cjvt Center za jezikovne vire in tehnologije

Center za jezikovne vire in tehnologije Univerze v Ljubljani

Simon Krek

Ljubljana, 20. 10. 2021

Zgodovina

Sporazum – december 2012

- Sporazum o sodelovanju pri oblikovanju Centra za jezikovne vire in tehnologije Univerze v Ljubljani
  - 17. decembra 2012
- Podpisnice
  - Univerza v Ljubljani (rektorat)
  - Fakulteta za družbene vede (dr. **Monika Kalin Golob**)
  - Filozofska fakulteta
  - Pedagoška fakulteta
  - Fakulteta za elektrotehniko
  - Fakulteta za računalništvo in informatiko
- „Podpisnice soglašajo, da se v okviru Mreže raziskovalnih infrastrukturnih centrov Univerze v Ljubljani (MRIC) oblikuje CJVT – Center za jezikovne vire in tehnologije Univerze v Ljubljani.“

CJVT UL (MRIC) – 2015-2021

- Mreža raziskovalnih infrastrukturnih centrov Univerze v Ljubljani
  - Administrativni sedež: Fakulteta za družbene vede
  - Prostori: Fakulteta za računalništvo in informatiko
  - Sedež od 2022: Fakulteta za računalništvo in informatiko
- Nova članica (od 2021)
  - Fakulteta za upravo
- Financiranje: infrastrukturni program ARRS (UL)
  - 0,3 FTE (+ materialni stroški in amortizacija)
  - Od 2018: **Veliki slovensko-madžarski slovar** (~38.000 EUR/leto)

Cilji in poslanstvo

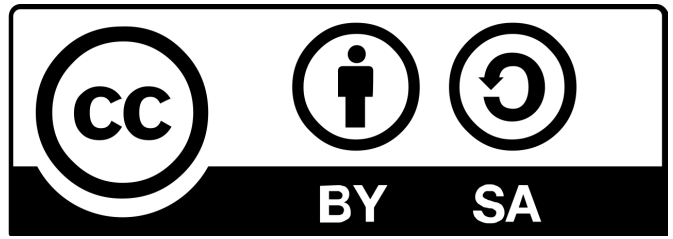
Cilji in poslanstvo CJVT UL

- Oblikovanje institucionalnega okvira, kjer poteka načrten in sistematičen dolgoročni **razvoj tehnologij, virov in orodij** za slovenski jezik.
- …le tako bo mogoče zagotoviti, da bo **slovenščina v digitalnem okolju** obdržala enakovreden status z drugimi nacionalnimi jeziki.
- Jezikovne vire in orodja želimo razvijati po **šestih stebrih:** jezikovni opis, standardizacija, tehnologije, terminologija, večjezičnost in govorci s posebnimi potrebami.
- Vizija rezultatov sodelovanja je urejena **jezikovna opremljenost** slovenščine.

- Več na spletni strani: https://www.cjvt.si/

Kako do cilja?

- Viri, orodja in servisi: besedilni korpusi, slovarji in leksikoni, spletni portali, jezikovnotehnološka orodja, na primer:
  - Referenčni korpus pisne standardne slovenščine Gigafida 2.0
  - Korpus šolskih pisnih izdelkov Šolar 2.0
  - Korpus govorjene slovenščine Gos 1.0
  - Slovenski oblikoslovni leksikon Sloleks 2.0
  - Slovar sopomenk sodobne slovenščine Sopomenke 1.0
  - Kolokacijski slovar sodobne slovenščine Kolokacije 1.0
  - Spletno orodje za strojno postavljanje vejic Vejice 1.0
  - Aplikacija za oceno težavnosti besedil v slovenščini Berljivost 1.0
- Sporazum: „Podpisnice si bodo prizadevale, da se za delovanje CJVT zagotovijo sistemska finančna sredstva na razpisih ARRS, razpisih različnih programov EU in drugih virov“



Dva primera – RSDO in ON

Razvoj slovenščine v digitalnem okolju

- Trajanje: **maj 2020 – februar 2023**
- Financer: **Evropski sklad za regionalni razvoj in Ministrstvo za kulturo**
- Sredstva: **4M EUR**
- Konzorcij: **12 partnerjev** (univerze, inštituti, podjetja)
- Koordinator: **Univerza v Ljubljani** (Center za jezikovne vire in tehnologije)
- Spletna stran: https://www.slovenscina.eu/

www.slovenščina.eu

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI STRUKTURNI IN INVESTICIJSKI SKLADI
NALOŽBA V VAŠO PRIHODNOST

Cilji in sklopi

- Cilj projekta je zadovoljiti potrebe po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik za raziskovalne organizacije, za podjetja in za širšo javnost.
- **Jezikovni viri** (besedilni korpusi itd.)
- **Govorne tehnologije** (razpoznava, sinteza govora)
- **Semantični viri in tehnologije**
- **Strojno prevajanje**
- **Terminološki portal**
- **Infrastrukturni center** (CLARIN.SI)

Online Notes (ON)

- Projekt: razvoj računalniškega sistema za avtomatizirano prevajanje slovenskih predavanj v izbrane tuje jezike (slovenščina-angleščina)
- Financiranje: Razvojni sklad Univerze v Ljubljani
- Izvajalec: Fakulteta za računalništvo in Center za jezikovne vire in tehnologije
- Trajanje: 2020-2024
- Stanje: poskusna predavanja na FRI, FF in FDV v **novembru 2021**

- Video o sistemu ON: https://youtu.be/hWRONPdHh3o

Hvala.

