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

Aleksandra Marković, Ranka Stanković, Natalija Tomić, Olivera Kitanović

FrameNet Lexical Database: Presenting a Few Frames Within the Risk Domain | Aleksandra Marković, Ranka Stanković, Natalija Tomić, Olivera Kitanović | Infotheca | 2021 | |

10.18485/infotheca.2021.21.1.1

http://dr.rgf.bg.ac.rs/s/repo/item/0005141

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.

Ризична_ситуација

Дефиниција:

Одређена Ситуација може (али не мора) да доведе до штетног догађаја који би задесио неку Вредност. Та Ситуација може бити неко стање, активност или неки кључни ентитет који мора бити схваћен као део неке шире Ситуације, која укључује и тај ентитет и Вредност. Иако је за разумевање овог оквира кључна идеја о штетном догађају, он не мора бити изражен као аргумент лексичких јединица у овом оквиру.

Да ли су климатске промене ОПАСНЕ по човечанство?
Купци се могу жалити на ШКОДЉИВЕ производе.
Највећа ОПАСНОСТ прети нашој инфраструктури.

Елементи оквира

Централни:

Вредност(Asset) [ass]	Нешто што се сматра вредним и пожељним, а постоји могућност да ће му штета бити нанета или да ће бити изгубљено. Мед није БЕЗБЕДАН за бебе.
Опасан ентитет (Dangerous entity) [dan ent]	Конкретан или апстрактан ентитет који може нанети штету Вредности или довести до њеног губитка. Том човеку је РИЗИК друго име!
Искључује: Ситуацију	
Ситуација (Situation) [sit]	Ситуација може довести до неког штетног догађаја. Већина људи се слаже да није БЕЗБЕДНО возити брже од 120 km/h.

Периферни:

Околности (Circumstances) []	Околности под којима је Вредност угрожена.
Степен (Degree) [deg]	Одредба која изражава одступање тренутног нивоа безбедности од очекиване вредности, узимајући у обзир Ситуацију и стање на које указује цифна ЈЦ. Терористи су наша највећа ПРЕТЊА.
Домен (Domain) [dom]	Домен у коме је Ситуација безбедна. Све наше сајтове је БЕЗБЕДНО користити у образовању.
Учесталост (Frequency) []	Колико често Вредност долази у *Ризичну_ситуацију*.
Место (Place) [pla]	Одређена локација на којој је Ситуација безбедна. Често се може закључити да карактеристике неке локације чине одређене Ситуације безбедним или небезбедним.
Време (Time) [tim]	Временски период током ког одређена Ситуација има прецизирани ниво сигурности.

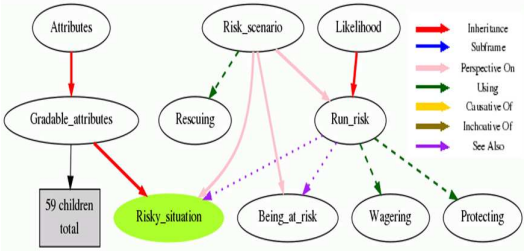


Figure 1. An illustration of the frame *Risky_ situation* and the related frames

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.

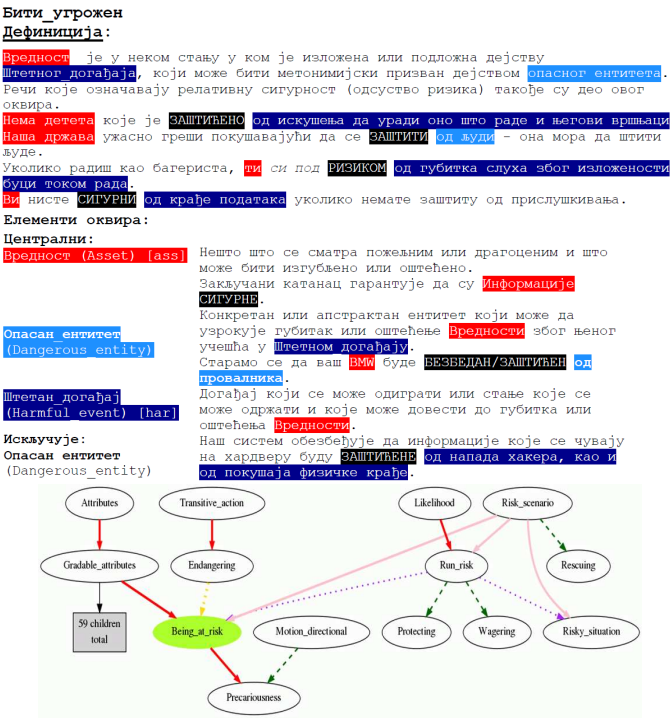


Figure 2. Semantic frame *Being_at_risk*

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.

