811.163.41'322.2
811.163.41’373.46:622
https://doi.org/10.18485/msc.2017.46.3.ch10

Иван ОБРАДОВИЋ,*
Александра ТОМАШЕВИЋ
Ранка СТАНКОВИЋ
Биљана ЛАЗИЋ
Рударско-геолошки факултет
Универзитета у Београду

УВОЂЕЊЕ ДОМЕНСКИХ И СЕМАНТИЧКИХ МАРКЕРА ЗА ОБЛАСТ РУДАРСТВА У СРПСКЕ ЕЛЕКТРОНСКЕ РЕЧНИКЕ

Семантички маркери у електронским речницима омогућавају постављање комплексних упита за екстракцију информација. Када су у питању упити везани за специфични домен, расположиви скуп лексичких маркера за тај конкретни домен од пресудног је значаја за квалитет одзива. У овом раду разматра се значај лексичких маркера у процесима проналажења и екстракције информација и предлаже проширивање скупа ових маркера за област рударства. Дат је и кратак опис развијеног корпуса текстова из области рударства за чије претраживање су новопредложени маркери изузетно значајни.

Кључне речи: рударство, корпус, лексички маркери, екстракција информација

1. Увод

Српски е-речници се уз одговарајуће трансдукторе користе за екстракцију информација (ЕИ) и означавање текста у оквиру различитих лингвистичких апликација (Крстев 2008). Сваки такав е-речник састоји се од листе уноса заједно са њиховим лемама, морфосинтактичким, семантичким, доменским и другим информацијама. Систем српских е-речника обухвата општу лексику, властита имена и неколико доменски специфичних термина са флективним облицима генерисаним од ~140.000 простих и ~18.000 сложених лема. Већина облика речи у српским морфолошким е-речницима има не само вредности граматичких категорија, већ и додатне маркере који су наслеђени

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

из лема из којих се ови облици генеришу. Ови маркери могу бити граматички, деривативни, доменски и семантички.

Проналажење информација (енг. *Information Retrieval*) подразумева издвајање из скупа текстова оних који одговарају информационој потреби (упиту) корисника (Манинг и др. 2008), док задатак екстракције информација обухвата анализирање информација садржаних у тексту, њихово издвајање, означавање и организовање у структуиране скупове података, као што су онтологије и базе података, ради даље обраде (Јурафски/Мартин, 2016). Ове две области, иако различите, неретко користе исте ресурсе, а често и алате.

Задатак ЕИ подразумева да се анализом текста екстрахују унапред прецизно дефинисане семантичке класе информација. У овом раду посебна пажња је посвећена текстовима из једног специфичног техничког домена, конкретно рударства. За српски језик је развијен систем за препознавање именованих ентитета (Крстев и др., 2013), заснован на правилима, који успешно препознаје различите типове именованих ентитета: имена особа, називе локација и организација, временске и нумеричке изразе. За текстове у области рударства од посебног интереса су организације, локације и временски изрази.

За различите области, односно домене је потребно допунити електронске речнике специфичном лексиком и дефинисати нове типове ентитета. Како семантички маркери, интегрисани у графове, обезбеђују постављање комплексних упита за екстракцију конкорданци, то је за специфичне, доменски зависне упите је потребно допунити скуп лексичких маркера.

Један од домена који су недавно уведени у српски морфолошки речник је рударство, а паралелно са тим развијан је и корпус текстова из области рударства. Анализа ових текстова показала је да појмови и терминологија специфични за рударски домен захтевају увођење нових доменских и семантичких маркера како би се обезбедила што прецизнија екстракција информација из овог корпуса.

Наредни одељак посвећен је значају лексичких маркера, док је трећи одељак посвећен лексичким маркерима из области рударства. У четвртом одељку дат је кратак опис корпуса рударских текстова.

2. Значај лексичких маркера за проналажење и екстракцију информација

Семантички и деривациони маркери носиоци су информација од изузетног значаја за проналажење и екстракцију информација из текстуелих ресурса. Семантички маркери наведени у српским е-речницима означавају домен употребе дате речи или их ближе одређују по неком значењском критеријуму, нпр.+Hum да ли је нешто живо или је +Org врста организације. Деривациони маркери указују на деривационе специфичности. Нпр. маркер +DerName означава именице изведене од личних имена (нпр. дарвинизам, реганизам...). Маркер +VN означава глаголске именице (нпр. копање, печење, реструкту-

рирање…). Значај маркера се огледа у томе што они у комбинацији са граматичким ознакама ближе дефинишу речи. Њима се прави финија дистинкција међу речима неопходна за прецизнију екстракцију информација.

