811.163.41'322.2:622
https://doi.org/10.18485/msc.2018.47.3.ch7

Милош Б. УТВИЋ*
Филолошки факултет
Универзитета у Београду
Иван М. ОБРАДОВИЋ**
Ранка М. СТАНКОВИЋ
Александра Ђ. ТОМАШЕВИЋ
Биљана Ђ. ЛАЗИЋ
Рударско-геолошки факултет
Универзитета у Београду

ИЗГРАДЊА СПЕЦИЈАЛНИХ КОРПУСА САВРЕМЕНОГ СРПСКОГ ЈЕЗИКА НА ПРИМЕРУ КОРПУСА ИЗ ОБЛАСТИ РУДАРСТВА

У овом раду се разматра изградња специјалног корпуса стручних текстова на српском језику из области рударства на Рударско-геолошком факултету Универзитета у Београду. Специјални корпус из области рударства је произашао из дигиталне библиотеке ROmeka@RGF, најпре као средство за унапређивање претраге дигиталне библиотеке захваљујући лингвистичкој анотацији, а потом и као ресурс за различита лингвистичка и терминолошка истраживања, укључујући екстракцију термина и друге задатке из области језичког инжењерства. У раду се пореде могућности неколико верзија корпуса језика струке из области рударства, односно коришћених софтверских пакета за креирање, управљање и претраживање корпуса.

Кључне речи: корпус, рударство, претраживање информација, екстракција термина

1. Увод

Поред општих корпуса, намењених истраживању језика у целини, у лингвистици, рачунарској лингвистици и обради природног језика користе се и специјал(изова)ни корпуси. Иако се по називу специјалних корпуса може закључити да сваки такав корпус има посебну намену, ипак се у пог-

* misko@matf.bg.ac.rs
** {ivan.obradovic, aleksandra.tomasevic, ranka.stankovic, biljana.lazic}@rgf.bg.ac.rs

леду примене специјалних корпуса могу уочити две важне заједничке карактеристике:

- специјални корпуси се користе као језички ресурси за изучавање специфичног лингвистичког феномена, па се тако разликују лексикографски корпуси, граматички корпуси, дијалекатски корпуси, регионални корпуси, нестандардни корпуси, корпуси језика као нематерњег, корпуси струке (енг. *domain specific corpora*) итд.
- специјални корпуси имају важну улогу у оквиру аутоматске обраде природног језика као помоћна средства током обуке софтверских алатки за решавање задатака из области језичког инжењерства (енг. *natural language engineering*), на пример као корпуси за обуку алатки за аутоматску морфосинтаксичку анализу текста, аутоматско препознавање и генерисање говора, екстракцију термина, анализу осећања итд.

Из наведених разлога постоји потреба да се и у оквиру области рударства, као сложене и мултидисциплинарне индустријске гране, изграде специјални корпуси који би се користили као језички ресурси и за потребе лингвистичких истраживања и за потребе истраживања у оквиру језичког инжењерства. Специјални корпус стручних текстова на српском језику из области рударства свакако пружа основу и за развој корпуса језика струке у оквиру других инжењерских области, као што су машинство, грађевина, енергетика, геологија.

У одељку 2 овога рада описан је систем за управљање рударском пројектном документацијом у оквиру којег је развијен Рударски корпус, специјални корпус стручних текстова на српском језику из области рударства. Одељак 3 даје преглед фаза у изградњи Рударског корпуса са посебним нагласком на анотацију корпуса. Одељак 4 илуструје могућности претраживања Рударског корпуса. Напослетку, одељак 5 наводи преглед закључака и смернице будућег развоја Рударског корпуса.

2. Систем за управљање рударском пројектном документацијом

У склопу истраживања на Рударско-геолошком факултету Универзитета у Београду осмишљен је систем за управљање рударском пројектном документацијом (у даљем тексту: систем) заснован на језичким технологијама. Систем покрива целокупан животни циклус рударских активности, почевши од фазе анализе, преко истраживања, процене добијених резултата, пројектовања, производње и затварања рудника. Тиме се постиже неопходан ниво сложености који омогућава позиционирање ризика и контролу пројектног циклуса и координацију рада на сложеним пројектима са више уговорних страна и неколико заинтересованих страна.