Simon Krek
simon.krek@cjvt.si

Center za
jezikovne vire
in tehnologije

Večna pot 113
1000 Ljubljana
Slovenija

www.cjvt.si
00386 14798299
info@cjvt.si

Univerza v Ljubljani

cjvt Center za jezikovne vire in tehnologije

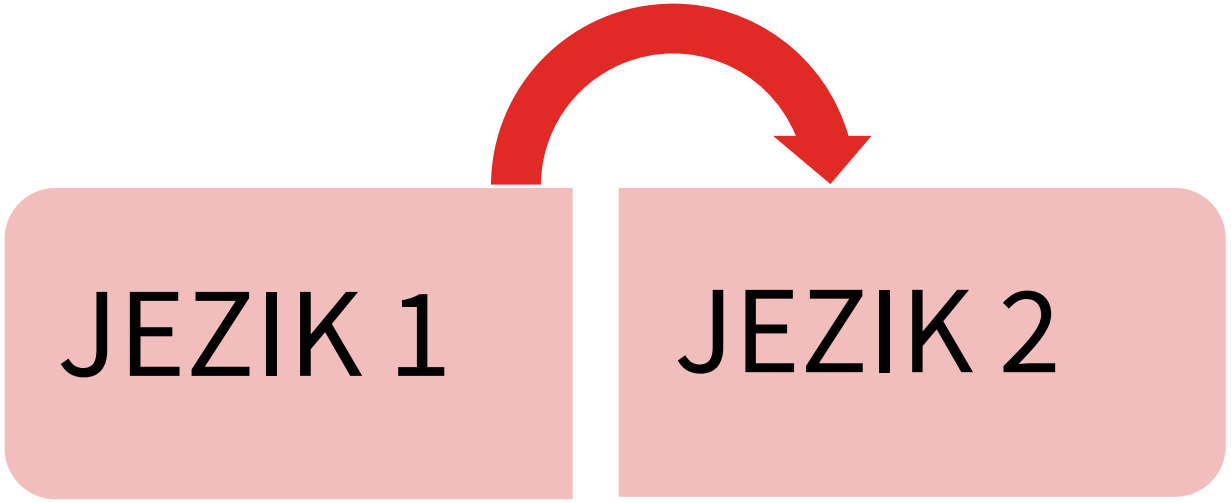
Veliki slovensko-madžarski slovar

Iztok Kosem, Júlia Bálint Čeh, Primož Ponikvar, Petra Zaranšek, Urška Kamenšek, Peter Koša, Annamária Gróf, Nándor Böröcz, Jolanda Harmat Császár, Imre Szíjártó, Borut Šantak, Polona Gantar, Simon Krek, Rebeka Roblek, Karolina Zgaga, Urban Logar, Eva Pori, Špela Arhar Holdt, Vojko Gorjanc

Osnovne informacije

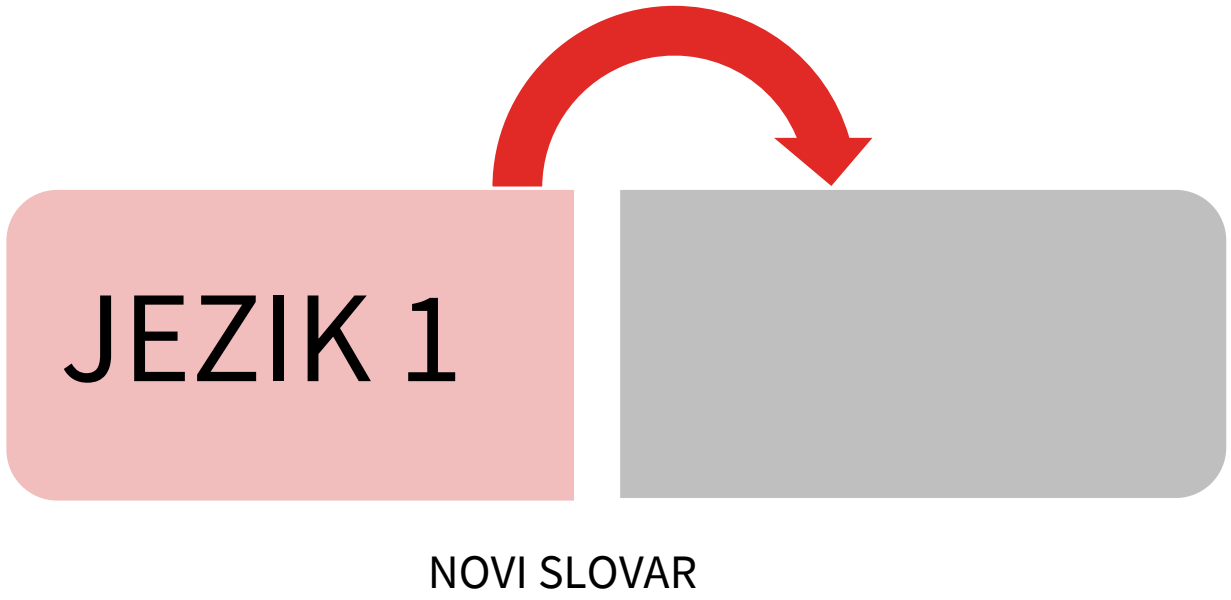
- Projekt poteka na Centru za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT UL)
- Okvir: Mreža infrastrukturnih centrov Univerze v Ljubljani
- Financer: ARRS (namenska sredstva)
  - 36.307 EUR letno (od septembra 2018)
  - Iztek financiranja konec 2021, vložena prijava za podaljšanje

