ISSN 1450-9687

Infotheca

infoteka.bg.ac.rs

Journal for Digital Humanities

Vol. 21, No. 1, September 2021

Лексичка база Фрејмнет: неколико примера оквира из домена ризика

УДК 81'322.2

САЖЕТАК: У раду се даје кратак приказ теорије семантике оквира (енгл. *Frame Semantics*), на којој је заснована лексичка база Фрејмнет (енгл. *FrameNet*). Представљена је концепција ове мреже, као и могућности њене примене. Представљена је и лексичка анализа која се примењује у пројекту израде Фрејмнета и указано на разлике између анализе засноване на оквиру у односу на анализу засновану на речи. Затим је приказано неколико повезаних оквира које призивају речи из домена ризика. У раду је представљена и платформа NLTK (енгл. *Natural Language Toolkit*), помоћу које се могу користити разни језички ресурси, међу њима и Фрејмнет. Завршно поглавље пружа анализу именице *ризик* на корпусу рударства. Представљени су најчешћи колокати ове именице, скица њене употребе, конкорданце за поједине моделе, проналажење синонима и повезаних речи у виду тезауруса, графички приказ фреквенција појединих колокација, као и облака речи.

КЉУЧНЕ РЕЧИ: Српски језик, семантика оквира, Фрејмнет, сценарио ризика, корпус рударства, обрада природних језика.

РАД ПРИМЉЕН: 15. јули 2021.
РАД ПРИХВАЋЕН: 6. септембар 2021.

Александра Марковић
aleksandra.markovic@isj.sanu.ac.rs
Институт за српски језик САНУ
Београд, Србија

Ранка Станковић
ranka.stankovic@rgf.bg.ac.rs
Универзитет у Београду
Рударско-геолошки факултет
Београд, Србија

Наталија Томић
ntomic@hotmail.com
Универзитет у Београду
Београд, Србија

Оливера Китановић
olivera.kitanovic@rgf.bg.ac.rs
Универзитет у Београду
Рударско-геолошки факултет
Београд, Србија

1. Увод

Теорија семантике оквира (енгл. *Frame Semantics*) је когнитивносемантички приступ Чарлса Филмора који експлицитно

повезује значења неке речи са синтаксичким контекстима у којима се та реч јавља (Atkins, Fillmore, and Johnson 2003, 254). Анализом значења речи баве се углавном лексикографи, као и они који се баве семантиком. Међутим, уколико је циљ представљање начина на који се нека реч заиста користи, анализа података из корпуса показује се као прилично компликован задатак, с обзиром на број конкорданци које савремени корпуси нуде за поједине кључне речи. Теорија семантике оквира, према оцени више аутора (Atkins 1994; Gildea and Jurafsky 2002; Atkins, Fillmore, and Johnson 2003; Pradhan et al. 2005; Boas and Dux 2017; Jurafsky and Martin 2020), представља поуздан, научно утемељен начин да се анализира и опише начин употребе неке речи. У основи овог приступа лежи идеја да се свако искуство које памтимо јавља у неком смисленом контексту и да га можемо и упамтити баш захваљујући томе што имамо когнитивну схему или оквир за разумевање искустава. Филмор сматра да се речи уче у таквим смисленим контекстима, а ти су контексти неопходни и за процес разумевања, када у сећању призивамо искуства кроз која смо научили поједине речи. Оквир идентификује тип искуства, даје му структуру и целовитост, односно даје значење предметима, догађајима и односима у том искуству (Fillmore 1976, 26).[1]

1.1 Концепција мреже Фрејмнет

Фрејмнет[2] је лексичка база података енглеског језика, заснована на анотацији примера употребе неке лексичке јединице (у даљем тексту ЛЈ; енгл. *lexical unit*) у аутентичним текстовима (дакле, у корпусу, а не у конструисаним примерима). Основна идеја своди се на то да се значења већине речи најбоље могу разумети на основу семантичког оквира, појмовне структуре налик на сценарио, који представља опис типа ситуације, догађаја, односа или ентитета, као и учеснике тих ситуација, догађаја и односа (Ruppenhofer et al. 2016, 7). На пример, ризиковање у типичном случају подразумева следећи скуп појмова: особу која је централна за сценарио РИЗИКА – *Протагонисту*, који свесно или несвесно доноси одлуку да ризикује или подлеже неком

1. Термин *оквир* Филмор користи као општи термин за скуп појмова који су у литератури о разумевању природних језика познати под називима *схема*, *сценарио*, *когнитивни модел*, *народна теорија* и неки други (Fillmore 1982, 111).

2. Пројекат се развија у Међународном институту за рачунарске науке у Берклију (International Computer Science Institute, Berkeley) од 1997. године.

ризику; могући *Лош исход* или штету; *Одлуку* која може довести до лошег исхода; *Циљ* који се жели постићи; *Околности* у којима се ризик јавља и *Вредност*, особу или предмет који су драгоцени Протагонисти и који су угрожени у датој ризичној ситуацији (Fillmore and S. Atkins 1994, 367).

1.2 Лексичка анализа заснована на оквиру

Лексичка анализа заснована на теорији семантике оквира подразумева анализу семантичког садржаја неке лексичке јединице, идентификацију њеног семантичког суседства, откривање граматичких конструкција у којима се јавља у корпусу, као и бележење свих конструкција у којима дата ЛЈ изражава пун семантички потенцијал. Сматра се да је неопходно описати све допуне и одредбе у тим конструкцијама. Посебна пажња поклања се речима које се не могу правилно употребити уколико се не знају конструкције у којима се јављају. Такве речи називају се речима које призивају оквире[3] (енгл. *frame-evoking words*), а то су на првом месту глаголи, али и именице, придеви и прилози (Atkins, Fillmore, and Johnson 2003, 252).[4]

Основне јединице лексичке анализе у Фрејмнету су оквир и ЛЈ, под којом се подразумева лексема у једном од својих значења (Fillmore et al. 2003, 297), (Ruppenhofer et al. 2016, 7).[5] За разлику од онога што је уобичајено у лексикографији, фокусирања на једну реч, односно лексему, и истраживања свих њених значења, тј. ЛЈ, у Фрејмнету се

3. Исти еквивалент енгл. глагола *to evoke* користи се и у преводу рада Fillmore 1982 у (Rasulić and Klikovac 2014, 79).

4. Теорија семантике оквира подстакла нас је да укажемо на потребу да се за четири основне класе речи које призивају оквире (именице, придеве, глаголе и прилоге) наводе релевантне конструкције у описним речницима српског језика (Марковић 2017, 34–41).

5. У српској лексиколошкој литератури, као и у синтаксичким радовима који се баве односом граматике и речника, користи се различита терминологија за оно што се у Фрејмнету назива лексичком јединицом (нпр. у универзитетском уџбенику лексикологије под лексичком јединицом подразумева се одредница или лема (Драгићевић 2007, 30)), док Поповић (2003, 202–203) указује на важност тога да се пажња у синтакси поклања конкретним лексемама, односно лексемама употребљеним у једном од више значења, које назива сублексемама. Ми смо се у овом раду определили за термин лексичка јединица, у складу с терминологијом мреже и приступа који представљамо.

у истом тренутку анализирају ЈЈ које припадају истом оквиру[6] (Fillmore et al. 2003). Тако су, на пример, у односу на оквир *Бити_ угрожен* (енгл. *Being_ at_ risk*) дефинисане именице *ризик*, *опасност*, *безбедност*, *рањивост*, придеви *несигуран*, *безбедан*, *сигуран*, *поуздан*, *рањив* и др.

Процес описивања ЈЈ у Фрејмнету представљен је у (Fillmore et al. 2003).[7] Он започиње неформалним описом оквира ком та јединица припада, прецизније описом типа ситуације или збивања представљених оквиром и стварања листе речи које се могу објаснити повезивањем с датим оквиром (299).[8] Затим се бира циљна ЈЈ, централни члан оквира (у терминологији Фрејмнета *target*), ЈЈ у односу на коју се ради анотација; то је типично једна реч, али може бити и вишечлана реч или фразеологизам (Ruppenhofer et al. 2016, 21) и испитује њена употреба екстраховањем из корпуса реченица које је садрже.

Свој увид у значења изабране лексеме, добијен на основу посматрања примера из корпуса, лексикограф који ради на мрежи Фрејмнет пореди са значењима дате лексеме у неком од референтних речника.[9] Када стекне јаснију представу о значењима, покушава прецизније да опише оквир у који спада изабрана ЈЈ. Затим пише дефиницију оквира – схематски опис типа ситуације која лежи у основи неке речи, заједно с називима улога учесника описане ситуације, које се називају елементима оквира (енгл. *frame elements*). У лексикографском смислу важан је начин на који елементи оквира добијају језички израз у реченицама у којима се јавља посматрана ЈЈ (Fillmore et al. 2003, 304–305).

6. Овде је у питању разлика између два приступа издвајању значења, први је заснован на речи (*word-based*), а други на оквиру (*frame-based*) (Atkins, Fillmore, and Johnson 2003, 254).

7. Иако се процес описује као низ одвојених корака, који се одвијају одређеним редом, аутори напомињу да се заправо у сваком тренутку може вратити на неки од претходних корака и ревидирати нека од одлука (Fillmore et al. 2003, 299).

8. Тај опис подразумева: 1) схематски опис типа ентитета или ситуације представљене оквиром; 2) бирање описних ознака за обележавање елемената оквира; 3) писање радне листе речи које припадају датом оквиру (припадност лексичких јединица истом оквиру указује на то да реченице које садрже те ЈЈ подлежу сличној семантичкој анализи) (297).

9. Пошто су анализи подвргли дефиниције глагола *ризиковати* у десет општих речника енглеског језика, Филмор и Еткинсова закључили су да чак ни речници сличних величина и намена не издвајају основна значења овог глагола, која припадају основном речничком фонду (Fillmore and S. Atkins 1994, 353).

1.3 Елементи оквира

Елементи оквира често се могу посматрати као примери општијих семантичких улога (као агенс, доживљавач, пацијенс), али се дефинишу на начин који је специфичан за поједине оквире (енгл. *frame-specific*). Постоји више разлога за то, а најважнији је тај што омогућава да се направе прецизне дефиниције елемената оквира за одређену групу речи, без потребе за утврђивањем начина на који ће се оне уклопити с малим, унапред задатим скупом општијих семантичких улога (305).

Најпре се утврђују централни елементи оквира (енгл. *core frame elements*).[10] Централни елементи су они који представљају обавезне компоненте оквира и који неки оквир чине јединственим и различитим од других. Поред централних, постоје и периферни елементи, који се могу појавити у свим оквирима у којима неки агенс врши неку радњу (периферни елементи означавају појмове МЕСТА, ВРЕМЕНА, НАЧИНА, СРЕДСТВА).[11] Дешава се да неки од централних елемената не буде и језички изражен, иако је обавезан у појмовној структури оквира; такав случај назива се нултим појављивањем (енгл. *null instantiation*) и анотира се у бази (320). Пошто се одреде централни и периферни елементи, дефинише се оквир.[12]

Филмор и Еткинсова закључили су, након поменуте анализе глагола *ризиковати* у референтним речницима енглеског, да се допунама не

10. Постоје неке формалне особине које помажу у одређивању статуса елемента као централног (нпр. централни су они елементи који морају бити изражени, као и они који, уколико остану неизражени, добијају одређену интерпретацију (нпр. у реченици *John arrived* није изражено место на које је Џон стигао, као циљна локација, GOAL, али се то место разуме на основу контекста) (Ruppenhofer et al. 2016, 23–24).

11. Дистинкција централни/периферни елементи оквира одговара у грубим цртама подели на аргументе и одредбе у традиционалној граматичкој анализи (Fillmore et al. 2003, 310). Периферни елементи не могу бити у функцији субјекта или објекта циљног глагола и често су изражени прилозима или предлошко-падежним конструкцијама (319).

12. Ruppenhofer et al. (2016, 65) издвајају и додатне типове елемената оквира: они који се јављају у зависним клаузама имају периферни или екстратематски статус (енгл. *extra-thematic*, нпр. ВРЕМЕ, РАЗЛОГ). Осим тога, постоје и елементи којима се приписује ознака *Core-Unexpressed*; то су они елементи неког оквира који се понашају као централни неизражени, али који се можда неће наћи међу елементима оквира потомака датог оквира (24–25). У овом раду нећемо се бавити ни једнима ни другима.

поклања довољно пажње (иако су веома битне, посебно у речницима за учење језика), а да постоје и неки други реченични чланови које речници занемарују, а који морају бити јасно издвојени и описани како би се посматрани глагол употребио правилно. На пример, радња коју изводи особа која нешто ризикује (и која може имати различите синтаксичке реализације): Ризиковала је свој живот *покушавајући да спасе дете које се давило*; затим, циљ који онај који ризикује има на уму кад се излаже ризику: Ризиковала је свој живот *како би спасла мој* (Fillmore and S. Atkins 1994, 362). Радња којом неко нешто ризикује спада у централне елементе оквира, док циљ ради ког се ризик предузима спада у периферне.

1.4 Односи између оквира – мрежа оквира

Пошто се дефинишу оквир и његови елементи, дати оквир се повезује са другим оквирима. На тај начин оквири, њихови елементи и с њима повезане ЛЈ добијају место у семантичком простору (Ruppenhofer et al. 2016, 79) и чине мрежу. Успостављање односа између оквира омогућава уочавање и бележење семантичких генерализација на основу типова учесника, догађаја и сл. Оквир може бити повезан са оквирима које разрађује или наслеђује, са онима који су његови подоквири, као и са онима које користи. Односи између оквира су усмерени или асиметрични: наиме, апстрактнији и независнији оквир се назива *Над_оквиром* (енгл. *Super_frame*) оквира који је зависнији и мање апстрактан и који се назива *Под_оквиром* (енгл. *Sub_frame*) (79).

Дефинисано је неколико врста односа, од којих су најважнији следећи (79–84):

– У односу *Наслеђивања* (енгл. *Inheritance relationship*) подоквир (подређени оквир) представља прецизнију разраду надоквира (надређеног оквира). Сви елементи оквира и подоквири надређеног оквира имају своје парњаке у подређеном оквиру, а он може имати додатне подоквире, елементе оквира и ограничења семантичког типа која се не јављају у Родитељу (Fillmore et al. 2003, 311). Нпр. оквир *Изложити се_ризику* (*Run_risk*) наслеђује оквир *Вероватноћа* (*Likelyhood*).
– Однос *Коришћења* (енгл. *Using*) постоји када се неки оквир на веома уопштен начин позива на структуру неког апстрактнијег, схематизованог оквира. На пример, оквир *Клађење* (*Wagering*) користи оквир *Изложити се_ризику* (*Run_risk*); *Брзина* (*Speed*)

користи оквир *Покрет* (*Motion*); оквир *Брбљивост* (*Volubility*) користи оквир *Комуникација* (*Communication*) (Ruppenhofer et al. 2016, 83).