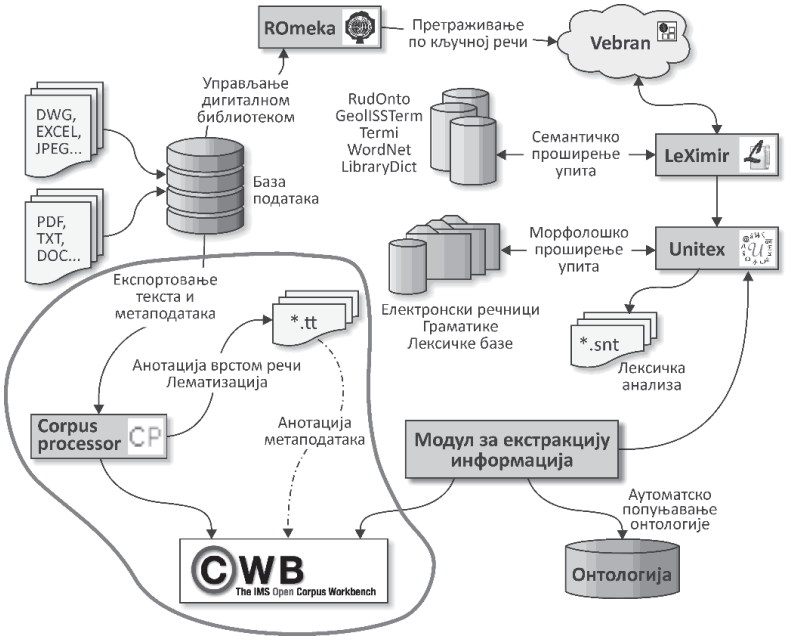
Основни елементи система за управљање документима у електронском облику помоћу језичких технологија су:

- дигитална библиотека — централни репозиторијум за складиштење рударске пројектне документације,
- корпус стручних текстова на српском језику из области рударства,
- лексички и терминолошки ресурси,
- онтологија.

Да би се у оквиру система за управљање рударском пројектном документацијом обезбедили проналажење и екстракција информација, развијена су два главна ресурса који садрже релевантну документацију: дигитална библиотека и пратећи корпус сакупљених и обрађених текстова из ове библиотеке. Поред тога, коришћени су бројни језички ресурси (Крстев и др. 2008), алати и технике (Крстев 2008), интегрисани у систем (Слика 1). Систем се састоји од различитих типова компоненти: развојне компоненте за припрему, побољшање и допуњавање ресурса (на пример Leximir, Corpus Preprocessor), као и компоненте које се користе за екстракцију и проналажење информација (на пример IMS Open Corpus Workbench, Unitex, ROmeka@RGF).

Слика 1 приказује систем за управљање рударском пројектном документацијом у целини, а затвореном линијом је издвојен део система који се односи се на концепт управљања самим документима, анотацију докумената и генерисање Рударског корпуса.

Слика 1. **Систем управљања документацијом заснован на језичким технологијама**



Систем за управљање рударском пројектном документацијом се одликује флексибилношћу и једноставношћу за коришћење и примену. Подржава претраживање кључних речи датих у упиту у свим граматичким формама, независно од писма (ћирилица, латиница). Поред тога, систем омогућава проширење упита на синониме, као и претраживање помоћу лексичких маски. Систем омогућава систематизацију, структурирање и складиштење законских прописа заједно са пројектном документацијом како би се, у каснијој фази, обезбедило унакрсно упоређивање и провера усклађености сегмента пројектне документације са важећим прописима.

У овом раду се описује изградња и примена Рударског корпуса као елемента система за управљање рударском пројектном документацијом, док је интегрални систем описан у раду (Томашевић и др. 2018). Како је корпус рударских текстова сачињен од докумената из домена рударства и заштите животне и радне средине, може се рећи да је намењен специфичном доменском истраживању и да се сврстава у специјализоване корпусе, за разлику од општих корпуса који се формирају да би представљали основу за произвољно лингвистичко истраживање.

3. Фазе изградње корпуса

Процес изградње Рударског корпуса је обухватио уобичајене фазе претходне обраде (енг. corpus preprocessing) у креирању електронских корпуса (Утвић 2014а: 59):

- прикупљање, дигитализацију и класификацију текстова за корпус;
- конверзију корпусних текстова у одговарајући формат електронског текста;
- лингвистичку обрада и анотацију електронских текстова за корпус:
- индексирање и компресију текстова корпуса.

У првој фази претходне обраде је искоришћена дигитална библиотека ROmeka@RGF[1], централни репозиторијума за складиштење рударске пројектне документације, тако што су сви текстуални документи дигиталне библиотеке, као и њихови одговарајући метаподаци, директно преузети. Укупно је преузето 172 документа, подељених по колекцијама и потколекцијама дигиталне библиотеке (Табела 1). У Рударском корпусу су тренутно заступљене три колекције: законска регулатива (правилници, закони, уредбе, смернице и стратегије: 94 документа), литература (докторске дисертације, уџбеници и монографије: 40 докумената) и пројектна документација (пројекти и студије: 38 докумената). Текстуални документи дигиталне библиотеке су претходно креирани комбиновањем сканирања и оптичког препознавања текста, при

[1] http://romeka.rgf.rs/

чему су из докумената уклоњени делови на страном језику, слике, референце и хипервезе (Томашевић и др. 2017).

