



Екстракција термина у српском језику

Проф. др Ранка Станковић

Универзитет у Београду - Рударско-геолошки факултет

Семинар Друштва за језичке ресурсе и технологије

ЈЕРТЕХ 24.03.2016., Свечана сала 167, први спрат

Екстракција термина



- Зашто екстракција термина (ЕТ)
- Стандарди и стање у свету (и код нас)
- Коришћени ресурси
- Процес екстракције и лематизације термина
- Приказ развијеног решења
- Примери екстракције и евалуација
- Даље унапређење ЕТ

Стандарди



- Технички одбор за стандардизацију терминологије при ISO задужен је за доношење и вођење стандарда везаних за терминологију.
- Основни стандард везан за терминологију ISO 704:2009 утврђује принципе и методе
 - за припрему и прикупљање терминологије и описује везе између објеката, концепата и њихових терминолошких репрезентација
 - за формирање ознака и формулацију дефиниција
- Терминолошки стандарди не обухватају морфологију

Стандарди



Стандард ISO 10241:1992 *International terminology standards — Preparation and layout* ... издваја фазе:

- Креирање и прочишћавање листе термина
- Успостављање поља за опис концепата и креирање система концепата
- Формулисање дефиниција
- Повезивање и избор термина

Листа термина може да укључи термине за

- Концепте специфичне за конкретан домен
- Концепте специфичне за више домена;
- Позајмљене концепте;
- Концепте општег језика.

Коначан речник треба да укључи термине

- Који су специфични за конкретан домен
- И само ограничен број позајмљених концепата и концепата заједничких са другим доменима

Зашто екстракција термина?



Да ли нам је потребна аутоматска процедура за прикупљање терминологије на српском језику?

Брзе промене у многим доменима захтевају креирање нових термина, а како се нови термини непрестано додају и уводе у српски језик постаје све важнија аутоматизација њиховог препознавања, издвајања и укључивања у терминолошке речнике за српски

Због комплексне морфологије ово представља веома тежак задатак, тако да је потребно креирати и одговарајуће лексичке ресурсе у облику морфолошких електронских речника и граматика потребних за њихову примену

Прикупљање



Pre

- Лингвистичка обрада (претпроцесирање) улазног (чистог) текста користећи Unitex
<http://www-igm.univmlv.fr/~unitex>

ET

- Анализа непрепознатих речи и проширење речника DELAS (једночланих ЛЈ)
- Екстракција вишечланих ЛЈ

Post

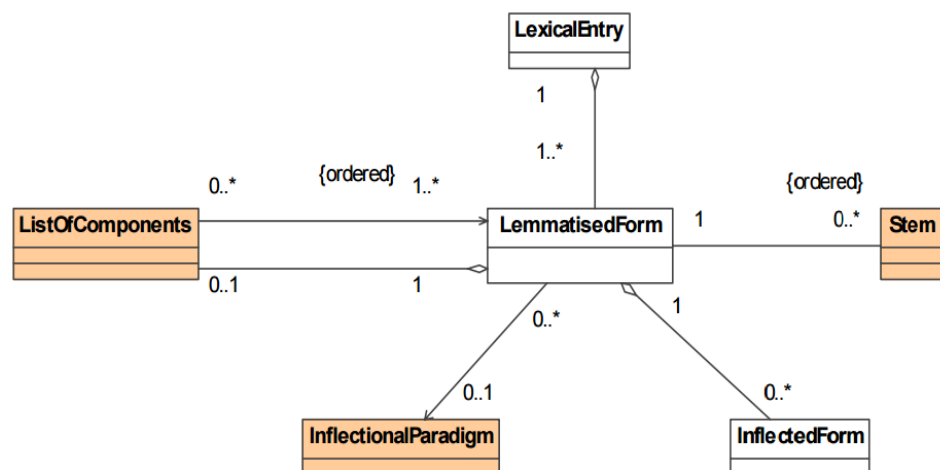
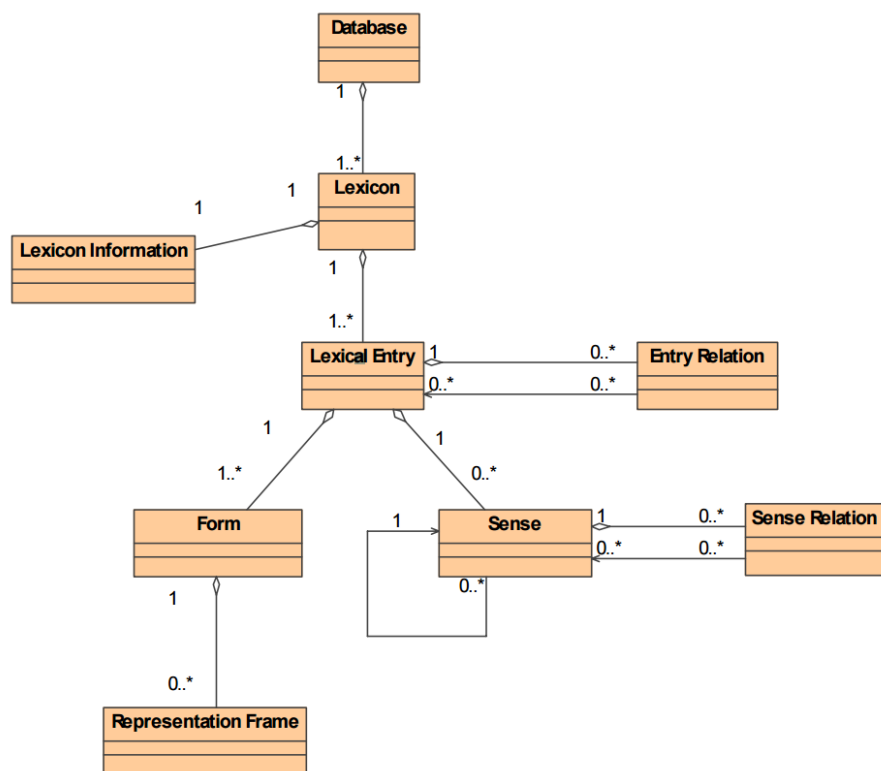
- Избор термина из листе ВЛЈ
- Проширење речника DELAC

Cvetana Krstev. *Processing of Serbian – Automata, Text and Electronic Dictionaries*, Faculty of Philology, Belgrade, 2008.

Стандарди



- LMF ISO-24613:2008 је стандард за NLP лексиконе и машински читљиве речнике (MRD), TC37/SC4



Стандарди



- TermBase eXchange (TBX), ISO 30042:2008 - размена термиолошких података сагласна са Оквиром за означавање терминологије TMF (Terminological Markup Framework), ISO16642:2003
- TBX и TMX се користе и у алатима за превођење заснованим на преводилачким меморијама (Translation memory, TM), SDL Trados, MemoQ и сл. директно подржавају поменуте стандарде.

Више детаља

- <http://www.tbxinfo.net/>
- Ranka Stanković, Ivan Obradović and Miloš Utvić. „Developing Termbases under the TBX Standard.” In: Pavlović-Lažetić, Gordana et al. (eds.), *Natural Language Processing for Serbian*, pp. 12-26. (Belgrade: University of Belgrade, Faculty of Mathematics, 2014):18-19.

