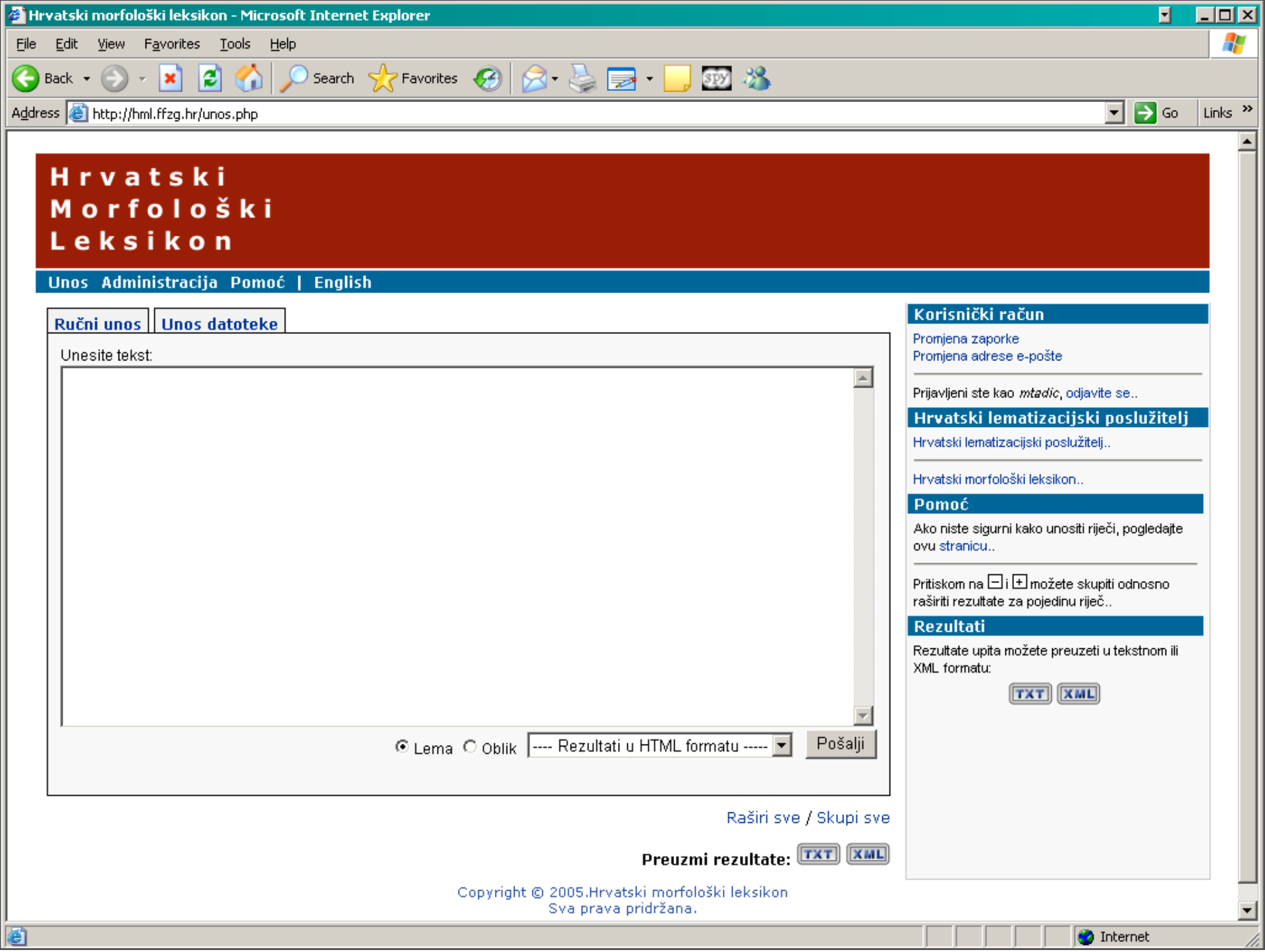

Hrvatski morfološki leksikon - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://hml.ffzg.hr/unos.php
Go Links
Hrvatski Morfološki Leksikon
Unos Administracija Pomoć | English
Ručni unos
Unos datoteke
Unesite tekst:
glava
Lema Oblik
---- Rezultati u HTML formatu -----
Pošalji
Korisnički račun
Promjena zaporke
Promjena adrese e-pošte
Prijavljeni ste kao mtadic, odjavite se..
Hrvatski lematizacijski poslužitelj
Hrvatski lematizacijski poslužitelj..
Hrvatski morfološki leksikon..
Pomoć
Ako niste sigurni kako unositi riječi, pogledajte ovu stranicu..
Pritiskom na ⊟ i ⊞ možete skupiti odnosno raširiti rezultate za pojedinu riječ..
Rezultati
Rezultate upita možete preuzeti u tekstnom ili XML formatu:
TXT XML
Copyright © 2005.Hrvatski morfološki leksikon
Sva prava pridržana.
Done
Internet


Hrvatski morfološki leksikon - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://hml.ffzg.hr/unos.php
Go Links
ovu stranicu..
Pritiskom na ⊟ i ⊞ možete skupiti odnosno raširiti rezultate za pojedinu riječ..
Rezultati
Rezultate upita možete preuzeti u tekstnom ili XML formatu:
TXT XML
Lema Oblik ---- Rezultati u HTML formatu ----- Pošalji
Raširi sve / Skupi sve
⊟ glava
Lema Oblik MSD ?
glava glava Ncfpg ?
glava glava Ncfsn ?
glava glavama Ncfpd ?
glava glavama Ncfpi ?
glava glavama Ncfpl ?
glava glave Ncfpa ?
glava glave Ncfpn ?
glava glave Ncfpv ?
glava glave Ncfsg ?
glava glavi Ncfsd ?
glava glavi Ncfsl ?
glava glavo Ncfsv ?
glava glavom Ncfsi ?
glava glavu Ncfsa ?
Google pretraga
Copyright © 2005.Hrvatski morfološki leksikon
Sva prava pridržana.
Done
Internet