Klasični pristop

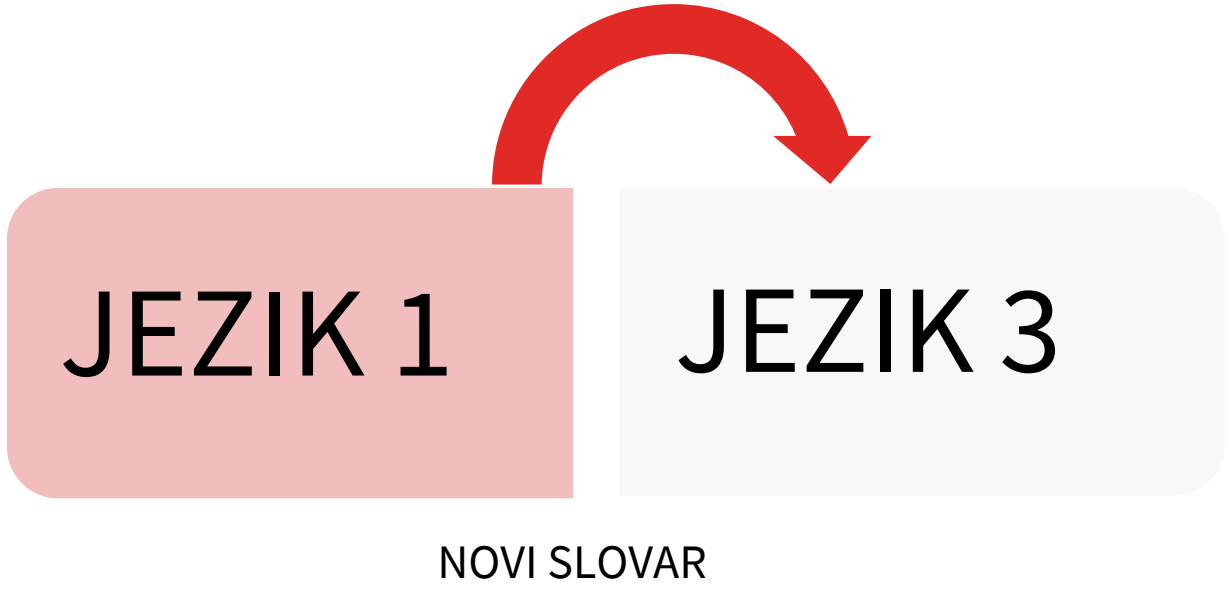


ŽE DOKONČANI SLOVAR

Klasični pristop

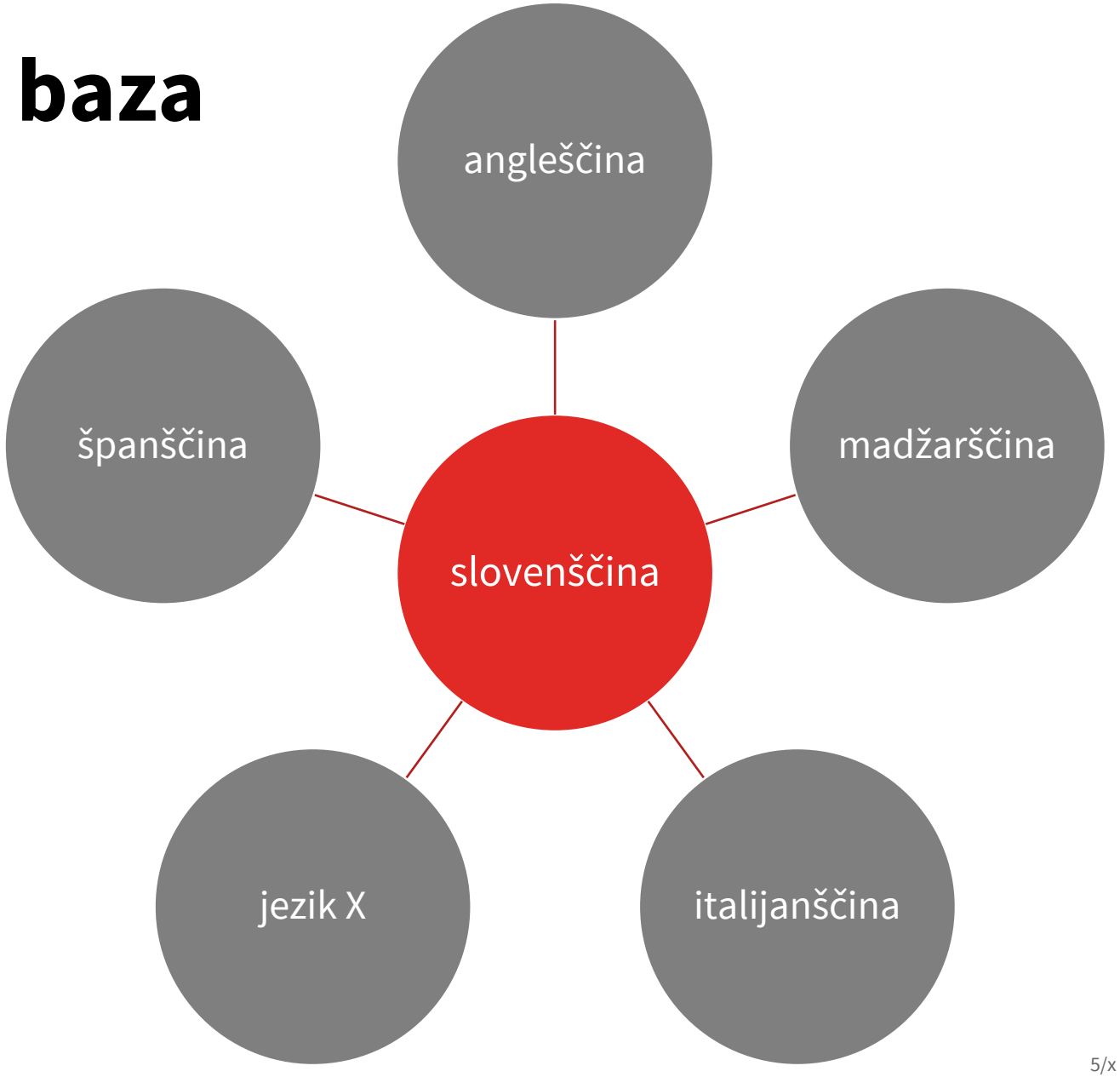


Klasični pristop



- POTENCIALNI PROBLEMI:
  - stroški odkupa (razen če slovar nastaja pri istem izdajatelju)
  - omejitve dostopnosti novega slovarja zaradi avtorskih pravic
  - zastarelost jezikovnega opisa
  - manjkajo kontrastivno zanimivi podatki