Српски речник простих речи у LADL формату се састоји од два речника (или пописа, листи): DELAS– је речник канонских облика (лема) који служи за генерисање другог речника – DELAF – који је речник подређених облика или речник флективних облика и само овај речник се користи у аутоматској обради текста. (Крстев, 2008).

Један пример записа у речнику DELAS је: дреглајн, N1001+DOM=Mining+Instrum, где је:

дреглајн	канонски облик (лема)
N	врста речи (именица)
(N)1001	флективна класа која генерише све флективне облике
+DOM=Mining	доменски маркер – припада области рударства
+Instrum	семантички маркер – опрема (или део опреме)

Маркере је могуће користити кроз регуларне изразе, за неке једноставније упите над корпусом. Други вид употребе јесте кроз постављање упита конструкцијом аутомата са сложенијим захтевима. Пример истраживања у коме су коришћени маркери у циљу екстракције глагола из кулинарског домена за српски језик дат је у (Крстев/Лазић, 2015). Један други пример употребе маркера јесте систем за екстракцију именованих ентитета NER осета (назива установа, личних имена, улица...) (Крстев и др., 2016) (Крстев и др., 2014). Употребе маркера могућа је за потребе обраде текста ради екстракције доменске терминологије, мерних јединица, прављење релација међу речима, итд.

3. Доменски, поддоменски и семантички маркери за област рударства

Рударство је веома сложена индустријска грана, са активностима које се међусобно прожимају и надопуњују, али се истовремено и веома разликују. У (Крстев, 2008) дефинисан је општи домен рударства DOM=+Mining. У међувремену се показало да је потребно увести финију поделу домена на гране рударства, као и специфичне семантичке маркере и тиме омогућити прецизнију екстракцију информација из текстова рударског домена. Како би свака од рударских активности била ближе описана и како би се омогућила прецизније екстракције информација из текстова рударског домена, дефинисани су нови доменски, поддоменски и семантички маркери. У табели 1 приказани су постојећи и новопредложени доменски маркери значајни за област рударства.

Табела 1. Доменски из области рударства

Маркер	Опис	Статус	Примери
+Mining	Рударство	Постојећи маркер	рудник, рудар, руда, експлоатација, угаљ, минерална сировина, руда бакра, етажа, јаловина, окно, бушење, цевовод, рударска мерења, вентилација, нафта, бушотина, одводњавање, млевење
+Mach	Машинство	Постојећи маркер	роторни багер, багер ведричар, камион, дозер, рипер, транспортна трака, цевополагач, железница, одлагач, транспортни мост, грејдер, комбајн, хидромонитор, депонијска машина
+Safety	Заштита на раду	Постојећи маркер са предлогом промене назива +Safe	вентилација, заштитна опрема, бука, вибрације, минерална прашина, штетни гасови, повреда на раду, професионално обољење, пожар, заштита од пожара, експлозија, служба спасавања
+Transport	Транспорт	Предложен нови маркер	транспортна трака, камион, дампер, хидраулички транспорт, извоз, извозно постројење, самоходно транспортно средство
+RockMech	Механика стена	Предложен нови маркер	притисак, смицање, напонско стање, носивост стене, физичко-механичке особине, деформабилност стена, деформабилност тла, стабилност косина, носивост тла
+Surveying	Геодезија	Предложен нови маркер	рудничка мрежа, јамски полигон, нивелмански влак, рударска висећа бусола, теодолит, жиротеодолит
+EnvProt	Заштита животне средине	Предложен нови маркер	аерозагађење, мониторинг животне средине, ремедијација, рекултивација

Ради прецизнијег описивања рударских термина предложена је и листа поддоменских маркера. Њима би се ближе дефинисале четири велике области унутар рударског домена (површинска експлоатација, подземна експлоатација, припрема минералних сировина и експлоатација нафте и гаса), које се разликују у толикој мери да оправдавају увођење додатних маркера. Поддоменски маркери би се придруживали доменском маркеру, па би тако, на пример, маркер +Mining+Surface означавао појмове који припадају домену рударства и поддомену површинска експлоатација. У табели 2 приказани су постојећи и новопредложени доменски маркери значајни за област рударства.

