cjvt Center za jezikovne vire in tehnologije

Priprava frazeoloških podatkov za računalniško obdelavo: aktualno stanje in izzivi

Polona Gantar

Filozofska fakulteta, Univerza v Ljubljani

Dvostransko frazeološko delovno srečanje, Ljubljana, 16. december 2022

Jezikovni viri CJVT, ki vključujejo večbesedne enote

- **Korpusi**
  - SSJ500k 2.3 → Training corpus SUK 1.0 http://hdl.handle.net/11356/1747
- **Leksikoni**
  - Leksikon večbesednih enot http://hdl.handle.net/11356/1421
  - SloIE http://hdl.handle.net/11356/1335
- **Slovarske baze in spletni slovarji**
  - Kolokacijski slovar sodobne slovenščine 1.0 → 2.0
  - Veliki Slovensko-madžarski slovar
- **Teorija in metodologija**

PARSEME Shared Task 1.1 (Savary et al., 2017)

- CILJ: izboljšati prepoznavanje glagolskih večbesednih enot (verbal MWE) v večjezičnem okolju.
  - izdelati univerzalno **terminologijo**
  - izdelati **učne korpuse** za **18 različnih jezikov**
  - izdelati **smernice** in **metodologijo** za označevanje korpusov

- Eden glavnih rezultatov tega prizadevanja je bil razvoj **smernic** (https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/) za označevanje, ki naj bi bile čim bolj **univerzalne**, a bi hkrati upoštevale **jezikovno specifične kategorije**.

- SAVARY, A. et al. (2017). The PARSEME Shared Task on Automatic Identificationof Verbal Multiword Expressions. Proceedings of MWE-17 Workshop, in conjunction with the European Association for Computational Linguistics 2017 Conference. Valencia, Spain.

- SAVARY , A. et al. (2018). PARSEME multilingual corpus of verbal multiword expressions. Multi-word expressions at length and in depth: Extended papers from the MWE 2017 workshop.

Učni korpus ssj500k

Glagolske večbesedne enote

- glagolski idiomi (VID: Verbal Idioms)
- zveze s pomensko oslabljenimi glagoli (LVC: Light Verb Constructior
- predložne glagolske zveze (IAV: Inherently Adpositional Verbs)
- inherentno povratni glagoli (IRV: Inherently Reflexive Verbs)

VID 724/457	LVC 303/103	IAV 710/154	IRV 1.672/345
plačati ceno	biti v dvomih	priti do	bati se
zravnati z zemljo	imeti mnenje	vplivati na	dati se
zgodba se ponavlja	biti v pomoč	skrbeti za	dogajati se
spati kot ubit	imeti v načrtu	temeljiti na	pobrati se
vedeti, koliko je ura	dati na voljo	zavzemati se za	lotiti se

GANTAR, Polona, ARHAR HOLDT, Špela, ČIBEJ, Jaka, KUZMAN, Taja. (2019) Structural and semantic classification of verbal multi-word expressions in Slovene. Prispevki za novejšo zgodovino.

GANTAR, Polona, ČIBEJ, Jaka, BON, Mija. (2019). Slovene multi-word units: identification, categorization, and representation. Computational and corpus-based phraseology: Third International Conference, Europhras 2019, Malaga, Spain.

