
ADMIRE 2 Serbian: Изградња српског мултимодалног скупа идиома

Данка Јокић^{1,*}[0000-0002-7621-5847], Срђан Шућур^{2,*}[0009-0003-6815-2887],
Јелена Марковић²[0000-0003-4962-3754]

¹ Универзитет у Београду, Србија, danka.jokic@rcub.bg.ac.rs

² Универзитет у Источном Сарајеву, Филозофски факултет Пале, Катедра за англистику,
{[srdjan.sucur](mailto:srdjan.sucur@ffuis.edu.ba)|[jelena.markovic](mailto:jelena.markovic@ffuis.edu.ba)}@ffuis.edu.ba

Научни рад

DOI: 10.18485/judig.2026.2.ch15

Апстракт

У овом раду представљен је српски мултимодални скуп идиома и процес његове изградње, који је реализован у оквиру *Admire 2* иницијативе. Размотрени су изазови припреме српских еквивалената за енглеске вишечлане идиоматске изразе и примера њихове употребе, као и њихове визуелне репрезентације помоћу генеративног модела за синтезу слика *Midjourney*. Кад год је то било могуће, примери употребе прикупљани су из језичких корпуса и других текстуалних извора како би се обезбедили аутентичност, природност и језичка валидност представљених израза. Посебна пажња посвећена је проналажењу одговарајућих превода идиома усклађених са српским културним наслеђем. Приликом генерисања слика уочена су ограничења генеративног модела у приказивању апстрактних појмова, сложених емоционалних стања и просторних односа, уз присуство различитих облика пристрасности у генерисаним приказима.

Кључне речи: идиоми, мултимодални корпуси, пристрасност АИ модела

1. Увод

Пројекат изградње српског мултимодалног корпуса идиома реализован је од маја до новембра 2025. године. Овај пројекат био је део веће *Admire 2* иницијативе у оквиру COST акције CA21167 – Universality, diversity and idiosyncrasy in language tech-

* Ови аутори су подједнако допринели раду.

nology (UniDive) ¹. У пројекту је учествовало преко 30 тимова из различитих земаља и обухваћено је преко 30 језика.

Цео подухват започет је развојем скупова за енглески и португалски језик, када је и направљен Admire мултимодални скуп идиома, коришћен за такмичење SemEval-2025 Task 1 (Pickard et al., 2025). Ови скупови били су намењени испитивању односа између значења идиома и његове визуелне репрезентације у одређеном контексту. Разумевање идиоматских израза притом се посматра као мултимодални проблем који, поред разликовања дословног и фигуративног значења, обухвата и процену степена семантичке усклађености визуелног садржаја са значењем израза. Због тога је поступак развоја корпуса организован тако да омогући систематско представљање различитих нивоа повезаности између идиома и његове визуелизације.

Касније се иницијатива проширила и на друге језике у оквиру Unidive акције, задржавајући исту методологију рада као и за португалски језик, што значи да је енглески језик коришћен као примарни извор идиома за које су онда тражени одговарајући еквиваленти са истим значењем на другим језицима. Један од циљева оба пројекта, а реализован кроз истраживачка такмичења (Pickard et al., 2025; Arslan et al., 2026), био је да се испита како генеративни модели разумеју идиоме и да ли могу да их препознају у тексту и на слици. Из тог разлога се ови скупови не деле јавно на мрежи да им модели не би имали приступ и тако „научили“ њихов садржај.

Изградња мултимодалног корпуса идиома одвијала се у три фазе:

1. Налажење еквивалената за енглеске вишечлане идиоматске изразе;
2. Генерисање примера за сваки од израза применом генеративног модела за слике;
3. Проналажење одговарајућих примера дословног и идиоматског значења за пронађене идиоме.

Полазни скуп садржао је 64 идиома на енглеском језику. Коначни скуп података садржи 89 јединствених идиоматских израза на српском језику екавског изговора и још 12 еквивалената

¹ <https://unidive.lisn.upsaclay.fr/>

ијекавског изговора, за идиоме где су се ови облици разликовали, што укупно чини 101 идиоматски израз. За два енглеска идиома није пронађен еквивалент на српском језику. Сваком изразу придружено је пет слика и 4 примера употребе.

Резултати Admire 2 пројекта представљени су на конференцији LREC 2026 ([Torunoğlu-Selamet et al., 2026](#)). Мултимодални корпус идиома на српском језику коришћен је на такмичењу спроведеном у склопу радионице која се бави темом вишечланих израза (Workshop on Multiword Expressions – MWE 2026) ([Arslan et al., 2026](#)).

