Razvoj slovenščine v digitalnem okolju

7. december 2022 (Simon Krek)

www.eu-skladi.si

www.slovenščina.eu

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI STRUKTURNI IN INVESTICIJSKI SKLADI
NALOŽBA V VAŠO PRIHODNOST

Osnovni podatki o projektu

- Trajanje: maj 2020 – februar 2023
- Višina sredstev: 4.000.000 EUR
- Financer: Ministrstvo za kulturo + ESRR
- Izvajalec: konzorcij (12 partnerjev)
  - UL, UM, UNG | IJS, ZRC SAZU, INZ | PS, STA | Aikwit, Alpineon, Amebis, Vitasis
- Koordinator: Univerza v Ljubljani
  - Center za jezikovne vire in tehnologije UL
  - Sodeluje 6 fakultet: FRI, FF, FE, FDV, PEF, FU
- Spletna stran: **slovenščina.eu**

Univerza v Ljubljani
Univerza v Mariboru
Institut "Jožef Stefan" Ljubljana, Slovenija
alpineon
AMEBIS
AI KW IT
ZRC SAZU
INŠTITUT ZA NOVEJŠO ZGODOVINO INSTITUTE OF CONTEMPORARY HISTORY
VITASIS
Pošta Slovenije
sta

www.slovenščina.eu

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI STRUKTURNI IN INVESTICIJSKI SKLADI
NALOŽBA V VAŠO PRIHODNOST

Nekaj statističnih podatkov

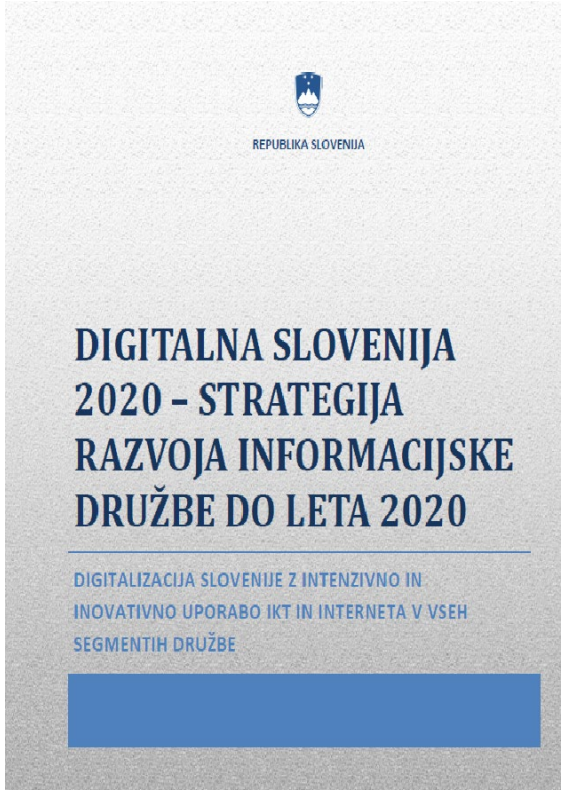
- Zaposlenih na projektu: **120** sodelavcev (R in STS)
- Delali smo **202.863,91** ur ali približno **100 človek-let**
  - 25.357,99 delovnih dni (8 urni delovnik)
- Zunanji izvajalci in drugi stroški: 133 računov
  - Največja nakupa
    - odkup avtorskih pravic (Avtorska agencija za Slovenijo): 195.200,00 EUR
    - strežnik NVIDIA DGX A100: 151.142,14 EUR
- Stroški od 1. 5. 2020 do 30. 11. 2022: €**3.289.519,05**
  - dobrih 100.000,00 eur mesečno

Cilj projekta

- 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**.
- Končni izdelki bodo na uporabniško prijazen način pomagali pri sporazumevanju, sodelovanju, poslovanju, izmenjavi znanja in udeleževanju v družabnih in političnih razpravah, in prispevali k premagovanju jezikovnih meja.
- **Od kod izhaja potreba po projektu?**

Razpisna dokumentacija: na podlagi …

- DIGITALNA SLOVENIJA 2020 – Strategija razvoja informacijske družbe do leta 2020 (marec 2016)
  - Uporaba slovenskega jezika in ohranjanje kulturne identitete
  - UKREP: Oblikovanje institucionalnega okvira za **načrten in sistematičen dolgoročen razvoj jezikovnih tehnologij, virov in orodij za slovenski jezik** ter izdelovanje le-teh.
- Resolucija o raziskovalni in inovacijski strategiji Slovenije 2011–2020 (ReRIS11-20)