Digitalna slovarska baza



Izhodišča

- Koncept slovensko-madžarskega slovarja (Kosem idr. 2018)
  - Ciljno-raziskovalni projekt ARRS (2015-2018)
- sodobna mednarodna leksikografska praksa
- analiza sodobnega slovenskega jezika
- uporaba najsodobnejše metodologije in orodij (npr. korpusni pristop, Sketch Engine)
- izdelava strojno berljive podatkovne baze
- uporabniška naravnanost
- odprta dostopnost podatkov

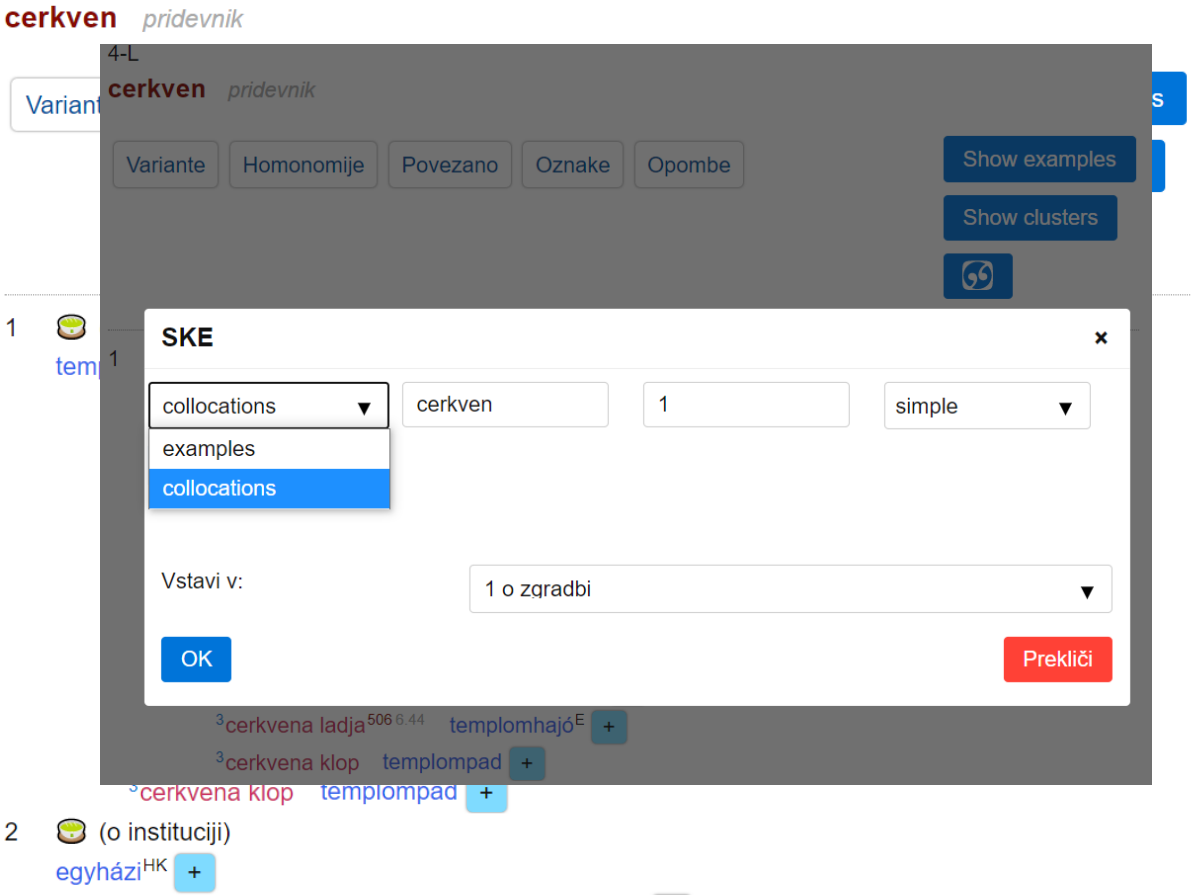
Ekipa

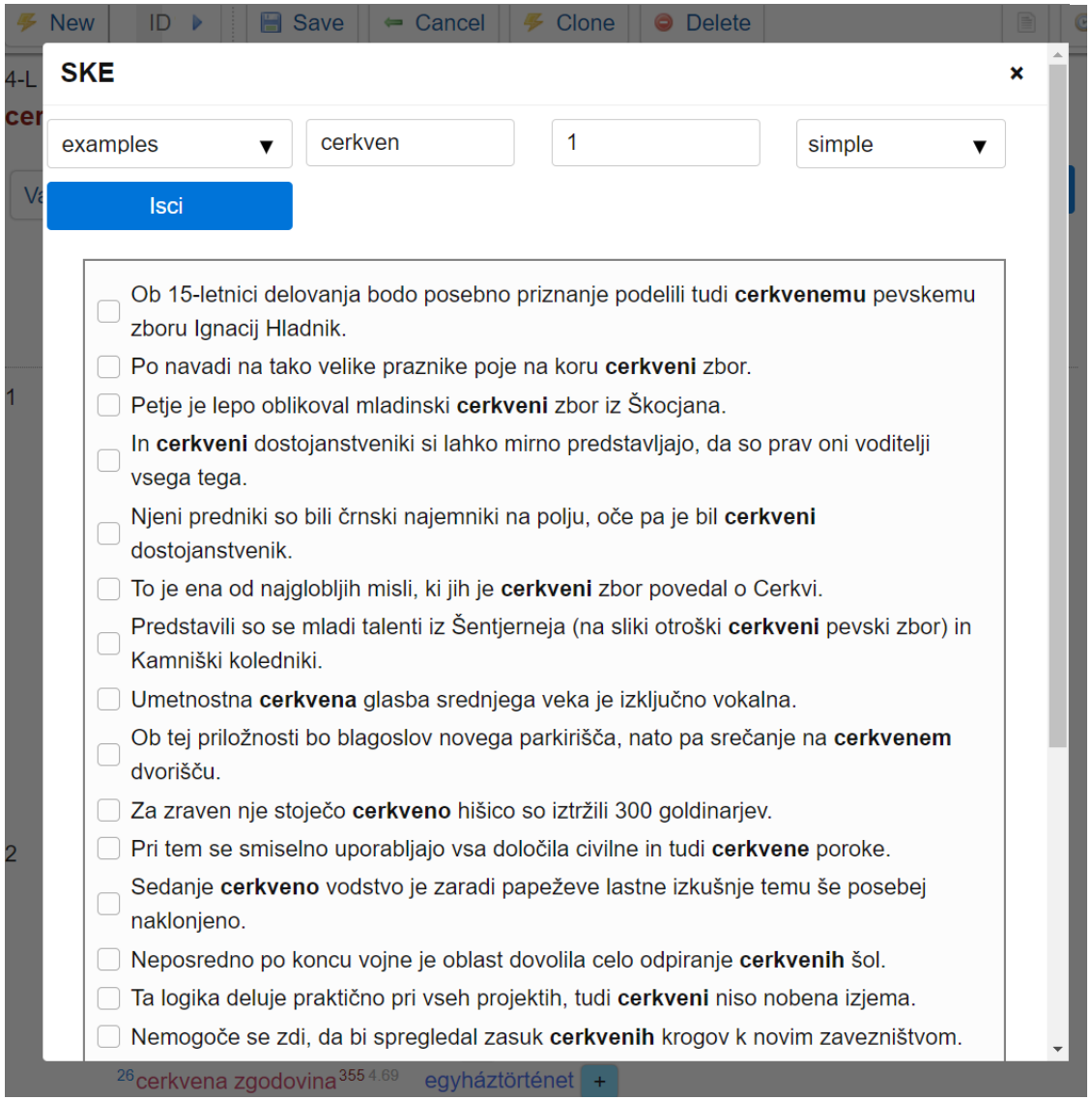
- Več kot 20 sodelavcev:
  - leksikografi
  - leksikografi-prevajalci
  - lektorji
  - terminologi
  - jezikoslovci
  - tehnični sodelavci
- Mednarodni strokovni svetovalci:
  - dr. Jelena Kallas (Inštitut za estonski jezik)
  - dr. Katalin P. Márkus (Univerza Reformirane cerkve Károli Gáspár, Budimpešta)
  - dr. Attila Mártonfi
  - Michael Rundell (Lexicography Masterclass)
  - dr. Carole Tiberius (Inštitut za nizozemski jezik)
  - dr. Tamás Váradi (Raziskovalni inštitut za jezikoslovje, Madžarska akademija znanosti)

Sodelovanja z inštitucijami

- Znanstveno-raziskovalna dejavnost