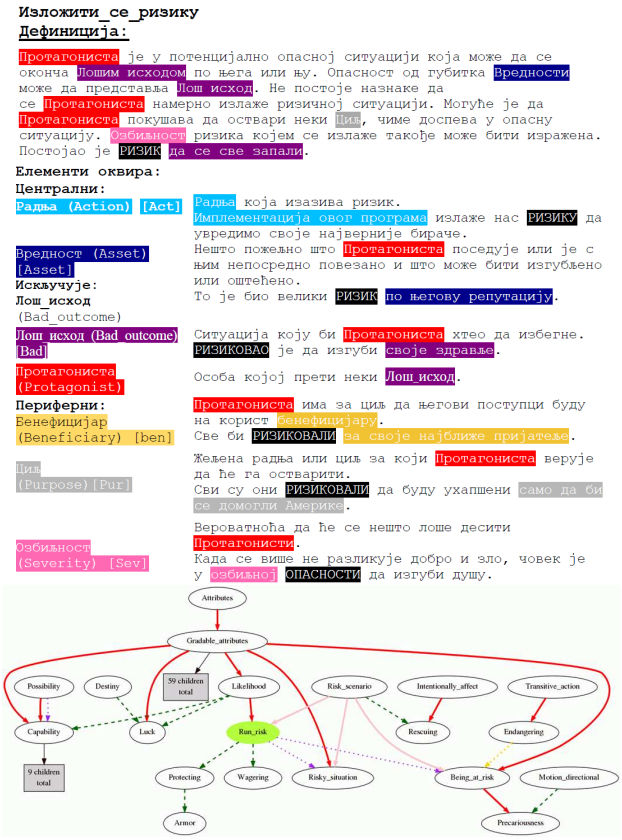


Figure 3. Semantic frame Run_risk

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.).

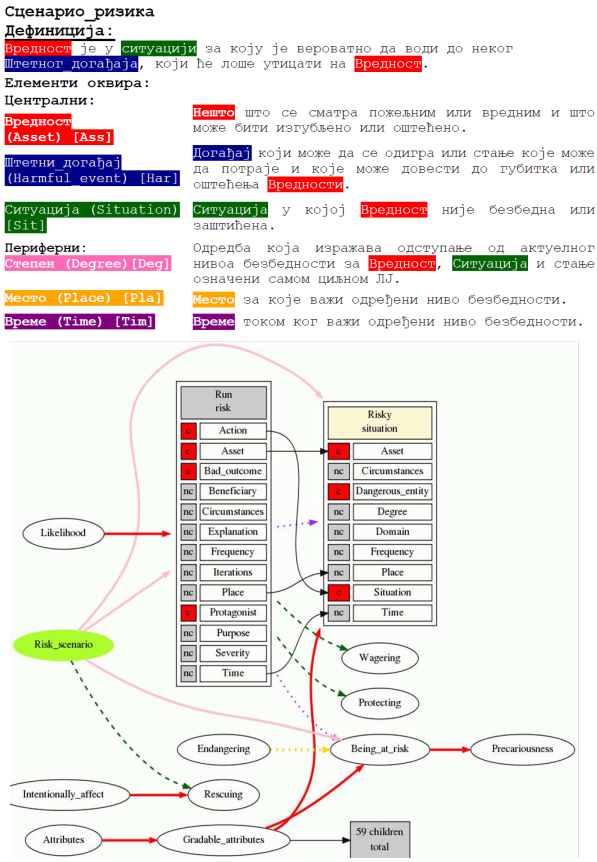


Figure 4. Semantic frame *Risk_ scenario* with a detailed view of two other related frames)

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:

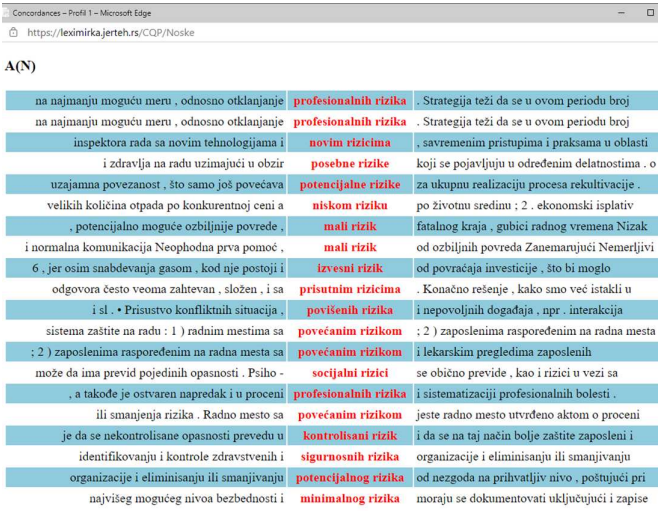
['Asset', 'Circumstances', 'Dangerous_entity', 'Degree', 'Domain',
'Frequency', 'Place', 'Situation', 'Time']
[<Parent=Gradable_attributes – Inheritance →
Child=Risky_situation>,
<MainEntry=Run_risk – See_also →
ReferringEntry=Risky_situation>,
<Source=Run_risk – ReFraming_Mapping →
Target=Risky_situation>,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

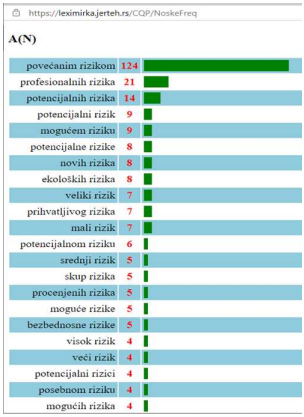


Figure 6. A histogram of frequencies for different inflectional forms of the noun *ризик*

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*).

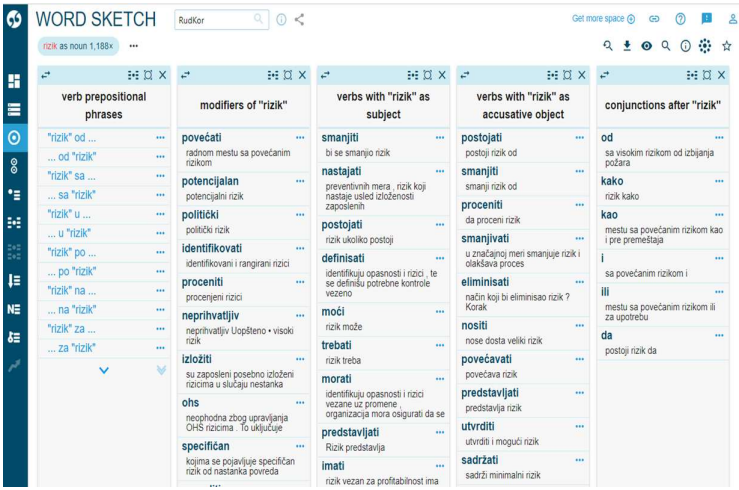


Figure 7. Sketch of the word *ризик* on *Sketch engine*

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.

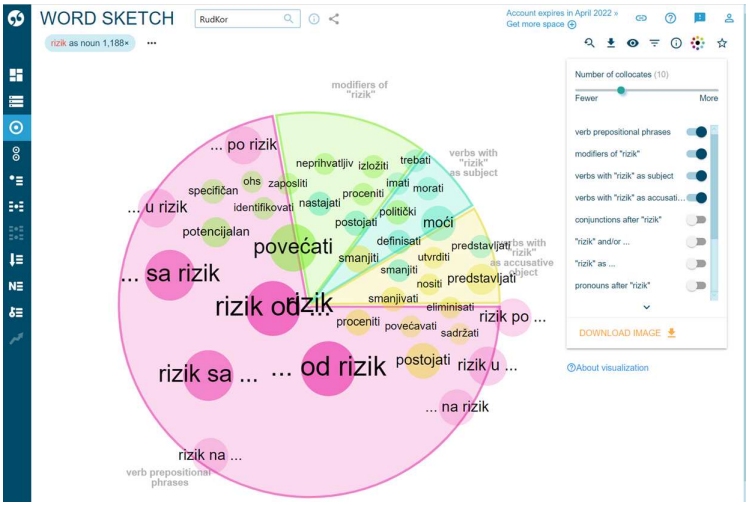


Figure 8. Illustration of collocations of the noun *ризик* in *Sketch engine*

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!=""]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