У овом раду представљен је начин креирања корпуса и размотрени су изазови припреме српских еквивалената за енглеске вишечлане идиоматске изразе и примера њихове употребе у контексту, као и проблеми који се јављају приликом њихове визуелне репрезентације употребом генеративног модела за синтезу слика.

Полазиште истраживања представља чињеница да су идиоми дубоко укорени у језику и култури одређене заједнице, због чега њихово превођење често превазилази оквире дословног преношења значења. Како истиче Бугарски ([Bugarski, 2014](#)), неки идиоми су непреводиви или само делимично преносиви у други језик, иако је у појединим случајевима могуће пронаћи културолошке еквиваленте.

2. Фаза 1: Налажење српских еквивалената за енглеске вишечлане идиоматске изразе

Приликом формирања српских еквивалената за енглеске вишечлане изразе коришћена су три приступа. Први се односио на директно преношење идиома у српски језик путем калкирања, што је било могуће код израза који имају сличну структуру и значење у оба језика. Пример таквог идиома представља нпр. „ђавољи адвокат“, чији је енглески еквивалент *devil's advocate*², где и енглески идиом заправо представља калк, тј. дослован

² One who debates from a viewpoint which one may not actually hold, usually to determine its validity or simply for the sake of argument (срп. *Онај који у расправи заступа становиште са којим се заправо можеда и не слаже, обично како би установио валидност тог становишта, или пак како би се расправа наставила*).

превод средњовековног латинског израза *advocatus diaboli*³ (Hunter et al., 1897, 1914). Следећи пример из ове групе јесте „прљав новац“ (енгл. *dirty money*). Овај пример показао се као нарочито захтеван за проналажење дословног значења (в. 4.). Дословно значење, захваљујући сервису books.google.com⁴, пронашли смо у једном чланку дигитализованог издања часописа *Народно здравље*, из друге половине XIX века: *Тако исто може не само дифтерија него и свака друга заразна болест да се развије код деце, ако по обичају турају прљав новац уста [...] (Стефановић, 1882, р. 232).*

Други приступ подразумевао је ослањање на интуицију изворних говорника и проверу предложених решења у релевантним фразеолошким и двојезичним речницима (Milosavljević and Vilijams-Milosavljević, 2007; Otašević, 2007), као и у речницима сленга (Gerzić and Janković, 2002). Организатори Admire иницијативе су уз идиоме приложили и дефиниције значења на енглеском језику, које су преузета из Wiktionary (Wikimedia Foundation). Српски еквивалент за енглески идиом *bread and butter*⁵, као и значајан број других еквивалената, пронашли смо у *Српско-енглеском речнику идиома*:

и ЛЕБА и с леба све што је потребно за живот – all the necessities of life ♦ bread and butter ♦ daily bread (Milosavljević and Vilijams-Milosavljević, 2007, р. 150).

³ The devil's advocate (*Advocatus diaboli*), the advocate who, when the canonization of a saint by the pope is proposed, is appointed to urge all that he justly can against the character of the person whom it is contemplated to honor, and give reasons why the canonization should not be carried out. The analogy between this functionary and the devil is that both are accusers of the brethren and that *diabolos* in Greek properly means a slanderer; the difference lies in their motives (срп. *Ђавољи адвокат (Advocatus diaboli), адвокат који се именује када папа предложи да неки светац буде канонизован, са задатком да изнесе све оправдане аргументе против репутације особе чија се канонизација разматра, те да понуди разлоге зашто би канонизацију требало спречити. Аналогија између ове улоге и ђавола лежи у томе да и један и други оптужују братију, а diabolos на грчком значи клеветник; но, њихови мотиви се разликују)* (Hunter et al., 1897, 1914).

⁴ Овај сервис настао је из Гуглове тежње да постане модерна Александријска библиотека (Schnapp and Presner, 2009, р. 6; Terras, 2012, р. 52). Циљ овог пројекта, изворно названог *Google Print*, био је да прикупи највећи корпус људског знања у дигиталном формату и да дигитализује све књиге на свету, индексира их и омогући њихово удаљено читање (Beaton, 2025).

⁵ That which is central or fundamental, as to one's business, survival, or income; a staple or cornerstone (срп. *оно што је кључно или фундаментално за нечије пословање, опстанак, приходе; неизоставан део, камен темељац*).

У овом речнику проналазимо и српске еквиваленте за идиоме *bear market*⁶, односно *bull market*⁷:

АКЦИЈЕ падају (1) деонице на берзи губе у вредности – *share prices are going down* ♦ *a bearish market* ♦ *a bear market* (Milosavljević and Vilijams-Milosavljević, 2007, p. 11).