Табела 1: Расподела докумената по колекцијама и потколекцијама

Колекција	**Потколекције**	**Број докумената**	**%**
Законска регулатива	Правилници	62	36,05%
	Закони	15	8,72%
	Уредбе	10	5,81%
	Смернице	5	2,91%
	Стратегије	2	1,16%
Литература	Докторске дисертације	28	16,28%
	Монографије	4	2,33%
	Уџбеници	5	2,91%
	Радови	3	1,74%
Пројектна документација	Пројекти	32	18,60%
	Студије	6	3,49%
Укупан број докумената		172	100,00%

Табела 2 илуструје величину Рударског корпуса изражену укупним бројем токена, типова, корпусних речи, корпусних типова[2], као и расподелу тих параметара корпусних текстова по колекцијама и потколекцијама којима припадају.

[2] „У корпусној лингвистици је уобичајено да се под *корпусном речју* подразумева низ карактера (корпусног) текста између два узастопна *сепаратора*, при чему се скуп сепаратора дефинише као скуп неалфанумеричких карактера. За корпусне речи и појединачне елементе скупа сепаратора (изузимајући белине [...заједнички назив за карактере размак, табулатор и знак за нови ред]) се користи заједнички назив *токени*, а репрезентација (корпусног) текста као низа узастопних токена се назива *токенизација*. Елементи скупа различитих корпусних речи, односно токена, називају се респективно *корпусни типови* (енг. *word types*), односно *типови* (енг. *types*)" (Утвић 2014б: 244).

Табела 2: Величина Рударског корпуса. Рачун је изведен на основу алата NoSketch Engine.

Врста текста	**Број корпусних речи**	**%**	**Број корпусних типова**	**%**	**Број различитих лема**	**%**
Пројектна документација	705.094	25,93%	35.362	35,22%	11.359	49,66%
Законска регулатива	459.912	16,91%	26.146	26,04%	9.968	43,58%
Докторске дисертације	976.667	35,92%	61.146	60,89%	16.102	70,39%
Литература	577.413	21,24%	45.609	45,42%	13.529	59,14%
Корпус	2.719.086	100,00%	100.414	100,00%	22.875	100,00%

У следећој фази текстуални документи дигиталне библиотеке су преведени у одговарајуће формате електронског текста, погодне за обраду коришћењем одговарајућих софтверских алата. Између осталог, коришћене су следеће конверзије текстова:

- конверзија из оригиналног кодног распореда полазног документа у кодну шему UTF-8,
- конверзија писма (из ћириличног у латинично, из латиничног у кодну шему *аурора*[3] и обратно),
- конверзија у формат вертикал(изова)ног текста.[4]

За све наведене конверзије је коришћен софтверски алат CorpusPreprocessor, развијен за претходну обраду текстова верзије СрпКор2013 Корпуса савременог српског језика, (Утвић 2014а: 272).

Анотација Рударског корпуса је обављена на неколико нивоа који су обрађени у посебним пододељцима: лингвистичка нотација (пододељак 3.1), структурна анотација (пододељак 3.2) и анотација метаподацима (пододељак 3.4). Алати за индексирање корпуса су описани у пододељку 3.3. Редослед излагања је условљен тиме што лингвистичка и структурна анотација не зависе од избора алата за индексирање корпуса, за разлику од анотације метаподацима.

3.1 Лингвистичка анотација

Тренутна верзија Рударског корпуса поседује делимичну морфолошку анотацију у смислу да су сваком токену придружене информације о врсти речи и леми на начин описан у (Утвић 2011). Морфолошка анотација је обављена аутоматски применом алата за обележавање врстом речи (енг. part-of-speech tagger) TreeTagger (Schmid 1997, Schmid 1999). С обзиром да TreeTagger захтева вертикални текст као улазни формат, сви корпусни текстови су пре аутоматске нотације конвертовани у формат вертикалног текста. Скуп обележја коришћених за анотацију врсте речи има 16 могућих вредности (10 ознака за врсту речи на српском, као и скраћенице, римске бројеве, интерпункције, префиксе, суфиксе и једна категорија за све остало). Речник који консултује TreeTagger приликом аутоматске анотације, који се у доку-