– *Угао_гледања* (енгл. *Perspective_on*) је релација слична општијој релацији *Коришћења*, али повезане оквире ограничава у већој мери (82). Примена ове релације указује на постојање бар два могућа угла гледања на један неутралан оквир. На пример, оквир *Сценарио_ризика* (*Risk_scenario*) је неутралан, док су оквири *Ризична_ситуација* (*Risky_situation*), *Бити_угрожен* (*Being_at_risk*) и *Изложити_се_ризику* (*Run_risk*) перспективизовани; оквири *Запошљавање* (*Hiring*) и *Добити_посао* (*Get_a_job*) перспективизују оквир *Почетак_радног_односа* (*Employment_start*) са становишта послодавца, односно запосленог.

Пошто се у базу унесу дефиниције оквира и његових елемената, оквиру на ком се ради додаје се ЈЈ (на пример, оквиру *Бити_угрожен* додаје се ЈЈ *ризик*). Затим се за дату ЈЈ уносе информације о врсти речи, значењу, формалном саставу (једна реч или вишечлани израз). Након тога прецизирају се претраге које ће омогућити да се из корпуса[13] екстрахују оне реченице (поткорпус) које садрже лексему коју испитујемо (у овом примеру именицу *ризик*) и чија је граматичка форма таква да указује на оно значење дате именице (ЈЈ) које је везано за оквир *Бити_угрожен* (*Being_at_risk*). Циљ је да се из поткорпуса избаце све оне реченице у којима задата кључна реч не представља ЈЈ која се везује за оквир који се креира. Пошто се спецификују претраге важне за дату ЈЈ, низ аутоматских процеса генерише поткорпус спреман за анотацију. Након што се одбаце предуге и друге неадекватне реченице, бира се од три до пет примера за сваки модел – циљ је илустровати разноликост модела, а не постићи статистичку репрезентативност.

Кад се заврши с анотацијом, примењују се алати за испитивање анотираних реченица и валентних модела који су у њима реализовани. Постоје два типа извештаја у виду динамичких веб-страна (*LexUnit Report*, *Lexical Entry Report*; извештаји се аутоматски генеришу пошто се заврши анотација и доступни су и на јавној веб-страни Фрејмнета). Први показује све анотиране реченице за једну ЈЈ и у њему се наводе сви елементи оквира пронађени у актуелном оквиру (табела елемената

13. Fillmore et al. (2003, 304) користе Британски национални корпус (енгл. British National Corpus).

оквира); сваки елемент се боји одређеном бојом, а истом том бојом елементи су обележени и у анотираним реченицама. Други представља резиме синтаксичких реализација елемената оквира и валентних модела ЛЈ у две табеле (Fillmore et al. 2003, 326–328).

Пошто се у Фрејмнету бележе и елементи оквира (за оквир специфичне семантичке улоге), као и њихове језичке реализације, за овај опис важни су и појмови *валентне групе*, *валентног модела* и *описа валенце*.[14] Елемент оквира заједно са својом граматичком реализацијом (тип јединице и њена функција у реченици) чини валентну групу; скуп валентних група реализованих у једној реченици чини валентни модел; скуп свих валентних модела које реализује једна лексичка јединица чини опис валенце (Atkins, Fillmore, and Johnson 2003, 255–257).

1.5 Примене мреже Фрејмнет

Фрејмнет је доступан на адреси https://framenet.icsi.berkeley.edu/. Може се претраживати и прелиставати онлајн, али и преузети и користити. Као што је на сајту истакнуто, може се користити у различите сврхе: као речник за учење језика (с обзиром на то да садржи више од 13.000 ЛЈ); као речник валенце; као скуп података за обуку програма који обележавају семантичке улоге,[15] што га чини значајним дигиталним језичким ресурсом за истраживања у области обраде природних језика (садржи више од 200.000 ручно обележених реченица повезаних са више од 1.200 семантичких оквира).

Фрејмнет је настао као лексичка база података за енглески, а касније су развијане базе за друге језике (француски, кинески, португалски, немачки, шпански, јапански и др.), кроз различите и независне пројекте, али уз исти формализам и концепцију развоја. Покренут је пројекат поравнавања датих база података за разне језике.

14. Својство глагола да за себе „везује“ зависне елементе (аргументе) назива се валенца. У зависности од броја аргумената које везују за себе, глаголи могу бити једновалентни (кад се могу употребити само са субјектом), двовалентни (са субјектом и објектом) итд.

15. У подтачки 1.6 представићемо нека од истраживања у којима се на разне начине користе Фрејмнет и програми за бележење семантичких улога у хрватском, словеначком и српском.

1.6 Претходна истраживања

У овом одељку осврнућемо се на истраживања која су се бавила бележењем семантичких улога (анотација семантичким улогама) у српском и њему сродним језицима и испитивањем значења именице *ризик* и глагола *ризиковати*.

У раду Gantar et al. (2018) представљен је модел за обележавање семантичких улога у словеначком и хрватском, развијан у оквиру међународног билатералног пројекта *Semantic Role Labeling in Slovene and Croatian*. Циљ је био развијање ручно анотираног корпуса који би се користио као база података за обуку за системе надгледаног машинског учења. Описан је и експеримент аутоматског обележавања семантичких улога заснован на надгледаном машинском учењу. За сваки корпус представљени су најфреквентнији глаголи, семантичке улоге и типични семантичко-синтаксички обрасци за најфреквентније глаголе. У оба корпуса најфреквентнији је глагол *бити* и семантичка улога пацијенса, а затим агенса (95–96). У раду су сематичке улоге бележене на стабилним семантичко-синтаксичким моделима (96–97), али остаје питање оправданоси таквог поступка будући да се семантичке улоге и оквири везују за ЈЈ, дакле за (глаголске) лексеме у једном од значења.

У раду Brač and Anić (2019) описан је пројекат чији је циљ развој методологије за бележење семантичких улога у језику специјализованих домена (конкретно авијације) који би се могао користити за све посебне области. Ауторке разматрају да ли је боље користити општије семантичке улоге или улоге специфичне за поједине глаголе и оквире, што је карактеристика Фрејмнета. Њихов закључак је да превелики број специфичних семантичких улога успорава анотацију а не доприноси значајно бољитку терминолошких ресурса, док се скуп општијих семантичких улога мора незнатно допунити (545).

У раду Wasserscheidt and Hrstić (2020) спроведено је интересантно истраживање о томе да ли лексеме које се користе и као термини (конкретно правни термини) и као речи општег лескичког фонда имају различита значења (или призивају различите оквире) у српском и хрватском језику (као језичким варијантама). Идеја је потекла од контрадикције у литератури посвећеној семантици оквира; наиме, у изворној Филморовој варијанти теорије указује се на разлике у оквирима које постоје између појединих говорника, друштвених група и култура; на те се разлике у радовима других аутора заборавља, па се оквири посматрају као универзалне структуре, независне од језика (88–89). Аутори су испитивали да ли *одредба* као правни термин

и реч општег лексичког фонда призива различите оквире у српском и хрватском, а испитивање је спроведено на корпусима ових језика. Коришћена је дистрибуциона анализа, чија је основна идеја да се значење речи може разумети на основу њеног контекста, али је тој анализи подвргнут и сам контекст. У анализи контекста примењена је и теорија семантике оквира (Wasserscheidt and Hrstić 2020, 90). На основу ове двоструке дистрибуционе анализе аутори су закључили да нема значајне разлике у значењу *одредбе* у посматраним корпусима, као и то да се метод двоструког кластеровања може користити у сложеној семантичкој анализи, која се пак може уклопити у појмовне структуре Фрејмнета[16] (108).

Иако није од непосредног значаја за наш рад (јер представљена анализа није заснована на теорији семантике оквира), поменућемо истраживање у ком се најпре примећује да је ризик постао важна тема истраживања у друштвеним наукама, али да се, када се говори о ризику, о значењу саме речи *ризик* не зна довољно (Hamilton, Adolphs, and Nerlich 2007, 164). Аутори истраживања, вођени поменутим запажањем, подвргавају анализи значење именице *ризик* и глагола *ризиковати* у корпусу разговорног језика (*the Cambridge and Nottingham Corpus of Discourse in English*, скраћено CANCODE). Испитујући семантичку склоност поменутих лексема, као и њихову семантичку прозодију, аутори истраживања закључују да на анализиране лексеме утиче контекст у ком се јављају (нпр. постоје разлике између њихових колоката и семантичке прозодије у интимној комуникацији, између чланова породице или партнера, у односу на комуникацију између професора и студената).

2. Неколико примера оквира из домена ризика

Као што је поменуто и на крају т. 1.3 овог рада, Филмор и Еткинсова разматрали су ограничења која су лексичкој анализи наметнута у традиционалном приступу лексикографији и у папирним издањима речника (Fillmore and B. T. Atkins 1992, 100–101), (Fillmore and S. Atkins 1994, 350–363). Пошто су упоредили анализе глагола *ризиковати* и именице *ризик* у референтним једнојезичним речницима с налазима о овим лексемама до којих су дошли анализом корпуса, схватили су да

16. Анализа је указала на то да се *одредба* јавља у чак 12 различитих оквира (Wasserscheidt and Hrstić 2020, 108).

ниједан од прегледаних речника не пружа задовољавајућу анализу и не бележи многе податке које су уочили прегледом примера из корпуса. Закључак је био да папирни речник, у ком је приказ линеаран, не може да представи тако сложен опис какав би био потребан да би се представили сви подаци важни за употребу неке речи. Због тога су осмислили онлајн речник заснован на оквирима, уместо папирног речника заснованог на лексемама, у ком је могуће представити резултате овакве анализе.

Овако замишљен онлајн речник омогућава да се прикажу поједини елементи оквира и њихове различите синтаксичке реализације, дакле све оно што спада у опис валенце (коју смо поменули у т. 1.4), као и односи између различитих оквира.

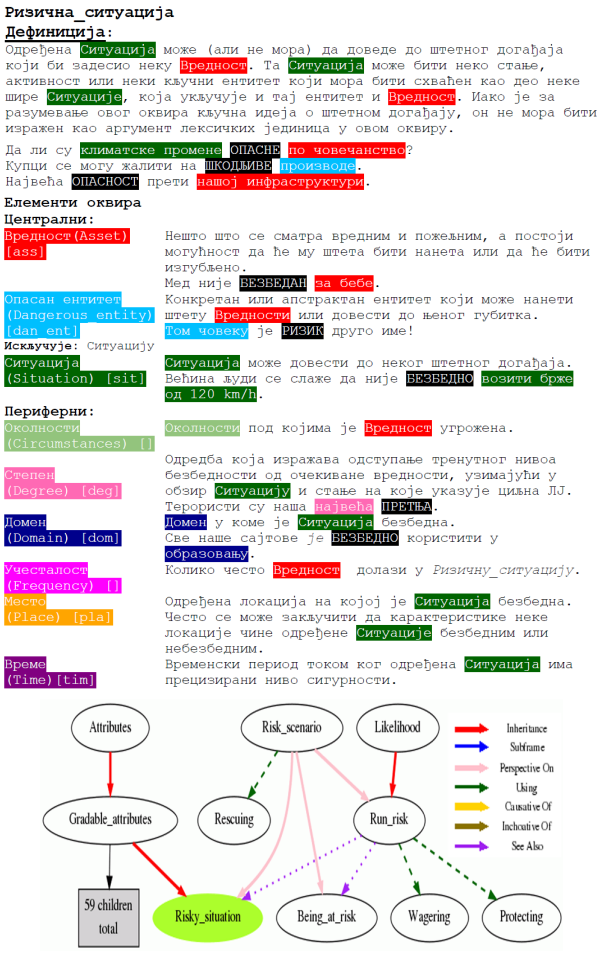
Алат за визуелизацију односа између оквира и елемената оквира (*FrameGrapher*)[17] пружа могућност да се изабере почетни оквир и систематски истраже везе између оквира. Слике 1–4 у наставку текста генерисане су управо коришћењем алата *FrameGrapher*.

2.1 Оквир *Ризична_ ситуација*

У наставку је представљен оквир *Ризична_ ситуација*.[18] Након дефиниције дају се примери, а потом централни и периферни елементи оквира. Као што је поменуто, сваком елементу оквира додељује се посебна боја, тако да се и у тексту дефиниције елементи, ако се појаве, боје одговарајућим бојама. Лексичке јединице које призивају оквир *Ризична_ ситуација* су: *опасност.n*, *опасан.a*, *ризик.n*, *рискантно.adv*, *ризичан.a*, *безбедан.a*, *безбедно.adv*, *небезбедан.a*, *претња.n*, *шкодљив/несигуран.a* (*n* је скраћеница за именице, *a* за придеве, *adv* за прилоге, *v* за глаголе). ЈЈ које призивају оквир маркирају се црном позадином у анотираном тексту. За сваки елемент оквира наводи се дефиниција и један обележен пример употребе.

17. FrameGrapher

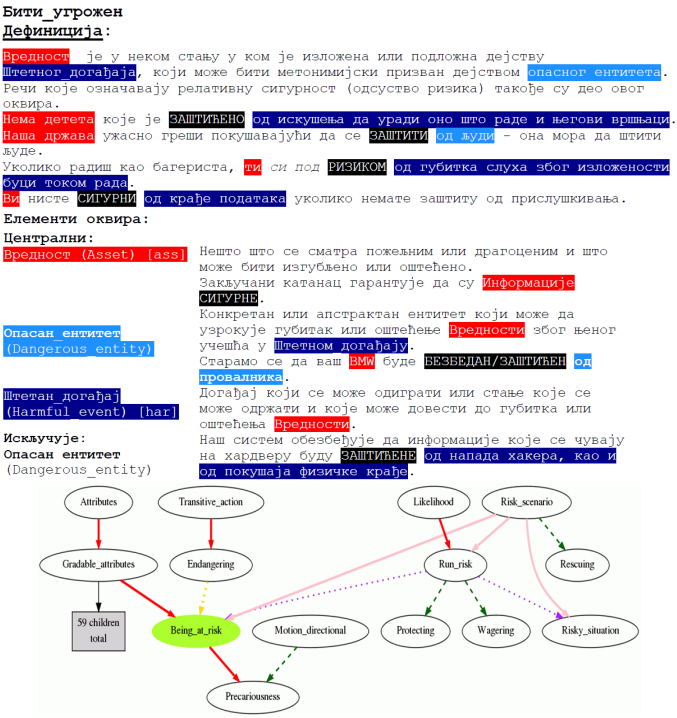
18. За потребе овог рада превели смо опис неколико оквира и њихових елемената, уз напомену да је реч о опису везаном за енглески, са примерима из корпуса енглеског језика, прилагођеним српском језику ради илустрације принципа. Надамо се да ћемо опис оквира заснован на подацима из корпуса српског језика представити ускоро.