АКЦИЈЕ скачу деонице на берзи добијају у вредности – *share prices are going up* ♦ *a bullish market* ♦ *a bull market* (Milosavljević and Vilijams-Milosavljević, 2007, p. 12).

Неколико примера српских еквивалената енглеских идиома пронашли смо и у *Речнику англо-америчког сленга*:

*beauty sleep*⁸ ићи на спавање пре поноћи (Gerzić and Janković, 2002, p. 26);

*beer and skittles*⁹, not all GB Nije sve samo zabava i uživanje; ne ide sve uvek glatko (Gerzić and Janković, 2002, p. 27).

Трећи приступ заснивао се на анализи значења идиома и проналажењу одговарајућих српских израза у речницима и збиркама народних пословица (Karadžić, 1969). Овај поступак је омогућио не само да се пронађу одговарајући преводи већ и да се оживе поједини архаични или мање познати фразеологизми који се ретко користе у савременом језику. Неколико примера таквих

⁶ A stock market where a majority of investors are selling ("bears"), causing overall stock prices to drop. Often operationally defined as a market value drop of 20% or more on some specific stock market index. (срп. берзанско тржиште где већина инвеститора продаје акције („медведи”), што узрокује пад укупних цена акција. Најчешће се оперативни дефинише као пад тржишне вредности од 20% или више на одређеном берзанском индексу)

⁷ A stock market where a majority of investors are buying ("bulls"), causing overall stock prices to rise (срп. берзанско тржиште где већина инвеститора купује акције („бикови”), што узрокује раст укупних цена акција).

⁸ Originally, sleep taken before midnight, on the belief that early sleep hours conduce to beauty and health; now (chiefly humorous), sleep at any time needed by one to stay beautiful; (countable) an instance of such sleep (срп. Изворно значење: одлазак на спавање пре поноћи, засновано на веровању да рани одлазак на спавање доприноси лепоти и здрављу; садашње значење (углавном шаљиво), спавање у било које доба дана у циљу очувања лепоте).

⁹ Fun times; pleasure and leisure (срп. Забава; ужитак и разонода).

фразеологизама, који имају значење приближно енглеским идиомима *armchair critic*¹⁰ и *bleeding heart*¹¹ наводимо у наставку:

Свакој тици бити кобац. Мијешати се у свашто (Karadžić, 1969, p. 252).

Дерати се уз туђа носила. Кад ко без узрока жали за туђом ствари (Karadžić, 1969, p. 86).

Посебан изазов представљали су идиоми који су снажно повезани са специфичним културним контекстом. У таквим случајевима било је неопходно ослонити се на познавање локалне културе и популарне традиције како би се пронашли функционални еквиваленти који ће корисницима српског језика бити разумљиви и препознатљиви. Такође су узете у обзир разлике између екавског и ијекавског изговора, па су оба изговора укључена тамо где је то било потребно. Тако за енглески идиом *armchair critic*, имамо између осталих и ова два еквивалента на српском језику: „не ници где те не сеју“ и „не ници ће те не сију“, на екавском и ијекавском изговору редом.

Веома значајан изазов у проналажењу српских еквивалената за енглеске вишечлане идиоматске изразе представљао је постављени формални критеријум идиоматичности, а то је да се фразеологизми састоје од најмање две речи. У српском језику пронашли смо неколико еквивалената енглеских идиома који се састоје од једне речи, а које за ауторе чланка имају суштински статус идиома, попут: *bad apple* – муњак, *brain teaser* – мозгалица, *cheat sheet* – пушкица. Прецизније, наведени примери не састоје се од више речи, али састоје се од више значењских јединица на нивоу зависних или независних морфема, тако да их можемо сматрати морфемским фраземама¹². Предвидевши такву могућност,

¹⁰ A person who makes criticisms on a subject they have little knowledge of or practical experience in (срп. *Особа која износи критике на рачун области о којој има врло мало знања или у којој има врло мало практичног искуства*).

¹¹ A person who is overly sympathetic to the plight of the underprivileged, exploited or helpless, or to casualties of circumstance (срп. *Особа која је превише саосећајна због недаћа оних који су угрожени, изабљивани, или беспомоћни; или су жртве околности*).