---

[3] Кодна шема *аурора* се користи као интерни кôд ради неутралисања утицаја различитих писама (ћирилица, латиница) у бројним дигиталним језичким ресурсима које је развила Група за језичке технологије Универзитета у Београду и Друштво за језичке ресурсе и технологије (JeРТех). Први пут је употребљена у истоименом програмском систему Душка Витаса (Витас 1981). „У коду аурора се карактери *Ш* (*Š*), *ш* (*š*), *Ж* (*Ž*), *ж* (*ž*), *Ћ* (*Ć*), *ћ* (*ć*), *Ч* (*Č*), *ч* (*č*), *Ђ* (*Đ*), *ђ* (*đ*), *Љ* (*Lj*), *љ* (*lj*), *Њ* (*Nj*), *њ* (*nj*), *Џ* (*Dž*), *џ* (*dž*) редом представљају следећим записима: *Sx*, *sx*, *Zx*, *zx*, *Cx*, *cx*, *Cy*, *cy*, *Dx*, *dx*, *Lx*, *lx*, *Nx*, *nx*, *Dy* и *dy*." (Утвић 2014б: 244).

[4] Формат вертикал(изова)ног текста је формат анотације који неретко користе софтверски алати за анотацију корпуса за свој улазни и излазни формат, као и алати за индексирање корпуса за свој улазни формат. „Вертикални формат користи запис по колонама, обично раздвојеним табулатором. У првој колони се наводе токени реченице, по један у сваком реду, док се у осталим колонама, у одговарајућем реду, записују придружене информације, при чему сваком нивоу анотације одговара по једна колона." (Утвић 2014а: 129).

ментацији алата TreeTagger још назива параметарска датотека језика (енг. language parameter file), дериват је система морфолошких електронских речника српског језика (у даљем тексту: СМР) чији су аутори Цветана Крстев и Душко Витас (Крстев 2008). С обзиром да је параметарска датотека за српски конструисана на основу верзије СМР-а која користи кодну шему *аурора*, а за финално писмо корпусних текстова је изабрана латиница, процес анотације латиничних текстова је обављен у три фазе:

- конверзија латиничног корпусног текста у кодну шему *аурора*;
- аутоматска морфолошка анотација конвертованог корпусног текста алатом TreeTagger;
- конверзија анотираног корпусног текста из кодне шеме *аурора* у латинично писмо.

3.2 Структурна анотација

Ако се у узме у обзир обим рударске пројектне документације у реалним системима и дужина века експлоатације током које се та документација непрекидно увећава, неопходно је обезбедити:

- брзо и ефикасно проналажење докумената и њихових метаподатака,
- ефикасну екстракцију релевантних сегмената текста из документа,
- директне (хипер)везе које међусобно повезују делове документе и пружају шири контекст у односу на ектраховане сегменте (на пример, директна веза између ектрахованог текста неког члана правилника и наслова виших структурних нивоа — одељака, поглавља итд. — који садрже екстраховани члан).

Смернице Иницијативе за обележавање текста (енг. Text Encoding Initiative Guidelines, скр. TEI Guidelines), прецизније пети предлог тих Смерница — TEI P5 (Бернард-Бауман 2018), представљају незванични стандард за анотацију произвољних типова докумената. Елементи и атрибути анотације коју пружа TEI P5 су дефинисани помоћу проширивог језика за обележавање (енг. Expanded Markup Language, скр. XML), па се као скраћени назив предлога Смерница често користи TEI P5 / XML. Подскуп најчешће коришћених или најбитнијих елемената предлога TEI P5 / XML, познат као „TEI, лака верзија" (енг. TEI-Lite) је популаран у пракси као незванични стандард за анотацију докумената, посебно као формат за структурну анотацију докумената.

Корисници нису обавезни да употребе све елементе и атрибуте које нуди, TEI P5 / XML, односно TEI-Lite, већ могу да додатно издвајају и дефинишу своје подскупове неопходних елемената и атрибута, чак и да дефинишу нове ако за тиме постоји потреба. За потребе структурне анотације Рударског корпуса определили смо се за минималан скуп елемената и атрибута:

- Најважнији структурни елементи су `div1`, `div2`, `div3`, `div4` (структурни нивои текста од највишег ка најнижем, на пример до-

кумент, глава, поглавље, члан), `head` (наслов и поднаслов), `p` (пасуси), `seg` (сегмент, односно реченица).