Слика 1. Графички приказ односа између оквира *Ризична_ ситуација* (енгл. *Risky_ situation*) и с њим повезаних оквира

2.2 Оквир *Бити_угрожен*

Оквир *Бити_угрожен* (енгл. *Being_at_risk*) призивају следеће лексичке јединице: *опасност.n*, *несигуран.a*, *ризик.n*, *безбедан.a*, *сигуран.a*, *безбедност.n*, *поуздан.a*, *рањивост.n*, *рањив.a*. Боје одговарајућих елемената оквира исте су као у претходном примеру; овај оквир садржи и додатни елемент – *Штетни_догађај* (*Harmful_event*).



Слика 2. Семантички оквир *Бити_угрожен* (енгл. *Being_at_risk*)

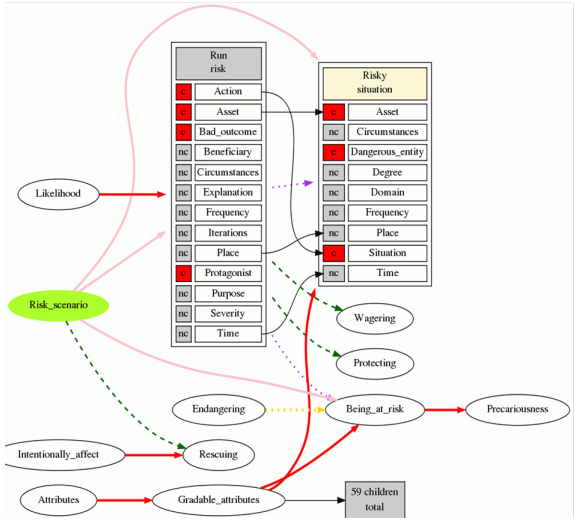
2.3 Оквир *Изложити се_ризику*

Оквир *Изложити се_ризику* (енгл. *Run_risk*) призивају следеће лексичке јединице: *угрожен.a*, *опасност.n*, *ризик.n*, *ризиковати.v*, *угрозити.v*. На слици 3 дати су дефиниција, примери и елементи овог оквира.

Слика 3. Семантички оквир *Изложити се_ризику* (енгл. *Run_risk*)

Сценарио_ризика
Дефиниција:
Вредност је у ситуацији за коју је вероватно да води до неког штетног догађаја, који ће лоше утицати на Вредност.
Елементи оквира:
Централни:

Вредност (Asset) [Ass]	Нешто што се сматра пожељним или вредним и што може бити изгубљено или оштећено.
Штетни_догађај (Harmful_event) [Har]	Догађај који може да се одигра или стане које може да потраје и које може довести до губитка или оштећења Вредности.
Ситуација (Situation) [Sit]	Ситуација у којој Вредност није безбедна или заштићена.
Периферни:	
Степен (Degree) [Deg]	Одредба која изражава одступање од актуелног нивоа безбедности за Вредност, Ситуација и стање означени самом циљном ЈЈ.
Место (Place) [Pla]	Место за које важи одређени ниво безбедности.
Време (Time) [Tim]	Време током ког важи одређени ниво безбедности.



Слика 4. Семантички оквир *Сценарио_ризика* (енгл. *Risk_scenario*)

2.4 Сценарио_ризика

Слика 4 представља оквир *Сценарио_ризика* (енгл. *Risk_scenario*), са детаљним приказом оквира *Изложити се_ризику* (*Run_risk*) и *Ризична_ситуација* (*Risky_situation*), за које су обележени централни (енгл. *core*, скр. *c*) и периферни (енгл. *non-core*, скр. *nc*) елементи. На десној страни слике дата је легенда с врстама релација које се успостављају између оквира, на пример: наслеђивање, угао гледања, коришћење (приказане су и неке од релација које нисмо помињали: подоквир, узрочник и др).

3. NLTK омотач за Фрејмнет

NLTK (енгл. *Natural Language Toolkit*) је платформа за развој програма *Python* за обраду текста на природним језицима, која је једноставна за коришћење, а садржи бројне корпусе и лексичке ресурсе чији се број континуално увећава. NLTK омогућава различите врсте обраде текста, међу којима су: класификација, токенизација (енгл. *tokenization*), стемовање (енгл. *stemming*), тагирање (енгл. *tagging*), парсирање (енгл. *parsing*) и семантичко закључивање (енгл. *semantic reasoning*). У оквиру NLTK система имплементирани су и омотачи (енгл. *wrappers*) за друге програмске библиотеке за обраду природних језика, као и за важне лексичке ресурсе. Један од расположивих ресурса у оквиру NLTK је и Фрејмнет, који прати програмска библиотека за претраживање овог ресурса и екстракцију информација из његових оквира.

Као што је поменуто у Уводу (одељак 1.1 овог рада), оквир је појмовна структура[19] која описује одређену врсту ситуације, ентитета или односа заједно са учесницима тих ситуација и односа. Физичка организација Фрејмнета у оквиру NLTK дистрибуције представља колекције XML (*Extensible Markup Language*) датотека разврстаних по каталозима: *frame*, *fulltext*, *lu*, *miscXML*, којима се приступа помоћу функција библиотеке или се могу директно претраживати и приказивати коришћењем XML докумената са XSL (енгл. *eXtensible Stylesheet Language*) трансформацијама: *frameIndex*, *luIndex*, *fulltextIndex*. У овом одељку ћемо приказати примере коришћења функција омотача Фрејмнета.

19. FrameNet и NLTK

Да би се добила листа свих оквира у Фрејмнету, може се користити функција `frames()`. Следећи код приказује иницијализацију рада са корпусом Фрејмнет и показује да верзија која је публикована уз NLTK има 1221 оквир.

```
from nltk.corpus import framenet as fn
len(fn.frames())
```

Да бисмо пронашли оквире који у називу садрже реч 'risk', користи се опција:

```
fn.frames(r'risk')
```

која као резултат даје:

[<frame ID=1560 name=Being_at_risk>,
<frame ID=378 name=Run_risk>].

Имајући у виду осетљивост на мала и велика слова, за упит:

```
fn.frames(r'Risk')
```

добија се другачији резултат:

[<frame ID=1763 name=Risk_scenario>,
<frame ID=1762 name=Risky_situation>]

Ако се зада регуларни израз '`(?i)risk`' функцији frame(), добија се листа од сва четири поменута оквира (т. 2.1–2.4), чија се имена подударају с прослеђеним обрасцем, јер '`(?i)`' указује на то да се не прави разлика између малих и великих слова.

Детаљи одређеног оквира могу се добити функцијом `frame()`, којој се проследи као параметар број оквира, рецимо `f=fn.frame(1762)`, што ће вратити податке за оквир Ризична_ситуација (*Risky_situation*).[20]

Уз то, може се прићи деловима оквира, рецимо називу оквира (`f.name`), његовој дефиницији (`f.definition`), елементима оквира (`f.FE`), лексичким јединицама (`f.lexUnit`), релацијама између оквира (`f.frameRelations`), што је илустровано следећим кодом:

```
f = fn.frame('Risky_situation')
print(sorted([e for e in f.FE]))
print([r for r in f.frameRelations])
```

20. Подаци за оквир *Ризична_ситуација*

који даје следећи резултат:

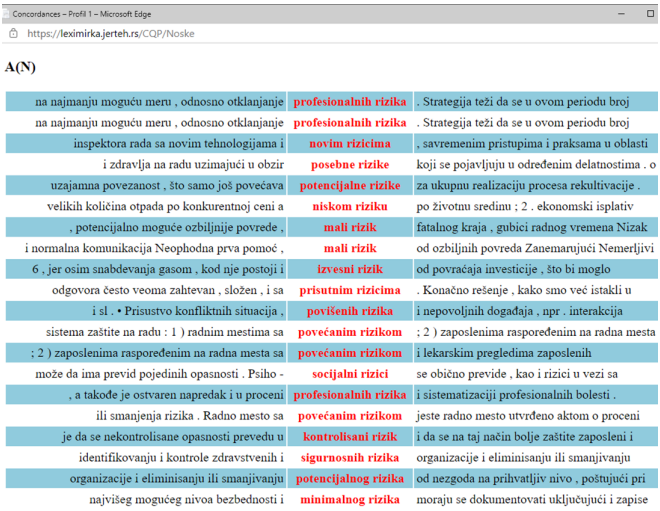
['Asset', 'Circumstances', 'Dangerous_entity', 'Degree', 'Domain', 'Frequency', 'Place', 'Situation', 'Time']
[<Parent=Gradable_attributes – Inheritance →
Child=Risky_situation>,
<MainEntry=Run_risk – See_also →
ReferringEntry=Risky_situation>,
<Source=Run_risk – ReFraming_Mapping →
Target=Risky_situation>,
<Neutral=Risk_scenario – Perspective_on →
Perspectivized=Risky_situation>]

4. Анализа речи *ризик* на корпусу рударства

Развој једнојезичног корпуса из домена рударства почео је као део пројекта везаног за управљање рударском пројектном документацијом коришћењем језичких технологија (Tomašević et al. 2018, 996). Тада је једнојезични корпус обухватао текстове из рударског домена и сродних истраживања, са укупно 172 документа (на српском језику) и 2,7 милиона речи у првом издању (997). Током даљих истраживања (Kitanović 2021) проширен је са додатна 63 документа. Тренутна верзија садржи 4,1 милион речи. Класификован према изворима, корпус обухвата пројектну документацију (26%), законодавство (11%), докторске дисертације (31%), уџбенике и осталу рударску литературу (32%) (Kitanović et al. 2021, 8).

Резултат претраге у CQL-у[21] (енгл. *Corpus Query Language*) анализира се на различите начине: листа фреквенција, колокације, конкорданце са ужим и ширим контекстом. Слика 5 приказује конкорданце, добијене из веб-апликације Лексимирка[22] за управљање електронским речницима (Stanković et al. 2018), за образац придев–именица за именицу *ризик*, док се на слици 6 може видети хистограм са фреквенцијама препознатих облика за исти образац, екстрахованих из рударског корпуса, доступног на корпусној платформи отвореног кода *NoSketch Engine* (Kilgarriff et al. 2004).[23] Инстанцу на локалном серверу одржавају чланови Друштва за језичке ресурсе и

21. Corpus Querying
22. Лексимирка
23. NoSketch на JePTexy, NoSketch Engine



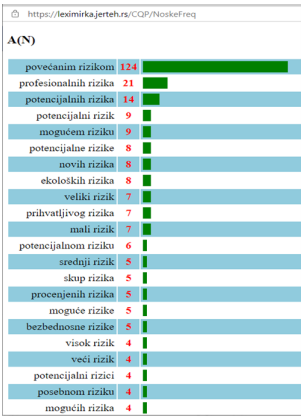
Слика 5. Конкорданце обрасца придев–именица за именицу *ризик*

технологије JePTex.[24] За тагирање корпуса обучен је *Treegger* модел за српски (Krstev and Vitas 2005; Utvić 2011), (Stanković et al. 2020, 3957), коришћењем ручно анотираног корпуса и Српских морфолошких речника (Krstev 2008).

Рударски корпус је публикован и на платформи *Sketch Engine*[25] (Kilgarriff et al. 2014), на којој се могу поставити различити упити. Рецимо, можемо извући конкорданце изабране леме или израза, потом колокације за лему, тезаурус сродних речи, затим такозвану скицу речи (енгл. *Word Sketch*) или видети разлику у употреби две сродне речи (енгл. *Word Sketch Difference*). Модул екстракције скице речи, који су развили Kilgarriff et al. (2004), помаже у изградњи Фрејмента и сличних ресура, убрзава доношење одлука о издвајању различитих значења вишезначних лексема (Baker 2012, 274).

Скица речи даје брзи преглед понашања изабране лексеме тако што прикупља информације из хиљада и милиона примера њене употребе и пружа сажетак категоризованих колокација на једној страници, с везама до појединачних примера. Слика 7 приказује пример скице речи за именицу *ризик* – један поглед на дату страницу довољан је да се

24. JePTex
25. Sketch Engine



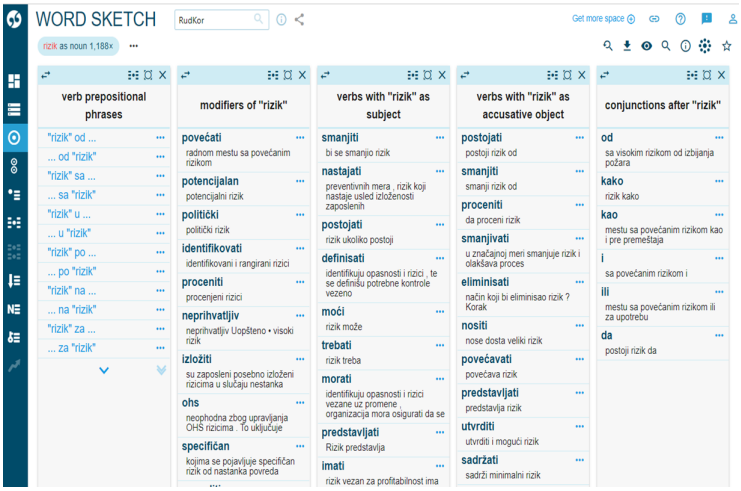
Слика 6. Хистограм са фреквенцијама облика обрасца придев–именица за именицу *ризик*

схвати како се та реч користи. Приказан је део скице речи: у првој колони видимо оно што се у српском језику назива предлошко-падежним конструкцијама (у енглеској граматичкој терминологији предлошким фразама):[26] *ризик од/са/у/по/на/за...*, за које је даље могуће доћи до конкорданци које одговарају конкретном обрасу кликом на "...". У другој колони представљени су модификатори речи *ризик*: трпни придев глагола (*повећан/идентификован/процењен/ ...* ризик) или придев (*потенцијалан/политички/неприхватљив/ ... ризик*). У трећој колони налазе се глаголи уз које се *ризик* јавља као субјекат, на пример: *смањити се/настајати/постојати/...* Даље следе обрасци у којима се *ризик* јавља као објекат уз глаголе: *смањити/проценити/ ...*

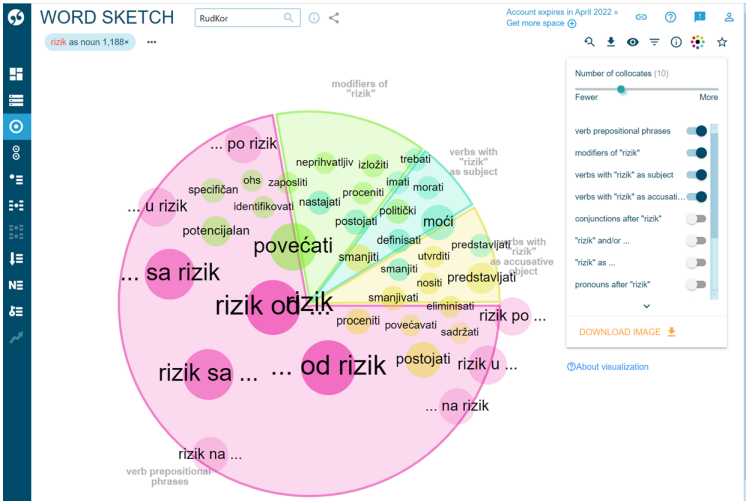
Приказ колокација у виду динамичког дијаграма дат је на слици 8. Може се видети да већину колоката чини образац предлошко-падежне конструкције (на слици: *verb prepositional phrases*). Десни део слике приказује могућности подешавања: који ће обрасци бити приказани и која је минимална фреквенција колокација да би биле укључене на граф.