  - Programske skupine na Univerzi v Ljubljani
    - Jezikovni viri in tehnologije za slovenski jezik (šifra ARRS: P6-0411)
    - Slovenski jezik - bazične, kontrastivne in aplikativne raziskave (šifra ARRS: P6-0215)
  - Raziskovalni inštitut za jezikoslovje, Madžarska akademija znanosti
- Govorne tehnologije:
  - eBralec (Amebis, Alpineon)
  - Profivox HMM TTS (Univerza za tehnologijo, Budimpešta)
- Slovarska in korpusna orodja:
  - ELEXIS (Evropska leksikografska infrastruktura, Institut Jožef Stefan)

Veliki slovensko-madžarski slovar



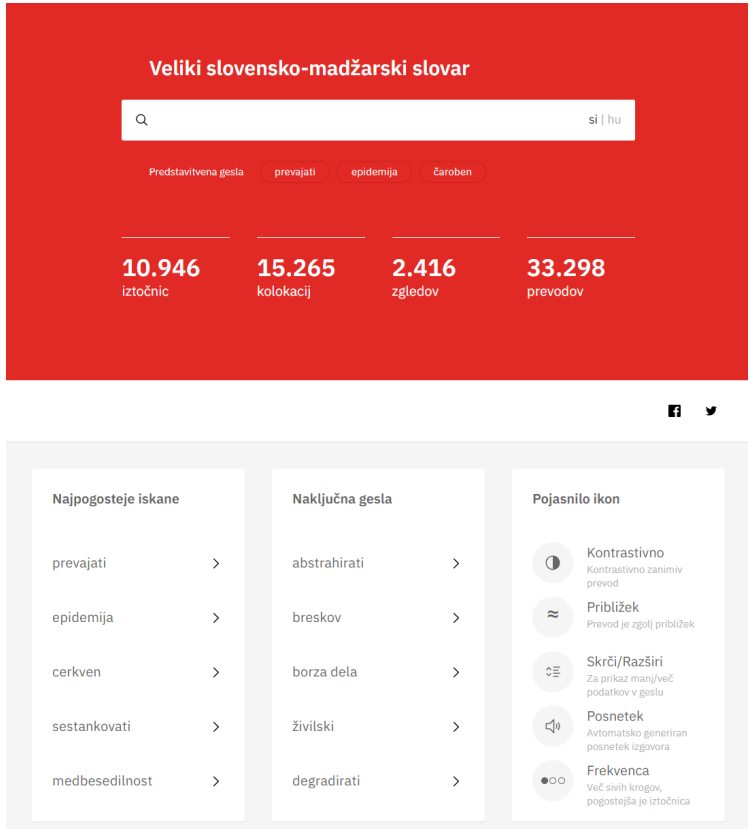


Gradivo

- Slovenščina
  - Leksikalna baza za slovenščino
  - Kolokacijski slovar sodobne slovenščine
  - Slovar sopomenk sodobne slovenščine
  - Korpusi:
    - Gigafida 2.0
    - JANES (*bukov*, *živčnež*)
    - KAS (*akumulativnost*)
    - Korpus šolskih učbenikov (*stotisočica*, *sulfoniranje, zenitne padavine, ces, fes*)
  - Posvetovalno: obstoječi enojezični viri
- Madžarski prevodi
  - Tezaver
  - Enojezični korpus Hungarian Web (huTenTen)
  - Razni enojezični viri
  - Dvojezični
  - Posvetovalno: obstoječi dvojezični viri