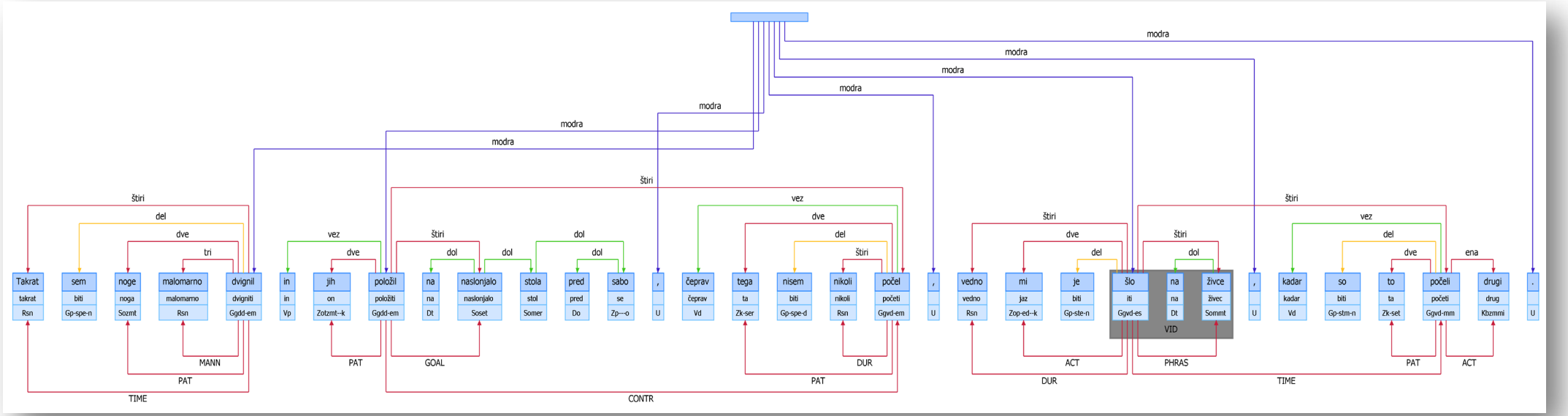
GANTAR, Polona, ARHAR HOLDT, Špela, ČIBEJ, Jaka, KUZMAN, Taja, KAVČIČ, Teja. (2018). Glagolske večbesedne enote v učnem korpusu ssj500k 2.1. Zbornik konference Jezikovne tehnologije in digitalna humanistika.

Učni korpus ssj500k

Glagolske večbesedne enote

- glagolski idiomi (VID: Verbal Idioms)
- zveze s pomensko oslabljenimi glagoli (LVC: Light Verb Constructions)
- predložne glagolske zveze (IAV: Inherently Adpositional Verbs)
- inherentno povratni glagoli (IRV: Inherently Reflexive Verbs)

VID 724/457	LVC 303/103	IAV 710/154	IRV 1.672/345
plačati ceno	biti v dvomih	priti do	bati se
zravnati z zemljo	imeti mnenje	vplivati na	dati se
zgodba se ponavlja	biti v pomoč	skrbeti za	dogajati se
spati kot ubit	imeti v načrtu	temeljiti na	pobrati se
vedeti, koliko je ura	dati na voljo	zavzemati se za	lotiti se



SloIE

- SloIE je ročno označena **elektronska zbirka podatkov**, ki vsebuje **29.400 stavkov**, ki vsebujejo **75 različnih frazemov**, izbranih na podlagi Leksikalne baze za slovenščino, ki se v stavkih pojavljajo tako v svojem **dobesednem** kot **frazeološkem pomenu.**
- Stavki, ki vsebujejo frazeme, so bili izluščeni iz korpusa **Gigafida 2.0.**

ŠKVORC, Tadej; GANTAR, Polona, ROBNIK-ŠIKONJA, Marko. (2020) Dataset of Slovene idiomatic expressions SloIE, Slovenian language resource repository CLARIN.SI

ŠKVORC, Tadej, GANTAR, Polona, ROBNIK-ŠIKONJA, Marko. (2022) MICE: mining idioms with contextual embeddings. Knowledge-based systems. 2022, vol. 235, str. 1-11.

#Doma bodo barvali jajčka s prav posebnimi barvami.

3 4 5 7 8

Doma	*	barvati kaj s črnimi barvami
bodo	*	barvati kaj s črnimi barvami
barvali	NE	barvati kaj s črnimi barvami
jajčka	NE	barvati kaj s črnimi barvami
s	NE	barvati kaj s črnimi barvami
prav	*	barvati kaj s črnimi barvami
posebnimi	NE	barvati kaj s črnimi barvami
barvami.	NE	barvati kaj s črnimi barvami

#S Svojimi črnimi odtenki bodo barvala dogajanja, ki se bodo odvijala v našem svetu.