Истраживање колокација од изузетног је значаја (на пример, у лексикографији, важно је навести најчешће колокате одређених ЛЈ; колокације су кључне и у учењу језика, али и у различитим задацима

26. Важно је напоменути да алати и аутоматско рангирање нису сасвим прилагођени за српски језик и да се јављају грешке, које се морају накнадно исправљати, али су свакако врло драгоцени.

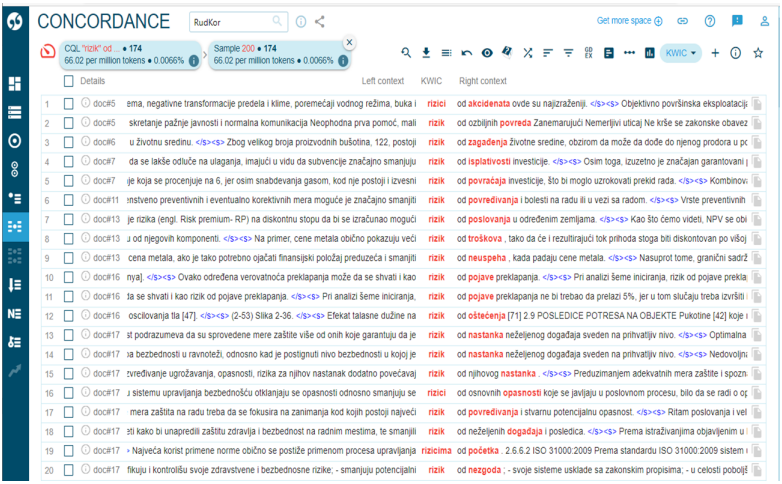


Слика 7. Скица речи *ризик* у алату *Sketch engine*



Слика 8. Графички приказ колокација за именицу *ризик* у алату *Sketch engine*

обраде природних језика). Полазећи од скице речи за облик *ризик* и колокацију *ризик од*, могу се приказати одговарајуће конкорданце (слика 9).



Слика 9. Конкорданце у алату *Sketch engine*

Скица речи даје брзи преглед за унапред припремљена правила, али се корисни резултати могу добити и CQL упитима. На пример, на питање *(Од) чега је ризик* (тј. *од чега постоји ризик*)?, одговор бисмо могли наћи следећим CQL упитом `[lemma="ризик"][tag="N"]`. Одговор на питање *Какав је ризик?* дао би упит `[tag="A"][lemma="ризик"]`; или, ако желимо да дозволимо размак између речи (не више од 5), онда бисмо написали упит на следећи начин: `[tag="V"][word!="."]0,5[lemma="ризик"]`.

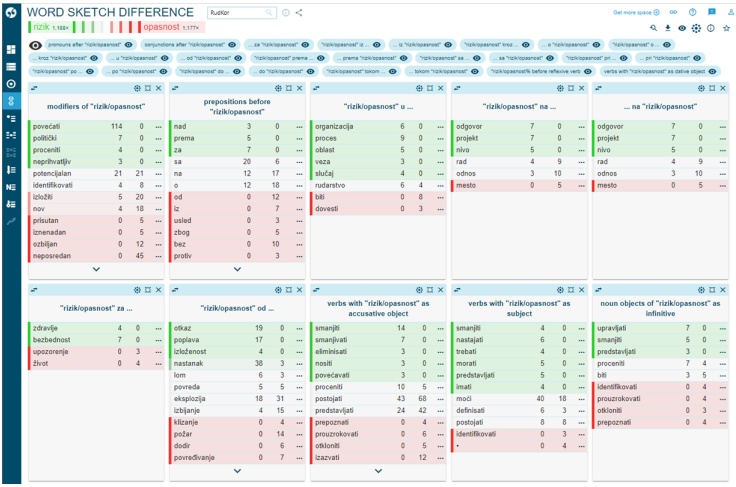
Осим што се може дати испис статистичких фреквенција, оне се могу приказати и визуелно, линијама на десној страни. Слика 10 приказује фреквенције појављивања именице *ризик* уз поједине придеве и глаголе.

Слика 11 приказује панел разлика у скицама речи (енгл. *Word Sketch Difference*). Панел разлика представља проширење модула скице речи. Генерише скице за две речи и упоређује их, што омогућава да се уоче разлике у њиховој употреби. Функција је посебно корисна за блискозначнице, за антониме и речи из истог семантичког поља. Са

		Lemma	Frequency ↓	Relative ?	
1	☐	povećati rizik	114	43.26	···
2	☐	potencijalan rizik	18	6.83	···
3	☐	velik rizik	11	4.17	···
4	☐	visok rizik	8	3.04	···
5	☐	specifičan rizik	7	2.66	···
6	☐	politički rizik	6	2.28	···
7	☐	mali rizik	6	2.28	···
8	☐	moguć rizik	6	2.28	···
9	☐	sav rizik	4	1.52	···
10	☐	nov rizik	4	1.52	···
11	☐	značajan rizik	4	1.52	···
12	☐	postojeći rizik	4	1.52	···
13	☐	izložiti rizik	4	1.52	···
14	☐	glavni rizik	4	1.52	···
15	☐	zaposliti rizik	3	1.14	···
16	☐	geološki rizik	3	1.14	···
17	☐	ekonomski rizik	3	1.14	···
18	☐	identifikovati rizik	3	1.14	···
19	☐	ekološki rizik	3	1.14	···
20	☐	izvestan rizik	3	1.14	···

Rows per page: 20 1–20 of 93 1 / 5

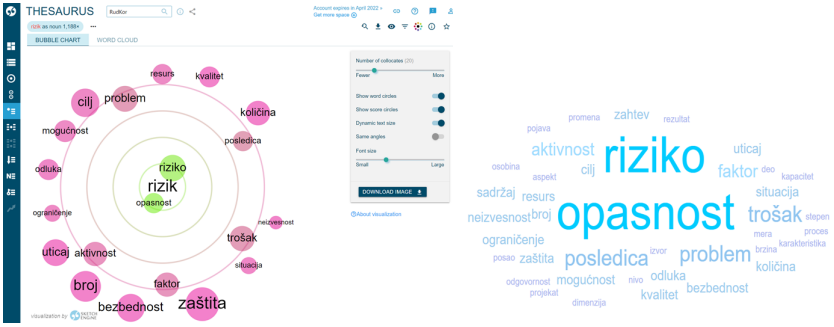
Слика 10. Фреквенције колокација за именицу *ризик* у алату *Sketch engine*



Слика 11. Скица разлика употребе речи *ризик* и *опасност*

слике 10 се, на пример, може видети да се уз *ризик* користе колокати *повећати*, *политички*, *проценити*, *прихватљив*, док се уз *опасност* јављају *непосредан*, *озбиљан*, *изненадан*.

Аутоматско генерисање тезауруса за задату реч проналази синониме или речи које припадају истој категорији (истом семантичком пољу) и приказује их у виду табеле, с могућношћу позивања одговарајућих конкорданци, скице речи, разлика, тезауруса. Могуће је и графичко приказивање тезауруса, као на слици 12, на којој се могу видети аутоматски генерисане речи које припадају истој категорији (семантичком пољу) као задата реч *ризик*, лево у виду графикона с мехурићима (енгл. *bubble graph*), а десно у виду облака речи. Листа речи тезауруса направљена је на основу контекста у којем се речи појављују у изабраном текстуалном корпусу, ослањајући се на теорију дистрибуционе семантике, која укратко каже да су речи које се појављују у истом контексту сличног значења. Да би се одредили синоними, упоређују се скице речи свих речи исте врсте, а оне које деле највећи део колоката наводе се као сличне речи. Оцена[27] дата за сваки синоним указује на проценат заједничких колоката.



Слика 12. Графички приказ тезауруса речи *ризик*

27. Статистичке формуле које се корите у алату *Sketch engine*: статистика/формуле

5. Закључак

У раду су приказани резултати прелиминарних истраживања у вези с могућностима примене теорије семантике оквира и принципа коришћених у изради семантичке мреже Фрејмнет на примерима из домена оквира ризика прилагођених српском језику. Уз то, представљена је платформа NLTK, погодна за коришћење разних језичких ресурса, као и систем *Sketch Engine* за корпусну анализу текста.

Показано је да анализа у оквиру Фрејмнета нуди детаљну, структурирану интерпретацију, која се даље може користити у различитим апликацијама за обраду природног језика, посебно у проналажењу и смисленом организовању текстова и у природнијој интеракцији између човека и рачунара, нарочито у све популарнијим дијалог-системима (енгл. *chatbot*). Систем дијалога који комуницира с корисницима мора бити у стању да препозна различите речи које призивају исту ситуацију или се односе на исти ентитет да би се извршило успешно препознавање намере.

Погодно је то што се енглески лексикон Фрејмнет може попунити лексичким подацима из других језика, рецимо српског (уз задржавање оних информација о неком оквиру које су заједничке за два језика и модификацију оних које су језички специфичне), тако да је модел погодан за вишејезичне ресурсе и апликације.

Представљена истраживања тек наговештају могућности прилагођавања Фрејмнета за српски језик и његовог поравнавања с таквим базама података за друге језике у будућности. Планирана даља истраживања треба да обухвате и могућности поравнања и заједничког коришћења српског Ворднета и Фрејмнета, као и њихово узајамно допуњавање. У том послу руководићемо се препорукама које дају Tonelli and Pighin (2009).

Овим истраживањима подстиче се и развој корпусне лексикографије српског језика, као и модернизација његовог граматичког и лексикографског описа. Тој модернизацији свакако би допринеле и студије случаја у којима би се поредиле обраде неких вишезначних лексема у Речнику САНУ с њиховим описом из угла теорије семантике оквира.

Могућности за будућа истраживања у овој области су велике. Примена теорије семантике оквира и методологије која се користи у изради Фрејмнета, као и представљених алата на српски језик захтеваће решавање бројних питања; на основу овог рада сматрамо да

ће бити веома изазовно сагледати уз помоћ појма нултог појављивања (*null-instantiation*) допуне транзитивних глагола које се не морају експлицирати, али се подразумевају (*кувати*, *писати* и сл.), као и начин на који се у описним речницима представљају такви случајеви. Уз то, сматрамо да би случајеве нултог појављивања (а у методологији Фрејмнета су препозната три таква случаја) било добро увести у српску граматику.

Захвалност

Овај рад је подржало Министарство просвете, науке и технолошког развоја Републике Србије према уговору бр. 451-03-9/2021-14, који је склопљен са Институтом за српки језик САНУ. Приступ алату SketchEngine је обезбеђен пројектом ELEXIS који се финансира из Horizon 2020 истраживачког и иновационог програма Европске уније, грант број 731015.

Литература

Atkins, Beryl T. S. 1994. "Analyzing the verbs of seeing: a frame semantics approach to corpus lexicography." In *Annual Meeting of the Berkeley Linguistics Society,* 20:42–56. 1.

Atkins, Sue, Charles J Fillmore, and Christopher R Johnson. 2003. "Lexicographic relevance: Selecting information from corpus evidence." *International Journal of lexicography* 16 (3): 251–280.

Baker, Collin F. 2012. "FrameNet, current collaborations and future goals." *Language Resources and Evaluation* 46 (2): 269–286.

Boas, Hans C., and Ryan Dux. 2017. "From the past into the present: From case frames to semantic frames." *Linguistics Vanguard* 3 (1): 20160003. https://doi.org/doi:10.1515/lingvan-2016-0003.

Brač, Ivana, and Ana Ostroški Anić. 2019. "From concept definitions to semantic role labeling in specialized knowledge resources." In *Proceedings of the 13th International Conference of the Asian Association for Lexicography,* 604–611.

Fillmore, Charles J. 1976. "Frame semantics and the nature of language." In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech,* 280:20–32. 1. New York.

Fillmore, Charles J. 1982. "Frame semantics." *Linguistic society of Korea (ed.), Linguistics in the morning calm,* 111–137.

Fillmore, Charles J, and Beryl T Atkins. 1992. "Toward a frame-based lexicon: The semantics of RISK and its neighbors." *Frames, fields and contrasts: New essays in semantic and lexical organization* 75:102.

Fillmore, Charles J, and Sue Atkins. 1994. "Starting where the Dictionaries Stop: The Challenge of Corpus Lexicography." In *Computational Approaches to the Lexicon,* edited by Sue Atkins and Antonio Zampolli, 349–393. Oxford: OUP.

Fillmore, Charles J, Miriam RL Petruck, Josef Ruppenhofer, and Abby Wright. 2003. "FrameNet in action: The case of attaching." *International journal of lexicography* 16 (3): 297–332.

Gantar, Polona, Kristina Štrkalj Despot, Simon Krek, and Nikola Ljubešic. 2018. "Towards semantic role labeling in Slovene and Croatian." In *Proceedings Conference on Language Technologies and Digital Humanities in Ljubljana,* 93–98.

Gildea, Daniel, and Daniel Jurafsky. 2002. "Automatic labeling of semantic roles." *Computational linguistics* 28 (3): 245–288.

Hamilton, Craig, Svenja Adolphs, and Brigitte Nerlich. 2007. "The meanings of 'risk': A view from corpus linguistics." *Discourse & Society* 18 (2): 163–181.

Jurafsky, Dan, and James H Martin. 2020. "Semantic Role Labeling and Argument Structure." Chap. 19 in *Speech and Language Processing,* 3rd ed. December 30, 2020 draft.

Kilgarriff, Adam, Vıt Baisa, Jan Bušta, Miloš Jakubıček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychl, and Vıt Suchomel. 2014. "The Sketch Engine: ten years on." *Lexicography* 1 (1): 7–36.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. "Itri-04-08 the sketch engine." *Information Technology* 105 (116).