Искуство у свету и код нас



- ЕТ углавном се заснива на статистичком приступу уз примену различитих статистичких мера (Kenneth Church и други 1990; Frank Smadja 1993)
- ЕТ независну од језика развили Pantel & Lin 2001 и применили на енглески и кинески
- Chen и други 2006: систем ЕТ заснован на повезаним сегментима текста и скупова докумената
- Tadić&Šojat, 2003. користе статистичку меру MI^3 (mutual information) за ЕТ

- ЕТ заснован на правилима, коришћењем каскада трансдуктора
 - Арапски (научни и технички корпус), Ammar и др. 2015
 - Пољски (економски домен), Savary и др. 2012
- Хибридни приступ: комбиновање статистичког и заснованог на правилима
 - Арапски, лингвистичким правилима се парсирају документи, а потом извлачи листа кандидата, примењују статистичке мере и рангирају кандидати (Bounhas&Slimani, 2009)
 - Сличан приступ за бугарски (Koeva, 2007), за словеначки (Vintar, 2010) и пољски (Broda et al., 2008)

Искуство у свету и код нас



- Кандидати ВЛЈ се препознају у различитим флективним облицима, које треба повезати и представити једним нормализованим обликом
- За енглески језик није битно, тако да су ретка истраживања овог проблема
- Неопходна минимална нормализације је да се урадили лематизација „реч по реч“
- За језике са комплексним флективним правилима је важно креирање исправне леме због евалуације, аутоматског креирања записа у е-речнику, који је неопходан јер се у корпусу обично не налазе сви флективни облици

Искуство у свету и код нас



- Лематизација за пољски
 - Статистички приступ, лична имена, Piskorski и други 2007
 - Систем заснован на правилима, именовани ентитети, Małyszko и други 2015
- Лематизација ВЛЈ за српски заснована на правилима се ослања на
 - лексичке ресурсе и морфолошке е-речнике и граматике (Vitas, D., Popović, Lj., Krstev, C., Obradović, I., Pavlović-Lažetić, G. and Stanojević, M. (2012). The Serbian Language in the Digital Age. Berlin; Springer-Verlag).



Cvetana Krstev, Ranka Stanković, Ivan Obradović, Biljana Lazić “Terminology Acquisition and Description Using Lexical Resources and Local Grammars”, in *Proceedings of the 11th Conference on Terminology and Artificial Intelligence*, Granada, Spain, 2015, eds. Thierry Poibeau and Pamela Faber, LexiCon (Universidad de Granada), pp. 81-89, CEUR Workshop Proceedings, ISSN 1613-0073, urn:nbn:de:0074-1495-6, http://ceur-ws.org/Vol-1495/paper_13.pdf



- Анализа структуре термина термилошких база из области геологије, рударства, библиотекарства и Ворднета за српски језик (<http://sm.jerteh.rs/>)
- Структура полилексичких јединица Морфолошких речника за српски језик (Крстев и Витас).
- Корпус савременог српског језика од 22М речи (<http://korpus.matf.bg.ac.rs>) – рефератан корпус за одређивање доменске специфичности екстрахованог термина

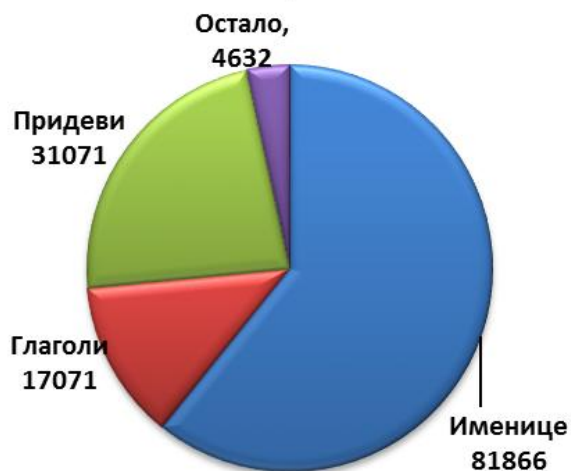


Анализа ресурса

POS	лема		флексивне класе	
Именице	81.866	61%	372	44%
Глаголи	17.071	13%	372	44%
Придеви	31.071	23%	69	8%
Остало	4.632	3%	41	5%
Укупно	134.640		854	

- Morphological e-dictionaries of Serbian are in DELA format (Courtois, 1990)
- DELAS ~135,000 lemmas generating more than 5 million forms (DELAF)
- DELAC ~ 13,000 MWUs

Леме е-речника



Флексивне класе



Анализа ресурса



- Анализа структуре термина у терминолошким базама
- Број речи у терминима 3 терминолошке базе
- Примери у презентацији из домена: библиотекарства, рударства, геологије, математике

Речник	Дужина термина (по броју речи)				
	1	2	3	4	≥5
GeolISS Term geoliss.mre.gov.rs/term	1436	2356	749	305	243
RNBS rbi.nb.rs/en/home.html	3302	6180	2062	806	415
RudOnto rudonto.rgf.bg.ac.rs	1004	1351	1350	1031	2341

Развијено је 14 трансдуктора који одговарају најпродуктивнијим класама (TIA 2015, LREC 2016)

Анализа ресурса



- Вишечлане лексичке јединице су класификоване према синтаксичкој структури, флективним и другим својствима
- Имена класа одговарају коначним трансдукторима за ту класу
- На пример: придев (**A**) за којим следи именица (**N**), који се слаже у роду, броју, падежи и аниматности припада **AXN** класи, где **X** означава компоненту која се не мења или сепаратор (размак или цртицу)

Анализа ресурса



- Именичке ВЛЈ у Српском:
 - 14 класа покрива 98% именичких ВЛЈ
- Број компоненти
 - 4 су са 2 компоненте,
 - 5 је са 3 и
 - 4 су са 4 компоненте.

Граф	Број ВЛЈ	%
NC_AXN	5893	45,91
NC_AXN3	2169	16,90
NC_N4X	625	4,87
NC_N2X	589	4,59
NC_2XN	417	3,25
NC_N2X2	385	3,00
NC_2XN3	350	2,73
NC_2XN1	295	2,30
NC_N4X1	238	1,85
NC_AXN2X	236	1,84
NC_NXN	180	1,40
NC_AXAXN	163	1,27
NC_N6X	157	1,22
NC_AXAXN1	106	0,83
NC_NXN2	93	0,72
NC_N6X1	71	0,55
NC_NXN2m	71	0,55
NC_2XAXN	65	0,51
NC_AXN4X	64	0,50
NC_AXN2X1	42	0,33
NC_N8X	36	0,28
NC_2XN2	30	0,23

Екстракција термина



- Прва верзија екстракције на пољопривредном корпусу (Весна Пајић, Цветана Крстев, Сташа Вујичић-Станковић, Ранка Станковић)
- Само екстракција и нормализација (лематизација реч по реч), отежана евалуација и допуна е-речника
- Појединачно извршавање графова
- Статистика ручно (коришћењем ексела)
- Поређење са корпусом општег језика апроксимативно

→ Неопходно урадити

Проширити и модификовати скуп графова
Лематизацију ВЛЈ
Проширити скуп статистичких мера
Урадити аутоматизацију поступка

ID	Syntactic pattern ^a	Description	Example
1	AXN	Both components inflect and agree in gender, number and case	<i>žitni kombajn</i> 'wheat harvester'
2	NXA	Both components inflect and agree in gender, number and case	/
3a	NXNg	Only the 1 st component inflects; the 2 nd is always in genitive	<i>obrada zemljišta</i> 'soil tillage', lit. 'tillage of soil'
3b	NXNi	Only the 1 st component inflects; the 2 nd is always in instrumental	<i>snabdevanje vodom</i> 'water supply', lit. 'supply with water'
3c	NXN	Both components inflect and agree in case and number; neither component is in instrumental or genitive	<i>traktor guseničar</i> 'tractor caterpillar'
4	N-N	The 1 st component does not inflect (and it is never used as an independent word in Serbian); the 2 nd component inflects	<i>eko-sistem</i> 'ecosystem'
5	NXAgXNg	The 1 st component inflects; the 2 nd and 3 rd are always in genitive	<i>tok vegetacionog perioda</i> 'course of vegetation period'
6	NXAiXNi	Only the 1 st component inflects; the 2 nd and 3 rd are always in instrumental	<i>oranje raonim plugovima</i> 'plowshare plowing'
7	NXNgXAg	Only 1 st component inflects; the 2 nd and 3 rd are always in genitive	/
8	NXNn(-)Nn	Only 1 st component inflects; 2 nd and 3 rd are always in nominative and are optionally separated by a hyphen	<i>stepen iskorišćenja energije</i> 'level of energy utilization'
9	NXPrepXN	Only 1 st component inflects; the 2 nd and 3 rd agree in case	<i>sistem za navodnjavanje</i> 'system for irrigation'