6 7 1 3 4

S	DA	barvati kaj s črnimi barvami
Svojimi	*	barvati kaj s črnimi barvami
črnimi	DA	barvati kaj s črnimi barvami
odtenki	DA	barvati kaj s črnimi barvami
bodo	*	barvati kaj s črnimi barvami
barvala	DA	barvati kaj s črnimi barvami
dogajanja,	DA	barvati kaj s črnimi barvami
ki	*	barvati kaj s črnimi barvami
se	*	barvati kaj s črnimi barvami
bodo	*	barvati kaj s črnimi barvami
odvijala	*	barvati kaj s črnimi barvami
v	*	barvati kaj s črnimi barvami
našem	*	barvati kaj s črnimi barvami
svetu.	*	barvati kaj s črnimi barvami

Leksikon večbesednih enot

- **Metodologija:** luščenje stavkov, ki vključujejo FE, iz korpusa Gigafida 2.1

- **Izhodiščni seznam FE:** CC BY-SA (CLARIN.SI)
  - Leksikalna baza za slovenščino (Gantar et al. 2013)
  - Slovar slovenskih frazemov (Keber 2011)
  - ssj500k učni korpus (Krek et al. 2020)

- **5.241 frazeoloških enot**

GANTAR, Polona, KREK, Simon. (2022) Creating Lexicon of Multi-Word Expressions for Slovene. Euralex 2022. Mannheim Nemčija.

poleteti na krilih česa	86	LBS
odrezati krila komu	87	LBS
pristriči krila komu	87	SSF
spodrezati krila komu	87	ssj500k
držati se mame za krilo	88	?
držati se maminega krila	88	?
skrivati se za čigavim krilom	88	LBS
biti korak pred časom	89	?
biti sama kost in koža	90	?
sama kost in koža je koga	90	ssj500k
debela koža	91	?
dobiti debelo kožo	91	?
dobiti trdo kožo	91	?
imeti debelo kožo	91	LBS
imeti slonjo kožo	91	LBS
imeti trdo kožo	91	LBS

Del izhodiščnega seznama FE v kanonični obliki

Zapis kanonične oblike

Osebek	Povedek	Neposredni predmet -1	Neposredni predmet - 2	Posredni predmet	Prislovno določilo/odvisni stavek
čas	je	denar			
	zamašiti	usta	komu		
	barvati	kaj		s črnimi barvami	
kaj	pade		komu	v naročje	kot zrela hruška
kaj	pade			v čigavo naročje	
	ne priplavati				po kisli juhi
kaj	ne pade		komu		na kraj pameti
	izkusiti	kaj			na lastni koži

Posebnosti pri t. i. besedilnih in pragmatičnih FE:
da padeš dol – da dol padeš; da pade/padeš/padete dol;
Prede težka komu – težka prede komu;
Povedati komu kar komu/mu gre;
Ne moči verjeti svojim očem – kdo ne more verjeti svojim očem

GANTAR, Polona. 2021. Zapis kanonične oblike frazeoloških enot v Leksikonu večbesednih enot za slovenščino. *Nova slovnica sodobne standardne slovenščine: viri in metode*.