Kitanović, Olivera. 2021. "Ontološki model upravljanja rizikom u rudarstvu." PhD diss., Univerzitet u Beogradu, Rudarsko-geološki fakultet. https://uvidok.rcub.bg.ac.rs/bitstream/handle/123456789/4305/Doktorat.pdf?sequence=1.

Kitanović, Olivera, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić, Ivan Babić, and Ljiljana Kolonja. 2021. "A Data Driven Approach for Raw Material Terminology." *Applied Sciences* 11 (7): 2892.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology of the University of Belgrade.

Krstev, Cvetana, and Duško Vitas. 2005. "Corpus and Lexicon-Mutual Incompletness." In *Proceedings of the Corpus Linguistics Conference,* 14:17.

Pradhan, Sameer, Wayne Ward, Kadri Hacioglu, James H Martin, and Dan Jurafsky. 2005. "Semantic role labeling using different syntactic views." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05),* 581–588.

Rasulić, Katarina, and Duška Klikovac. 2014. *Jezik i saznanje: hrestomatija iz kognitivne lingvistike.* Univerzitet u Beogradu, Filološki fakultet.

Ruppenhofer, Josef, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended theory and practice.* Revised November 1, 2016. https://framenet.icsi.berkeley.edu/fndrupal/the_book.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic dictionaries-from file system to lemon based lexical database." In *Proceedings of the 11th LREC - W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018), LREC 2018, Miyazaki, Japan, May 7-12, 2018,* 48–56.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proceedings of The 12th LREC – Language Resources and Evaluation Conference,* 3954–3962.

Tomašević, Aleksandra, Ranka Stanković, Miloš Utvić, Ivan Obradović, and Božo Kolonja. 2018. "Managing mining project documentation using human language technology." *The Electronic Library,* https://doi.org/10.1108/EL-11-2017-0239.

Tonelli, Sara, and Daniele Pighin. 2009. "New features for framenet-wordnet mapping." In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009),* 219–227.

Utvić, Miloš. 2011. "Annotating the Corpus of Contemporary Serbian." *INFOtheca* 12, no. 2 (December): 36a–47a.

Wasserscheidt, Philipp, and Andrea Hrstić. 2020. "Legal Variation? A Frame Analysis of Croatian and Serbian in the Domain of Law." *Mediterranean Language Review* 27:87–112.

Драгићевић, Рајна. 2007. *Лексикологија српског језика.* Београд: Завод за уџбенике.

Марковић, Александра. 2017. "Однос граматике и речника – граматика инхерентна описним речницима српског језика." *Наш језик* XLVIII (1-2): 27–43.

Поповић, Љубомир. 2003. "Интегрални речнички модели и њихов значај за лингвистички опис и анализу корпуса." *Научни састанак слависта у Вукове дане* 31 (1): 201–220.

FrameNet Lexical Database: Presenting a Few Frames Within the Risk Domain

UDC 81'322.2
DOI 10.18485/infotheca.2021.21.1.1

ABSTRACT: This paper gives a short overview of the frame semantics theory that forms the theoretical basis of the Berkeley *FrameNet* project. We present the basic concepts of this database, as well as the possibility of implementing it in Serbian. We also take a close look at the lexical analysis used in the FrameNet development project and point out the differences between the frame-based lexical analysis and its word-based counterpart. This is followed by an illustration of a couple of related frames evoked by words from the risk domain. FrameNet data is also readily available through the Python API included in the NLTK (*Natural Language Toolkit*) suite, which provides a good natural language processing resource. The last chapter shows a corpus search of the noun *risk* in a mining-themed corpus. We also present its most common collocates, word sketch, individual pattern concordances, thesaurus entry of its synonyms and related words, collocation frequency graphs. A word cloud for the word *risk* is also included.

KEYWORDS: Serbian language, frame semantics, FrameNet, risk scenario, mining corpus, natural language processing.

PAPER SUBMITTED: 15 July 2021
PAPER ACCEPTED: 6 September 2021

Aleksandra Marković
aleksandra.markovic@isj.sanu.ac.rs
Institute for Serbian Language, SASA
Belgrade, Serbia

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
University of Belgrade
Faculty for Mining and Geology
Belgrade, Serbia

Natalija Tomić
ntomic@hotmail.com
University of Belgrade
Faculty for Mining and Geology

Olivera Kitanović
olivera.kitanovic@rgf.bg.ac.rs
University of Belgrade
Faculty for Mining and Geology
Belgrade, Serbia

1 Introduction

Charles Fillmore's Frame Semantics Theory is a cognitive theory of meaning that links word meanings to the syntactic context in which they occur (Atkins, Fillmore, and Johnson 2003, 254). Word sense analysis is tra-

ditionally left to lexicographers and those interested in semantics. However, if the aim is to show the manner in which a word is actually used, an analysis of corpus data proves to be a fairly complicated task, in view of the number of concordances proposed by contemporary corpora for certain key words. Frame semantics theory, as cited by the following authors (Atkins 1994; Gildea and Jurafsky 2002; Atkins, Fillmore, and Johnson 2003; Pradhan et al. 2005; Boas and Dux 2017; Jurafsky and Martin 2020), gives a reliable, scientifically valid way of approaching word usage analysis and description. The basis of this approach is the idea that every experience that we memorize occurs in some meaningful context and our ability to memorize those experiences stems from the existence of mental schemas that we possess giving meaning to objects, relationships and events. Fillmore argues that words are learned within such meaningful contexts, and that context is also essential to the process of comprehension, when we evoke specific experiences through which we learned the meaning of a word. A frame identifies the type of experience and provides its structure and coherence, lending meaning to entities, events and relations that make it up (Fillmore 1976, 26).[1]

1.1 The design of FrameNet

FrameNet[2] is a lexical database of English based on annotated examples of how a *lexical unit* (hereinafter abbreviated as LU) is used in an actual7 texts. The basic premise comes down to the fact that most LUs are best defined through semantic frames, a conceptual structure that provides a description of the type of situation, relation or entity and the participants involved in it (Ruppenhofer et al. 2016, 7). For example, taking a risk typically involves the following: a person taking the risk that is central to the RISK scenario or the Protagonist. The Protagonist takes a risk willingly or otherwise or runs the risk; possible Bad outcome or Harmful event; the Decision which may lead to a bad outcome; a Purpose; an Action; certain Circumstances in which the protagonist stands; an Asset (a person or an object), perceived by the Protagonist as desirable, all of which is compromised in the RISK scenario (Fillmore and S. Atkins 1994, 367).

1. The term *frame* in Fillmore's usage denotes a general signifier that can be referred to as schema, scenario, cognitive model, folk model, etc. (Fillmore 1982, 111).

2. The project has been in development at The International Computer Science Institute in Berkeley since 1997.

1.2 Frame semantics lexical analysis

Frame semantics based lexical analysis comprises an analysis of the meaning of an LU, its lexical surroundings, phrases and grammatical constructions in which it appears in the corpus, the context in which it is used provided by corpus examples, as well as all the phrases in which the LU fulfills its full semantic potential. This approach consists of listing all LU arguments and adjuncts crucial to describing its meaning. Special attention is given to words that cannot be defined outside of the frames they are associated with. Those words are called frame-evoking words and are primarily verbs, but they also include nouns, adjectives and adverbs (Atkins, Fillmore, and Johnson 2003, 252).[3]

The basic units of a FrameNet analysis are frame and LU, a lexeme used in one of its senses (Fillmore et al. 2003, 297), (Ruppenhofer et al. 2016, 7).[4] In contrast to the standard lexicographic practice, which includes listing all the senses of a word in as much detail as possible, the LU in FrameNet is defined together with other LUs that belong to the same frame (Fillmore et al. 2003, 299)).[5] That is how, when we have defined the *Being_at_risk* frame, we can then define the nouns *risk*, *danger*, *safety*, *vulnerability*; adjectives *insecure*, *safe*, *secure*, *susceptible*, *vulnerable*, etc. with reference to the frame in question.

The process of describing a LU in FrameNet is defined in (Fillmore et al. 2003).[6] It begins with an informal description of the frame which a LU

3. The Frame semantic theory inspired us to point out the necessity of citing relevant constructions alongside the description of word meaning in the descriptive dictionaries of Serbian for all of the four most common frame-evoking word classes (nouns, adjectives, verbs and adverbs) (Марковић 2017, 34–41).

4. In Serbian lexicographic literature, as well as in syntax papers that explore the relationship between grammar and dictionaries, different terminology is used for what is referred to as lexical unit within FrameNet (e.g. in a university textbook of lexicology, that what is called a lexical unit refers to a lemma or a vocabulary entry, (Драгићевић 2007, 30), while Lj. Popović insists on shifting the focus to individual word senses and a lexeme used in one of its meanings is dubbed a sublexeme in his terminology (Поповић 2003, 202–203). In this paper, we decided to use the term *lexical unit* in order to stay within the framework's terminology.

5. Here we are referring to two approaches to describing lexical meaning, one that is *word-based* and the other *frame-based* (Atkins, Fillmore, and Johnson 2003, 254).

6. Although the process is described as an ordered sequence of steps, the authors still call for revising the data at any point and going back and correcting it if necessary (Fillmore et al. 2003, 299).

belongs to, a description of the situation or event represented by the frame and creating a list of words whose meaning would be described with reference to that frame (Fillmore et al. 2003, 299).[7] After that a target LU for which annotation is being done is chosen; that is typically one word but can be a multi-word unit or a phrase (Ruppenhofer et al. 2016, 21) and its use is looked into by extracting sentences, which contain it, from the corpus.

A lexicographer working in FrameNet compares his or her insight into the meaning of a target lexeme, based on corpus examples, to the meaning given in descriptive dictionaries.[8] Once he gets a clearer idea of its meaning, the lexicographer tries to describe the frame the LU belongs to more closely. After that, he writes the definition of the frame – a schematic description of an event which is central to a word, along with the names of participant roles called frame elements. The way in which frame elements are expressed in sentence examples of the target LU is lexicographically relevant (Fillmore et al. 2003, 304–305).

1.3 Frame elements

Frame elements have often been viewed as an extension of semantic roles (agent, experiencer, patient), but they are defined as *frame-specific*. This stems from a multitude of reasons, the most prominent being the ability to create a detailed definition of frame elements, which is not afforded when trying to fit the role into a predefined set (305).

First, the central elements of the frame (*core elements*) need to be identified.[9] Core elements are essential as they identify the frame as unique and set it apart from other frames. Alongside the core elements, there are *non-core*

7. That description entails:1) a schematic description of entity types or situation illustrated by the frame; 2) choosing descriptive labels for describing the frame; 3) drawing up a draft list of words that belong to the frame (if an LU belongs to a frame, it means that it can be subjected to the same analysis as other LUs in the frame) (Fillmore et al. 2003, 297).

8. Having analyzed the definition of the verb to risk in ten general-use dictionaries of English, Fillmore and Atkins concluded that even dictionaries of a similar size and purpose do not feature the basic meanings of the verb, which are part of basic vocabulary (Fillmore and S. Atkins 1994, 353).

9. There are some formal characteristics that help determine element centrality (e.g. core elements need to be expressed and so do those that have an interpretation even though they are not expressed (e.g. in the sentence *John arrived* the place where John arrived, the GOAL element, is not expressed but is still interpreted in the context (Ruppenhofer et al. 2016, 23–24).

elements that appear in all the frames in which an agent performs an action (they usually denote Place, Time, Manner, Instrument).[10] The situations where core elements are not linguistically expressed also occur, but they are still mandatory in the conceptual structure of the frame; this is called *null-instantiation* and is also annotated in the database (320). (Fillmore et al. 2003, 320). After the core and non-core elements are identified, we can move on to defining the frame itself.[11]

After analyzing the verb *to risk* in descriptive English dictionaries, Fillmore and Atkins discovered that not enough attention is given to its arguments (although they are very important for describing the word's meaning and essential in L2 English dictionaries) and that there are other sentence constituents that are completely overlooked in dictionaries, but need to be singled out and well-described in order to demonstrate correct verb usage. For example, an action performed by a person who is risking something (and can be syntactically expressed in multiple ways): She risked her life *trying to save a drowning child*; an objective someone has when putting themselves at risk: She risked her life *in order to save mine* (Fillmore and S. Atkins 1994, 362). An action by means of which someone takes a risk is one of the core elements of the frame, while the objective because of which they are taking it is non-core.

1.4 Frame-frame relations – FrameNet

After a frame and its elements are defined, a frame is connected to other frames. In that way frames, their elements and LUs belonging to them are placed in the semantic space (Ruppenhofer et al. 2016, 79) and make up a network. Creating frame-to-frame relations allows us to see and record semantic generalizations based on the type of participants, events, etc. A frame can be connected to frames it inherits from, has a perspective on, is perspectivized in, its subframes as well as the ones it uses. Frame-to-frame

10. The core/non-core distinction in the broadest terms corresponds to arguments and adjuncts in the traditional grammatical analysis (Fillmore et al. 2003, 310). Non-core elements cannot function as subject or object of the target verb and are often expressed by using an adverb or a prepositional phrase (319).

11. Ruppenhofer et al. (2016, 65) define other frame elements as well: elements that appear in subordinate clauses are non-core or extra-thematic e.g. TIME, MOTIVE. In addition to these, there are core-unexpressed elements that are considered core but do not have to be inherited by a child-frame (24–25). This paper does not get into detail about either of them.

relations are directed or asymmetrical: the more abstract and independent frame is called *Super_frame* and the more dependent and less abstract frame is called *Sub_frame* (Ruppenhofer et al. 2016, 79).

A list of frame-to-frame relations has been defined with the following ones being the most important (79–84):

- Within the *Inheritance relation* the Sub_frame is a more specific version of a more abstract parent frame. All the frame elements of the parent have a specified mapping with the frame elements of the child, while the child can have Sub_frames, FEs and semantic constraints specific only to itself (Fillmore et al. 2003, 311). For instance, the frame Run_risk Inherits from the frame Likelihood.
- The *Using relation* exists when a frame makes a general reference to the more abstract frame. An illustration of this would be the following frames: Wagering which uses the frame Run_risk; Speed which uses the frame Motion; Volubility which uses the frame Communication (Ruppenhofer et al. 2016, 83).
- *Perspective_on is* a relation similar to the broader relation of Using, but it puts greater constraints on the frames bound by it (82). In order for this relation to be possible, there need to be at least two perspectives for viewing a neutral frame. For instance, the frame *Risk_scenario* is a neutral frame, while the frames *Risky_situation*, *Being_at_risk* and *Run_risk* are all perspectivized; the situation is viewed from the perspective of one of the participants. The frames *Hiring* and *Get_a_job* are both perspectives on a neutral form of *Employment_start*, from employer and employee perspective.