- Најважнији атрибути елемента `head` (опционо и елемената `div1`, `div2`, `div3`, `div4`) су `type` (као тип структурног нивоа: документ, глава, поглавље, члан и сл.) и `n` (као нумеричко обележје структурног нивоа, на пример, 1.2.3 као трећи члан другог одељка прве главе документа).[5]

Ради поједностављења обраде корпусних текстова, у овом случају је погодније да структурна анотација корпусног текста претходи његовој лингвистичкој анотацији. Приликом конверзије текста у вертикални формат структурне етикете се налазе у истој колони као и текст. Приликом аутоматске лингвистичке анотације структурно анотираног теста само се токенима текста придружују колоне које одговарају морфолошкој анотацији (врста речи и лема), не и структурним етикетама. Слика 2 илуструје одломак из вертикалног корпусног текста са лингвистичком и структурном анотацијом (Закон о рударству и геолошким истраживањима). Примећује се да се садржина документа (обичан текст) појављује у првој колони, друга колона представља ознаку врсте речи (N као именица, V као глагол, PREP као предлог, PRO као заменица, NUM као број, SENT као ознака краја реченице, итд.).

Структурна анотација се обавља комбиновањем различитих приступа (ручно, полуаутоматски, аутоматски) у зависности од формата улазног текста. Аутоматска структурна анотација је уобичајена за структурне нивое реченица (сегмената) и, у појединим случајевима, пасуса, с обзиром на њихову учесталост. Аутоматско означавање реченица се реализује уз помоћ софтверског алата Unitex и одговарајуће локалне граматике за XML-документе коју је за српски језик прилагодила Цветана Крстев (Помије 2016, Крстев 2008). Што се тиче пасуса, у случају када је текст форматиран тако да се знак за крај линије умеће искључиво на крају пасуса (тј. једна линија текста представља један пасус), означавање пасуса се постиже простим уметањем одговарајућих отворених и затворених етикета елемента `p` на почетак и крај линије респективно.

С обзиром на могућности аутоматске структурне анотације, тренутна верзија Рударског корпуса није структурно анотирана. Неколико корпусних текстова, махом закона и правилника, структурно је анотирано и искоришћено за креирање малог поткорпуса Рударског корпуса на коме се тестирају могућности претраге корпуса по структурним обележјима (в. одељак 4).

---

[5] Иако TEI-Lite нуди и елемент `div` који може да се користи за означавање произвољног структурног нивоа, при чему се на основу вредности атрибута `type` и `n` тог елемента одређује тип и дубина нивоа, поједини алати за индексирање корпуса (нпр. IMS CWB описан у пододељку 3.3) не подржавају рекурзивно угњеждавање истог елемента. Из тог разлога смо користили алтернативу коју пружа TEI-Lite, тј. нумерисане варијанте елемента `div` (`div1`, `div2`, `div3`, `div4`).

Слика 2: **Одломак из вертикалног корпусног текста са лингвистичком и структурном анотацијом (Закон о рударству и геолошким истраживањима)**

```
<div4>
<head type="clan" n="1.2.3">
<seg>
Član    N   Član
3   NUM 3
.   SENT    .
</seg>
</head>
<p>
<seg>
Pojmovi N   pojam
upotrebljeni    V   upotrebiti
u   PREP    u
ovom    PRO ovaj
zakonu  N   zakon
imaju   V   imati
sledeće A   sledeći
značenje    N   značenje
:   SENT    :
</seg>
</p>
```

3.3 Алати за индексирање и претрагу корпуса

У развоју Рударског корпуса до сада су коришћена три различита софтверска система за конструкцију, управљање и претрагу корпуса, у сврху два различита сценарија коришћења. Први сценарио се односи на кориснике који раде са копијом Рударског корпуса на свом рачунару и решавају различите задатке језичког инжењерства (попут проналажења и екстракције информација, екстракције термина) развојем електронских речника и локалних граматика уз помоћ програмског система Unitex (Помије 2016).

Потоњи сценарио се односи на онлајн коришћење корпуса за потребе различитих лингвистичких истраживања где се као резултат упита заданог помоћу регуларних израза добијају конкорданце. За потребе оваквог сценарија користи се информациони систем који комбинује веб сучеље (енг. web interface), као предњи део (енг. front-end) система, и програмски систем за креирање, управљање и претрагу корпуса, као задњи део (енг, back-end) информационог система. У раду (Томашевић и др. 2018) су већ описана и упоређена како два поменута сценарија, тако и два одговарајућа алата: Unitex

(за први сценарио), односно и CQPweb (Харди 2012) и IMS CWB (Еверт-Харди 2011), као одговарајући предњи и задњи крај информационог система за други сценарио. У овом раду упоредићемо могућности два слична информациона система који се користе за потребе другог сценарија: с једне стране, већ поменути CQPweb и CWB, а са друге стране — NoSketch Engine (Рихли 2007). NoSketch Engine такође има предњи и задњи крај система, веб сучеље Bonito и алат за управљање корпусима Manatee.