Zapis skladenjske strukture

structure type – @type

human-readable code – @label

structure ID – @id

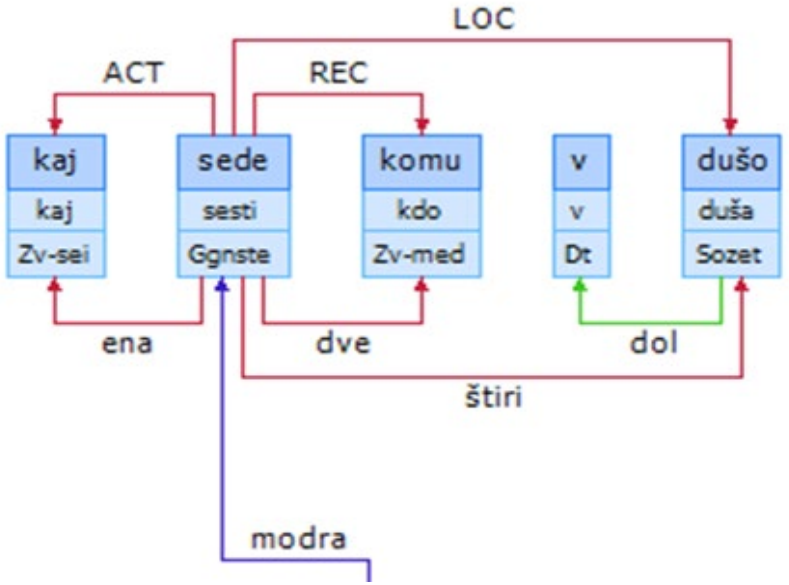
<syntactic_structure type="other" label="z-l-gg-s2-z" id="122">

kaj *ne* *da* *miru* *komu*

Zapis skladenjskih razmerij

Skupina povezav	Tip povezave	Kaj povezuje
Prvi nivo – **besedna zveza**	dol	Jedro in določilo besedne zveze
	del	Deli zloženega povedka
	prir	Jedra prirednih zvez
	vez	Besede, ločila v vezniški vlogi
	skup	Nepolnopomenske besede
Drugi nivo – **stavčni člen**	ena	**Osebek stavka**
	dve	Predmet stavka
	tri	Prislovno določilo lastnosti
	štiri	Prislovno določilo kraja, časa …
Tretji nivo	modra	Hierarhično najvišje pojavnice: ločila …

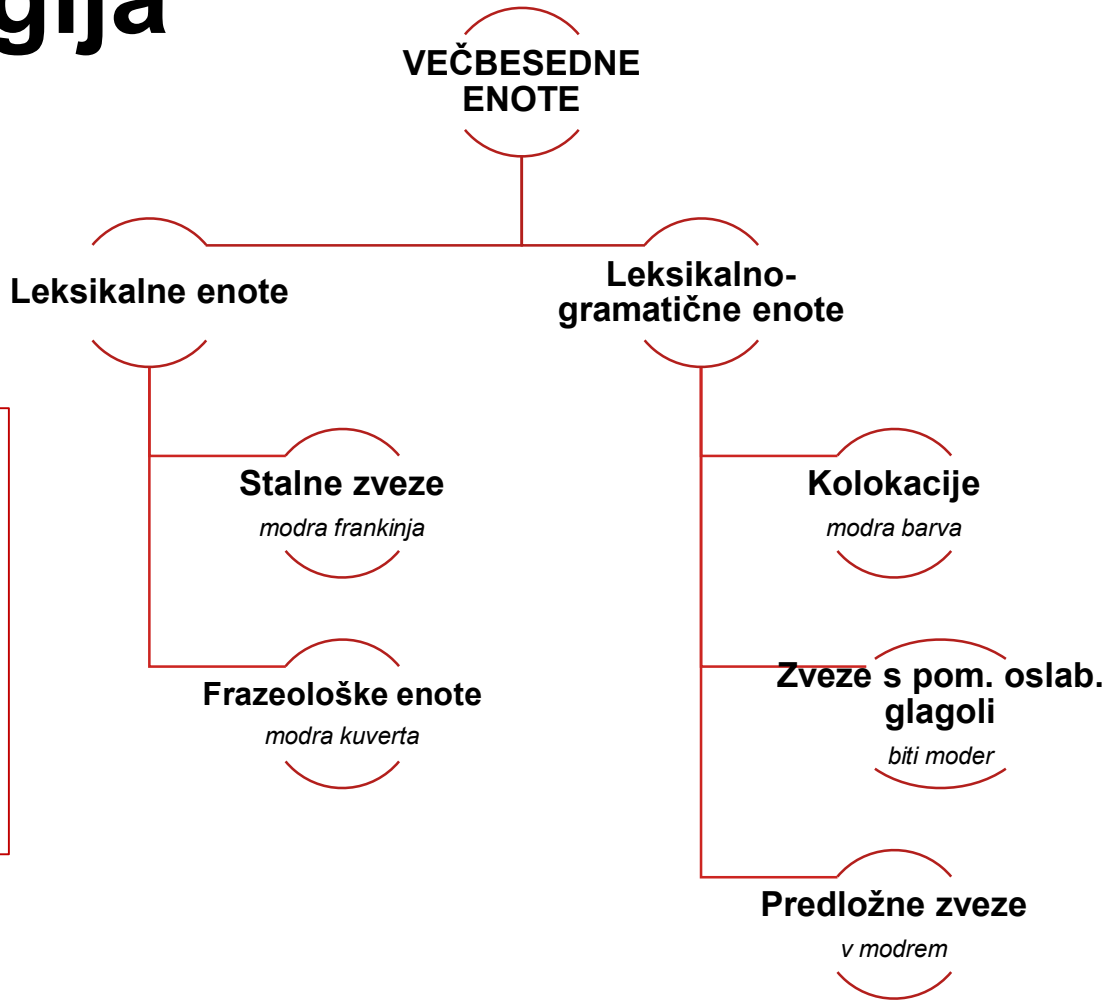
Vrste skladenjskih povezav v označevalnem sistemu JOS.