Табела 2. Поддоменски маркери из области рударства

Маркер	Опис	Примери
+Surface	Површинска експлоатација лежишта минералних сировина	површинска експлоатација, површински истражни радови, етажа, косина, завршна косина, стабилност косина, откривка, коефицијент откривке, БТО систем, депонија, минирање, бушење, граница копа, јаловина, рекултивација, нагиб косине, багер
+Underground	Подземна експлоатација лежишта минералних сировина	подземна експлоатација, подземни истражни радови, окно, ходник, ускоп, нископ, сипка, бункер, заштитни стуб, широко чело, вентилација
+MinProcess	Припрема минералних сировина	кретање масе, узорак, уситњавање, дробљење, сито, млевење, млин, класирање, концентрација, флотација, флотацијска пулпа, згушњавања, лужење, центрифугирање, депоновање
+Petroleum	Експлоатација нафте и гаса	нафта, гас, угљоводонични флуид, бушотина, цевовод, динамика протока, гас лифт, гасоводни систем

Како би се омогућила екстракцију специфичних концепата и релација међу концептима креирањем лексичких маски, предложени су и нови семантички маркери значајни за област рударства. Листа ових маркера, са описом и примерима приказана је у табели 3, заједно са примерима употребе постојећих маркера.

Табела 3. Семантички маркери за област рударства

Маркер	Опис	Статус	Примери
+MinStatus	Статус рудник	Предложен нови маркер	активан рудник, неактиван рудник, затворен рудник, напуштен рудник, конзервиран рудник, рудник у развоју
+Ore	Минералне сировине	Предложен нови маркер	угаљ, лигнит, мрки угаљ, камени угаљ, руда гвожђа, хематит, магнетит, руда бакра, халкопирит, ковелин, руда олова, галенит, нафта, гас, геотермална вода, шљунак, песак, камен
+Activity	Рударска активност	Предложен нови маркер	пројектовање рудника, минирање, бушење, експлоатација минералне сировина, вентилација рудника, одлагање откопаног материјала, дробљење руде, млевење руде, уситњавање руде
+Object	Рударски објекти	+Mining+Object	окно, ходник, ускоп, нископ, бункер, усек, поткоп, засек, јама, дробилана, бушотина, хоризонт, сипка, јамска просторија
+Prof+Hum	Професије	+Mining+Prof +Hum	рударски инжењер, рудар, геолог, мерач, рударски техничар, геолошки техничар, багериста, копач, минер, површинац, подземљаш, припремаш, помоћни радник
+Org	Организације	+Mining+Org	Колубара, Костолац, Рудник, РТБ Бор, Дрмно, ЕПС
+Instrum	Рударска опрема	+Mining+Instrum	багер, камион, транспортна трака, дозер, утоварач, скрепер, концентратор, теодолит, бушаћи чекић, дробилица, одлагач, дреглајн, додавач
+Exploration	Истраживање минералних сировина	Предложен нови маркер	експлоатационо истраживање, површински истражни радови, подземни истражни радови, истражно бушење
+Conc+Fashion	Средства личне заштите	Постојећи маркери	лампа, самоспасилац, маска, шлем, чизме, заштитна опрема, заштитно одело
+Text	Типови рударских докумената	+Mining+Text	рударска пројектна документација, геолошки елаборат, претходна студија оправданости, студија оправданости, рударски пројекат, технички пројекат

Семантички маркери дефинисани су по угледу на EarthResourceML[1], стандард за размену дигиталних информација заснован на XML-у за минералне појаве, ресурсе и резерве, потом руднике и рударске активности, као и производњу концентрата, излазних производа, и рударског отпада.

Речници који су коришћени за креирање семантичких маркера су: ClassificationMethodUsed, CommodityCode, EarthResourceExpression, EarthResourceForm, EarthResourceMaterialRole, EarthResourceShape, EndUsePotential, EnvironmentalImpact, ExplorationActivityType, ExplorationResult, MineStatus, MineralOccurrenceType, MiningActivity, ProcessingActivity, RawMaterialRole, ReserveCategory, ResourceCategory, UXFCValue, WasteStorage. Ови речници су јавно доступни као RDF и XML датотеке, а постоји и SPARQL приступна тачка за RESTfulAPIs.