Олакшавајућа околност приликом припреме корпусних текстова за индексирање алатима CQPweb+CWB, односно NoSketch Engine, јесте што оба алата очекују вертикализовани текст, истоветни формат лингвистичке (у овом случају морфолошке) и структурне анотације (Слика 2). Оно по чему се поменути алати разликују, осим назива и синтаксе команди за индексирање, јесте начин на који су метаподаци придружени корпусним текстовима.

3.4 Метаподаци

Обе верзије Рударског корпуса, креиране помоћу алата CQPweb+CWB, односно NoSketch Engine, користе исте метаподатке. У зависности од примењеног алата, метаподаци се на различит начин придружују одговарајућим корпусним текстовима. NoSketch Engine захтева да сваки корпусни текст буде добро формирани XML-документ и да корени XML-елемент (који дефинише корисник) садржи метаподатке као вредности произвољно дефинисаних XML-атрибута. Слика 3 илуструје корени елемент `doc` корпусног текста са следећим метаподацима:

- наслов документа (атрибут `title`),
- аутори документа (атрибут `author`[6]),
- година настанка или публиковања документа (атрибут `year`),
- језик документа (атрибут `language`),
- листа кључних речи документа (атрибут `subject`),
- порекло (тип) документа (атрибут `provenance`[7]).

[6] У случају да документ има више аутора, имена различитих аутора се раздвајају вертикалном цртом.

[7] Вредност атрибута `provenance` представља назив одговарајуће потколекције дигиталне библиотеке којо припада оригинални документ, односно изведени корпусни текст (правилници, закони, уредбе, смернице, стратегије, докторске дисертације, монографије, уџбеници, радови, пројекти, студије). Имена атрибута одговарају називима истоимених колона у табели метаподатака коју користи дигитална библиотека.

Слика 3: **Представљање метаподатака у заглављу корпусног текста који је улаз за NoSketch Engine.**

```
<doc
n="2" file="udzben0114.tt"
title="Transport u pripremi mineralnih sirovina"
author="Kolonja Božo|Knežević Dinko"
provenance="udžbenici"
subject="transport,priprema mineralnih sirovina"
year="2000" language="srpski">
```

С друге стране, CQPweb+CWB такође очекује да корпусни текст буде добро формирани XML-документ, али не и да корпусни текст садржи своје метаподатке као део анотације, већ је неопходно да се метаподаци преко веб сучеља морају или ручно унети или увести као претходно припремљена датотека[8], при чему се прихваћени метаподаци чувају у релационој бази података. Сам корпусни текст као корени елемент садржи XML-елемент `text` који има само један атрибут (`id`) чија је вредност идентификатор текста (Слика 4).

Слика 4: **Заглавље корпусног текста „Транспорт у припреми минералних сировина" који је улаз за CQPweb+CWB.**

```
<text id="udzben0114">
```

4. Претраживање корпуса

Могућности претраге коју нуде алати CQPweb+CWB, односно NoSketch Engine, приближно су исте, чак добрим делом користе и истоветну синтаксу упита, односно два дијалекта упитног језика CQL (оригинално име упитног језика у CQPweb+CWB-верзији гласи CQP[9] Query Language, односно у верзији алата NoSketch Engine — Corpus Query Language), развијеног 1990. године на Универзитету у Штутгарту у оквиру Групе за корпусе и речнике (енг. IMS Textcorpora and Lexicon Group) која је и развила прву верзију алата CWB.

С обзиром на важност ефикасне екстракције релевантних сегмената текста из обимне пројектне докуменатације, овде ћемо размотрити само упите

[8] Формат претходно припремљене датотеке са метаподацима корпусних текстова одговара типичном улазном формату који користе системи за управљање релационим базама података када увозе податке у табелу (линије датотеке одговарају редовима табеле, док су поља, односно колоне, раздвојени табулатором).

[9] Скраћени назив за командни интерпретатор који се користи за претрагу корпуса направљеног помоћу алата CWB. Пуни енглески назив је Corpus Query Processor. За више детаља видети упутство на адреси http://cwb.sourceforge.net/files/CQP_Tutorial.pdf.

који се ослањају на структурну анотацију текста, као што је пример упита за екстракцију из корпуса сваког члана произвољног документа корпуса таквог да нумерација члана почиње ознаком 1.3. и завршава се произвољним бројем. Упит А (Табела 3) екстрахује наслов члана користећи чињеницу да је наслов обележен елементом `head` чији атрибути `type` и `n` тим редом садрже информације о структурном нивоу документа (`clan` као ознака за члан) и нумерацији (`1.3.3`, `1.3.4`, итд.). С обзиром да је члан чији наслов одговара упиту А обележен XML-елементом `div4`, Упит Б (Табела 3) издваја његову садржину. Слика 5 илуструје део резултата добијених помоћу алата NoSketch Engine.