Prikaz skladenjskih razmerij v orodju Q-Cat (Brank 2022).

Teorija in metodologija

- **Definicija VBE**
  - **Tipologija in terminologija**



GANTAR, Polona, KREK, Simon, KOSEM, Iztok. 2021. Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. Kolokacije v slovenščini.

KOSEM, Iztok, KREK, Simon, GANTAR, Polona. 2020. Defining collocation for Slovenian lexical resources. Collocations in lexicography: existing solutions and future challenges.

GANTAR, Polona, COLMAN, Lut, PARRA ESCARTÍN, Carla, MARTÍNEZ ALONSO, Héctor. 2019. Multiword expressions: between lexicography and NLP. International journal of lexicography.

Spoznanja in izzivi

- Potrebe po računalniškem razumevanju jezika narekujejo razvoj semantičnih tehnologij: **semantika** in govorjeni jezik sta **ključni področji** za razvoj orodij in aplikacij
- **Univerzalni koncepti** – medjezikovna povezljivost: semantične ontologije (Framenet, Wordnet), baze znanja (Wikipedia, BabelNet, Dbpedia), podatkovni modeli (slovarji, leksikoni, Digitalna slovarska baza)
- Računalniško procesiranje naravnega jezika zahteva **kategorizacijo jezika** – t. i. pripisovanje vrednosti, zato je pomembno, da so kategorije določene z **jezikoslovnim konsenzom**.
- **VBE** enote predstavljajo **problem**, pri računalniškem procesiranju jezika.

  Sag et al. 2002: „Multiword Expressions: A Pain in the Neck for NLP“
- **Enotna klasifikacija VBE** in **terminologija**
- Izdelava **odprtih jezikovnih virov**, ki vsebujejo VBE s čim več **strukturiranega dodatnega znanja** (ang. rich structured or semi-structured knowledge sources)
- **UniDive** “Universality, diversity and idiosyncrasy in language technology”: ttps://www.cost.eu/actions/CA21167/

  WG1:
  - Unification and enhancement of cross-lingual annotation guidelines for morpho-syntax and MWEs
  - Construction of annotated corpora

Izbrana literatura na temo večbesednih enot v strojno procesljivih jezikovnih virih za slovenščino