4. Опис корпуса рударских текстова

Новопредложени маркери значајни су пре свега за екстракцију информација из корпуса текстова везаних за рударство. Прикупљање корпуса са текстовима из рударског домена је почело 2014, након чека је урађена прва допуна електронских речника простих речи рударским и геолошким термина (Крстев и др., 2015), да би следећа анализа и екстракција вишечланих израза урађена током 2015 године (Станковић и др., 2016).

Током 2016. године прикупљени су, разврстани по типу и адекватно обрађени текстови из области рударства. Сви текстови су пречишћени, уклоњени су делови на страном језику, табеле, слике, референце и линкови. Обухватају комплетну законску регулативу из области рударства (законе, правилнике, уредбе, стратегије развоја), докторске дисертације, пројектну документацију рађену од 2009. године на Катедри за планирање и пројектовање површинских копова на Рударско-геолошком факултету и литературу из области рударства.

Ради заједничке обраде сви текстови су спојени и формирана је једна текстуална датотека величине 39 MB, са 6200 страна текста формата A4. Након обраде текста добијено је: 150.365 реченица, 2.719.086 (100.414 различитих) простих речи. У припреми је 22.875 речи специфичних за области рударства, безбедности и заштите на раду и процене ризика за укључење у електронски речник, након чега следи екстракција вишечланих термина, њихова лематизација, обележавање маркерима и укључење у продукциони скуп речника.

Преглед резултата обраде корпуса рударских текстова дат је у Табели 4.

[1] http://resource.geosciml.org/vocabulary/earthresourceml/2016

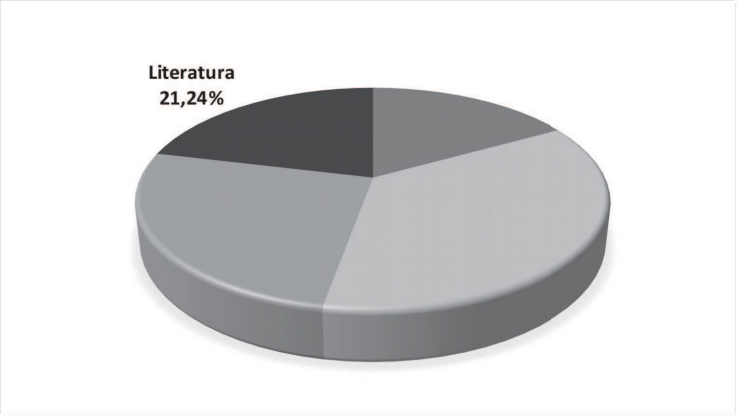
Табела 4. Резултати обраде текстова за рударски корпус

Тип текстова	Број текстова	Број реченица	Број речи	Број различитих речи	Број различитих лема	Величина текста (MB)	Индекс ЛР (у %)
Законска регулатива	94	25.250	459.912	26.146	9.968	6,6	2.17
Дисертације	28	47.033	976.667	61.146	16.102	14,2	1.65
Пројектна документација	38	46.792	705.094	35.362	11.359	10,1	1.61
Литература	12	31.405	577.413	45.609	13.529	8,2	2.34
Цео корпус	172	150.365	2.719.086	100.414	22.875	39,1	

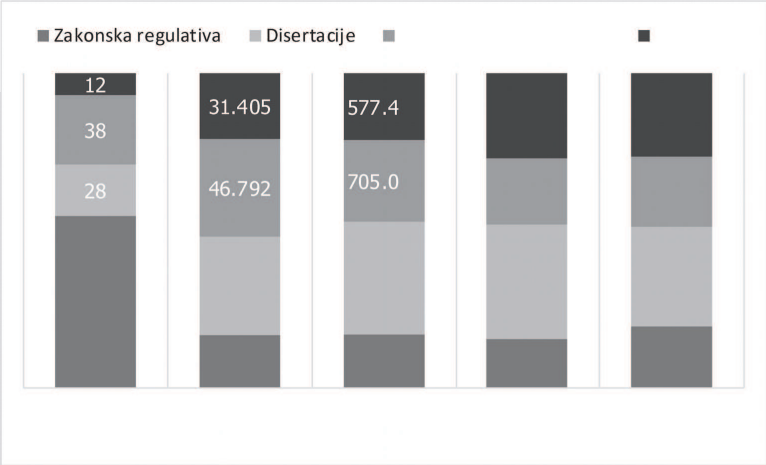
Структура корпуса према броју речи за одређени типа текста дата је на слици 1.