Veliki slovensko-madžarski slovar

- Temeljni dvojezični slovar (ARRS OSIC)
- Gesla izdelana povsem na novo
- Rastoči slovar
- Različica 1.0 na https://viri.cjvt.si/slovensko-madzarski/slv/:
  - 10.946 iztočnic
  - 15.265 kolokacij
  - 2.416 zgledov
  - 33.298 prevodov
- Uradna objava: 20. 10. 2021
- Namenjen različnim uporabniškim skupinam, tako slovenskim kot madžarskim uporabnikom
- Baza na voljo v repozitoriju CLARIN.SI pod licenco 4.0 CC-BY-SA (http://hdl.handle.net/11356/1453)



Vsebina slovarja

- Leksikografske novosti:
  - Večbesedne iztočnice
  - Pomenska členitev izhaja iz slovenščine
  - Vsak pomen je opremljen z indikatorjem
  - Nov sistem oznak oz. kvalifikatorjev
- Raznolikost iztočnic/pomenov oz. besedišča
  - Od enobesednega do večbesednega (stalne zveze, frazeologija)
  - Od splošnega do strokovnega
    - 90 različnih področnih oznak (zoologija, kemija, botanika ipd.)
  - Od standardnega do nestandardnega
    - neformalno (npr. *cincati*, *fejst*, *čekiranje*), ljudsko (*babji zob* → *šipek*)
  - Od nezaznamovanega do zaznamovanega
    - izraža negativen odnos (*butelj*, *golobnjak*, *cviliti*), vulgarno (npr. *jajca*)
  - Od občnega do lastnoimenskega
    - imena držav, prebivalcev, zgodovinskih osebnosti

Prevodne rešitve

- Neposredni prevedki (praviloma do trije)
  - Izjeme odobrijo redaktorji, npr. *cviliti*

1 **o ljudeh in živalih** ▸ **visít** **visítozik** **vonyít** | **vonít**

- Prevedki + razlagalna dopolnila, npr. *čekan*

1 **zob pri živalih** pogosto v množini ▸ ≈ **fog** [állatfog]

- Približki, npr. *kisik*

2 **pogoj bivanja ali razvoja** izraža odobravanje ▸ ≈ **létfeltétel**

- Razlage (ko ni ustreznega prevedka), npr. *EMŠO*

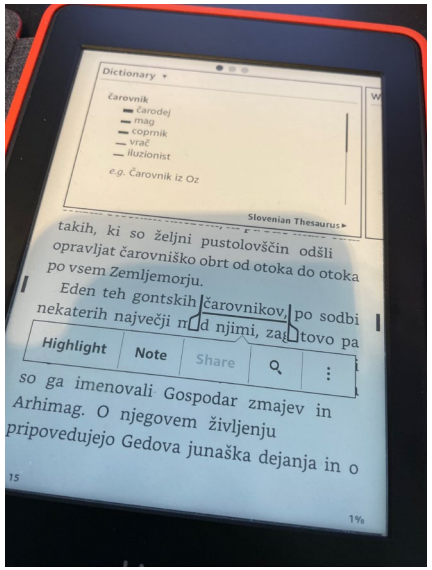
1 **enotna matična številka občana** ▸ *személyi azonosító szám*