Екстракција термина



- **AXN** — придев именица који се слажу у роду, броју, падежу, аниматности; **геолошка карта, минерална сировина, векторски простор, информациони систем, библиотечки материјал**
- **2XN** — испред именице реч која се не мења (префикс или прилог изведен од придева, а сепаратор је најчешће цртица), **анкер-мрежа, дацито-андезит, библиотечко-информациони систем, аудио-запис, научно-истраживачки рад**
- **N2X** — иза именице реч која не подлеже флексији (именица у генитиву или инструменталу); **патрона експлозива, база података, врста експлозива**

Александра Тртовац, **Дескриптори метаподатака и дескриптори садржаја у проналажењу информација у дигиталним библиотекама**, докторска дисертација (у процедури за одбрану), Ментор Проф. Цветана Крстев



- **N4X** – иза именице две речи које се не мењају:
 - NNgi - **откопавање широким челом, обрада природних језика, софтвер отвореног кода, систем морфолошких речника**
 - NprepNr – **ланац са грабуљама, рачунарство у облаку, лиценца за рад**
- **AXN2X** – испред именице се налази придев који се са њом слаже, а иза именице реч која се не мења; **геолошко картирање терена, семантичка мрежа речи, први извод функције, оптичко препознавање карактера**
- **NXN** – слажу се, **багер кашикар, алеврит-песак, библиотека целина**
- **AXAXN** – придеви се слажу са именицом; **површински истражни радови, морфолошки електронски речник**



NNgiPrepNr - припрема лежишта за откопавање, извод функције у тачки, конверзија текста у говор

**NNgiNgiNgi - истраживање лежишта минералних сировина, примена
методе коначних елемената, решење система диференцијалних
једначина**

AXN4X : испред именице придев који се слаже, а иза именице две речи које се не мењају

ANPrepNr – гравитацијска концентрација у води, бела књига о језицима, интеграбилна функција на сегменту

ANNgINgi – површинска експлоатација минералних сировина, предикатски рачун првог реда

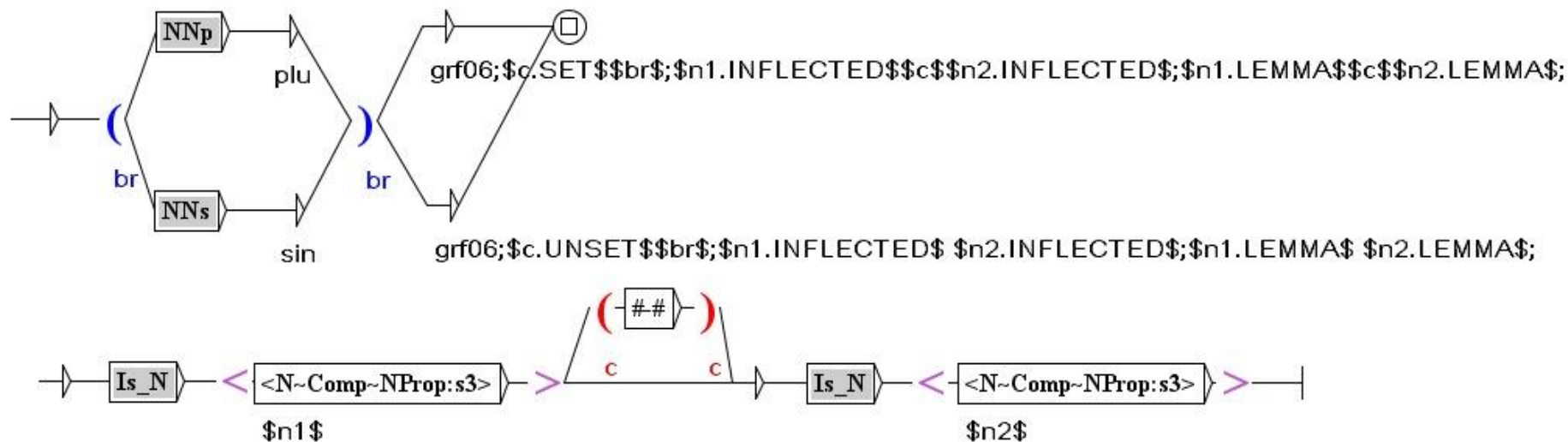
2XAXN : придев и именица који се слажу, а испред њих реч која се не мења, магматско-еруптивни масив, информационо-комуникационе технологије, узрочно-последична веза, диференцијално-алгебарска функција

Екстракција термина NXN



- FST за препознавање NXN: подграфови NNp и NNs препознају облике јединине и множине
- Речничке променљиве \$n1.LEMMA\$ и \$n2.LEMMA\$ на излазу врше нормализацију, односно једноставну лематизацију препознатих облика речи \$n1\$ и \$n2\$

Recognizes MWTs with structure NXN

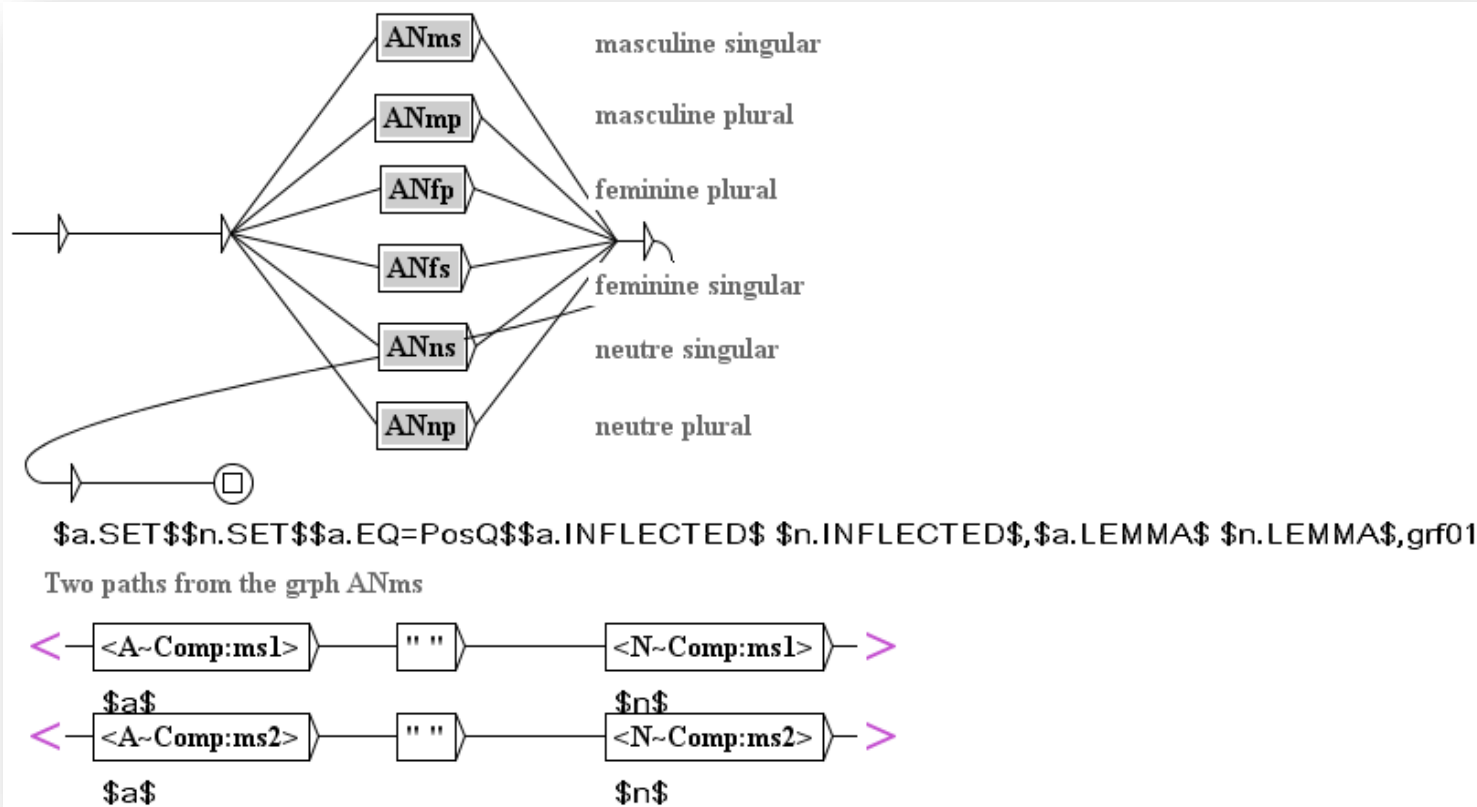


багер кашикар, багера кашикара, багеру кашикару, багерима кашикарима,....