Vendar tudi …

- Resolucija o Nacionalnem programu za jezikovno politiko 2014–2018 (ReNPJP14–18)
- Svet za spremljanje razvoja jezikovnih virov in tehnologij (2017-2018)
  - Vlada RS je 23. marca 2017 ustanovila Svet za spremljanje razvoja jezikovnih virov in tehnologij, **koordinacijsko telo** za podporo celovitim rešitvam **na področju digitalizacije slovenskega jezika**.
    - Pod vodstvom **ministra za kulturo** Antona Peršaka bo skrbel za razvoj digitalizacije slovenskega jezika ter strateške usmeritve na področju razvoja jezikovnih virov in tehnologij.
- Prioritetna področja

Prioritetna področja (delovni sklopi RSDO)

1. Vzdrževanje in nadgradnja korpusov (jezikovni viri)
2. Govorne tehnologije (razpoznava govora)
3. Semantični viri in tehnologije (razumevanje naravnega jezika)
4. Strojno prevajanje
5. Terminološki portal
6. Vzdrževanje infrastrukturnega centra (CLARIN.SI)
7. Koordinacija in informiranje

Evropska unija, jezikovne tehnologije in slovenščina

- Projekt: European Language Equality (ELE)
  - 52 ustanov iz 32 držav
  - Slovenija: Institut “Jožef Stefan”
  - *Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap*
- Resolucija Evropskega parlamenta (2018)
  - *Language equality in the digital age*

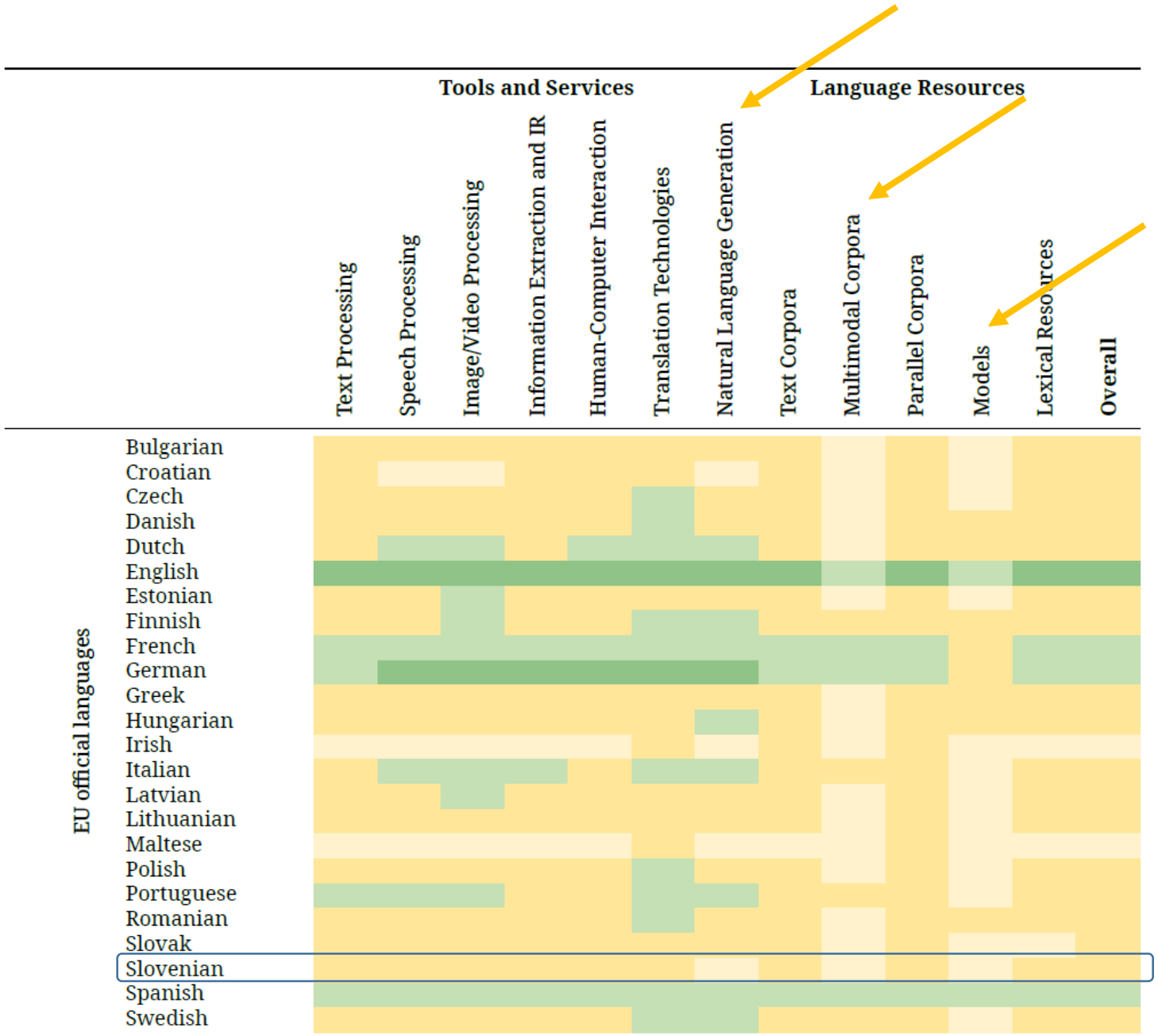
European Language Equality dashboard (pregled stanja)