After the definitions of the frames and their elements have been entered into the database, LUs can be added to the frames (in the case of *Being_at_risk*, the LU *risk* would be added). This is followed by the information on word class, meaning, formal composition (whether it is a single word or a multiword expression), after which instructions are given on how the corpus [12] can be searched in order to extract the concordances (subcorpus) that contain the exact lexeme we are looking for (in our case the noun risk) whose grammatical form points to the LU which belongs to the frame *Being_at_risk*. The aim is to weed out all the instances in which the searched keyword does not represent the LU that belongs to the frame which is being created. After the suitable searches for the desired LU have been specified, a number of

12. Fillmore et al. (2003, 304) use British National Corpus.

automated processes generate a subcorpus ready for annotation. This subcorpus is then cleaned of sentences that are too long or in any other way inadequate, and from those three to five sentences are chosen for each pattern with the aim of illustrating the variety of existing patterns rather than their statistical representativeness.

When the annotation is over, tools for analyzing the annotated sentences and the valence patterns instantiated within them are used. There are two types of reports in the form of dynamic web-pages (*LexUnit Report* and *Lexical Entry Report*) which are automatically generated after the annotation is finished and are available on the FrameNet website. The first report shows all the annotated sentences for an LU. Moreover, all the elements found in the current frame are listed (in a table of frame elements) and each element is color coded in the table, as well as in the annotated sentence. The second report gives an overview of the syntactic realizations of the frame elements and LU valence patterns in two tables (Fillmore et al. 2003, 326–328).

Since FrameNet also annotates frame elements (for frame-specific semantic roles) and their lexical realizations, terms like *valence group*, *valence pattern* and *valence description* are also important.[13] A frame element, together with its grammatical realization (unit type and its role in a sentence) constitutes a valence group, a set of valence groups used in a sentence makes up a valence pattern and the set of all valence patterns that a particular LU uses makes up a valence description (Atkins, Fillmore, and Johnson 2003, 255–257).

1.5 Different applications of FrameNet

FrameNet is available on the website. It can be searched and scrolled through online, but also downloaded and used locally. As the website states, it can be used for different purposes: as a dictionary for language learning (since it contains more than 13,000 LUs); as a valence dictionary; as a training dataset for semantic role labeling[14] which makes it a rich digital language resource (with over 200,000 manually annotated sentences linked to over 1,200 semantic frames).

13. The property of verbs to take arguments is called *valence*. Depending on the number of arguments they take, verbs can be: *monovalent* (when they require a subject), *divalent* (when they require a subject and an object), etc.

14. Subsection 1.6 will give an overview of some of the research done on the use of FrameNet and semantic role labeling programs for Croatian, Slovenian and Serbian.

FrameNet was conceived as a lexical database of English, which incorporates the databases subsequently developed for other languages (French, Chinese, Portuguese, German, Spanish, Japanese etc.) as part of various independent projects, applying the same formal structure and concepts. A project for aligning the data created for different languages has also been launched.

1.6 Previous research

In this section we will look into the research done in the field of semantic role labeling for Serbian and the languages related to it, as well as into the research devoted to the meaning of the noun risk and the verb to risk in discourse.

In the paper (Gantar et al. 2018) a model of semantic role labeling for Slovenian and Croatian was presented that they had developed as part of the international bilateral project *Semantic Role Labeling in Slovene and Croatian*. The objective was to develop a manually annotated corpus that would be used as a training dataset for supervised machine learning systems. An automatic semantic role labelling experiment, based on supervised machine learning is also described in the paper. The most frequent verbs, semantic roles and typical semantic-syntactic patterns of the most frequent verbs were presented for each of the corpora. The verb *to be* and the semantic role of patient were the most frequent in both corpora, while the second place went to the role of agent (95–96). In the paper, semantic roles were labeled in stable semantic-syntactic models (96–97), but the question of whether this is a valid method remains because semantic roles and frames are formed around a LU, a (verb) lexeme in one of its senses.

The paper Brač and Anić (2019) showcases a project aimed at developing a methodology for semantic-role labeling in a domain-specific language (in their case the domain of aviation) that could also be used in other fields. The authors of the paper examined whether it would be better to use more general semantic roles or verb-specific and frame-specific roles, typical of FrameNet. They came to the conclusion that too many specific semantic roles slow down the annotation process, but do not, in turn, contribute significantly to the improvement of terminology resources, although they noted that the list of broader semantic role labels needed to be slightly expanded (545).

The paper Wasserscheidt and Hrstić (2020) presents interesting research done for Serbian and Croatian (viewed as varieties of one language) on lexemes that both enter the general lexicon and form part of a certain professional domain (in this case legal terminology). It focused on whether or not

they take different meaning (evoke different frames) in Serbian and Croatian. The idea came from the authors noting a contradictory stance in the literature on frame semantics. Namely, Fillmore's works point to a difference in frames that individual speakers, social groups and cultures have, but later papers by other authors overlook this fact and treat frames as universal language-independent structures (88–89). The authors of the paper explored the meaning of the word *odredba* (section of a legal act) within the legal framework and the general lexicon (where it can be used as a synonym for a legal act as a whole) in both Serbian and Croatian corpus data. They used distributional analysis whose main tenet is that word meaning can be defined based on the context in which the word appears, and additionally applied the analysis on the context itself. Frame semantics theory was used to analyze the context (90). In view of the findings of these two distributional analyses, the authors concluded that there was no significant difference in the meaning of *odredba* in the corpora under examination and that the method of double clustering can be used in complex semantic analyses, which can then be represented through FrameNet structures (108).[15]

Although not directly related to our topic of FrameNet, we would still like to mention a paper that notes that *risk* has become a prominent topic in social science research with the research into the meaning of the word itself remaining vaguely defined (Hamilton, Adolphs, and Nerlich 2007, 164). Guided by this notion, the authors continue to analyze the meaning of the noun *risk* and the verb *to risk* using *the Cambridge and Nottingham Corpus of Discourse in English*, abbreviated as CANCODE. Analyzing the semantic tendencies of these lexemes and their semantic prosody, they conclude that the target lexemes are influenced by the context in which they appear (for example, there is a difference between their collocations and semantic prosody in a more intimate setting between family members and partners as opposed to student-professor exchanges).

2 A Couple of Instances from the Risk Domain

As cited above, at the end of Subsection 1.3, Fillmore and Atkins discuss the constraints on lexical analysis put by the traditional approaches to lexicography and the form of descriptive dictionaries (Fillmore and B. T. Atkins 1992, 100–101), (Fillmore and S. Atkins 1994, 350–363). After they juxta-

15. The analysis indicated that *odredba* is part of as much as 12 frames (Wasserscheidt and Hrstić 2020, 108).

posed the analyses done for the verb *to risk* and the noun *risk* in monolingual dictionaries and corpus data, they concluded that the dictionaries do not give a comprehensive enough description, with a lot of the meanings found through corpus search not even being mentioned. The finding was that printed dictionaries, with a linear approach to meaning, cannot represent a complex description needed to provide all the data of significance for the ways in which a word is used. This was the motivation for creating an online dictionary whose entries are frames rather than lexemes, as found in paper dictionaries, providing a notation better suited to such a complex system.

Conceived in such a manner, an online dictionary allows for representation of individual frame elements and their diverse syntactic realizations and therefore a full description of an element's valence (described in Subsection 1.4) as well as the relations between frames.

A visualization tool for viewing the relations between frames and their FEs (*FrameGrapher*)[16] makes it possible to choose the target frame and explore its relations to other frames. Figures 1–4 in this paper have been generated using this tool.

2.1 Frame Risky_situation (*Ризична_ситуација*)

The frame Risky_situation is shown below.[17] After giving a definition, we see illustrative examples in the form of sentences, as well as core and non-core elements of the frame. As mentioned above, all the FEs are color coded, with the same color that is used in the FE list appearing in the definition. The LUs evoking the frame Risky_situation are: *опасност.n* (*danger.n*), *опасан.a* (*dangerous.a*), *ризик.n* (*risk.n*), *рискантно.adv* (*riskily.adv*), *ризичан.a* (*risky.a*), *безбедан.a* (*safe.a*), *безбедно.adv* (*safely.adv*), *небезбедан.a/шкодљив.a* (*unsafe.a*), *претња.n* (*threat.n*). Frame-evoking LUs in the annotated example sentences are highlighted in black. A definition is given for each FE and followed by an example of its use.

16. FrameGrapher

17. For the purpose of this paper, we took original English frames and their elements, based on the data from English language corpora, and translated them into Serbian in order to illustrate the way of presenting data in FrameNet. It is our hope that we will soon get a chance to illustrate frames using Serbian corpus data.

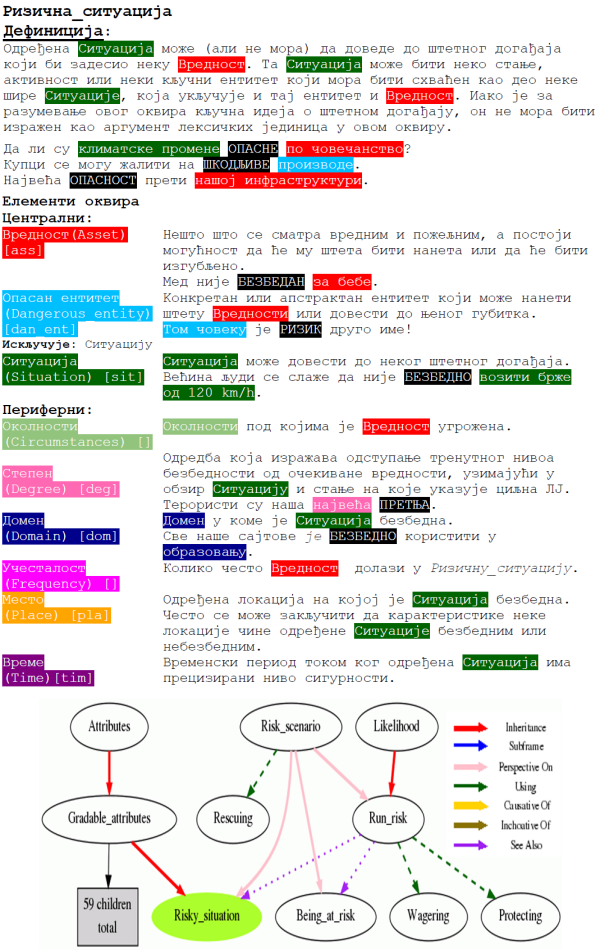


Figure 1. An illustration of the frame *Risky_ situation* and the related frames

2.2 Frame Being_at_risk (*Бити_угрожен*)

The LUs which evoke the frame Being_at_risk are: *опасност.n* (*danger.n*), *несигуран.a* (*insecure.a*), *ризик.n* (*risk.n*), *безбедан.a* (*safe.a*), *сигуран.a* (*secure.a*), *безбедност.n* (*safety.n*), *поуздан.a* (*reliable.a*), *рањивост.n* (*susceptibility.n*), *рањив.a* (*susceptible.a*). This frame contains the same FEs as the previous frame with the addition of Harmful_event (*Штетан_догађај*) and has the same color coding.

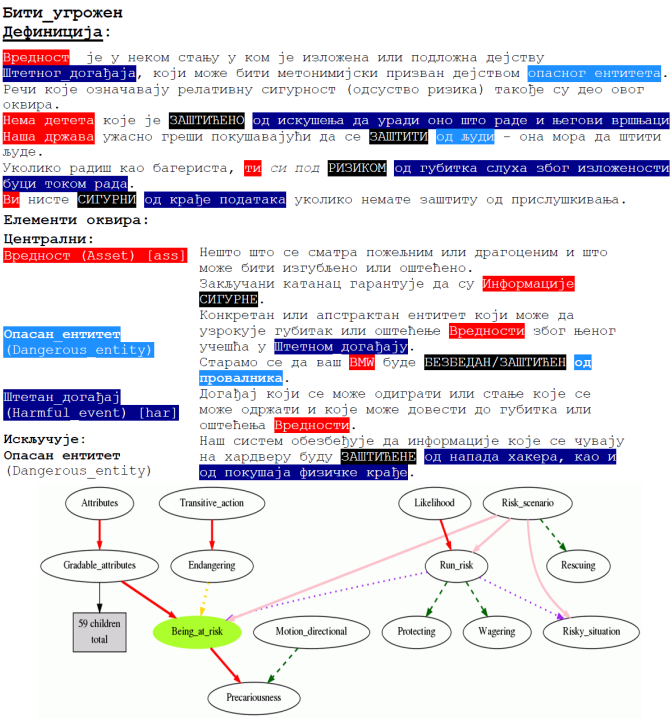


Figure 2. Semantic frame *Being_at_risk*

2.3 Frame Run_risk (*Изложити_се_ризику*)

The LUs evoking the frame Run_risk are: *угрожен.a* (*endangered.a*), *опасност.n* (*peril.n*), *ризик.n* (*risk.n*), *ризиковати.v* (*risk.v*), *угрозити.v*

(*endanger.v*). The definition, examples and FEs of the frame are given in Figure 3.

Figure 3. Semantic frame Run_risk

2.4 Frame Risk_scenario (*Сценарио_ризика*)

Figure 4 illustrates the relations between the frame Risk_scenario (*Сценарио_ризика*) and frames Run_risk (*Изложити се_ризику*) and Risky_situation (*Ризична_ситуација*) whose characteristics are shown in detail with their core (abbreviated as c) and non-core (abbreviated as *nc*)

elements listed. On the right-hand side there is a legend showing different types of frame-to-frame relations e.g. Inheritance, Perspective on, Using (as well as some of the relations we did not mention: Causative of, Subframe, etc.).