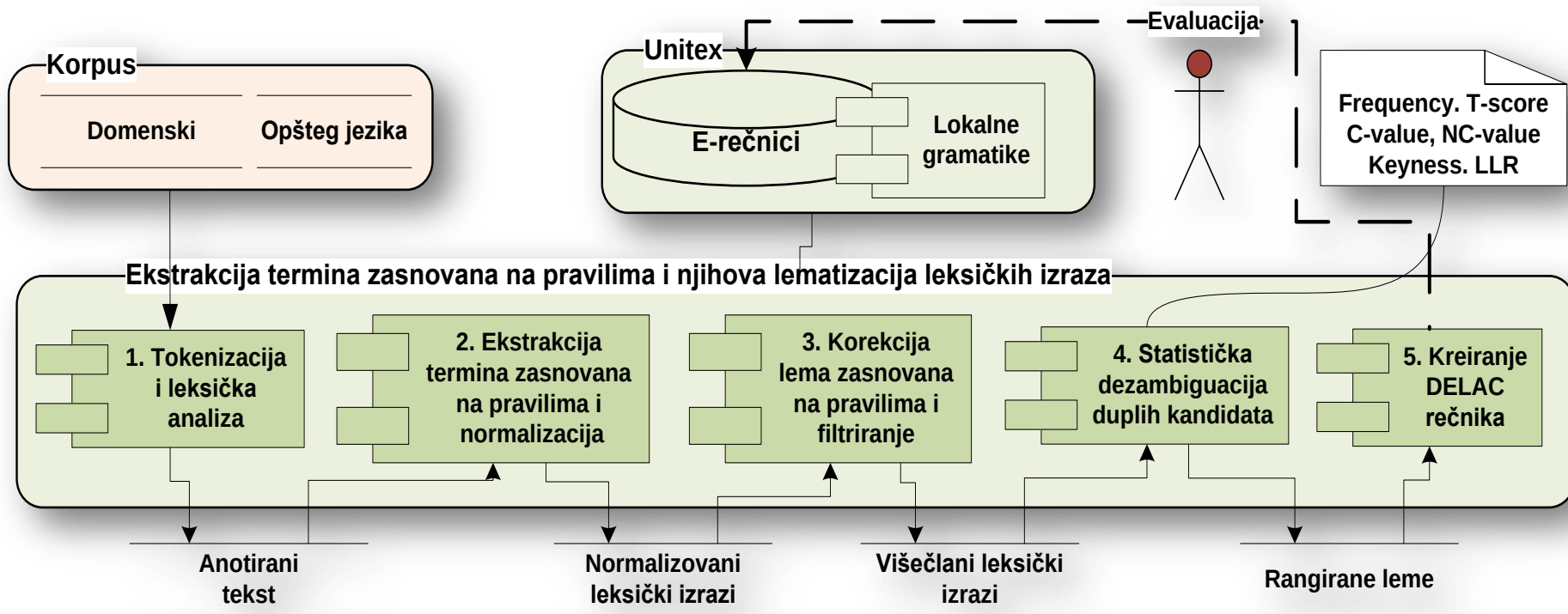
Екстракција термина AXN



- Подграфовима се обезбеђује слагање придева и именица
- Речничка променљива се користи за FST излаз у облику $\$a.LEMMA\$$ враћа лему препознате речи $\$a\$$ чиме остварује просту лематизацију речи



Процес екстракције и лематизације термина



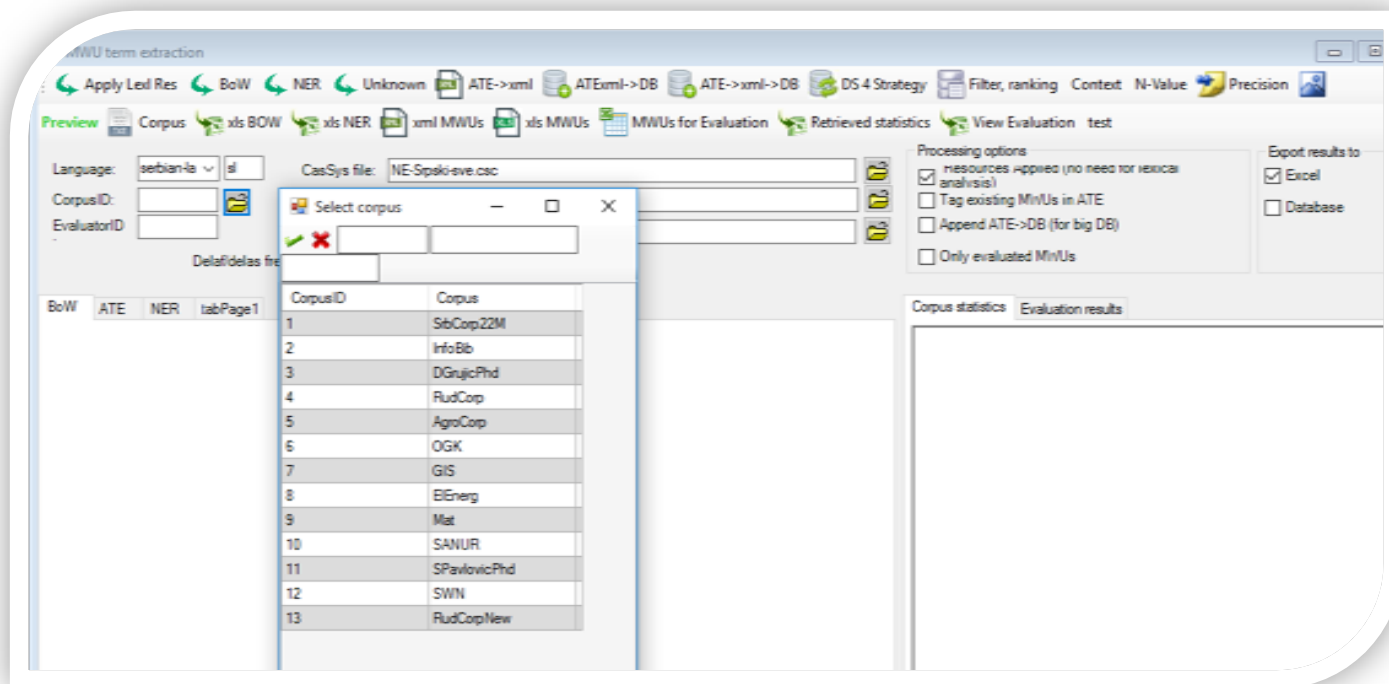
Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić and Aleksandra Trtovac. „Rule based automatic multi-word term extraction and lemmatization.” In: *10th edition of the Language Resources and Evaluation Conference (LREC)*, 23-28 May 2016, Portorož (Slovenia) (у штампи)

- Архитектура софтверског решења заснована је на веб сервисима, па омогућава другим апликацијама да користе неке од ових сервиса (индексирање или проналажење информација у документу) за потребе екстракције термина
- Апликација која се тренутно користи развијена је и тестирана у окружењу Windows док је нова веб апликација тренутно у изради, а она треба да понуди екстракцију термина из различитих домена ширем кругу корисника

Приказ развијеног решења



- Графови су интегрисани у алат Лексимир, где се врши руковање корпусима за специфичне домене (математика, геологија, библиотекарство, енергетика,...)



Приказ развијеног решења



- ... руковање резултатима екстракције, рачунање различитих мера за рангирање кандидата, филтрирање резултата, генерисање комплетних лема и припрема за евалуацију.