¹² Фразеологија је традиционално усмерена на лексемске фраземе (енг. *lexemic phrasemes*), тј. вишечлане изразе (идиоми, колокације), које су, према Melčuk (2021), тек један тип фразема. Наиме, поред њих, Мелчук препознаје још два типа, и то: морфемске фраземе (енг. *morphemic phrasemes*), тј. морфофраземе (подразумевају фразеолошка ограничења унутар речи) и синтаксичке фраземе (енг. *syntactic phrasemes*)(подразумевају синтаксичке обрасце). Категорија

Torunoğlu-Selamet et al. (2026) наводе категорију *Циљни језик користи идиом сачињен од једне лексичке јединице (није вишечлани израз)*¹³, те тако једночлане преводне еквиваленте нпр. за енглески идиом cheat sheet илуструју португалским (cola), грчким (skonaki), бугарским (pishtov) и мађарским примером (puska). По тоњи, мађарски, еквивалент (puska) и претходно наведени српски (пушкица) несумњиво имају сличну етимологију.

Дефиницију пушкице у значењу израженом у овом истраживању не налазимо у *Речнику српскога језика* (2011), где се у одредници за ову лексему наводи да је деминутив и хипокористик од лексеме пушка. Nišavić et al. (2024) истичу да се претрагом интернета могу пронаћи савети за прављење и употребу пушкица, али не и поузданији подаци о њиховој дефиницији, етимологији или од када се користе. Зечевић (2013), пак, наводи да пушкице у жаргону студената означавају помоћно, али и недозвољено средство на испиту, које се припрема исписивањем градива на папирићима ради преписивања.

Порекло и употребу бугарске лексеме pishtov (пищов) и мађарске puska потражили смо у етимолошким речницима¹⁴. Бугарска лексема пиштов потиче од турског назива за пиштољ; piştov, који је изведен од француског pistole. Поред основног значења, пиштов се у бугарском жаргону користи и за папирић посебно припремљен за преписивање¹⁵. Глагол стрељати (буг. стрелям) такође се жаргонски употребљава у значењу „преписивање на испиту“¹⁶ (Анастасов et al., 2002, р. 279). Сличан семантички развој бележи и мађарска лексема puska. Према трећем тому *Историјско-етимолошког речника мађарског језика* (Benkő, 1976), puska потиче од немачког Büchse, што означава ручно ватрено оружје, а првобитно футролу или кутију за држање барута. Речник наводи пет значења лексеме puska, од којих се последње односи на класичне (грчко-латинске) текстове које су ученици ко-

морфемских фразама може да послужи за будућу интерпретацију идиоматичности српских једночланих фразама.

¹³ Target language uses idiom made up of single lexical item (not an MWE).

¹⁴ Издада Етимолошког речника српског језика је у току (Бјелетић, 2024), а 4. свеска (Бл-Бљ) објављена је 2023. године.

¹⁵ „листче, специјално приготвено за преписивање“ (Анастасов et al., 2002, р. 279).

¹⁶ „преписам на испит“ (Анастасов et al., 2002, р.279).

ристили као помоћно средство при превођењу¹⁷. Реч је, дакле, о својеврсном подсетнику или мнемоничком средству, уз напомену да није било дозвољено као средство за учење. Ово значење повезано је са глаголом *puskáz*, који је, на основу идиома гађати стрелама из туђег тобољца¹⁸, тј. китити се туђим перјем¹⁹, развио значење завирити у нешто, или преузети нешто однекуд (Benkő, 1976, pp.315–316). Бугарски и мађарски примери показују, дакле, да се називи за ватрено оружје, као и у српском језику метафорички преносе на средства која ученицима омогућавају да током испита „потегну“ за туђим знањем, или претходно припремљеним белешкама.

3. Фаза 2: Генерисање визуелних примера

У другој фази за сваки од идиома генерисане су одговарајуће слике помоћу генеративног модела Midjourney²⁰, намењеног искључиво за стварање визуелног садржаја. Овај модел изабран је због расположивости „невидљивог“ (енг. *stealth*) режима, који омогућава да слике и промптови буду скривени од јавности, што је био неопходан услов да би се обезбедило да скуп остане ван домета генеративних мултимодалних модела.

За сваки од идиома било је потребно генерисати пет типова слика које представљају: (1) идиоматско значење (енг. *strongly figurative*), (2) значење блиско идиоматском (енг. *mildly figurative*), (3) значење блиско дословном (енг. *mildly literal*), (4) дословно значење (енг. *strongly literal*) и (5) нерелевантан садржај који ни на који начин није повезан са идиомом (енг. *distractor*). Овакав избор произлази из описане методологије мултимодалног представљања идиома и омогућава да се визуелна повезаност са фигуративним и дословним значењем не посматра само као бинарна опозиција, већ и кроз различите степене семантичке сличности. Прва четири типа слика формирају континуум од изразито фигуративног до изразито дословног представљања значења идиома, док нерелевантна слика има улогу негативног примера, односно контролне категорије која није семантички повезана ни са јед-

¹⁷ „klasszikus (görög-latin) szövegek fordítása mint a diákok segédeszköze“ (Benkő, 1976, pp.315–316)

¹⁸ „más tegzéből vagy a máséből lövöldöz“ (Benkő, 1976, p. 316).