Hrvatski morfološki leksikon - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://hml.ffzg.hr/unos.php
Go Links
Hrvatski Morfološki Leksikon
Unos Administracija Pomoć | English
Ručni unos
Unos datoteke
Unesite tekst:
žene
Lema
Oblik
---- Rezultati u TXT formatu -----
---- Rezultati u HTML formatu -----
Pošalji
Korisnički račun
Promjena zaporke
Promjena adrese e-pošte
Prijavljeni ste kao mtadic, odjavite se..
Hrvatski lematizacijski poslužitelj
Hrvatski lematizacijski poslužitelj..
Hrvatski morfološki leksikon..
Pomoć
Ako niste sigurni kako unositi riječi, pogledajte ovu stranicu..
Pritiskom na ⊟ i ⊞ možete skupiti odnosno raširiti rezultate za pojedinu riječ..
Rezultati
Rezultate upita možete preuzeti u tekstnom ili XML formatu:
TXT
XML
Copyright © 2005.Hrvatski morfološki leksikon
Sva prava pridržana.
Done
Internet


Hrvatski morfološki leksikon - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://hml.ffzg.hr/unos.php
Go Links
Hrvatski lematizacijski poslužitelj
Hrvatski lematizacijski poslužitelj..
Hrvatski morfološki leksikon..
Pomoć
Ako niste sigurni kako unositi riječi, pogledajte ovu stranicu..
Pritiskom na ⊟ i ⊞ možete skupiti odnosno raširiti rezultate za pojedinu riječ..
Rezultati
Rezultate upita možete preuzeti u tekstnom ili XML formatu:
TXT XML
Lema Oblik ---- Rezultati u HTML formatu ----- Pošalji
Raširi sve / Skupi sve
⊟ žene
Oblik Lema MSD ?
žene žena Ncfpa ?
žene žena Ncfpn ?
žene žena Ncfpv ?
žene žena Ncfsg ?
žene ženiti Vmip3p ?
Google pretraga
Preuzmi rezultate: TXT XML
Copyright © 2005.Hrvatski morfološki leksikon
Sva prava pridržana.
Done
Internet


Hrvatski morfološki leksikon - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://hml.ffzg.hr/unos.php?upload=file
Go Links
Hrvatski Morfološki Leksikon
Unos Administracija Pomoć | English
Ručni unos
Unos datoteke
Unesite datoteku:
C:\temp4\HNK_v20dio002132.txt
Browse...
Lema
Oblik
oblik + lema + MSD (sve)
ništa
Pošalji
Korisnički račun
Promjena zaporke
Promjena adrese e-pošte
Prijavljeni ste kao mtadic, odjavite se..
Hrvatski lematizacijski poslužitelj
Hrvatski lematizacijski poslužitelj..
Hrvatski morfološki leksikon..
Pomoć
Ako niste sigurni kako unositi riječi, pogledajte ovu stranicu..
Pritiskom na ⊟ i ⊞ možete skupiti odnosno raširiti rezultate za pojedinu riječ..
Rezultati
Rezultate upita možete preuzeti u tekstnom ili XML formatu:
TXT
XML
Copyright © 2005.Hrvatski morfološki leksikon
Sva prava pridržana.
Internet


Hrvatski morfološki leksikon - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://hml.ffzg.hr/unos.php
Go Links
Hrvatski Morfološki Leksikon
Unos Administracija Pomoć | English
Ručni unos
Unos datoteke
Unesite tekst:
Lema Oblik
---- Rezultati u HTML formatu -----
Pošalji
Raširi sve / Skupi sve
Preuzmi rezultate: TXT XML
Copyright © 2005.Hrvatski morfološki leksikon
Sva prava pridržana.
Korisnički račun
Promjena zaporke
Promjena adrese e-pošte
Prijavljeni ste kao mtadic, odjavite se..
Hrvatski lematizacijski poslužitelj
Hrvatski lematizacijski poslužitelj..
Hrvatski morfološki leksikon..
Pomoć
Ako niste sigurni kako unositi riječi, pogledajte ovu stranicu..
Pritiskom na ⊟ i ⊞ možete skupiti odnosno raširiti rezultate za pojedinu riječ..
Rezultati
Rezultate upita možete preuzeti u tekstnom ili XML formatu:
TXT XML
Internet

HML: lemmatization and PoS tagging

a	a	a	C				
s	s	s	S				
druge	druge	drug	N	druga	N	drugi	A
je	je	biti1	V	on	P		
strane	strane	stran	A	strana	N		
moguće	moguće	moguć	A				
istodobno	istodobno	istodoban	A	istodobno	R		
biti	biti	biti1	V	biti2	V		
u	u	u	S				
starosnoj	starosnoj	starosan	A				
mirovini	mirovini	mirovina	N				
i	i	i	C				
biti	biti	biti1	V	biti2	V		
član	član	član1	N	član2	N		
uprave	uprave	uprava	N	upraviti	V		
trgovačkog	trgovačkog	trgovački	A				
društva	društva	društvo	N				
,	,						
tumače	tumače	tumač	N	tumačiti	V		
u	u	u	S				
Hrvatskoj	Hrvatskoj	Hrvatska	N	hrvatski	A		
obrtničkoj	obrtničkoj	obrtnički	A				
komori	komori	komora	N				
i	i	i	C				
stoga	stoga	stog	N	stoga	C	stoga	R
traže	traže	trag	N	tražiti	V		
ocjenu	ocjenu	ocjena	N				
ustavnosti	ustavnosti	ustavnost	N				
sporne	sporne	sporan	A				
odredbe	odredbe	odredba	N				
Zakona	Zakona	zakon	N				
o	o	o	S	o	Y		
mirovinskom	mirovinskom		m				
osiguranju	osiguranju	osiguranje	N				
,	,						
koja	koja	koji	P				
govori	govori	govor	N	govoriti	V		
o	o	o	S	o	Y		
obveznom	obveznom	obvezan1	A				
osiguranju	osiguranju	osiguranje	N				
obrtnika	obrtnika	obrtnik	N				
i	i	i	Ccs				

Homography

- *žene*

Homography

- *žene*

žena	N	G sg
		N pl
		A pl
		V pl

- two types of homography
  - internal or grammatical (IH)
    - word-form belong to the same lemma
    - consequently to the same PoS