Jezikovne tehnologije (tehnološki dejavniki)

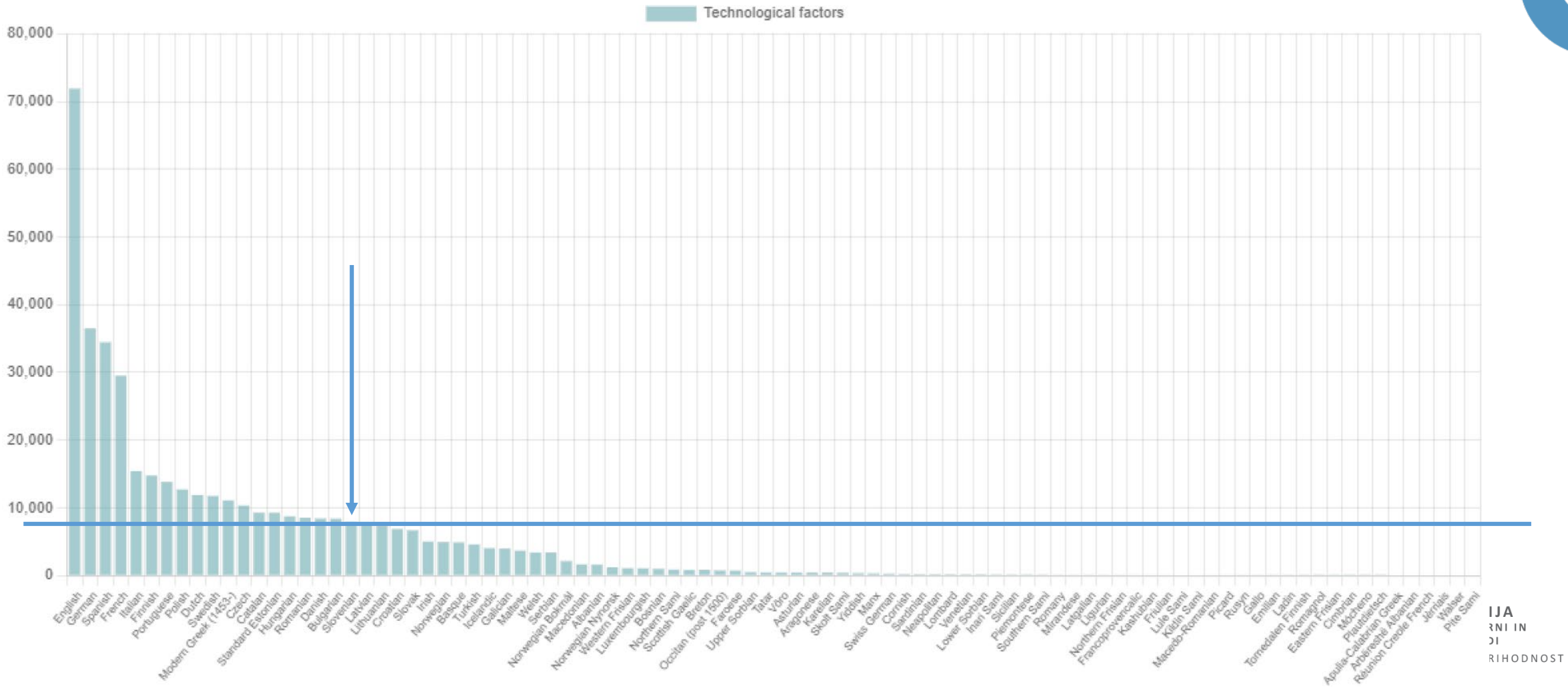
- Analiza besedil (Text Analysis)
  - Prepoznavanje in označevanje jeziko(slo)vnih informacij, ki jih vsebujejo besedila v naravnem jeziku, npr. prepoznavanje besed, zvez, stavkov, oblikoslovnih ali besedotvornih lastnosti, skladenjskih ali semantičnih vlog itd.
- Govorne tehnologije (Speech processing)
  - Omogočajo glasovno sporazumevanje z elektronskimi napravami: razpoznava/sinteza govora, prepoznavanje govorcev itd.
- Strojno prevajanje (Machine Translation)
  - Avtomatizirano prevajanje iz enega naravnega jezika v drugega.
- Luščenje informacij in informacijsko poizvedovanje (Information Extraction and Information Retrieval)
  - Luščenje strukturiranih informacij iz nestrukturiranih besedil: prepoznavanje imenskih entitet, luščenje relacij (relation extraction) itd.
- Tvorjenje naravnega jezika (Natural Language Generation)
  - Avtomatizirano tvorjenje besedil: avtomatsko povzemanje, poenostavljanje besedil, parafraziranje itd.
- Komunikacija med človekom in strojem (Human-Computer Interaction)
  - Sistemi za komunikacijo z računalniki v naravnem jeziku (besedilno, govorno ali neverbalno): klepetalniki itd.