Табела 3: Упити за екстракцију наслова и целог члана закона са нумерацијом која почиње ознаком 1.3. и завршава се произвољним бројем.

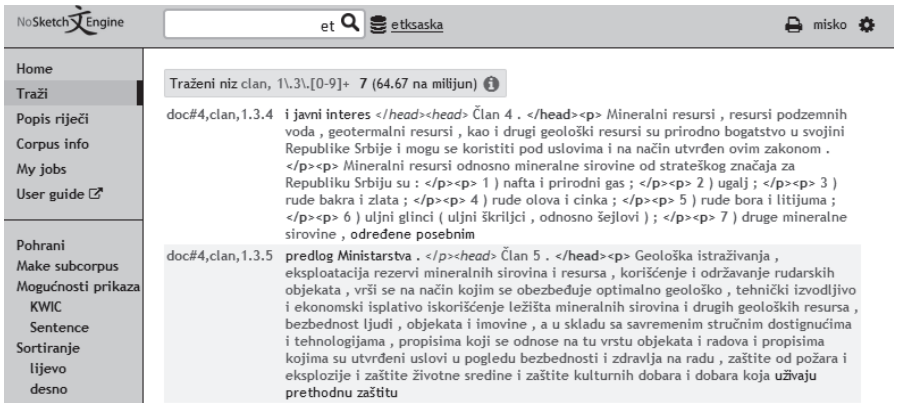
Алат	**Упит који користи структурну нотацију**	
CQPweb +CWB	А	/region[head, a] :: a.head_type=»clan» & a.head_n=»1\.3\.[0-9]+»
	Б	/region[head,a] :: a.head_n=»1\.3\..+» & a.head_type=»clan» expand to div4
NoSketch Engine	А	<head (type=»clan» & n=»1\.3\.[0-9]+») />
	Б	<div4/> containing <head (type=»clan» & n=»1\.3\.[0-9]+») />

Упоређивањем упита А и Б (Табела 3) за алате CQPweb+CWB, односно NoSketch Engine, примећује се да је синтакса коју користи NoSketch Engine једноставнија и разумљивија. CQPweb+CWB користи ознаку (енг. label) `a` у својству променљиве помоћу које приступа атрибутима наслова, као и операторе :: (праћене условима који сужавају скуп резултата) и `expand to` (како би се резултат проширио са елемента `head` на елемент `div4`), што додатно усложњава синтаксу упита. С друге стране, NoSketch Engine користи уобичајен начин означавања близак синтакси XML-а (`<div4/>`), заграде, оператор логичке конјункције & (као и CQPweb+CWB) и оператор `containing` којим захтева да XML-елемент са леве стране оператора садржи елемент са његове десне стране (у духу употребе истоимене речи у енглеском језику).

Додатне предности алата NoSketch Engine у односу на CQPweb+CWB су одлична документација, као и могућност локализације веб-сучеља на српски језик.

Иако оба алата слободно нуде свој изворни кôд, NoSketch Engine је већ прерастао у комерцијални производ Sketch Engine и питање је хоће ли се и у будућности та два алата равноправно развијати. Предност алата CQPweb+CWB је могућност креирања корисничких налога и профила, док NoSketch Engine препушта администратору да или дозволи слободан приступ свима без налога или да искористи механизме ауторизације коју нуде веб-сервери.

Слика 5: **Резултати упита за екстракцију члана са нумерацијом која почиње почиње ознаком 1.3. и завршава се произвољним бројем (NoSketch Engine).**



5. Закључак

Обе верзије Рударског корпуса, креиране помоћу алата CQPweb+CWB, односно NoSketch Engine, доступне су онлајн на привременим адресама и захтевају ауторизацију корисника (Табела 4).

Рад на унапређењу система за морфолошку анотацију корпусних текстова, засновану на морфолошким речницима за српски језик један је од најважнијих циљева даљег истраживања, уз континуиране допуне морфолошких и терминолошких електронских речника специфичних за области рударства. Такође, као што је то већ учињено у дигиталној библиотеци ROmeka@RGF, неопходно је имплементирати претраживање корпуса са подршком веб сервиса за проширивање упита тако да се резултати обухвате и семантичке лексичке релације (синонимију, антонимију, хипонимију, хиперонимију, итд.).

Табела 4: Адресе верзија Рударског корпуса креираних помоћу алата CQPweb и NoSketch Engine

Алат којим је креирана верзија корпуса	Адреса верзије Рударског корпуса на вебу
CQPweb	http://147.91.181.179/~cqp/cqpweb/
NoSketch Engine	http://147.91.181.179/~cqp/noske/

ЛИТЕРАТУРА

Бернард-Бауман 2018: Lou Burnard and Syd Bauman (eds), *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Version 3.3.0. Last updated on 31st January 2018, revision f4d8439. TEI Consortium, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/.