Homography

- *žene*

žena	N	G sg N pl A pl V pl
ženiti	V	3p pl pres

- two types of homography
  - internal or grammatical (IH)
    - word-form belong to the same lemma
    - consequently to the same PoS
  - external or lexical (EH)
    - word-form belong to two or more different lemmas
    - not necessarily different PoS

Homography

- *žene*

žena	N	G sg N pl A pl V pl
ženiti	V	3p pl pres

- two types of homography
  - internal or grammatical (IH)
    - word-form belong to the same lemma
    - consequently to the same PoS
  - external or lexical (EH)
    - word-form belong to two or more different lemmas
    - not necessarily different PoS
- EH characteristical for analytical Ls (en, fr…)
  IH characteristical for synthetic Ls (Slavic…)

Homography: statistic

- homographs appear in
  - lexicon
  - corpus
- in lexicon: HML (v 4.4)
  - lemmas: 38,388
  - different word-forms: 2,174,441
  - unique word-forms: 678,883
  - avg. homography: 3.20 MSDs per WF
  - internal homographs: 431,266 (63.55%)
  - external homographs: 23,396 (3.45%)

Homography: statistics 2

- in corpus (100 Kw newspaper corpus, *Croatia Weekly*)
  - tokens: 101,793
  - types: 23,170
  - tokens not in lexicon (NIL): 3,959 (3.89%)
  - non-homographs: 26,399 (25.93%)
  - internal homographs: 53,174 (52.24%)
  - external homographs: 17,783 (17.47%)
    - excl. 'je' → biti1 / on 14,240 (13.99%)
- comparison of homography

	lexicon	corpus
internal:	63.55%	52.24%
external:	3.45%	17.47%

- some frequent functional words = external homographs
  - can be treated as stop-words for many NLP tasks

Existing resources: lexica 4

- HML coverage tested on a 46 Mw hr daily newspaper corpus
  - 96.4% tokens known
  - 3.6% tokens unknown
    - unknown words, typos, foreign names etc.
- all words queried are stored in logs
  - collecting unknown WFs for manual updating of HML
  - also automatic updating of HML (Oliver-Tadić LREC2004, Bekavac-Šojat 2005)

Additional value: search engine

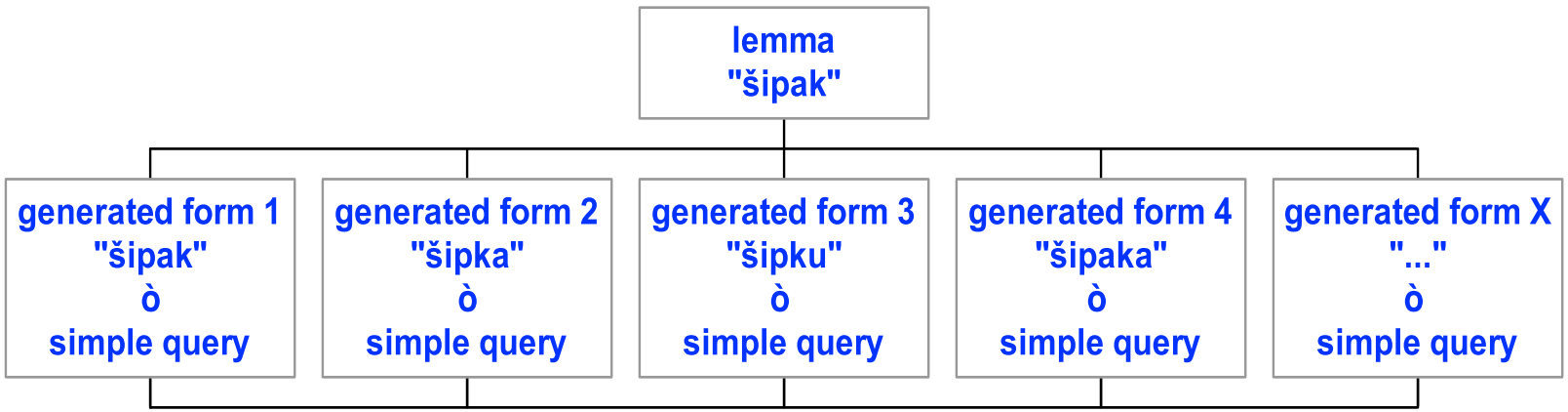
- web search engines (SE) = everyday commodity
  - global: Google, Bing, Yahoo,…
  - local: portals, web sites
  - “hidden web”: vast quanitites of data accessible only by queries
- generally well suited for queries in English
- what about other languages?
  - should the SE be multilingually sensitive?
  - should the SE be sensitive to linguistic structures?
- problem of document retrieval for documents written in inflectionally rich languages
  - e.g. German, Finnish, Slavic languages, Arabic…
  - elementary linguistic fact: in these languages lexemes appear in many different word forms (WF)