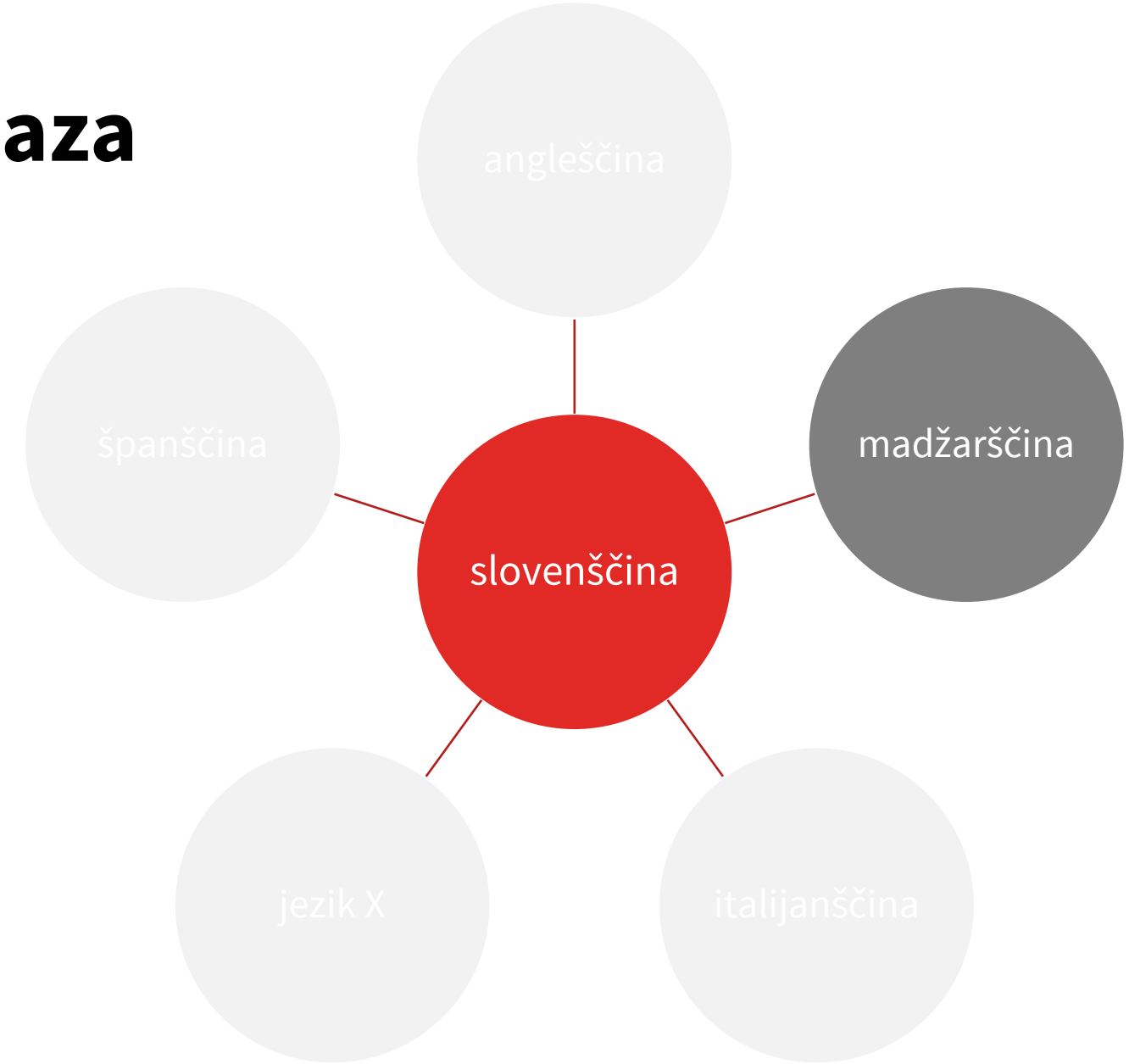
- Kontrastivno zanimivi prevodi, npr. *deževati*

rahlo deževati ▸ szemerkél az eső

Digitalna slovarska baza

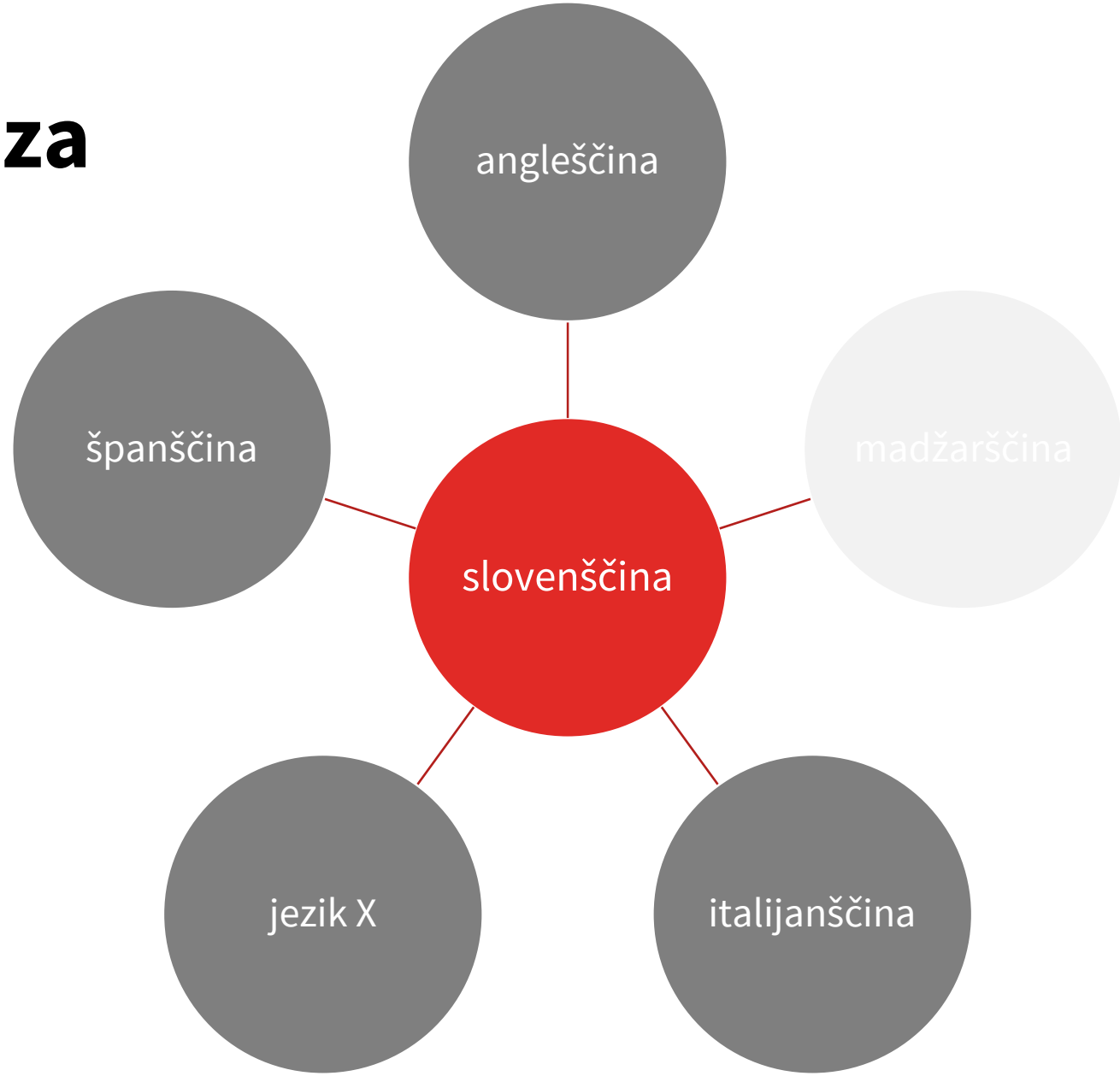
- Koristi za par slovenščina – madžarščina:
  - didaktične aplikacije
  - jezikovne tehnologije: igre, orodja za pomoč pri pisanju/branju (npr. Kindle), strojni prevajalniki