Jezikovni podatki (tehnološki dejavniki)

- Besedilni korpusi
  - Standardni pisni jezik, spletni korpusi, korpusi družabnih omrežij, korpusi akademskega jezika, parlamentarni korpusi, zgodovinski in drugi korpusi
- Multimodalni korpusi (avdio, video)
  - Govorni korpusi, dialoški korpusi, multimodalni (video) korpusi
- Dvo- ali večjezični / vzporedni korpusi
- Leksikalni oz. konceptualni viri
  - Sloleks (oblikoslovni), sloWNet (WordNet), eno- in večjezični slovarji
- Jezikovni modeli in (formalne) slovnice
  - Modeli (fastText, RoBERTa)

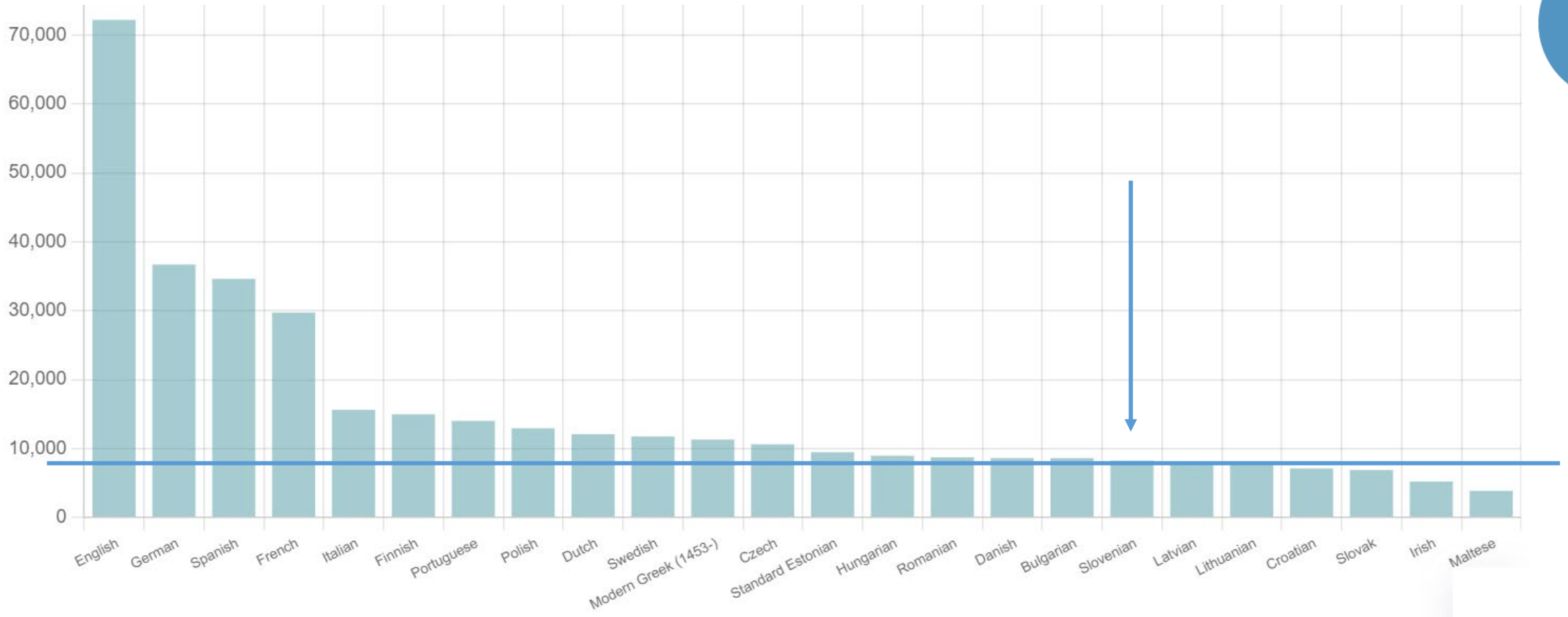


Tehnološki dejavniki



IJA
RNI IN
OI
RIHODNOST

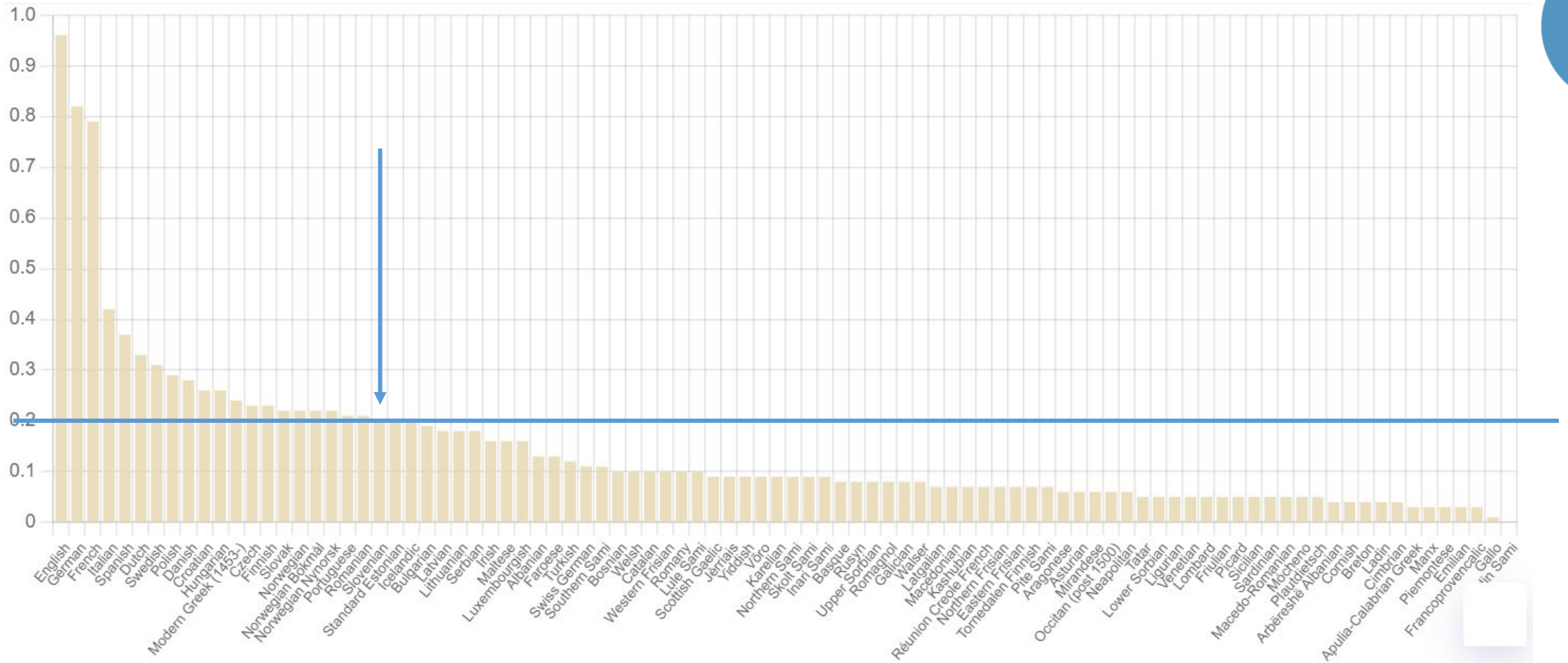
Tehnološki dejavniki – uradni jeziki EU



Kontekstualni dejavniki

- Gospodarstvo: obseg gospodarstva, obseg IKT sektorja