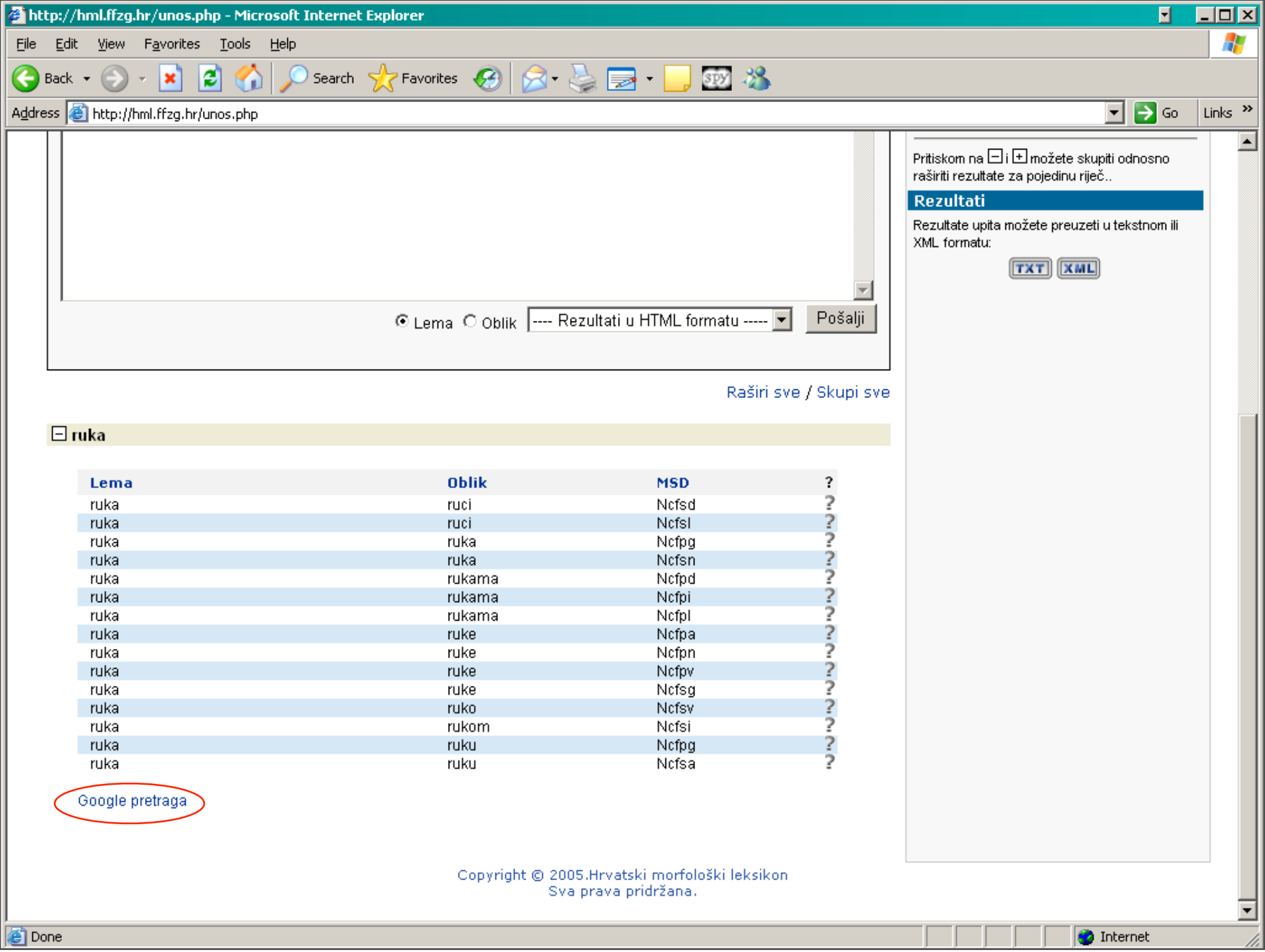
Additional value: search engine 2

- which WF does the speaker of inflectionally rich language use in web queries? usually a lemma
- user's intuition: lemma covers/represents all WFs of a lexeme
- e.g. how do you input the hr noun in google.hr?
  - nominative singular!
  - accusative and genitive case in hr are more frequent than nominative!
- result
  - all documents with lemma occuring are being hit
  - all documents without lemma, but with some other (more frequent!) WFs, are being missed
- this problem should be solved in order to
  - build more user friendly / user language sensitive SE
  - get better recall without decrease of precision in doc. retrieval

Additional value: search engine 3

- could we use the HML for generation of a web search engine query?
  - a list of all WFs of a lemma reduced to
  - a list of unique WFs connected with Boolean OR operator




http://hml.ffzg.hr/unos.php - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://hml.ffzg.hr/unos.php
Go Links
Lema Oblik
---- Rezultati u HTML formatu -----
Pošalji
Pritiskom na ⊟ i ⊞ možete skupiti odnosno raširiti rezultate za pojedinu riječ..
Rezultati
Rezultate upita možete preuzeti u tekstnom ili XML formatu:
TXT XML
Raširi sve / Skupi sve
⊟ ruka
Lema Oblik MSD ?
ruka ruci Ncfsd ?
ruka ruci Ncfsl ?
ruka ruka Ncfpg ?
ruka ruka Ncfsn ?
ruka rukama Ncfpd ?
ruka rukama Ncfpi ?
ruka rukama Ncfpl ?
ruka ruke Ncfpa ?
ruka ruke Ncfpn ?
ruka ruke Ncfpv ?
ruka ruke Ncfsg ?
ruka ruko Ncfsv ?
ruka rukom Ncfsi ?
ruka ruku Ncfpg ?
ruka ruku Ncfsa ?
Google pretraga
Copyright © 2005.Hrvatski morfološki leksikon
Sva prava pridržana.
Done
Internet

Address http://www.google.com/search?hl=hr&q=ruka%20OR%20ruci%20OR%20ruka%20OR%20rukama%20OR%20ruke%20OR%20ruko%20OR%20rukom%20OR%20ruku8 Go Links

Prijava

Google

Web Slike Grupe Imenik

ruka OR ruci OR ruka OR rukama OR ruke OR r Traži Napredno pretraživanje Postavke

Pretraži Web Pretraži stranice na hrvatskom jeziku

Web Rezultati **1 - 10** od približno **1.710.000 hrvatskom** stranica za **ruka OR ruci OR ruka OR rukama OR ruke OR ruko OR rukom OR ruku**. (**0,14** sekundi)

Ranko Marinkovic: **Ruke**
A desna ti je **ruka** veæe dugo le¾ala na onom kamenu ¹to je pritiskao hlaèe i **...** No, u stvari, i nema¹ karata u **rukama** nego si ra¹irio prste lijeve **ruke** pa **...**
www.yurope.com/books/yu/Mar_Ruke.HTM - 344k - Spremljeno - Slične stranice

Vaše zdravlje :: Pregled članka - Čiste **ruke** - pola zdravlja
Infektologija / Javno zdravstvo Čiste **ruke** - pola zdravlja Autor:Rosana Svetić-Čišić **...** Čiste **ruke** - pola zdravlja. Ispravna tehnika pranja **ruku** smanjuje i **...**
www.vasezdravlje.com/izdanje/clanak/421/ - 26k - Spremljeno - Slične stranice

Vaše zdravlje :: Pregled članka - Zašto je dobro prati **ruke**?
Javno zdravstvo Zašto je dobro prati **ruke**? Autor:Marina Gradinac, dr. med. objavljeno u broju 40 (2/05) Pranje **ruku** vrlo je učinkovita strategija brige o **...**
www.vasezdravlje.com/izdanje/clanak/589/ - 23k - Spremljeno - Slične stranice
[Više rezultata za www.vasezdravlje.com]