The screenshot shows the 'MWU term extraction' application window. The interface includes a menu bar with options like 'Apply Lex Res', 'BoW', 'NER', 'Unknown', 'ATE->xml', 'ATExml->DB', 'ATE->xml->DB', 'DS 4 Strategy', 'Filter, ranking', 'Context', 'N-Value', and 'Precision'. Below the menu bar is a toolbar with icons for 'Preview', 'Corpus', 'xls BOW', 'xls NER', 'xml MWUs', 'xls MWUs', 'MWUs for Evaluation', 'Retrieved statistics', and 'View Evaluation test'. The main workspace is divided into several sections:

- Language:** serbian-la, sl
- CorpusID:** 13
- EvaluatorID:** (empty)
- CasSys file:** NE-Sipki-sve.csc
- Project folder:** D:\Cvetana\MojUnitex\Serbian-Latin\Corpus\RudCorp
- Main file name (.txt, .snt):** RudCorp.txt
- Delatidelas frequency threshold:** 0, 7
- Processing options:**
 - ☒ resources Applied (no need for lexical analysis)
 - ☐ Tag existing MWUs in ATE
 - ☐ Append ATE->DB (for big DB)
 - ☐ Only evaluated MWUs
- Export results to:**
 - ☒ Excel
 - ☐ Database
- BoW ATE NER tabPage1**
 - Path with graphs for term extraction:** D:\Cvetana\MojUnitex\Serbian-Latin\Graphs\TermExtraction\
 - ☒ All NE Categories
 - ☒ 01AXNIA-PosQ_N.fst2
 - ☒ 02ZXNINepoz_N.fst2
 - ☒ 03N2XN2X.fst2
 - ☒ 04N4XN4X.fst2
 - ☒ 05AXN2XAXN2X.fst2
 - ☒ 06N0XN0Kontekst_N.fst2
 - ☒ 07AXXNAA-PosQ_N.fst2
 - ☒ 08N6XN6X.fst2
 - ☒ 09AXN4XAXN4X.fst2
 - ☒ 10ZXAXN2XAXN.fst2
- Duplicate elimination strategy:**
 - ☒ Without elimination of duplicates
 - ☐ Duplicates Elimination by Graph Order
 - ☐ Duplicates Elimination by Frequency
- Corpus statistics Evaluation results:**
 - Token number: 1657953
 - Word number: 625105
 - Sentence number: 32633
 - Domain: 10 knjiga iz rударства, 2 projekta 51 rad iz Podzemnih radova

Корпуси



- За поређење општи речника коришћен корпус 22 милиона речи из Корпуса савременог српског језика.
- Корпус из области библиотекарства (Александра Тртовац)
 - 4 књиге и 72 чланка из часописа „Инфотека” (22.948 реченица и 480.588 облика простих речи)
 - За евалуацију задржане леме са фреквенцијом >3
 - Од 2.655 лема, одбачено је 72 дупликата, и 2.583 лема покрива 24.699 препознатих облика који су задржани. За те облике прорачунате су мере и они су према њима рангирани и прослеђени на евалуацију.

Корпуси



- Корпус из области рударства (Биљана Лазић)
 - 10 књига, 2 пројекта и 51 рад из часописа „Подземни радови” (32,633 реченица и 625,105 облика простих речи)
 - За евалуацију задржане леме са фреквенцијом >7
 - Од 3,498 лема, одбачено је 1,540 примењујући правила за дезамбигуацију, и преосталих 1,958 лема покрива 33,153 препознатих облика који су задржани, од чега 4,067 различитих.
 - За те термине прорачунате су мере и они су према њима рангирани и прослеђени на евалуацију.
- Корпус из области математике (Марија Радојичић)
 - Прикупљене 243 тезе на српском и хрватском, 15 књига и 212 скрипти, области математика и информатика
 - Подкорпус само математика, српски, латиница: 1,802,519 облика простих речи од којих 118,027 различитих.
- Корпус из области електричне енергије (Тања Ивановић), просторног планирања (Милена Милинковић), геологије (Биљана Лазић), е-учење (материјал са ВАЕКТЕЛ пројекта)

Приказ развијеног решења



Врећа речи

Lema	PC	Frequ	Lema	PC	Frequ
funkcija	N	17303	skup	A	5834
tačka	N	10526	ravan	A	3978
skup	N	6847	prav	A	3557
red	N	5801	dat	A	3097
prostor	N	5124	sam	A	2835
jednačina	N	5113	kriv	A	2829
integral	N	4653	neprekidan	A	2703
broj	N	4254	drugi	A	2540
vrednost	N	4189	sledeći	A	2426
teorem	N	3880	jednak	A	2111
teorema	N	3866	prvi	A	2092
ravan	N	3843	odnosni	A	1983
uslov	N	3673	oblast	A	1907
niz	N	3616	diferencijalan	A	1795
			diferencijalni	A	1795
			isti	A	1758
			prethodan	A	1754
			proizvoljan	A	1600

Непрепознате речи

token	freq
tenzor	423
arcsin	330
arctan	326
akko	308
neprazan	205
dualnosti	177
tenzora	169
sinh	169
cosh	167
podniz	155
arctg	149
martingal	149
raširenje	140
kvadrature	139
celobrojnog	129
izometrijska	123
arccos	118
izometrijskih	116
podskupova	115

ET: рударство



		sin		plu	
Graph		Lemma	Forms	Lemma	Forms
grf01	AXN	570	11,845	481	8,200
grf02	2XN	6	55	5	53
grf03	N2X	701	11,330	613	6,931
grf04a	N4X	148	2,113	107	1,359
grf04b	N4X	122	2,002	106	1,171
grf05	AXN2X	63	800	33	205
grf06	NXN	299	4,249	134	1,751
grf07	AXAXN	5	37	4	36
grf08a	N6X	6	77	6	46
grf08b	N6X	8	103	8	72
grf08c	N6X	9	103	7	81
grf09a	AXN4X	17	195	10	124
grf09b	AXN4X	9	86	7	47
grf10	2XAXN	8	62	6	100
Total		1,971	33,057	1,527	20,176

Total (plu+sin)	3,498	53,233
-----------------	-------	--------

Термини (леме) које
су екстраховани
графовима

Пример екстракција термина



Graph	Num	Recognized form	Frequency	Temporary lemma	Lemma
grf01	plu	električne energije	85	električni energija	električne energije
grf01	sin	električna energija	10	električni energija	električna energija
		električne energije	85		
		električnom energijom	8		
		električnu energiju	5		
DELAC		električna{električni.A2:aefs1g) energija{energija.N600:fs1q),NC_AXN			
DELACF		električnoj energiji,električna energija.N:fs7q			
		električne energije,električna energija.N:fp1q			
		električnih energija,električna energija.N:fp2q			
		električnim energijama,električna energija.N:fp3q			

Филтрирање листе кандидата предикцијом класе



Правила за одређивање флективне класе ВЛЈ,
<http://poincare.matf.bg.ac.rs/~cvetana/biblio/CLA-2011-fin.pdf>
(ТЕИ документ)

```
<CompRules ID="CompoundsStat7_Ver5" StratType="T2">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Strategija za detekciju svojstava slozenih reci - imenice i pridevi</title>
        <respStmt>
          <resp>autor</resp>
          <name>Cvetana Krstev</name>
        </respStmt>
        <respStmt>
          <resp>formalizacija XML pravila</resp>
          <name>Ranka Stanković</name>
        </respStmt>
      </titleStmt>
      <publicationStmt>
      <sourceDesc>
      <revisionDesc> Verije različitih strategija okt.2007-maj.2008.
    </fileDesc>
  </teiHeader>
  <CondItems>
    <CondItem Code="$FLXN" Condition="FLX,N679,N681,N683,N685,N687,N613a,N743"/>
    <CondItem Code="$PRE" Condition="startsWithUpperCase"/>
  </CondItems>
  <Rules WordNo="2">
<!--
    <Rule ID="1" CFLX="NC AXN3F" CflxGroup="NC AXNF" Status="false" Desc="pridev/imenica; ne menja s ebroy; imenica menja broj">
    <Rule ID="2" CFLX="NC AXN3" CflxGroup="NC AXN" Desc="pridev imenica; ne menja se broj">
    <Rule ID="3" CFLX="NC AXNF" CflxGroup="NC AXN" Desc="pridev/imenica; imenica menja rod">
    <Rule ID="4" CFLX="NC AXN" CflxGroup="NC AXN" Desc="pridev/imenica">
      <RuleGenCond Example="nagazna mina">
        <Word ID="1" POS="A" Flex="true" Case="1" Anim="$a" Gen="$g" Num="s"/>
        <Word ID="2" POS="N" Flex="true" Case="1" Anim="$a" Gen="$g" Num="s" Cond="!$FLXN"/>
      </RuleGenCond>
    </Rule>
    <Rule ID="5" CFLX="NC NXN2m" CflxGroup="NC NXN" Desc="imenica/imenica; slazxu se po zxivo; ne slazxu po rodu; druga imenica se manja po padezxu ili ne">
      <RuleGenCond ID="1" Example="Kraljevina Maroko" Desc="preuzima rod od druge imenice, a prva imenica mora da ima drugaciji rod.">
        <Word ID="1" POS="N" Flex="true" Case="1" Anim="$a" Gen="!$g" Num="s" Cond="$PRE"/>
        <Word ID="2" POS="N" Flex="true" Case="1" Anim="$a" Gen="$g" Num="s" Cond="!$FLXN"/>
      </RuleGenCond>
    </Rule>
  </Rules>
</-->
```