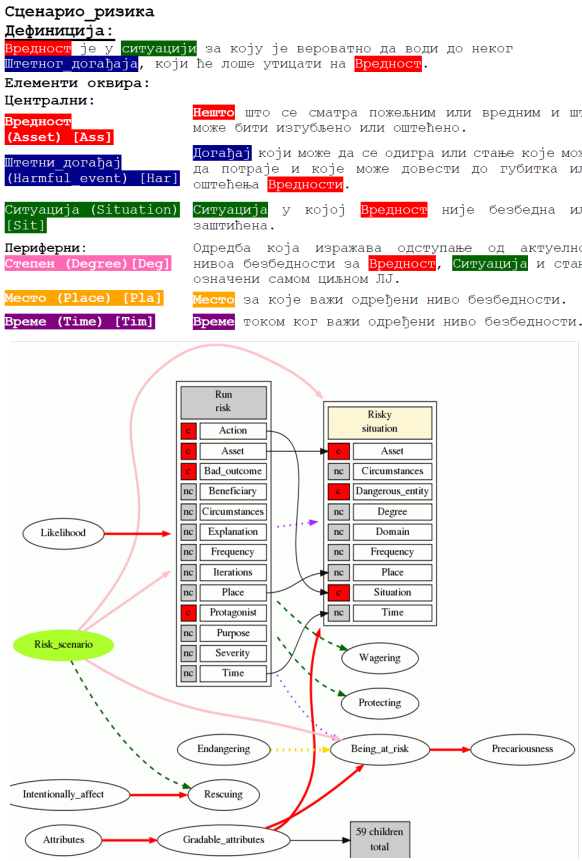


Figure 4. Semantic frame *Risk_ scenario* with a detailed view of two other related frames)

3 NLTK FrameNet Wrappers

NLTK (*Natural Language Toolkit*) is an easy-to-use natural language processing Python suite that accesses continually increasing number of corpora and lexical resources. NLTK offers different types of text processing, amongst which are: classification, tokenization, stemming, tagging, parsing and semantic reasoning. The NLTK system uses wrappers for other Python natural language processing and lexical resource libraries. One of the APIs available within NLTK is FrameNet and the accompanying program library designed for searching this resource, as well as for extracting information from it.

As mentioned in the Introduction (Section 1.1 of this paper), a frame is a conceptual structure describing a type of situation, entity or relation together with its participants. The structure of FrameNet within the NLTK framework is comprised of a collection of XML (*Extensible Markup Language*) files catalogued as: *frame*, *fulltext*, *lu*, *miscXML*, which are accessed through the library's commands or can be directly searched and visualized by means of XML files using XSL (*eXtensible Stylesheet Language*) transformations: *frameIndex*, *luIndex*, *fulltextIndex*. In this section, we will show the use of the FrameNet wrapper.

The function `frames()` lists all the frames contained in FrameNet. The following lines of code illustrate the initialization of working with FrameNet and return the information that the FrameNet version available in NLTK contains 1221 frames.

```
from nltk.corpus import framenet as fn
len(fn.frames())
```

In order to find all frames that contain the word *risk*, we use the command:

```
fn.frames(r'risk')
```

which outputs the following information:

[<frame ID=1560 name=Being_at_risk>,
<frame ID=378 name=Run_risk>].

Since the query is case-sensitive, we need to do a second search in order to find all the instances in which *risk* appears:

```
fn.frames(r'Risk')
```

which outputs a different result:

[<frame ID=1763 name=Risk_scenario>,
<frame ID=1762 name=Risky_situation>]

If the function `frame()` is given a regular expression '`(?i)risk`' as an argument, we get a combined list of the two, containing all four frames (sections 2.1–2.4), whose names correspond to the given pattern because '`(?i)`'expresses that the case of the letter is irrelevant.

The details of a frame can be listed through the command `frame()`, which is given the number of the frame as an argument, for instance `f=fn.frame(1762)`, returns all the data of the frame Risky_situation.[18]

Individual components of the frame can be accessed separately through the commands like: `f.name` giving the name of the frame, `f.definition` giving its definition, `f.FE` listing the elements of the frame, `f.lexUnit` giving frame LUs, `f.frameRelations` giving frame relations, as shown in the following example:

```
f = fn.frame('Risky_situation')
print(sorted([e for e in f.FE]))
print([r for r in f.frameRelations])
```

that outputs:

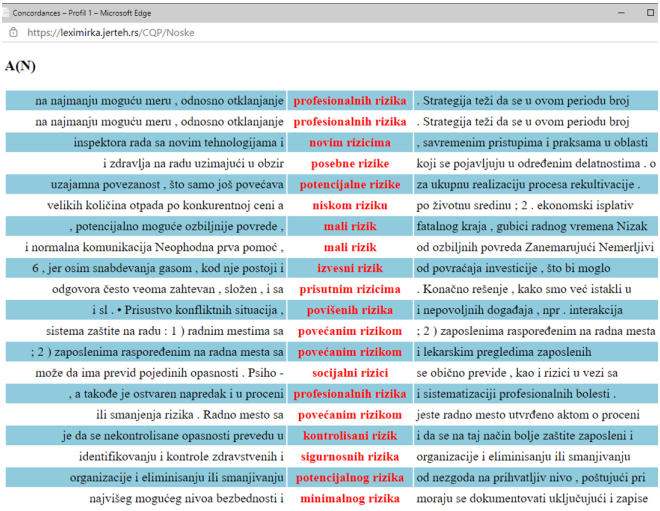
['Asset', 'Circumstances', 'Dangerous_entity', 'Degree', 'Domain',
'Frequency', 'Place', 'Situation', 'Time']
[<Parent=Gradable_attributes – Inheritance →
Child=Risky_situation>,
<MainEntry=Run_risk – See_also →
ReferringEntry=Risky_situation>,
<Source=Run_risk – ReFraming_Mapping →
Target=Risky_situation>,
<Neutral=Risk_scenario – Perspective_on →
Perspectivized=Risky_situation>]

4 Lexical Analysis of the Word *Risk* in a Mining-related Corpus

The development of a monolingual corpus in the domain of mining started as part of a mining project documentation management project using language

18. Data for the frame *Risky_situation*

technologies (Tomašević et al. 2018, 996). Back then, the corpus contained texts from the domain of mining and similar research areas with a total of 172 documents (in Serbian) and 2.7 million words in the first iteration (997). In the course of further research, 63 documents have been added (Kitanović 2021). The current version contains 4.1 million words. It comprises project documentation (26%), legislation (11%), doctoral dissertations (31%), textbooks and other mining literature (32%) (Kitanović et al. 2021, 8).


Concordances – Profil 1 – Microsoft Edge

https://leximirka.jerteh.rs/CQP/Noske

A(N)

na najmanju moguću meru , odnosno otklanjanje	**profesionalnih rizika**	. Strategija teži da se u ovom periodu broj
na najmanju moguću meru , odnosno otklanjanje	**profesionalnih rizika**	. Strategija teži da se u ovom periodu broj
inspektora rada sa novim tehnologijama i	**novim rizicima**	, savremenim pristupima i praksama u oblasti
i zdravlja na radu uzimajući u obzir	**posebne rizike**	koji se pojavljuju u određenim delatnostima . o
uzajamna povezanost , što samo još povećava	**potencijalne rizike**	za ukupnu realizaciju procesa rekultivacije .
velikih količina otpada po konkurentnoj ceni a	**niskom riziku**	po životnu sredinu ; 2 . ekonomski isplativ
, potencijalno moguće ozbiljnije povrede ,	**mali rizik**	fatalnog kraja , gubici radnog vremena Nizak
i normalna komunikacija Neophodna prva pomoć ,	**mali rizik**	od ozbiljnih povreda Zanemarujući Nemerljivi
6 , jer osim snabdevanja gasom , kod nje postoji i	**izvesni rizik**	od povraćaja investicije , što bi moglo
odgovora često veoma zahtevan , složen , i sa	**prisutnim rizicima**	. Konačno rešenje , kako smo već istakli u
i sl . • Prisustvo konfliktnih situacija ,	**povišenih rizika**	i nepovoljnih događaja , npr . interakcija
sistema zaštite na radu : 1) radnim mestima sa	**povećanim rizikom**	; 2) zaposlenima raspoređenim na radna mesta
; 2) zaposlenima raspoređenim na radna mesta sa	**povećanim rizikom**	i lekarskim pregledima zaposlenih
može da ima previd pojedinih opasnosti . Psiho -	**socijalni rizici**	se obično previde , kao i rizici u vezi sa
, a takođe je ostvaren napredak i u proceni	**profesionalnih rizika**	i sistematizaciji profesionalnih bolesti .
ili smanjenja rizika . Radno mesto sa	**povećanim rizikom**	jeste radno mesto utvrđeno aktom o proceni
je da se nekontrolisane opasnosti prevedu u	**kontrolisani rizik**	i da se na taj način bolje zaštite zaposleni i
identifikovanju i kontrole zdravstvenih i	**sigurnosnih rizika**	organizacije i eliminisanju ili smanjivanju
organizacije i eliminisanju ili smanjivanju	**potencijalnog rizika**	od nezgoda na prihvatljiv nivo , poštujući pri
najvišeg mogućeg nivoa bezbednosti i	**minimalnog rizika**	moraju se dokumentovati uključujući i zapise

Figure 5. Concordances for adjective-noun pattern containing the noun *ризик*

The results of a CQL[19] (*Corpus Query Language*) query are analyzed for: frequency lists, collocations, concordances with a narrower and broader context. Figure 5 shows the concordances extracted from the Leximirka[20] digital dictionary management web app (Stanković et al. 2018) of the adjective-noun pattern containing the noun ризик (risk), while in Figure 6 there is a histogram of frequencies for different inflected forms of the same pattern taken from a mining corpus, available on the open-source platform *NoSketch Engine* (Kilgarriff et al. 2004).[21]. The version on the local servers is maintained

19. Corpus Querying
20. LeXimirka
21. NoSketch at JeRTeh, NoSketch Engine

by members of the JeRTeh Society for Language Resources and Technologies.[22] A *Treegger* model for Serbian was trained for tagging (Krstev and Vitas 2005; Utvic 2011), (Stanković et al. 2020, 3957) using a manually annotated corpus of Serbian morphological dictionaries (Krstev 2008).

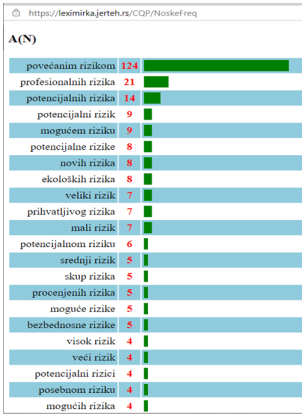


Figure 6. A histogram of frequencies for different inflectional forms of the noun *ризик*

The mining corpus is published in *Sketch Engine*[23] too (Kilgarriff et al. 2014), a platform that provides the option of different types of searches. For instance, we can extract concordances for a target lemma or multi-word expression, collocates of a lemma, related-word thesaurus, *Word Sketch* or *Word Sketch Difference* for two related words. The word sketch approach, developed by Kilgarriff et al. (2004), helps build FrameNet and similar resources and speeds up the process of sense disambiguation of polysemous words (Baker 2012, 274).

Word sketch gives a quick overview of the behavior of the target lexeme by gathering information from thousands or millions of examples of its use and summarizes collocates by category, with links to individual examples. Figure 7 illustrates the word sketch for the noun *ризик* – one look at the page gives a clear idea of the word's use. The first column shows prepositional phrases (in Serbian linguistic terminology referred to

22. JeRTeh
23. Sketch Engine

as *предлошко-падежна конструкција*):[24] *risk of/with/in/on/for...* (*ризик од/са/у/по/на/за...*), and if we clicked on "..." we would get the concordances for each individual phrase. The second column features the modifiers of the word, in this case passive participles of verbs: *increased* (*повећан*)/ *identified* (*идентификован*)/ *assessed* (*процењен*)/... *risk* (*ризик*) or adjectives: *potential* (*потенцијалан*)/ *political* (*политички*)/ *unacceptable* (*неприхватљив*)/ ... *risk* (*ризик*). The third column contains the verbs with which *ризик* appears as the subject e.g. *to decrease* (*смањити*)/ *to arise* (*настајати*)/ *to exist* (постојати)/.... What follows are the expressions in which *ризик* appears as an object: to *decrease* (*смањити*)/ to assess (*проценити*)/ ... *risk* (*ризик*).

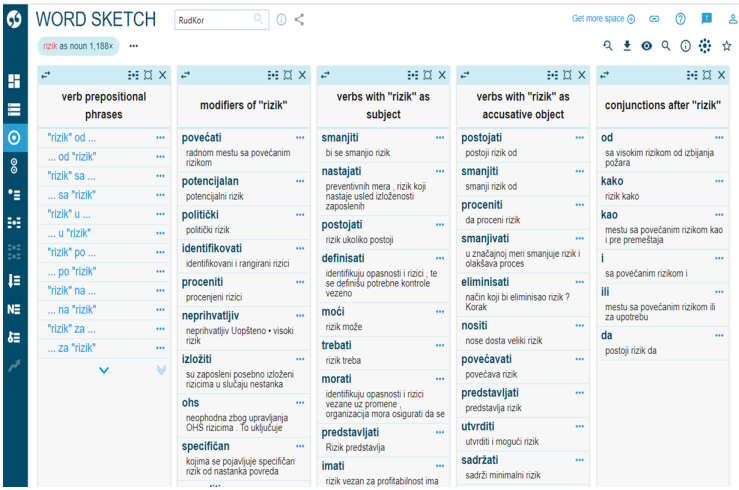


Figure 7. Sketch of the word *ризик* on *Sketch engine*

Figure 8 shows a dynamic diagram of the collocations. It is clear that most of the collocations are prepositional phrases. On the right-hand side of the picture there is the settings option allowing to choose which patterns are to be shown and the minimal frequency requirement that collocations have to meet in order to be included in the diagram.

24. It should be mentioned that the tools and automatic detection are not that well-suited for Serbian but are nevertheless valuable. Namely, mistakes are found that need to manually be corrected.

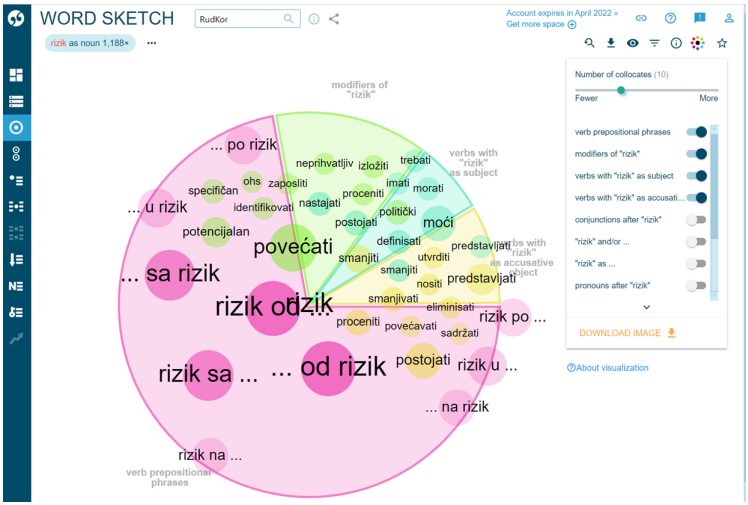


Figure 8. Illustration of collocations of the noun *ризик* in *Sketch engine*

Collocations research is very important (for example, in lexicography, it is important to list the most frequent collocates of a LU; collocations are crucial not only in language learning, but also in different natural language processing tasks). Using the word sketch and the collocation *risk of* (*ризик од*) as a starting point, a detailed view of the concordances can be shown (Figure 9).