Corner Slobodni kut
vijek mi kod ljudi koje tek upoznajem pogled spontano klizne k **rukama**. Jer, **ruke** govore o osobi više od lica. **Ruke** imaju svoj život. **...**
www.corner.hr/kolumne/slobodni.asp?id=2812 - 49k - Spremljeno - Slične stranice

Ranko Marinkovic: **Ruke**. [Borut's Literature Collection]
A desna ti je **ruka** već dugo ležala na onom kamenu što je pritiskao hlače i **...** No, u stvari, i nemaš karata u **rukama** nego si raširio prste lijeve **ruke** pa **...**
www.borut.com/library/texts/marinkovic/**ruke**/index.htm - 302k - Spremljeno - Slične stranice

Vesna Parun: Ti koja imas **ruke** nevinije od mojih. [Borut's **...**
Ti koja imaš **ruke** nevinije od mojih. i koja si mudra kao bezbrižnost. **...** Ja neću nikad voditi za **ruku**. njegovu djecu. I priče. koje za njih davno pripremih **...**
www.borut.com/library/texts/parun/poetry/**ruke**.htm - 6k - Spremljeno - Slične stranice

More Comics - Stripovi - Alahova **ruka**
Zagor extra: Alahova **ruka** cijena: 35.00 kn broj izdanja: 21 crtač: Franco Donatelli, Gallieno Ferri pisac: Guido Nolitta kreator likova: Guido Nollita **...**
www.morecomics.hr/item/1436 - 12k - Spremljeno - Slične stranice

Oliver Dragojevic - Tvoje **Ruke** LYRICS
Oliver Dragojevic LYRICS,Tvoje **Ruke** ,Oliver Dragojevic Tvoje **Ruke** Lyrics, Oliver Dragojevic Tvoje **Ruke** Song Lyrics.
www.lyricsdownload.com/ oliver-dragojevic-tvoje-**ruke**-lyrics.html - 24k - Spremljeno - Slične stranice

Existing resources: lexica

- other monolingual machine readable dictionaries
  - general language: Anić (^{4}2003) published on CD
    - digitalised, planned storing in LMF
- EUROVOC thesaurus translated to hr in 2000
  - http://www.hidra.hr
- CROVALLEX: valency lexicon
  - Mikelić-Preradović (2008)
  - 1739 the most frequent Croatian verbs
  - already used in disambiguation of our chunker output
  - http://cal.ffzg.hr/crovallex
- Croatian Wordnet (CroWN) is under development
  - Raffaelli et al. (2008)
  - adding lexical semantic information to the subset of the HNK
  - http://rmjt.ffzg.hr/p3_en.html

Existing tools

- tokenizer
- sentence splitter
  - in hr: ordinal numbers written with dot (“5.” means ‘the fifth’)
  - 24% also end of the sentence
- inflectional generator (described)
  - lemmatization & PoS/MSD tagging at unigram level
- CroTag: stochastic tagger combined with HML into hybrid sys.
  - Agić-Tadić(-Dovedan) (2006, 2007, 2008, 2009, 2010...)
  - inspired by TnT (Brants 2000) & open source reimplementation HunPos (Halacsy et al. 2007)
  - trigram / second order HMM tagging paradigm with linear interpolation
  - suffix trie and successive abstraction for unknown word handling, input and output formats identical to TnT

Existing tools: tagging and lemmatisation

- combining HML and CroTag into a hybrid tagger
  - to improve overall MSD tagging accuracy
- HML encoded as miminal FSA
- using HML as a handler for words unknown to stochastic tagger
- overall tagging accuracy (measured and trained on the same genre: newpaper)
  - only PoS: 99.31%
  - full MSD: 97.51%
- this combinaton also used for disambiguation in lemmatisation (measured and trained on the same genre: newpaper)
  - accuracy: 98.25%

Existing tools

- normalizator
  - conflates WF of all lemmas with the same root
  - computationally efficient, linguistically not completely valid
  - Šnajder et al. (2008)
  - KTLab, Faculty of Electrical Engineering and Computing, UniZg
  - http://ktlab.fer.hr, also http://rmjt.ffzg.hr/p5_en.html
- environments for development of (local) grammars
  - regular grammars: INTEX
  - regular & CF grammars: NooJ
  - Silberztein (2005, 2006)
  - http://www.nooj4nlp.net
  - Croatian module started in 2007

Existing tools: NERC system

- developed NERC system
  - Bekavac (2005), Bekavac-Tadić (ACL2007)
  - rule-based system
  - gazetteers of personal and location names
  - local regular grammars for names, dates, numbers, measures, values, percentages...
  - modelled in INTEX development environment
  - 0.9 F-measure score at hr newspaper texts

Večernji list, 2005-02-17, gospodarstvo

Hrvatski izvoz još na niskim razinama

90 posto tvrtki uopće ne izvozi!

Autor Piše ***Josip Bohutinski***

Hrvatski izvoz napokon je **prošle godine** počeo rasti brže od uvoza te je, prema podacima za **prvih 11 mjeseci 2004. godine**, izvoz u kunama rastao **15,7 posto** a uvoz **5,7 posto**. Iz **Hrvatske** je izvezeno robe u vrijednosti nešto manjoj od **44 milijardi kuna** ili **7,25 milijardi američkih dolara**, dok je vrijednost uvoza bila **91,19 milijardi kuna** ili više od **15 milijardi dolara**.

No podaci o izvozu po glavi stanovnika upozoravaju da je hrvatski izvoz još na niskim razinama u usporedbi s drugim i sličnim zemljama. Prema podacima udruge Hrvatski izvoznici, u **2003. godini** vrijednost hrvatskog izvoza po glavi stanovnika bila je samo **1106 dolara**.

Koliko je je to mala vrijednost, govori podatak o slovenskom izvozu po glavi stanovnika od čak **4774 dolara**. **Irska** na svakog svoga stanovnika izveze **22.119 dolara** roba i usluga. Amerikanci, pak, po glavi stanovnika izvezu robe u vrijednosti **2360 dolara**.