Екстракција термина рударство



Graph		sin		plu	
		Lemma	Forms	Lemma	Forms
grf01	AXN	568	11,828	47	740
grf02	2XN	6	55		
grf03	N2X	668	10,993	8	34
grf04a	N4X	143	2,065	2	22
grf04b	N4X	122	2,002	1	9
grf05	AXN2X	63	800	3	28
grf06	NXN	265	3,903		
grf07	AXAXN	5	37	1	8
grf08a	N6X	6	77		
grf08b	N6X	6	83	1	7
grf08c	N6X	9	103		
grf09a	AXN4X	17	195		
grf09b	AXN4X	8	78		
grf10	2XAXN	8	62	1	24
Total		1,894	32,281	64	872

Total (plu+sin)	1,958	33,153
-----------------	-------	--------

Термини (леме) које
су прослеђене на
ручну евалуацију

Статистичке мере



N - величина корпуса, f_A – број појављивања речи A , f_B – број појављивања речи B , f_{AB} – број ко-појављивања речи A и B

- Релативна фреквенција $\frac{f_{AB}}{f_A} \cdot 100$
- Кључност речи (Keyness) - Релевантност за домен, користи фреквенција нормализована на милион речи, n је параметар глаткости, подразумевано 1
Kilgariff: "Getting to know your corpus". Proceedings of Text, Speech and Dialogue 2012, Lecture Notes in Computer Science. Springer, 2012.

$$\frac{f p m_{\text{домен}} + n}{f p m_{\text{општи}} + n}$$

- Log-likelihood, Dunning, Accurate Methods for the Statistics of Surprise and Coincidence, Computational Linguistics 19:1 1993

$$2 \cdot (x l x(f_{AB}) + x l x(f_A - f_{AB}) + x l x(f_B - f_{AB}) + x l x(N) + x l x(N + f_{AB} - f_A - f_B) - x l x(f_A) - x l x(f_B) - x l x(N - f_A) - x l x(N - f_B)), \text{ где } x l x(f) \text{ је } f \ln(f)$$

Статистичке мере



C-value

$$termhood(a) = f(a) - \sum_{b \in T_a} f(b)$$
$$C-value(a) = \begin{cases} \log_2 |a| \cdot f(a) & a \text{ is not nested,} \\ \log_2 |a| (f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)) & otherwise \end{cases}$$

$$NC-value(a) = 0.8 C-value(a) + 0.2 \sum_{b \in C_a} f_a(b) weight(b)$$

$weight(b) = t(w)/n$; w је контекстна реч (N, A, V), а $t(w)$ број термина у којима се појављује, n је укупан број термина који се анализирају

LLR

Поређење 2 корпуса (доменски и општи)

$$E1 = N2 * (f1(a) + f2(a)) / (N1 + N2)$$

$$E2 = N1 * (f1(a) + f2(a)) / (N1 + N2)$$

$$LLR = 2 * (f1(a) * \ln(f1(a)/E1) + f2(a) * \ln(f2(a)/E2))$$

$N1, N2$ – број биграма у корпусима, $f1(a), f2(a)$ фреквенција у корпусима

Статистичке мере



Graph	Num	Lemma	Eliminates lemma	Measures					Rank				
				Freq	CValue	TScore	LLR	Keyness	Freq	CValue	TScore	LLR	Keyness
grf01	sin	mineralna sirovina	mineralne sirovine	736	736	27129	5144	707	1	1	1	1	1
grf01	sin	površinski kop	površinski kopovi	305	230	17464	1925	161	2	4	2	3	17
grf01	sin	toplotni tok		258	249	16062	1803	335	4	3	4	4	3
grf01	sin	toplotna provodljivost	toplotne provodljivosti	236	222	15362	1649	312	5	5	5	5	4
grf03	sin	kvalitet uglja	kvaliteti uglja	297	289	17234	2076	375	3	2	3	2	2
grf03	sin	nivo buke	nivo buka;nivoi buke	118	118	10863	825	171	15	18	15	13	11
grf01	sin	površinska povreda	površinske povrede	116	104	10770	811	169	17	27	17	15	13
grf09a	sin	površinska povreda u predelu	površinske povrede u predelu	24	48	4899	168	39	272	90	272	262	223
grf05	sin	površinska povreda potkolenice	površinske povrede potkolenic	8	13	2828	56	14	1660	867	1660	1544	1521
grf01	sin	električna energija	električne energije	121	121	11000	128	4	13	15	13	383	1739
grf01	sin	električna provodljivost	električne provodljivosti	46	46	6782	321	72	97	94	97	87	80
grf03	sin	vibracija šake	vibracije šake	54	34	7348	377	83	74	171	74	66	62
grf04b	sin	vibracija šake ruke		54	63	7348	377	83	75	56	75	67	63
grf08b	sin	sindrom vibracija šake ruke	sindromi vibracija šake ruke	16	32	4000	112	26	539	190	539	514	496
grf04b	sin	sindrom vibracija šake	sindromi vibracija šake	16	0	4000	112	26	551	1690	550	496	489
grf04b	sin	merenje vibracija šake	merenja vibracija šake	15	0	3873	105	25	577	1691	575	557	53
grf08b	sin	merenje vibracija šake ruke	merenja vibracija šake ruke	15	30	3873	105	25	584	204	584	547	