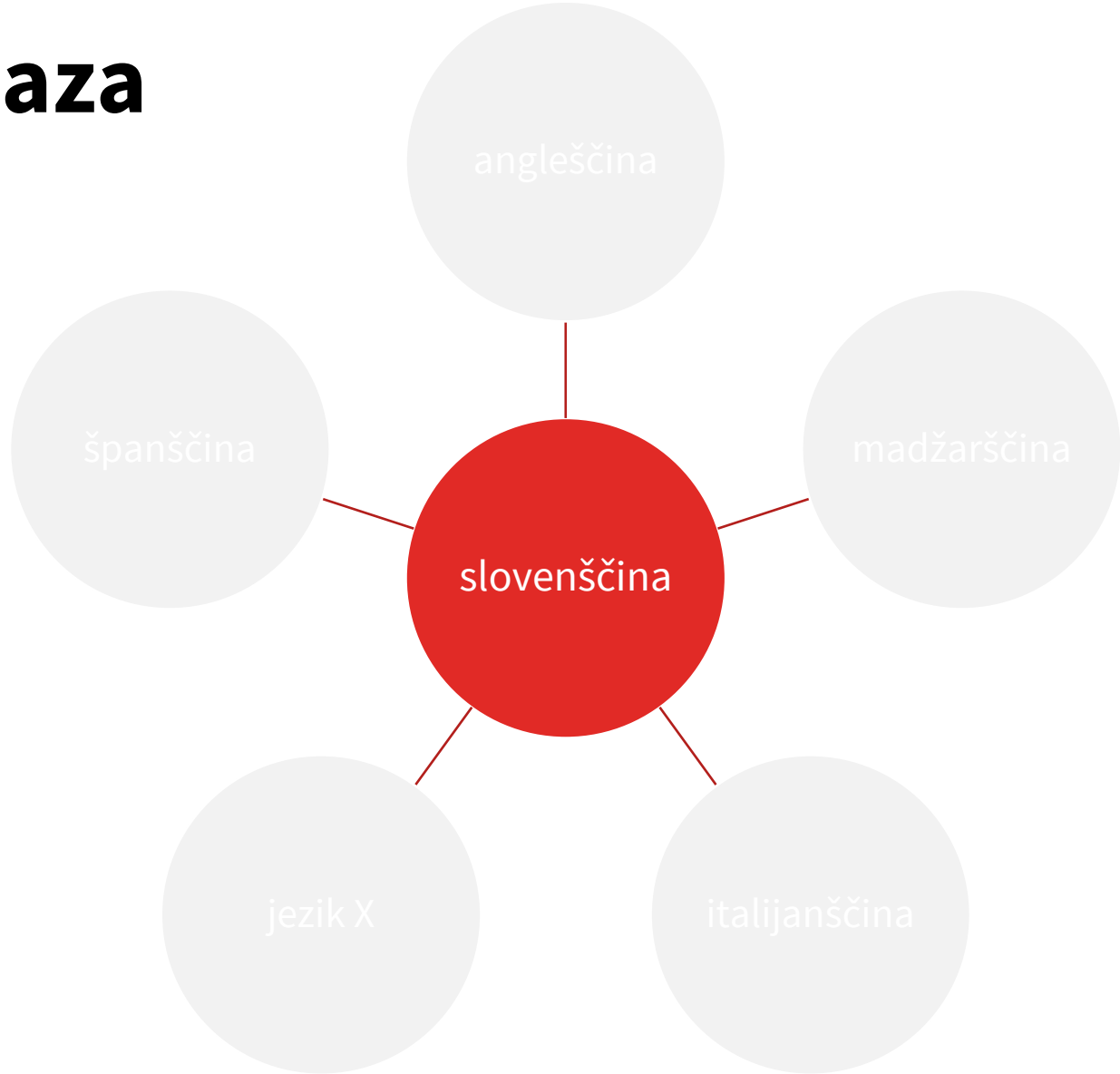
Digitalna slovarska baza

- Koristi za dvojezične vire:
  - v bazi so na voljo dodatne kolokacije in zgledi
  - možnost izdelave slovarjev in podobnih virov za pare slovenščina – tuji jezik



Digitalna slovarska baza

- Koristi za enojezične vire:
  - Pomenski koncepti
  - Indikator za vsak pomen
  - Zgledi, kolokacije
  - Označevanje pomenov s semantičnimi tipi (ontologija SLONEST)



Načrti za naprej

- Zagotovitev nadaljnjega financiranja
- Izdelava novega slovarskega urejevalnika (pohitritev dela)
- Vključitev novih metod, ki izhajajo iz znanstvenega raziskovanja sodelavcev CJVT:
  - Spremljanje aktualne rabe jezika (pomenski premiki ipd.)
  - Avtomatsko prevajanje kot pomoč leksikografom
  - Avtomatsko zaznavanje pomenskih odtenkov (ang. word sense induction)
- Izboljšava uporabniškega vmesnika (na podlagi uporabniških raziskav)
- Povezovanje z drugimi jezikovnimi viri (projekt ELEXIS)

Zahvale

Infrastrukturni program št. I0-0022 (Mreža raziskovalnih infrastrukturnih centrov Univerze v Ljubljani) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.