На основу резултата датих у табели 4, чији је део графички приказан на слици 2, може се оценити лексичка разноврсност текстова зависно од њиховог типа. Лексичка разноврсност у најширем смислу дефинише се као опсег различитих речи коришћених у тексту (Макарти/Џарвис, 2010). Лексичка разноврсност текстова може се изразити на неколико начина, па је тако у Табели 4 дат индекс LR рачунат као проценат броја различитих лема у односу на број речи. Уочава се да је највећи код литературе, потом следи законска регулатива, дисертације и коначно пројектна документација, што је и очекиван резултат.

Слика 1. Проценат речи у корпусу према типу текстова

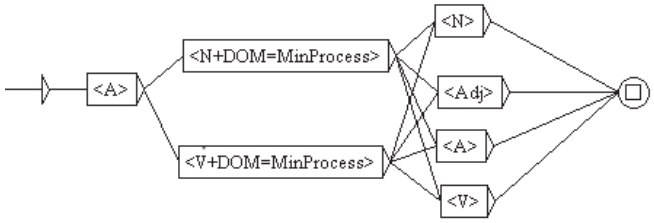


Слика 2. Лексичка разноврсност текстова за рударски корпус



На слици 3 приказан је граф који екстрахује леви и десни контекст уз именицу или глагол који су означени доменским маркером „Припрема минералних сировина". Услов за леви контекст је постављен у виду придева, десни контекст могу бити именица, прилог придев или глагол. Резултат упута над корпусом рударских текстова јесу 42 појављивања. Неке од добијених колокација јесу: *директно флотирање је*, *добра аерисаност земљишта*, *контролно флотирање бакра*, *независно флотирање песка*. Пример добијених конкорданци дат је на слици 4.

Слика 3. Граф за екстракцију колокација уз глагол/именицу из домена Припреме минералних сировина



Слика 4. Резултати примене графа са слике 3 над рударским корпусом

ija može da bude direktna i obrnuta.{S} Direktno flotiranje je proces hidrofobizacije i prelask
flotiranje, i • obrnuto flotiranje. {S}Direktno flotiranje podrazumeva flotiranje korisnih min
preminska gustina se često posmatra kao dobra aerisanost zemljišta i širenje korenja. {S}Fizičk
korisni i nekorisni proizvod.{S} Tokom dodatnog flotiranja menjaju se parametri procesa (udeo
icioner kontrolno flotiranje bakra K/Pb grubo flotiranje bakra otok ciklusa bakra grubo flotira
rolno) i prečišćavanje (slika 8.11). {S}Grubo flotiranje je prva operacija flotiranja čiji je p
procesa flotiranja dat je na slici 3.6. Grubo flotiranje predstavlja proces inicijalnog prolask
Nikola Tesla B" Namenski projektovani i izrađeni spigot predstavlja sistem cevi složenih u obli
e olova prečišćavanje pumpa kondicioner kontrolno flotiranje bakra K/Pb grubo flotiranje bakra
Slika 3.6. Šema faza procesa flotiranja Kontrolno flotiranje predstavlja nastavak flotiranja ot
ektivno, • kolektivno-selektivno, • kao nezavisno flotiranje peska i mulja, • kao flotiranje sa
a kolektivno-selektivnog flotiranja. {S}Nezavisno flotiranje peska i mulja primenjuje se kod fl
šine izabran je slučaj izbora mašine za osnovno flotiranje rude bakra.{S} Izbor je vršen između
3.11. data je uprošćena, tipična, šema pojedinačnog flotiranja peska i mulja. {S}U ciklus J ko
toj najsitnijoj, glinovitoj, klasi.{S} Uspešno flotiranje podrazumeva prethodno klasiranje i r
toj najsitnijoj, glinovitoj, klasi.{S} Uspešno flotiranje podrazumeva prethodno klasiranje i r
svaka 2 m dubine po jedan poduzorak.{S} Uzeti poduzorci mogu se analizirati pojedinačno kako bi
svaka 2 m dubine po jedan poduzorak.{S} Uzeti poduzorci mogu se analizirati pojedinačno kako bi