¹⁹ „tollakkal ékeskedik’ szólás” (Benkő, 1976, p. 316).

²⁰ <https://www.midjourney.com/home>

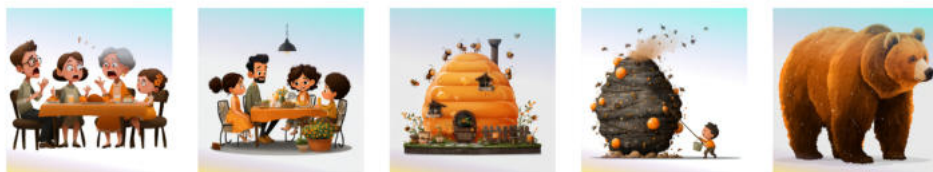
ним од ова два значења. Овако формиран скуп слика омогућава да се приликом евалуације испита не само да ли модел разликује дословно од фигуративног значења већ и колико успешно препознаје различите степене повезаности визуелног садржаја са значењем идиома.

Да би се слике генерисале, било је потребно најпре написати промптове, односно описе садржаја за сваку од слика као упутство генеративном моделу шта је потребно да се прикаже на слици. Промптови су писани на енглеском језику, а модел генерише четири слике за сваки промпт који му се пошаље од којих се бира једна за потребе представљања значења. Примери генерисани за идиом „дирнути у осињак”, српски еквивалент енглеског идиома *can of worms*, приказани су на Слици 1. За генерисање ових слика коришћени су следећи промптовима на енглеском језику према претходно наведеном редоследу значења:

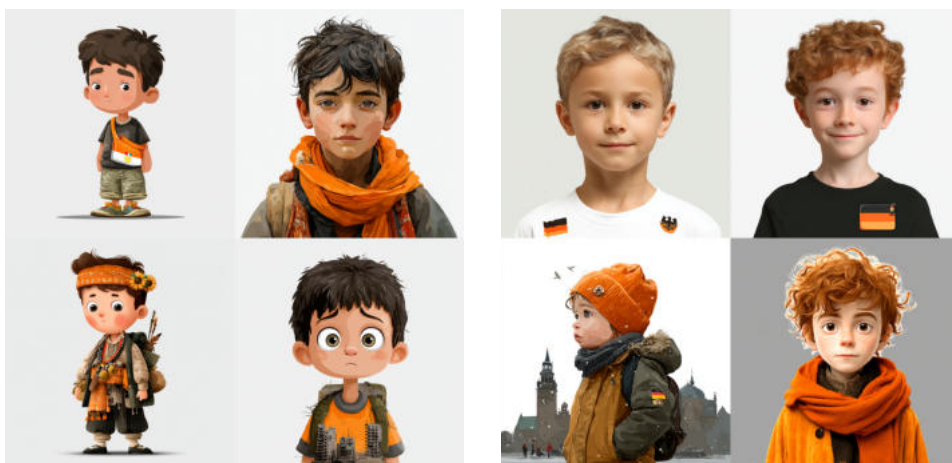
1. people arguing at the family dinner (срп. *људи се расправљају током породичне вечере*);
2. family dinner (срп. *породична вечера*);
3. a beehive (срп. *кошница*);
4. a person pokes a wasp nest (срп. *особа дира осиње гнездо*);
5. a bear (срп. *медвед*).

У одређеним случајевима резултати су указивали на ограничења генеративног модела приликом визуализације идиоматских и апстрактних појмова. Услед тога, било је неопходно итеративно прилагођавати текстуалне захтеве (промптове) како би се у генерисаним сликама успешно постигле жељене сцене. Резултати показују да модел често има потешкоће у приказивању сложених емоционалних стања, нарочито када је потребно приказати више особа са различитим емоцијама у истој сцени. Даље је уочена тенденција ка дословном тумачењу апстрактних појмова, што доводи до стереотипних и поједностављених визуелних решења. Поред тога, модел показује ограничене способности у приказивању просторних односа, количинских одредница и сложених интеракција између људи и објеката.

Анализа генерисаних слика указала је на присуство различитих облика пристрасности у неким од примера, што је у складу



Слика 1: Визуелни примери за идиом „дирнути у осиње гнездо”.