- GANTAR, Polona, COLMAN, Lut, PARRA ESCARTÍN, Carla, MARTÍNEZ ALONSO, Héctor. 2019. **Multiword expressions: between lexicography and NLP**. International journal of lexicography.
- GANTAR, Polona, KREK, Simon, KOSEM, Iztok. 2021. **Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino**. Kolokacije v slovenščini.
- GANTAR, Polona. 2021. **Zapis kanonične oblike frazeoloških enot v Leksikonu večbesednih enot za slovenščino**. Nova slovnica sodobne standardne slovenščine: viri in metode.
- GANTAR, Polona, KREK, Simon. 2022. **Creating Lexicon of Multi-Word Expressions for Slovene**. Euralex 2022. Mannheim Nemčija.
- GANTAR, Polona, ARHAR HOLDT, Špela, ČIBEJ, Jaka, KUZMAN, Taja. 2019. **Structural and semantic classification of verbal multi-word expressions in Slovene**. Prispevki za novejšo zgodovino. https://ojs.inz.si/pnz/article/view/325
- GANTAR, Polona, ČIBEJ, Jaka, BON, Mija. 2019. **Slovene multi-word units: identification, categorization, and representation**. Computational and corpus-based phraseology: Third International Conference, Europhras 2019, Malaga, Spain.
- GANTAR, Polona, ARHAR HOLDT, Špela, ČIBEJ, Jaka, KUZMAN, Taja, KAVČIČ, Teja. 2018. **Glagolske večbesedne enote v učnem korpusu ssj500k 2.1**. Zbornik konference Jezikovne tehnologije in digitalna humanistika.
- KOSEM, Iztok, KREK, Simon, GANTAR, Polona. 2020. **Defining collocation for Slovenian lexical resources**. Collocations in lexicography: existing solutions and future challenges.

Izbrana literatura na temo večbesednih enot v strojno procesljivih jezikovnih virih za slovenščino

- KOSEM, Iztok, KREK, Simon, GANTAR, Polona, ARHAR HOLDT, Špela, ČIBEJ, Jaka, LASKOWSKI, Cyprian Adam, et al. 2018. **Collocations dictionary of modern Slovene**. *Proceedings of the 18th EURALEX International Congress, [17-21 July 2018, Ljubljana]*.
- KOSEM, Iztok, KREK, Simon, GANTAR, Polona, ARHAR HOLDT, Špela, ČIBEJ, Jaka, LASKOWSKI, Cyprian Adam. 2018. **Kolokacijski slovar sodobne slovenščine**. *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 20. september - 21. september 2018, Ljubljana, Slovenija*
- KOSEM, Iztok, GANTAR, Polona, KREK, Simon. **Sense menus in collocations dictionary of Slovene**. *Electronic lexicography in the 21st century: lexicography from scratch*. Leiden, Nizozemska.
- KREK, Simon, GANTAR, Polona, KOSEM, Iztok. 2022. **Extraction of collocations from the Gigafida 2.1 corpus of Slovene**. Proceedings of the XX EURALEX International Congress, 12-16 July 2022, Mannhein, Germany.
- KREK, Simon, GANTAR, Polona, KOSEM, Iztok, DOBROVOLJC, Kaja. 2021. **Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa**. Nova slovnica sodobne standardne slovenščine: Znanstvena založba Filozofske fakultete.
- RAMISCH, Carlos, GANTAR, Polona, et al. *2018.* **Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions**. *Proceedings: LAW-MWE-CxG*. The 12th Linguistic Annotation Workshop (LAW XII) and the 14th Workshop on Multiword Expressions (MWE 2018). Santa Fe.
- ŠKVORC, Tadej; GANTAR, Polona, ROBNIK-ŠIKONJA, Marko. 2020. **Dataset of Slovene idiomatic expressions SloIE**, Slovenian language resource repository CLARIN.SI
- ŠKVORC, Tadej, GANTAR, Polona, ROBNIK-ŠIKONJA, Marko. 2022. **MICE: mining idioms with contextual embeddings**. Knowledge-based systems.
- ŠKVORC, Tadej, GANTAR, Polona, ROBNIK ŠIKONJA, Marko. 2021. **Strojno prepoznavanje idiomov z globokimi nevronskimi mrežami**. *Nova slovnica sodobne standardne slovenščine.* Ljubljana: Znanstvena založba Filozofske fakultete.

Hvala.

Polona Gantar
apolonija.gantar@ff.uni-lj.si