The sketch gives a quick search with preset rules, but a custom search can be executed with CQL queries. If we wanted to see where the risk was coming from we would get an answer with the following query `[lemma="ризик"][tag="N"]`. The query `[tag="A"][lemma="ризик"]` would give an answer to the question what type of risk it is; or, if we allowed the result to contain examples in which no more than 5 words divide our target word and the verb we would write the following query: `[tag="V"][word!="ѝ"]0,5[lemma="ризик"]`.

The frequencies of collocations can be both listed and presented visually with bars as shown in the picture below. Figure 10 shows the frequencies of the collocations containing the noun ризик.

Figure 11 illustrates Word Sketch Difference (an extension of Word Sketch). It generates a sketch of two target words and compares them, which allows for a clear overview of the differences in their use. This option is particularly valuable for similar meaning words, for antonyms and

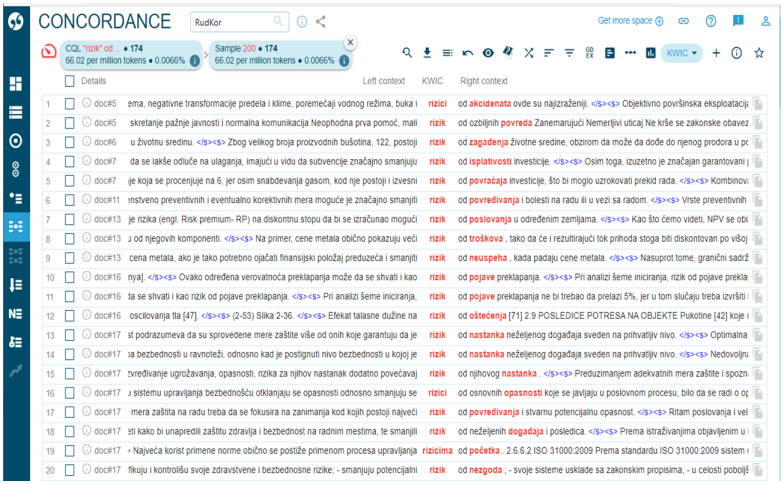


Figure 9. Concordances for *ризик од* in *Sketch engine*

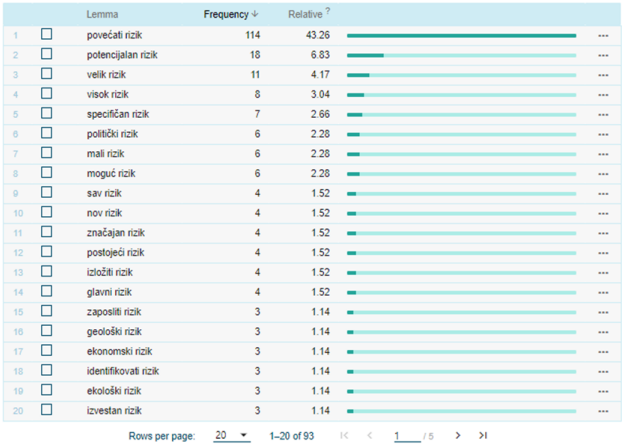


Figure 10. Collocation frequencies for the noun *ризик* in *Sketch engine*

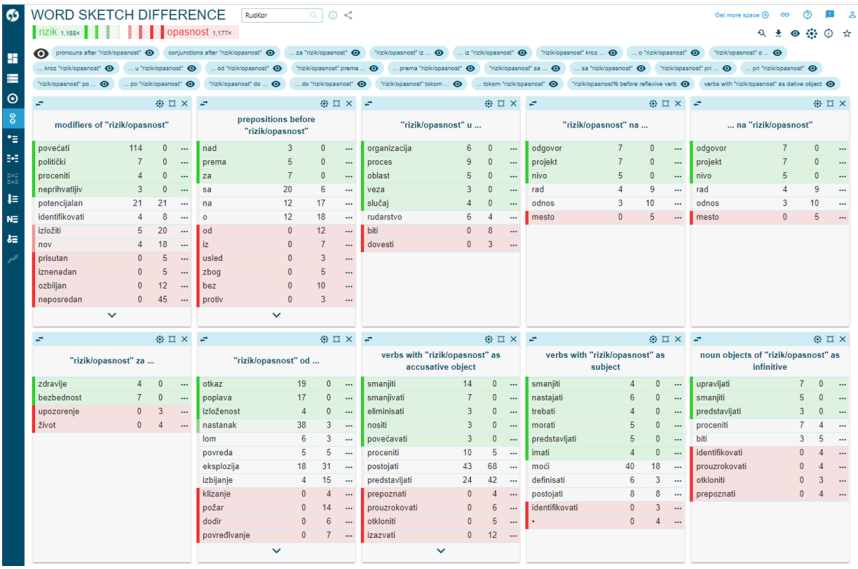


Figure 11. Word Sketch Difference of the words *ризик* and *опасност*

words from the same semantic field. It is shown in Figure 10 that the noun *risk* (*ризик*) has as its most frequent collocates: to *increase* (*повећати*), *political* (*политички*), *to assess* (*проценити*), *acceptable* (*прихватљив*), while the most frequent ones of the noun *danger* (*опасност*) are adjectives *непосредан* (*immediate*), *озбиљан* (*serious*), and *изненадан* (*sudden*).

The automatically generated thesaurus for the target word finds synonyms or words that fall in the same category (same semantic field) and lists them in a table with links to the sketches of individual words, concordances, word sketch differences and thesauruses. Figure 12 shows an illustration of the thesaurus which contains automatically retrieved words from the same semantic field as the target word *risk* (*ризик*), on the left-hand side in the form of a bubble graph and on the right-hand side as a word cloud. The thesaurus word list is created based on the context in which the searched word appears within a chosen corpus, relying on the distributional semantics theory, which, in short, postulates that words that appear in the same context have a similar meaning. In order to determine synonyms, word sketches for all words belonging to the same part of speech are compared and the words

that share the most collocates are paired as similar. The grade[25] given to each of the synonym points to the number of shared collocates.

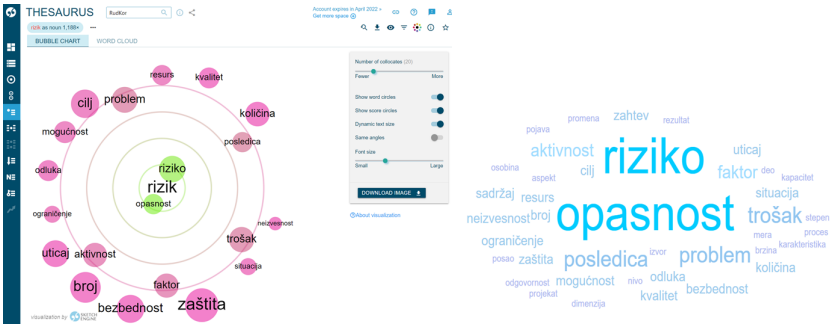


Figure 12. Illustration of the word's *ризик* thesaurus

5 Conclusion

This paper illustrates the results of preliminary research exploring the possibility of application of the frame semantics theory and the principles used in building the FrameNet semantic network using the examples from the risk domain adapted to Serbian. We also show the inner workings of the NLTK suite usable for many different language resources, as well as the Sketch Engine corpus analysis tool.

We have shown that FrameNet offers a detailed and structured mapping, which can then be used in different ways for language processing, especially in text extraction and organizing, as well as in an effort to make human-computer interaction more natural in applications like chatbots. A chatbot needs to be able to recognize different lexical units that evoke the same event or refer to the same entity in order to successfully recognize intent.

It is of great importance that the English FrameNet can be filled with entries from other languages e.g. Serbian (keeping frame information which is shared and adding language-specific material) therefore making it applicable to multilingual resources.

25. Статистичке формуле које се корите у алату *Sketch engine*: statistics/formulae

The research presented above only hints at the possibility of adapting FrameNet to Serbian and aligning that network with the FrameNet data in other languages. Future research intends to align the use of Serbian WordNet and Serbian FrameNet, joining them together. While working toward this aim, we will be following the recommendations given by Tonelli and Pighin (2009).

This research is also aimed at encouraging the growth of Serbian corpus lexicography efforts and modernization of the description of the grammar and lexicography of this language. A good step forward in the modernization process would be case studies that compare polysemous lexeme entries from the SASA (Serbian Academy of Sciences and Arts) dictionary to their description using the frame semantics analysis.

The possibilities for future research on this topic are vast. The implementation of frame semantics theory and methodology used in FrameNet, as well as the discussed tools, will pose a challenge for Serbian. Based on this paper, we speculate that it will be very challenging to use the concept of null instantiations to explore transitive verb complements which do not have to be overtly expressed and are, therefore, implicit (e.g. verbs *to cook*, *to write*, etc.), as well as to look into the ways in which descriptive dictionaries of Serbian deal with such phenomena. We also believe it would be useful to introduce this notion (three types of null-instantiation are defined in FrameNet) into Serbian grammar.

Acknowledgment

This paper was supported by the Ministry of Education, Science and Technological Development, of the Republic of Serbia, in accordance with Contract No. 451-03-9/2021-14, which was entered into with the SASA Institute for Serbian Language. Sketch Engine access is provided by the ELEXIS project funded by the European Union Horizon 2020 research and innovation program under grant number 731015.

References

Atkins, Beryl T. S. 1994. "Analyzing the verbs of seeing: a frame semantics approach to corpus lexicography." In *Annual Meeting of the Berkeley Linguistics Society,* 20:42–56. 1.

Atkins, Sue, Charles J Fillmore, and Christopher R Johnson. 2003. "Lexicographic relevance: Selecting information from corpus evidence." *International Journal of lexicography* 16 (3): 251–280.

Baker, Collin F. 2012. "FrameNet, current collaborations and future goals." *Language Resources and Evaluation* 46 (2): 269–286.

Boas, Hans C., and Ryan Dux. 2017. "From the past into the present: From case frames to semantic frames." *Linguistics Vanguard* 3 (1): 20160003. https://doi.org/doi:10.1515/lingvan-2016-0003.

Brač, Ivana, and Ana Ostroški Anić. 2019. "From concept definitions to semantic role labeling in specialized knowledge resources." In *Proceedings of the 13th International Conference of the Asian Association for Lexicography,* 604–611.

Fillmore, Charles J. 1976. "Frame semantics and the nature of language." In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech,* 280:20–32. 1. New York.

Fillmore, Charles J. 1982. "Frame semantics." *Linguistic society of Korea (ed.), Linguistics in the morning calm,* 111–137.

Fillmore, Charles J, and Beryl T Atkins. 1992. "Toward a frame-based lexicon: The semantics of RISK and its neighbors." *Frames, fields and contrasts: New essays in semantic and lexical organization* 75:102.

Fillmore, Charles J, and Sue Atkins. 1994. "Starting where the Dictionaries Stop: The Challenge of Corpus Lexicography." In *Computational Approaches to the Lexicon,* edited by Sue Atkins and Antonio Zampolli, 349–393. Oxford: OUP.

Fillmore, Charles J, Miriam RL Petruck, Josef Ruppenhofer, and Abby Wright. 2003. "FrameNet in action: The case of attaching." *International journal of lexicography* 16 (3): 297–332.

Gantar, Polona, Kristina Štrkalj Despot, Simon Krek, and Nikola Ljubešic. 2018. "Towards semantic role labeling in Slovene and Croatian." In *Proceedings Conference on Language Technologies and Digital Humanities in Ljubljana,* 93–98.

Gildea, Daniel, and Daniel Jurafsky. 2002. "Automatic labeling of semantic roles." *Computational linguistics* 28 (3): 245–288.

Hamilton, Craig, Svenja Adolphs, and Brigitte Nerlich. 2007. "The meanings of 'risk': A view from corpus linguistics." *Discourse & Society* 18 (2): 163–181.

Jurafsky, Dan, and James H Martin. 2020. "Semantic Role Labeling and Argument Structure." Chap. 19 in *Speech and Language Processing,* 3rd ed. December 30, 2020 draft.

Kilgarriff, Adam, Vıt Baisa, Jan Bušta, Miloš Jakubıček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychl, and Vıt Suchomel. 2014. "The Sketch Engine: ten years on." *Lexicography* 1 (1): 7–36.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. "Itri-04-08 the sketch engine." *Information Technology* 105 (116).

Kitanović, Olivera. 2021. "Ontološki model upravljanja rizikom u rudarstvu." PhD diss., Univerzitet u Beogradu, Rudarsko-geološki fakultet. https://uvidok.rcub.bg.ac.rs/bitstream/handle/123456789/4305/Doktorat.pdf?sequence=1.

Kitanović, Olivera, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić, Ivan Babić, and Ljiljana Kolonja. 2021. "A Data Driven Approach for Raw Material Terminology." *Applied Sciences* 11 (7): 2892.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology of the University of Belgrade.

Krstev, Cvetana, and Duško Vitas. 2005. "Corpus and Lexicon-Mutual Incompletness." In *Proceedings of the Corpus Linguistics Conference,* 14:17.

Pradhan, Sameer, Wayne Ward, Kadri Hacioglu, James H Martin, and Dan Jurafsky. 2005. "Semantic role labeling using different syntactic views." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05),* 581–588.

Ruppenhofer, Josef, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended theory and practice.* Revised November 1, 2016. https://framenet.icsi.berkeley.edu/fndrupal/the_book.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic dictionaries-from file system to lemon based lexical database." In *Proceedings of the 11th LREC - W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018), LREC 2018, Miyazaki, Japan, May 7-12, 2018,* 48–56.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proceedings of The 12th LREC – Language Resources and Evaluation Conference,* 3954–3962.

Tomašević, Aleksandra, Ranka Stanković, Miloš Utvić, Ivan Obradović, and Božo Kolonja. 2018. "Managing mining project documentation using human language technology." *The Electronic Library,* https://doi.org/10.1108/EL-11-2017-0239.

Tonelli, Sara, and Daniele Pighin. 2009. "New features for framenet-wordnet mapping." In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009),* 219–227.

Utvic, Milos. 2011. "Annotating the Corpus of Contemporary Serbian." *INFOtheca* 12, no. 2 (December): 36a–47a.

Wasserscheidt, Philipp, and Andrea Hrstić. 2020. "Legal Variation? A Frame Analysis of Croatian and Serbian in the Domain of Law." *Mediterranean Language Review* 27:87–112.

Драгићевић, Рајна. 2007. *Лексикологија српског језика.* Београд: Завод за уџбенике.

Марковић, Александра. 2017. "Однос граматике и речника – граматика инхерентна описним речницима српског језика." *Наш језик* XLVIII (1-2): 27–43.

Поповић, Љубомир. 2003. "Интегрални речнички модели и њихов значај за лингвистички опис и анализу корпуса." *Научни састанак слависта у Вукове дане* 31 (1): 201–220.