No vrijednost izvoza velikih zemalja po glavi stanovnika u pravilu je manja od izvoza malih zemalja zbog velikog domaćeg tržišta koje može apsorbirati veliki dio domaće prozivodnje. To potvrđuju i podaci o izvozu po stanovniku i "malih zemalja" poput **Belgije**, **Nizozemske** i **Finske**.

Uz malu vrijednost izvoza po glavi stanovnika, za **Hrvatsku** je nepovoljan i podatak o broju domaćih tvrtki čija godišnja vrijednost izvoza premašuje **milijun kuna**.

Njih je samo **pet posto** od ukupno aktivnih poduzeća. Naime, prema podacima Hrvatskih izvoznika, od 70-ak tisuća aktivnih kompanija u **Hrvatskoj**, svoje proizvode i usluge na strana tržišta izvozi samo njih 6700. Pritom je izvoznika čija vrijednost izvoza premašuje **milijun kuna** samo 3144. Ta grupa izvoznika, prema podacima udruge Hrvatski izvoznici, ostvaruje čak **96 posto** ukupnog hrvatskog izvoza.

Koliko je bitna uloga izvoznika u cjelokupnom hrvatkom gospodarstvu, potvrđuje podatak da 2688 izvoznika izdvaja **83 posto** ukupne dobiti u **Hrvatskoj**, odnosno 16,6 od **19,9 milijardi dolara**.

Upozoravajući na podatke o hrvatskom izvozu po glavi stanovnika, predsjednik Hrvatskih izvoznika **Darinko Bago**, prilikom prošlotjednog potpisivanja Sporazuma o suradnji s **Hrvatskom** bankom za obnovu i razvitak, najavio je sklapanje sličnih sporazuma s drugim udruženjima i institucijama koje mogu pridonijeti afirmaciji hrvatskog izvoza, bez kojeg, naglasio je **Bago**, **Hrvatska** nema budućnosti.

A velike zasluge za prošlogodišnji brži rast hrvatskog izvoza sigurno ima upravo **HBOR** i njegovi programi poticanja izvoza. Preko programa Kreditiranje priprema roba za izvoz i izvoza roba **lani** je odobreno 170 kredita u vrijednosti **1,25 milijardi kuna**, što je čak **448 posto** veći iznos nego **2003. godine** kada su odobrena 52 kredita, ukupno vrijedna nešto više od **279 milijuna kuna**.

I Program osiguranja izvoza zabilježio je **lani** veliki rast. U **2004. godini** osiguran je promet od **580 milijuna kuna**, što je povećanje**180 posto** prema **prethodnoj godini**, a odobreno je 357 zahtjeva, što je povećanje od **306 posto**. **Lani** je **HBOR** osigurao izvoz 67 izvoznika, za razliku od 35 u **2003. godini**. Od početka poslovanja **HBOR** je dosad isplatio 12 odšteta u iznosu **3,2 milijuna kuna**, a od toga je **lani** četvero izvoznika dobilo odštetu od **538.000 kuna**.

Predsjednik Uprave **HBOR-a Anton Kovačev**, potpisujući sporazum s Hrvatskim izvoznicima, rekao je da je **2004.** bila godina izvoza za njegovu banku te da se nada da će ova biti izvozna za cijelu **Hrvatsku**, čemu bi trebao pridonijeti i sporazum o suradnji **HBOR-a** i **HIZ-a**.

Kovačev je upozorio i da rast hrvatskog izvoza **lani** nije isključivo rezultat brodogradnje.

- Oko **90 posto** kredita koje smo dali za priremu roba za izvoz i izvoz roba odnosi se na prerađivačku industriju, poput prehrambene, metalske, farmaceutske i drvne industrije. A te industrije su ostvarile porast izvoza **6,5 posto**, što je veći rast od prosječnog ukupnog rasta od **15,7 posto** - rekao je **Kovačev**.

Legenda:

brojčani i postotni iznosi

vremenski izrazi

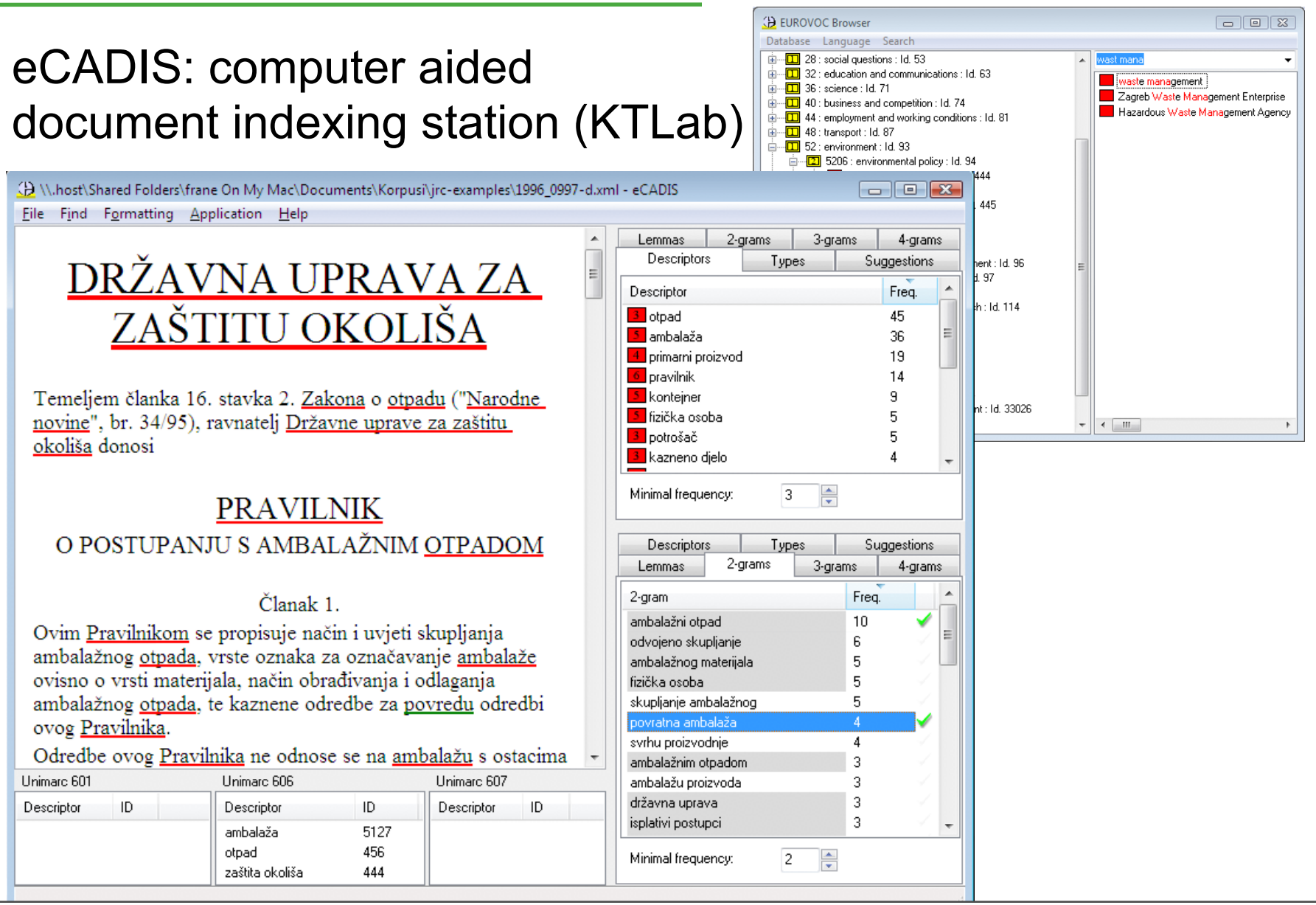
imena osoba

imena lokacija

imena organizacija

Existing tools: document indexing

- eCADIS: computer aided document indexing station (KTLab)



Existing tools: document classification

- TMT object-oriented text classification library
  - Šilić et. al. (2007)
  - developed within the projects AIDE and CADIAL
  - http://aide.hidra.hr and http://www.cadial.org
  - API library that includes modules for
    - tokenization
    - morphological normalization
    - lemmatisation
    - PoS/MSD tagging
    - feature building, selection and weighting (χ^2, MI,…)
    - classifier training (SVM, k-NN, Bayes, Winnow, Rocchio)
    - evaluation of classifiers (precision, recall, F1,…)
    - optimization of classifier parameters (grid search, simplex method, genetic algorithm, simulated annealing)
    - serialization of intermediate results

CLARA

How to shape a project?

Analysis of user needs

- be prepared for a series of meetings
- understand each other
  - define the main concepts at the beginning
  - fix the terminology
  - scientists and industrials talk different language
    - scientists are usually laymen in economy
    - industrials are usually laymen in linguistics/IT…
- make a general presentation of your research group
  - do not use PR talk (too much)
  - list the achievements references, i.e. sucessful projects so far
  - adapt the presentation to the level of your audience, i.e. use technical terms only when necessary
- expect the similar presentation from industrials
  - ask for people from development and not PR department

Analysis of user needs 2

- analyse the user needs regarding the needed LT expertise at
  - morphological level
  - phrasal level
  - syntactical level
  - semantic level
- check whether the necessary resources and tools exist for your language
  - BLARK grid (Krauwer 1998)
  - if they do exist, excellent
  - if they do not exist, use this business case to build it
  - build all resources and tools to be interoperable
    - follow the standards and/or reccomendations
  - use the existing solutions, do not build everything by yourself
    - speeding up the development time

Analysis of user needs 3

- user wants to build a system that will
  - put into relations the
    - people, locations, organisations
    - events

    in very large number of streamlined documents
  - automatically classify these documents according to the predefined classification schema (IPT classification of topics)
  - automatically add keywords extracted from the text to metadata
  - offer extensive querying system
  - visualise the results in different manners
- can you meet all these needs? do not be greedy!
  - FFZG could not deal with it all
  - so we invited another faculty – FER – to take care about the parts that we were not being able to produce ourselves

Analysis of user needs 4

- morphological level
  - lemmatisation task
  - needed for processing at other levels for inflectionally rich L
  - direct usage in search engine of the Hina system
- phrasal level
  - NERC task
  - keyword extraction task
  - chunker not used, only experiments in chunking Croatian
- syntactic level
  - sentence segmetation
  - needed in NERC module for detecting NE boundaries
- semantic level
  - level of “document semantics”, applying the vector space model
  - needed for document classification task

Analysis of user needs 4

- through the series of meetings and negotiations with Hina development department we agreed upon the system that is
  - modularly designed
    - each team is responsible for several modules
  - covers only the needs that we can deal with at that moment
  - allows the upgrading of modules
  - allows the addition of new modules
- the system was called Semantical Analysis of Text (SAT)
  - lemmatisation (FFZG)
  - NERC (FFZG)
  - document classification (FER)
  - keyword extraction (FER)
  - wrapper (FER)
  - control module for tracing the functioning of the system (FER)

Analysis of user needs 5

- FFZG and FER jointly applied to a public tender
  - beside our there was a competitive offer
    - by Siemens software department
  - but, we were not scared
    - be bold
    - trust in yourself and your capabilities
- Hina decided to organise the competition
  - starting with one task at the time
  - two pilot solutions for a single task (our and Simens')
  - measuring performance of offerers using the same conditions
  - each task was decisive, the offerer who is significantly performing worse, falls off
  - document classification task
    - FFZG/FER: 84% accuracy Siemens: 67%

CLARA

Adaptation of existing resources and tools

Adaptation of FFZG resources

- HML formed the basis for lemmatisation module
  - existed in FSA version
  - additional entries needed
    - analysis of Hina corpus for evidence and frequency of unknown lemmas
    - foreign names
    - possesive adjectives of names (people, locations)
    - …
  - FSA recompiled
  - adapting it to predefined input/output data format for communication between modules

Adaptation of FFZG tools

- NERC
  - existed in research prototype form (Bekavac PhD, 2005)
    - a set of cascading regular grammars that provided annotation in text for beginning/ending and classification of NEs
    - modelled in development environment INTEX
    - following the MUC7 specification for 7 NE types
  - it was not an industry strength application
  - the FSTs exported into C-tables
  - module completely rewritten in C++
  - adapting it to predefined input/output data format for communication between modules