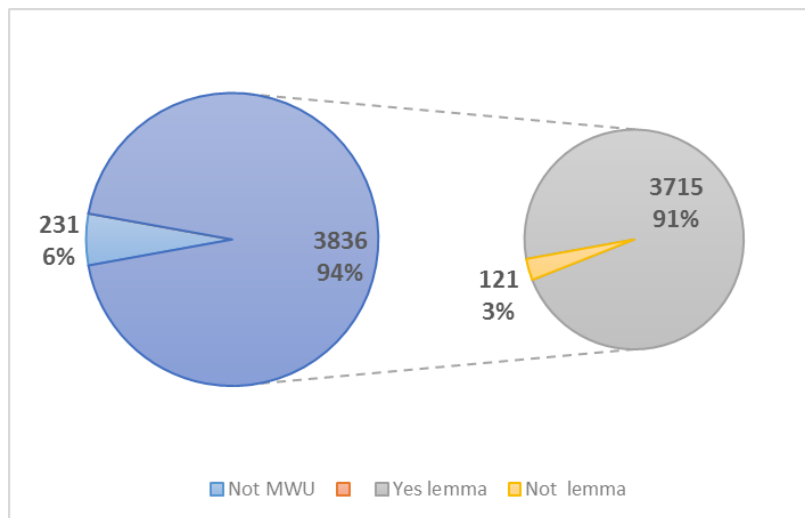
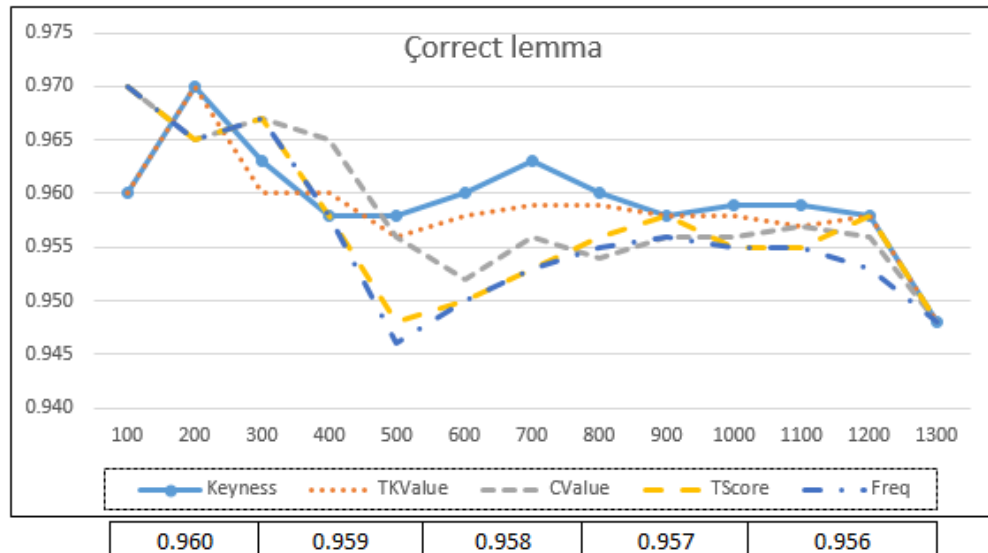
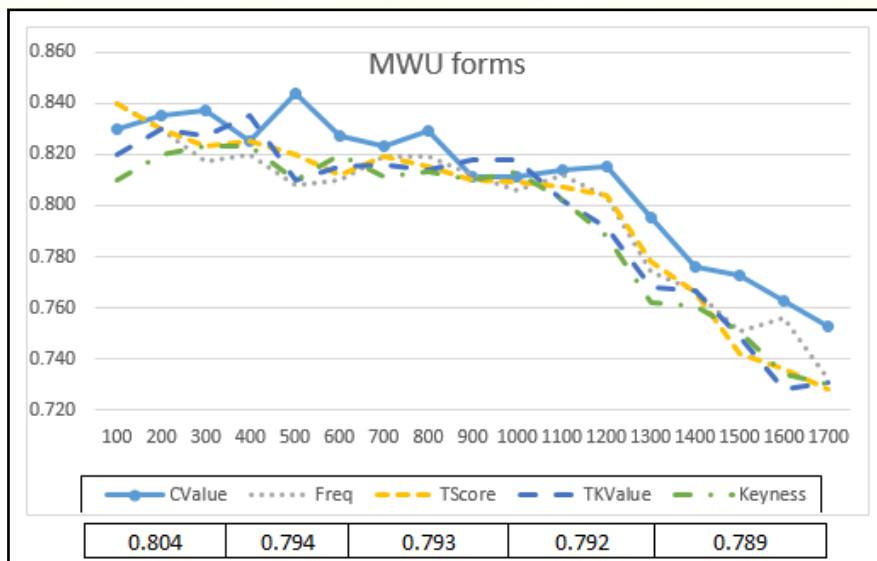
Екстракција термина рударство



			MWU OK		Lemma OK		
			Graph	No	%	No	%
plu	grf01	AXN	42	89	8	19	
	grf03	N2X	7	88	3	43	
	grf04a	N4X	1	50		0	
	grf04b	N4X	1	100		0	
	grf05	AXN2X	3	100		0	
	grf07	AXAXN	0	0		0	
	grf08b	N6X	1	100		0	
	grf10	2XAXN	1	100		0	
sin	grf01	AXN	553	97	552	100	
	grf02	2XN	6	100	5	83	
	grf03	N2X	608	91	599	99	
	grf04a	N4X	102	71	97	95	
	grf04b	N4X	111	91	106	95	
	grf05	AXN2X	58	92	58	100	
	grf06	NXN	29	11	13	45	
	grf07	AXAXN	4	80	4	100	
	grf08a	N6X	5	83	5	100	
	grf08b	N6X	6	100	5	83	
	grf08c	N6X	7	78	7	100	
	grf09a	AXN4X	10	59	10	100	
	grf09b	AXN4X	7	88	7	100	
	grf10	2XAXN	8	100	8	100	
	Total		1570	83	1487		

Термини (леме) које
су прошле ручну
евалуацију

Резултат евалуације



Генерисане пуне леме



based procedure for compound dictionary production

Prediction Disambiguation Export results Dictionary production DIF Debug file Selected rule Select all SIN2_FST FSD

File with compounds list
D:\WS4LR_Res\Delas\Compounds\CompoundsList Test.bt

Strategy
D:\WS4LR_Res\Delas\Compounds\CompoundsStrat7_Ver5-3.xml

Result selection preview
☒ Compounds witch are not in dictionary
☒ Compounds witch are in dictionary
☐ Only first prediction for each compound
☐ Open bt file with prediction results

Automatic term extraction
CorpusID:

Save prediction in DB
Read evaluated from DB

Results with rules details Results

MinU CFLX SIN2 OKNOTOK

Select	MinU	CFLX	SIN2	Count	CountNOTOK			
☐	visokozskolske bi...	visokozskolske[visokozskolski.A2.aefp1g] biblioteke[bibliote...	NC_AXN3	+SIN=AXN(plu)	OK	AXN(plu)	1	0
☐	visokozskolske u...	visokozskolske[visokozskolski.A2.aefp1g] ustanove[ustano...	NC_AXN3	+SIN=AXN(plu)	OK	AXN(plu)	1	0
☐	vremenski okvir	vremenski[vremenski.A2.aemp1g] okvir[okvir.N1.mp1q]	NC_AXN3	+SIN=AXN(plu)	OK	AXN(plu)	1	0
☐	vremenski periodi	vremenski[vremenski.A2.aemp1g] period[period.N1.mp1q]	NC_AXN3	+SIN=AXN(plu)	OK	AXN(plu)	1	0
☐	zakonski propisi	zakonski[zakonski.A2.aemp1g] propisi[propis.N1.mp1q]	NC_AXN3	+SIN=AXN(plu)	OK	AXN(plu)	1	0
☐	zavicyajni fondovi	zavicyajni[zavicyajan.A7.aemp1g] fondovi[fond.N81.mp1q]	NC_AXN3	+SIN=AXN(plu)	OK	AXN(plu)	2	0
☐	zavicyajni fondovi	zavicyajni[zavicyajni.A2.aemp1g] fondovi[fond.N81.mp1q]	NC_AXN3	+SIN=AXN(plu)	OK	AXN(plu)	2	0
☐	aerodomska bibli...	aerodomska[aerodomski.A2.aefs1g] biblioteka[biblioteka.N...	NC_AXN	+SIN=AXN(sjn)	OK	AXN(sjn)	1	0
☐	akademsko obra...	akademsko[akademski.A2.aens1g] obrazovanje[obrazova...	NC_AXN	+SIN=AXN(sjn)	OK	AXN(sjn)	1	1
☐	alfabetски каталог	alfabetски[alfabetски.A2.adms1g] каталог[katalog.N11.ms1q]	NC_AXN	+SIN=AXN(sjn)	OK	AXN(sjn)	1	0
☐	alfabetски принцип	alfabetски[alfabetски.A2.adms1g] принцип[princip.N1.ms1q]	NC_AXN	+SIN=AXN(sjn)	OK	AXN(sjn)	1	0
☐	alfabetски редослед	alfabetски[alfabetски.A2.adms1g] редослед[redosled.N1.ms1q]	NC_AXN	+SIN=AXN(sjn)	OK	AXN(sjn)	1	0
☐	analiticka obrada	analiticka[analiticky.A2.aefs1g] obrada[obrada.N600fs1q]	NC_AXN	+SIN=AXN(sjn)	OK	AXN(sjn)	1	0
☐	arhivska gradba	arhivska[arhivski.A2.aefs1g] gradba[gradba.N600fs1q]	NC_AXN	+SIN=AXN(sjn)	OK	AXN(sjn)	1	0
☐	automatska analiza	automatska[automatski.A2.aefs1g] analiza[analiza.N600fs1q]	NC_AXN	+SIN=AXN(sjn)	OK	AXN(sjn)	1	0
☐	automatska dodela	automatska[automatski.A2.aefs1g] dodela[dodela.N600fs1q]	NC_AXN	+SIN=AXN(sjn)	OK	AXN(sjn)	1	0
☐	automatska ekstr...	automatska[automatski.A2.aefs1g] ekstrakcija[ekstrakcija.N...	NC_AXN	+SIN=AXN(sjn)	OK	AXN(sjn)	1	0
☐	automatsko inde...	automatsko[automatski.A2.aens1g] indeksiranje[indeksiran...	NC_AXN	+SIN=AXN(sjn)	OK	AXN(sjn)	1	1