5. Закључак

Рад на развоју и допуњавању електронских речника је континуирани процес у коме значајно место заузима допуна речничких одредница лексичким маркерима. У раду је детаљно образложено због чега су лексички маркери значајни, посебно за екстракцију информација, а на примеру рударства илустровано је како се маркери могу допуњавати за одређене специфичне домене. Један број нових маркера је предложен и они ће бити имплементирани у наредној фази развоја српских е-речника. Након анализе рударског

корпуса који је такође описан у раду припремљено је 22.875 речи специфичних за области рударства, безбедности и заштите на раду и процене ризика за укључење у е-речник. Следи екстракција вишечланих термина, њихова лематизација, обележавање маркерима и укључење у продукциони скуп речника. На сличан начин, е-речнике треба обогаћивати речима специфичним за друге домене, уз паралелно дефинисање одговарајућих доменских и семантичких маркера и развој одговарајућих доменских корпуса.

ЛИТЕРАТУРА

Јурафски/Мартин, 2016: Daniel Jurafsky & James H. Martin, Speech and Language Processing, Draft of November 7, 2016.

Крстев 2008: Cvetana Krstev, Processing of Serbian – Automata, Texts and Electronic dictionaries Faculty of Philology, University of Belgrade, Belgrade.

Крстев и др., 2008: Cvetana Krstev, DuškoVitas, Gordana Pavlović-Lažetić, "Resources and Methods in the Morphosyntactic Processing of Serbo-Croatian", *Formal Description of Slavic Languages: The Fifth Conference*, Leipzig 2003, Zybatow, Gerhildetal. (eds.), Peter Lang: Frankfurt am Main, pp. 3–17.

Крстев и др., 2013: Cvetana Krstev, Ivan Obradović, Miloš Utvić, DuškoVitas, "A system for namedentity recognition based on local grammars", In: *J Logic Computation* 24 (2), pp. 473–489.

Крстев/Лазић, 2015: Цветана Крстев, Биљана Лазић, „Глаголи у кухињи и за столом", *Научни састанак слависта у Вукове дане*, 44/3, 117–136.

Крстев и др. 2015: Cvetana Krstev, Ranka Stanković, Ivan Obradović, Biljana Lazić "Terminology Acquisition and Description Using Lexical Resources and Local Grammars", In: *Proc. of the 11th Conferenceon Terminology and Artificial Intelligence*, Granada, Spain, eds. Thierry Poibeau and Pamela Faber, LexiCon (Universidad de Granada), pp. 81–89.

Станковић и др. 2016: Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, AleksandraTrtovac, "Rule-based Automatic Multi-word Term Extraction and Lemmatization", In: *Proc. of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, Portorož, Slovenia, pp. 23–28.

Крстев и др. 2016: Cvetana Krstev, Anđelka Zečević, Duško Vitas, and Tita Kyriacopoulou, "NERosettafor the Named Entity Multi-lingual Space", In: *Human Language Technology Challenges for Computer Science and Linguistics, LNCS*, pp. 327–340.

Манинг и др. 2008: Manning, C., Raghavan, P. I Schütze, H. Introduction to Information Retrieval (Volume 1). Cambridge, MA, USA: Cambridge University Press.

Макарти/Царвис, 2010: McCarthy, P.M., & Jarvis, S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment, Behavior Research Methods, 42(2), pp. 381–392.

Ivan Obradović, Aleksandra Tomašević, Ranka Stanković, Biljana Lazić

INTRODUCING DOMAIN AND SEMANTIC MARKERS FOR THE FIELD OF MINING IN SERBIAN ELECTRONIC DICTIONARIES

Summary

Semantic markers in electronic dictionaries allow for complex queries for information extraction. When it comes to domain-specific queries, the availablese to flexical markers for that specific domain is critical to the quality of there sponse. This paper discusses the importance of lexical markers in the processes of information retrieval and extraction, and proposesanexpansion of the set of the semarkers for the field of mining. A brief description of the developed corpus of texts from the field of mining is also given, for the search of which the proposed markers are extremely important.