Витас 1981: Duško Vitas, „Generisanje imeničkih oblika u srpskohrvatskom jeziku", *Informatica*, 3/81, 49–55.

Еверт-Харди 2011: Stefan Evert and Andrew Hardie, „Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millennium", In: *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham, University of Birmingham.

Крстев 2008: Cvetana Krstev, *Processing of Serbian — Automata, Texts and Electronic Dictionaries*. Belgrade: Faculty of Philology, University of Belgrade.

Крстев и др. 2008: Cvetana Krstev, Duško Vitas and Gordana Pavlović-Lažetić, „Resources and Methods in the Morphosyntactic Processing of Serbo-Croatian", In: Gerhild Zybatow et al. (eds.), *Formal Description of Slavic Languages: The Fifth Conference*, Frankfurt am Main: Peter Lang, 3–17.

Крстев и др. 2015: Cvetana Krstev, Ranka Stanković, Ivan Obradović and Biljana Lazić, „Terminology Acquisition and Description Using Lexical Resources and Local Grammars" In: Thierry Poibeau and Pamela Faber (eds.), *Proceedings of the 11th Conference on Terminology and Artificial Intelligence*, CEUR Workshop Proceedings (ISSN 1613-0073), 81–89.

Помије 2016: Sébastien Paumier, *Unitex 3.1. User Manual*. http://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf.

Рихли 2007: Pavel Rychlý, „Manatee/Bonito — A Modular Corpus Manager", In: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno : Masaryk University, 65–70.

Schmid 1997: Helmut Schmid, „Probabilistic Part-of-Speech Tagging Using Decision Trees", In Jones, D. B. *et al.* (eds.) *New Methods in Language Processing*, Routledge, 154–164.

Schmid 1999: Helmut Schmid, „Improvements in Part-of-Speech Tagging with an Application to German", In: Armstrong, S. *et al.* (eds.) *Natural Language Processing Using Very Large Corpora*, Dordrecht: Springer, 13–25.

Станковић и др. 2016: Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić and Aleksandra Trtovac, „Rule-based Automatic Multi-Word Term Extraction and Lemmatization", In: Nicoletta Calzolari *et al.* (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 507–514.

Томашевић и др. 2017: Александра Томашевић, Биљана Лазић, Далибор Воркапић, Михаило Шкорић и Љиљана Колоња, „Употреба веб платформе Омека за дигиталне библиотеке из домена рударства", *ИНФОтека*, 17/2, 27–51.

Томашевић и др. 2018: Aleksandra Tomašević, Ranka Stanković, Miloš Utvić, Ivan Obradović and Božo Kolonja, „Managing Mining Project Documen-

tation Using Human Language Technology", *The Electronic Library*, DOI 10.1108/EL-11-2017-0239 (rad prihvaćen za štampu).

Утвић 2011: Милош Утвић, „Анотација Корпуса савременог српског језика", *ИНФОтека*, 12/2, 39–51.

Утвић 2014а: Miloš Utvić, „Izgradnja referentnog korpusa savremenog srpskog jezika" (neobjavljena doktorska disertacija), Beograd: Filološki fakultet, https://fedorabg.bg.ac.rs/fedora/get/o:10061/bdef:Content/download.

Утвић 2014б: Милош Утвић, „Листе учестаности Корпуса савременог српског језика", *Научни састанак слависта у Вукове дане*, 43/3, 241–262.

Харди 2012: Andrew Hardie, „CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool", In: *International Journal of Corpus Linguistics*. 17/3, 380–409. https://doi.org/10.1075/ijcl.17.3.04har

Miloš B. Utvić
Ivan M. Obradović
Ranka M. Stanković
Aleksandra Đ. Tomašević
Biljana Đ. Lazić

THE CONSTRUCTION OF SPECIAL CORPORA OF CONTEMPORARY SERBIAN — AN EXAMPLE OF CORPUS FOR MINING DOMAIN

Summary

This paper explores construction of domain-specific corpus of texts in Serbian from the mining domain at the University of Belgrade, Faculty of Mining and Geology. Special linguistically annotated corpus for mining domain originated from digital library ROmeka@RGF, initially as a means to improve features of digital library search engine, later as a language resource to be used in various linguistic research and multiple tasks of language engineering (terminology extraction, information retrieval, computational lexicography etc.). Also, several versions of the same special linguistically annotated corpus for mining domain, along with software packages used for corpora creation, management and search, are compared related to their search features.