CLARA

Possible future developments

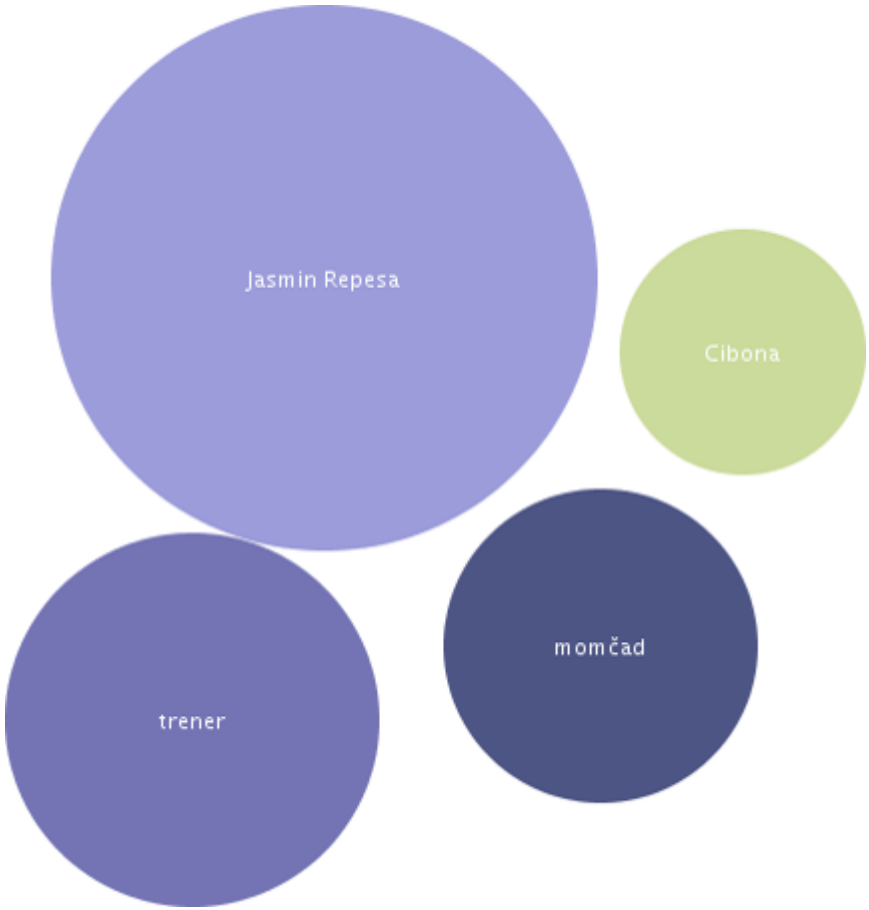
Possible future developments

- develop a system that really find relations between
  - people, locations, organisations
  - events
- in metadata, but also in text itself (at different levels)
- SVO triples
  - in morphologically rich language
  - skip over syntax (linguistic blasphemy?!)
  - go directly to semantic roles by brute-force mapping of cases of NP within the clause
    - N = agent, A = patient, D = beneficiary, I = instrument
  - e.g.
    - *Agrokor prodaje Frigokom.*
    - *Microsoft kupuje Skype.*
    - ...

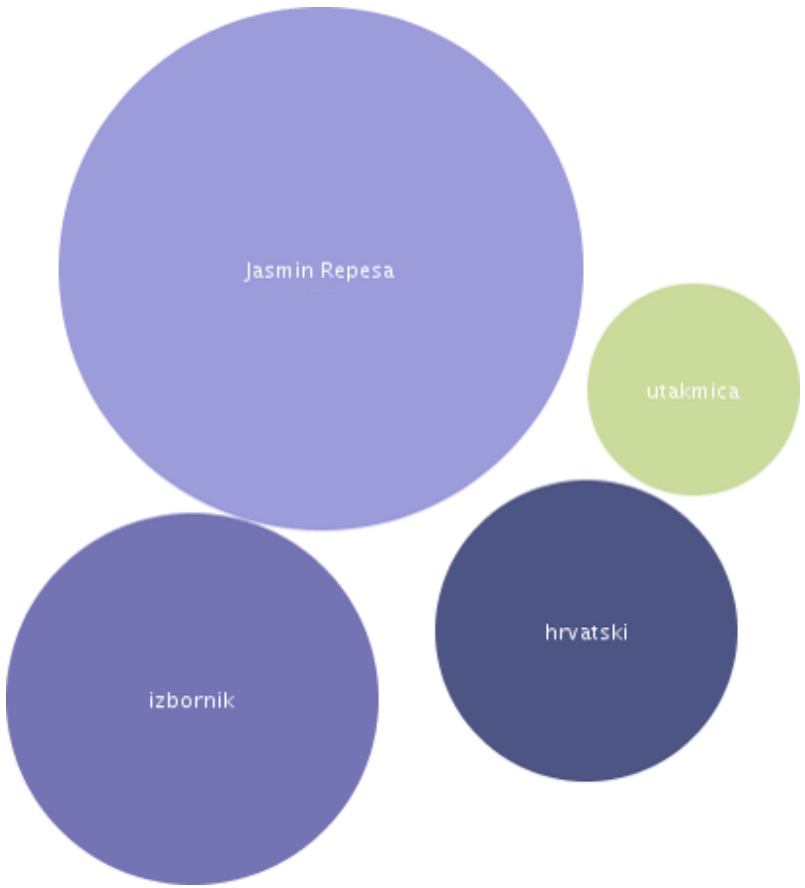
Possible future developments 2

- tracking the NE trough time

2001



2009



Possible future developments 3

- integration of resources and tools with search engine
- visualisation module
  - different ways to visualise findings
- location deduction
  - location NE triggers upper level locations
  - hierarchical ontology of locations
    - town
    - county
    - state
    - region
    - continent
  - gazetteer of all Croatian
    - populated places, mountains, rivers, lakes, seas, countries

Thank you for your attention.

CLARA is an Initial Training Network in the
Marie Curie Actions financed by the EC under FP7