Слика 2: Лево: дете из Србије. Десно: дете из Немачке.

са налазима других истраживања о генеративној вештачкој интелигенцији (Ferrara, 2024; Zhou et al., 2024). Најчешће су уочене родне и професионалне пристрасности, као и културни и расни стереотипи који се репродукују кроз визуелне приказе. На слици 2 приказани су резултати за промптове ”дете из Србије” на левом делу (енг. *a kid from Serbia*) и „дете из Немачке“ (енг. *a kid from Germany*) у десном делу. Може се уочити значајна разлика између ове две представе модела. Поред тога, модел показује склоност ка фаворизовању познатих личности и комерцијалних брендова, што указује на утицај података коришћених током обучавања. Један од упечатљивијих примера приказан је на левој страни слике 3, насталој као резултат промпта „патике“ (енг. *sneakers*), где се може приметити да је модел од четири слике генерисао два пара патика бренда Nike, и један пар бренда Converse.

Одређени резултати могу се довести у везу и са техникама ублажавања пристрасности које мењају начин представљања породичних и друштвених улога (Yue and Yang, 2024), као што је приказано на слици 3 десно која је настала као резултат промпта „породична вечера“ (енг. *a family dinner*). На крају, уочено је да системи за филтрирање садржаја онемогућавају генерисање појединих сцена које се сматрају осетљивим или потенцијално неприкладним, као у примеру „замршено као пасја црева“, што је у складу са налазима Marsoof et al. (2023).



Слика 3: Лево: патике. Десно: породична вечера.

4. Фаза 3: Проналажење примера дословног и идиоматског значења вишечланих израза

За сваки идиом било је потребно припремити по два примера употребе у идиоматском и дословном значењу. Премда су упутства за проналажење одговарајућих примера дословног и идиоматског значења за пронађене идиоме подразумевала и употребу алата генеративне вештачке интелигенције (Gemini Flash 2.5), или смишљање примера уз ослањање на интуицију²¹, настојали смо да пронађемо што више аутентичних примера употребе српских идиома из постојећих текстуалних ресурса. Том приликом смо, поред Корпуса савременог српског језика Srp-Kor2021 (Vitas et al., 2025) и колекције старих српских романа

²¹ Поједини тимови (чланови тимова) у значајном броју наводе управо такве примере дословног и идиоматског значења уз напомену *my own creation*.

srpELTeC (Stanković et al., 2022), интензивно користили и репозиторијум дигиталних издања books.google.com, вршећи ћириличне упите, те водили детаљну евиденцију о изворима одакле су примери (у неизмењеном облику) издвојени. Тако су за претходно споменути идиом „дирнути у осињак” пронађени примери „*Не дирај у осињак, да те осе не изапу.*”, преузет из Српског дијалектолошког зборника (Младеновић, 2004) и „*Већ на првом кораку његови потези наишли су на тиха гунђања челника политичких странака и вечитих синдикалних лидера, јер је дирнуо у осињак.*” из листа Политика²², који илуструју примену у дословном и у идиоматском значењу редом.

Отежавајућу околност у овој фази представљала је и чињеница да у неким случајевима није било могуће пронаћи дословну употребу неког идиоматског значења, будући да таква у српском језику у свакодневной употреби или не постоји или је изнимно ретка; на пример: банана република (енг. *banana republic*), црно тржиште (енг. *black market*, одлив мозга (енг. *brain drain*), замршено као пасја црева (енг. *brain surgery*). Индикативно је да су наведени примери, изузев потоњег, калкови енглеских идиоматских израза. У ретким случајевима, није било могуће пронаћи пример идиоматског значења за неки српски преводни еквивалент, попут постпорођајна депресија (енг. *baby blues*).

5. Закључак

Посматрано у целини, истраживање показује да је проналажење одговарајућих идиоматских еквивалената и примера њиховог коришћења сложен процес који захтева комбинацију лингвистичког знања, културолошког разумевања и консултовања релевантних фразеолошких извора. Истовремено, резултати указују на то да коришћени генеративни модел још увек има значајна ограничења у разумевању и визуелизацији апстрактних и културолошки условљених значења, као и на присуство различитих облика пристрасности који утичу на квалитет генерисаних садржаја. Потенцијални правац за ублажавање уочених ограничења може бити интеграција модела са релевантним лингвистичким и фразеолошким ресурсима, уз укључивање стручне процене и

²²

<https://www.politika.rs/sr/clanak/276976/pogledi/kome-smeta-ogledalo>

систематске провере генерисаног садржаја у погледу језичке и културолошке адекватности и присуства пристрасности.