Математика



УНИВЕРЗИТЕТ У БЕОГРАДУ
РУДАРСКО - ГЕОЛОШКИ ФАКУЛТЕТ

SRPSKI

Ažuriranje

- terminološki rečnik
- bibliotekarstvo
- geoinformatika
- inženjerstvo nafte i gasa
- jezički resursi i tehnologije
- matematika
 - algebra
 - analiza
 - granična vrednost niza
 - Tejlorova formula
 - aritmetika
 - geometrija
- Novi čvor
- Novi čvor2
- obrazovanje
- računarstvo i informatika
- rudarsko inženjerstvo
- statistika
- zaštita životne sredine

Status: Zaključan

Naziv: Tejlorova formula

Sinonimi: formula Tejlor, tejlorova formula u Šlemilih-Rošovom ostatku

Skracenica: Tejlorova f.

Opis:

Neka funkcija $f(x)$, neprekidna sa svim svojim izvodima do n -og reda, zaključno u nekoj okolini U tačke c , ima izvod $f^{(n+1)}$ -og reda u toj okolini. Ako je $x \in U$ i $p \in \mathbb{N}$, onda važi formula:

$$f(x) = f(c) + \frac{f'(c)}{1!}(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \dots + \frac{f^{(n)}(c)}{n!}(x-c)^n + R_n(x)$$

Neka funkcija $f(x)$, neprekidna sa svim svojim izvodima do n -og reda, zaključno u nekoj okolini U tačke c , ima izvod $f^{(n+1)}$ -og reda u toj okolini. Ako je $x \in U$ i $p \in \mathbb{N}$, onda važi formula:

$$f(x) = f(c) + \frac{f'(c)}{1!}(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \dots + \frac{f^{(n)}(c)}{n!}(x-c)^n + R_n(x)$$

$$R_n(x) = \left(\frac{x-c}{x-\xi}\right)^p \frac{(x-\xi)^{n+1}}{p \cdot n!} f^{(n+1)}(\xi),$$

za svako ξ koje je između c i x . Prikazana formula se naziva Tejlorovom formulom u Šlemilih-Rošovom ostatku.

```

<langSet xml:lang="sr">
  <ntig>
    <termGrp>
      <term>Tejlorova formula</term>
      <termNote type="termType">entryTerm</termNote>
    </termGrp>
    <descripGrp>
      <descrip type="definition">Neka funkcija  $f(x)$ , neprekidna sa svim svojim izvodima do  $n$ -og reda, zaključno u nekoj okolini  $U$  tačke  $c$ , ima izvod  $f^{(n+1)}(c) = f^{(n)}(c) + \frac{f^{(n)}(c)}{1!}(x-c) + \frac{f^{(n)}(c)}{2!}(x-c)^2 + \dots + \frac{f^{(n)}(c)}{n!}(x-c)^n + R_n(x)$ , gde  $R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}$  za svako  $\xi$  koje je između  $c$  i  $x$ . Prikazana formula se naziva Tejlorovom formulom u Šlemilih-Rošovom ostatku.</descrip>
    </descripGrp>
  </ntig>
</langSet>
<langSet xml:lang="en">
  <ntig>
    <termGrp>
      <term>Taylor's formula</term>
      <termNote type="termType">entryTerm</termNote>
    </termGrp>
    <descripGrp>
      <descrip type="definition">Let  $f(x)$ , and its first  $n$  derivatives be continuous, and let  $f^{(n+1)}(x)$  exist on the region  $U$  of point  $c$ . If  $x \in U$  and  $p \in \mathbb{N}$ , then:  $f(x) = f(c) + \frac{f'(c)}{1!}(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \dots + \frac{f^{(n)}(c)}{n!}(x-c)^n + R_n(x)$ , where  $R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}$  for all  $\xi$  between  $c$  and  $x$ . Hence this formula is named Taylor's formula in Scholmic-Roche residuum.</descrip>
    </descripGrp>
  </ntig>
</langSet>

```

Проблеми у препознавању



- Препознати делови термина који имају више од 4 компоненте, нпр.
 - мотор механизма за потискивање (N6X) и
механизам за потискивање кашике (N6X)
..... уместо правог термина
мотор механизма за потискивање кашике (N8X)
- Препознато „Народна библиотека“ у именованом ентитету „Народна библиотека Србије“
графови избацују речи са маркером +Nprop, за именоване ентитете се користи друга процедура

Проблеми у препознавању



Хомографија

- „облога траке“ (N2X, NNgi) компоненте су облога и трака, при чему су неки флективни облици хомоними са другим именицом облог
- Само облици „облого траке“ и „облогама траке“ су препознати у тексту и њима лема може бити придружена
- Али, да је „облоге траке“ препознато, онда би два обрасца (N2X, NXN) била придружена са различитим лемама. Уз то, ако би имали и „облози траке“ онда би се одбацио образац N2X и „облога трака“ за образац NXN

MWT inflected form	Grammatical values	Struc.	First constituent	Second constituents
obloga trake	s_nom+ p_gen	N2X	obloga_s_nom	traka_s/p_gen
			obloga_p_gen	traka_s/p_gen
			oblog_s/p_gen	traka_s/p_gen
		NXN	oblog_s_gen	traka_s_gen
obloge trake	s_gen+ p_nom/ acc/voc	N2X	obloga_s_gen+ p_nom/ acc/voc	traka_s/p_gen
			oblog_p_acc	traka_s/p_gen
		NXN	oblog_p_acc	trak_p_acc
			oblog_p_acc	traka_p_acc
			obloga_p_acc	trak_p_acc
			obloga_s_gen+ p_nom/acc	traka_s_gen+ p_nom/acc
oblozi trake	s_dat/loc	N2X	obloga_s_dat/ loc	traka_s/p_gen
		NXN	oblog_p_nom/vo c	traka_p_nom/ voc
oblogu trake	s_acc	N2X	obloga_s_acc	traka_s/p_gen
			oblog_s_dat/ loc	traka_s/p_gen
oblogotrake	s_voc	N2X	obloga_s_voc	traka_s/p_gen
oblogom trake	s_ins	N2X	obloga_s_ins	traka_s/p_gen
			oblog_s_ins	traka_s/p_gen
oblogama trake	p_dat/ ins/loc	N2X	obloga_p_dat/ ins/loc	traka_s/p_gen

Дискусија и закључак



- Представљена екстракција термина је заснована на лексичким ресурсима и правилима
- Приказани резултати показују да се екстракција термина и уопште вишечланих лексичких јединица може успешно аутоматизовати
- Уз то је могуће и одређивање флективне класе и генерисање записа у форми која одговара записима у електронском морфолошком речнику, што је тежак задатак за словенске језике, па и српски
- Представљеним решењем се значајно убрзава развој терминолошких и морфолошких речника за српски

Шта је још у плану?



- Развој веб апликације која треба да подржи цео презентовани процес, уз интеграција предикције флективне класе у веб апликацију
- Екстракција и производња записа у DELA формату и за глаголе, као што је урађено за именице
- Детаљна евалуација сваког корака како би се омогућило унапређење процедуре како би се смањила интервенција корисника и још више аутоматизовао процес прикупљања терминологије и њен опис
- Израда апликације за веб-евалуацију уместо садашње засноване на екселу