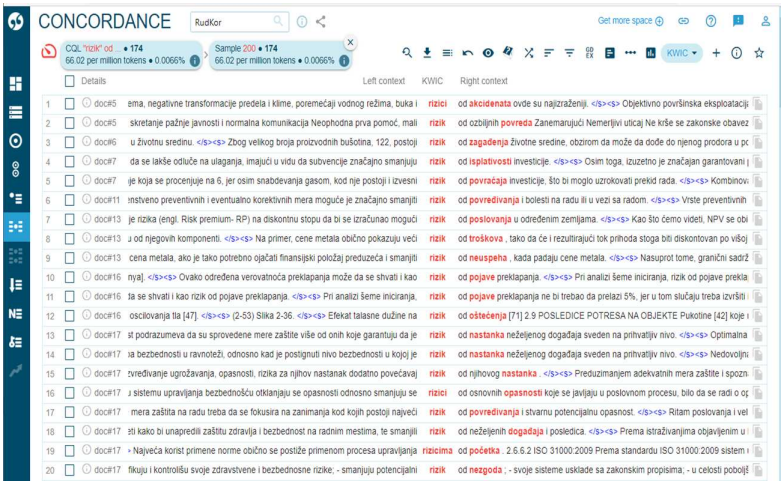


Figure 9. Concordances for *ризик од* in *Sketch engine*

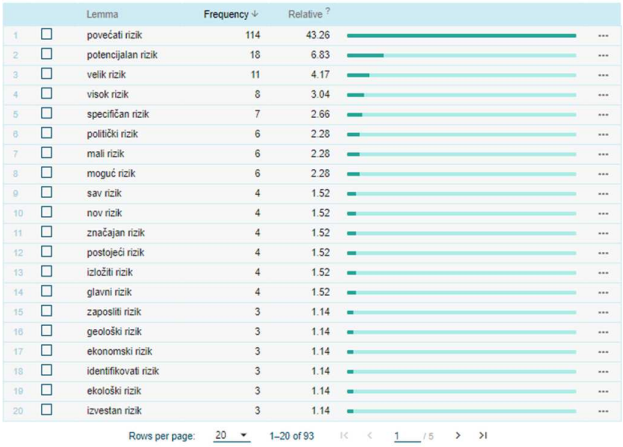


Figure 10. Collocation frequencies for the noun *ризик* in *Sketch engine*

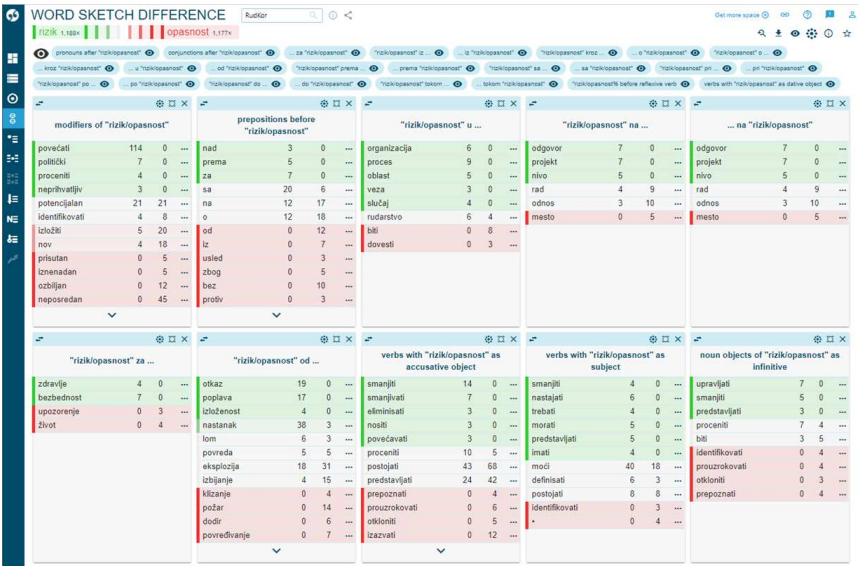


Figure 11. Word Sketch Difference of the words *ризик* and *опасност*

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.

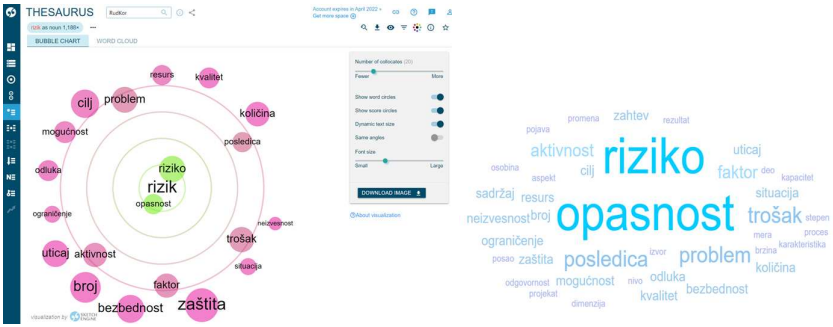


Figure 12. Illustration of the word's *ризик* thesaurus

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.