5.1. Ограничења истраживања и смернице за будуће радове

Овај чланак, како је наведено, проистекао је из Admire 2 иницијативе у оквиру COST акције CA21167 – Universality, diversity and idiosyncrasy in language technology (UniDive). Стога наше полазиште нису могли бити аутентични идиоми српског језика, већ идиоматски еквиваленти полазном скупу идиома енглеског језика. У циљу превазилажења поменутог ограничења истраживања, примери употребе прикупљани су доминантно из аутентичних језичких извора, иако су организатори COST акције прихватили и примере који су представљали адаптације или креативне употребе сарадника. Имајући у виду поменуто ограничење истраживања, аутори планирају накнадно проширење овог истраживања које би било засновано на одабраном сету аутентичних идиома српског језика екавског и ијекавског изговора.

Захвалност

Овај рад је подржан кроз COST акцију CA21167 UniDive, коју финансира COST (European Cooperation in Science and Technology).

Литература

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoglu Selamet, Thomas Pickard, Aline Villavicencio, Adriana Silvina Pagano, and Gülşen Eryiğit. 2026. [MWE-2026 Shared Task: AdMIRE 2 Advancing Multimodal Idiomaticity Representation](#). In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 276–287, Rabat, Morocco. Association for Computational Linguistics.

Toni Beaton. 2025. [Google Books: Mass Digitization and the Implications for Public Libraries](#). *Pathfinder: A Canadian Journal for Information Science Students and Early Career Professionals*, 5(1):1–16.

- Loránd Benkő, editor. 1976. *A magyar nyelv történeti-etimológiai szótára*, volume 3. Akadémiai Kiadó, Budapest. Ö-Zs.
- Ranko Bugarski. 2014. English idioms in theory and practice: A personal account In. In T Prčić and M Marković, editors, *Zbornik u čast Draginji Pervaz - Engleski jezik i anglofone književnosti u teoriji i praksi*, pages 81–89. Filozofski fakultet, Novi Sad.
- Emilio Ferrara. 2024. [Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies](#). *Sci*, 6(1):3.
- Borivoj Gerzić and Ksenija Janković. 2002. *Rečnik anglo-američkog slenga: preko 5000 žargonskih reči i izraza*. Istar.
- Robert Hunter et al. 1897. *The American Encyclopaedic Dictionary*, volume 5. R. S. Peale and J. A. Hill, Chicago and New York.
- Vuk Stefanović Karadžić. 1969. *Srpske narodne poslovice i druge različne kao one u običaj uzete riječi*. Nolit.
- Althaf Marsoof, Andrés Luco, Harry Tan, and Shafiq Joty. 2023. [Content-filtering AI Systems—Limitations, Challenges and Regulatory Approaches](#). *Inf. Commun. Technol. Law*, 32(1):64–101.
- Igor Melčuk. 2021. [Morphemic and syntactic phrasemes](#). *Yearbook of Phraseology*, 12(1):33–74.
- Boško Milosavljević and Margot Vilijams-Milosavljević. 2007. *Srpsko-engleski rečnik idioma*. Beograd: Zavod za udžbenike.
- Ivan Nišavić, Aleksandar Pavlović, Ana Lipij, Balša Delibašić, Đorđe Hristov, Igor Cvejić, Sekulović Milica, Predrag Krstić, and Tamara Plećaš. 2024. *Inventar obrazovanja*. Institut za filozofiju i društvenu teoriju, Univerzitet u Beogradu, Beograd.
- Đorđe Otašević. 2007. *Mali srpski frazeološki rečnik*. Istar.
- Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. [SemEval-2025 Task 1: AdMIRe – Advancing Multimodal Idiomaticity Representation](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2597–2609, Vienna, Austria. Association for Computational Linguistics.
- Jeffrey Schnapp and Todd Presner. 2009. [Digital Humanities Manifesto 2.0](#). Accessed: 28 August 2026.

- Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, Duško Vitas, Mihailo Skoric, and Milica Ikonić Nešić. 2022. [Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3337–3345, Marseille, France. European Language Resources Association.
- Melissa Terras. 2012. Digitization and Digital Resources in the Humanities. In Claire Warwick, Melissa Terras, and Julianne Nyhan, editors, *Digital Humanities in Practice*, pages 47–70. Facet Publishing, London.
- Dilara Torunoğlu-Selamet, Doğukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Vilavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amália Mendes, Anna Kanellopoulou, Antoni Brosa-Rodriguez, Baiba Valkovska, Beata Wojtowicz, Bollette Pedersen, Carlos Manuel Hidalgo-Ternero, Chaya Liebeskind, Danko Jokić, Diego Alves, Eleni Triantafyllidi, Erik Velldal, Fred Philippy, Giedre Valunaite Oleskeviciene, Ieva Rizgeliene, Inguna Skadina, Irina Lobzhanidze, Isabell Stinessen Haugen, Jauza Akbar Krito, Jelena M. Marković, Johanna Monti, Josue Alejandro Saucá, Kaja Dobrovoljc, Kingsley O. Ugwuanyi, Laura Rituma, Lilja Øvrelid, Maha Tufail Agro, Manzura Abjalova, Maria Chatzigrigoriou, María del Mar Sánchez Ramos, Marija Pendevska, Masoumeh Seyyedrezaei, Mehrnoush Shamsfard, Momina Ahsan, Muhammad Ahsan Riaz Khan, Nathalie Carmen Hau Norman, Nilay Erdem Ayyıldız, Nina Hosseini-Kivanani, Noémi Ligeti-Nagy, Numaan Naem, Olha Kanishcheva, Olha Yatsyshyna, Daniil Orel, Petra Gommarelli, Petya Osenova, Radovan Garabik, Regina E. Semou, Rozane Rebechi, Salsabila Zahirah Pranida, Samia Touileb, Sanni Nimb, Sarfraz Ahmad, Sarvinoz Sharipova, Shahar Golan, Shao-xiong Ji, Sopuruchi Christian Aboh, Srdjan Sucur, Stella Markantonatou, Sussi Olsen, Vahide Tajalli, Veronika Lipp, Voula Giouli, Yelda Yeşildal Eraydın, Zahra Saaberi, and Zhuohan Xie. 2026. [A Parallel Cross-Lingual Benchmark for Multimodal Idiomaticity Understanding](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 9434–9448, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

- Duško Vitas, Ranka Stanković, and Cvetana Krstev. 2025. [O familiji korpusa savremenog srpskog jezika srpkor](#). In *South Slavic Languages in the Digital Environment (JuDig)*, volume 1, pages 3–27. Filološki fakultet Univerziteta u Beogradu, Beograd, Srbija.
- Wikimedia Foundation. [Wiktionary: The Free Dictionary](#).
- Meiting Yue and Boyi Yang. 2024. The Gender Show in “Midjourney”: The Stubborn Biases and Gender Fantasies of AI Painting Tools. *Communications in Humanities Research*, 33:152–159.
- Mi Zhou, Vibhanshu Abhishek, Timothy P. Derdenger, Jaymo Kim, and Kannan Srinivasan. 2024. [Bias in Generative AI](#). *ArXiv*, abs/2403.02726.
- Васил Анастасов, Христина Дейкова, Лиляна Димитрова-Тодорова, Уте Дукова, Димитрина Михайлова, Мира Начева, Мария Рачева, Георги Риков, Людвиг Селимски, and Тодор Ат. Тодоров. 2002. *Български етимологичен речник*, volume 6. Академично издателство „Марин Дринов“, София. [пускам-словар2](#).
- Марта Бјелетић. 2024. Етимолошки одсек Института за српски језик САНУ – 40 година постојања. *Трибина библиотеке САНУ*, 12(12):222–225.
- Снежана Зечевић. 2013. Улога жаргона у формирању друштвеног идентитета припадника студентске популације. *Баштина*, 34:15–28.
- Александар Младеновић, editor. 2004. [Српски дијалектолошки зборник](#), volume 51. Српска академија наука и уметности и Институт за српски језик САНУ, Београд.
- Лазар Стефановић. 1882. Новци могу да пренесу болест. *Народно здравље*, 1(19):231–232.
- Милица Вујанић и други. 2011. *Речник српскога језика*. Матица српска, Нови Сад.

ADMIRE 2 Serbian: Building a Serbian Multimodal Idiom Dataset

Danka Jokić, Srđan Šućur, Jelena Marković

Summary

This paper presents a Serbian multimodal idiom dataset and the process of its construction, developed within the *Admire 2* initiative. It discusses the challenges involved in preparing Serbian equivalents of English multi-word idiomatic expressions and examples of their use, as well as their visual representations generated using the *Midjourney* generative image synthesis model. Whenever possible, usage examples were collected from language corpora and other textual sources to ensure the authenticity, naturalness, and linguistic validity of the presented expressions. Particular attention was devoted to identifying appropriate idiom translations aligned with Serbian cultural heritage. The image generation process revealed limitations of the generative model in depicting abstract concepts, complex emotional states, and spatial relationships, as well as various forms of bias in the generated images.

Keywords: idiomatic expressions, multimodal corpora, model bias

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