- Izobraževanje: število študentov JT/jezika, vključenost v izobraževanje
- Industrija: število JT podjetij
- Zakonodaja: pravni status in zaščita jezika
- Splet: Wikipedija
- RRI: sposobnost za inovacije, število (znanstvenih) člankov
- Družba: velikost jezikovne skupnosti, uporaba družbenih omrežij
- Tehnologija: dostop do interneta, digitalna povezanost

Kontekstualni dejavniki

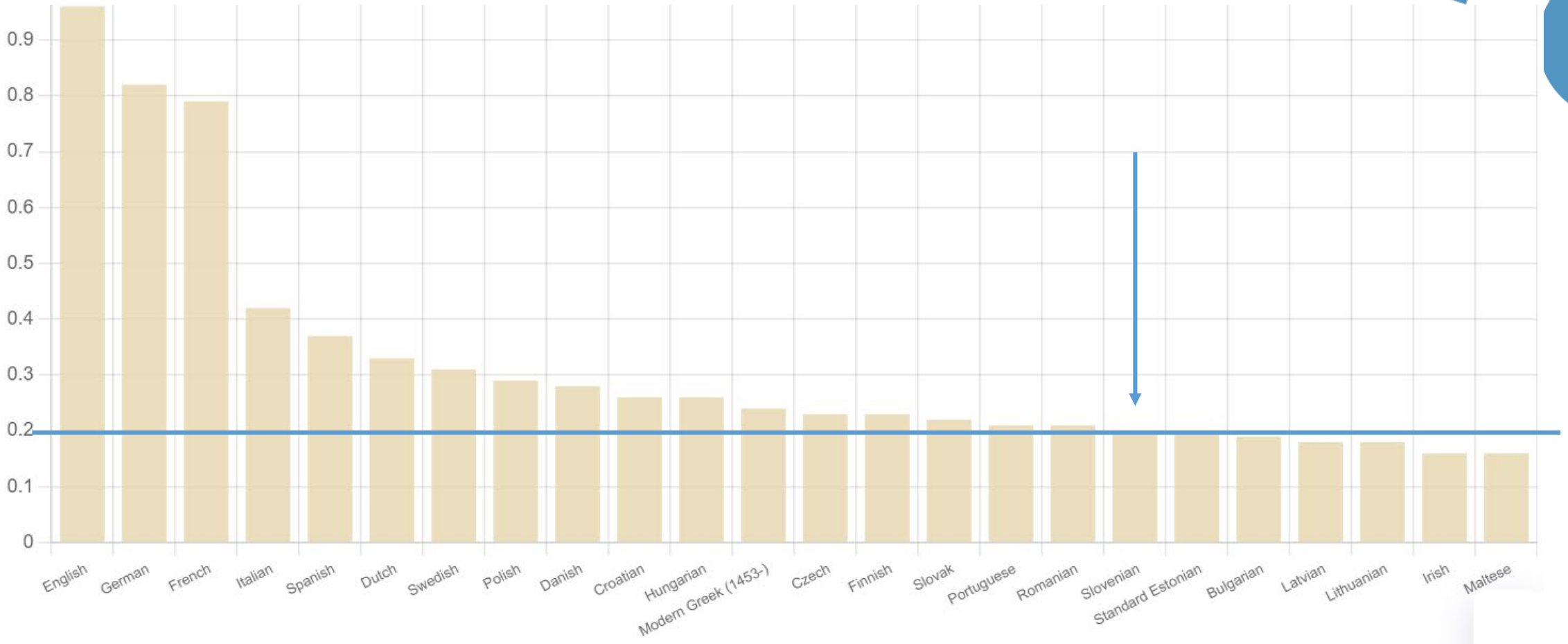


www.slovenščina.eu

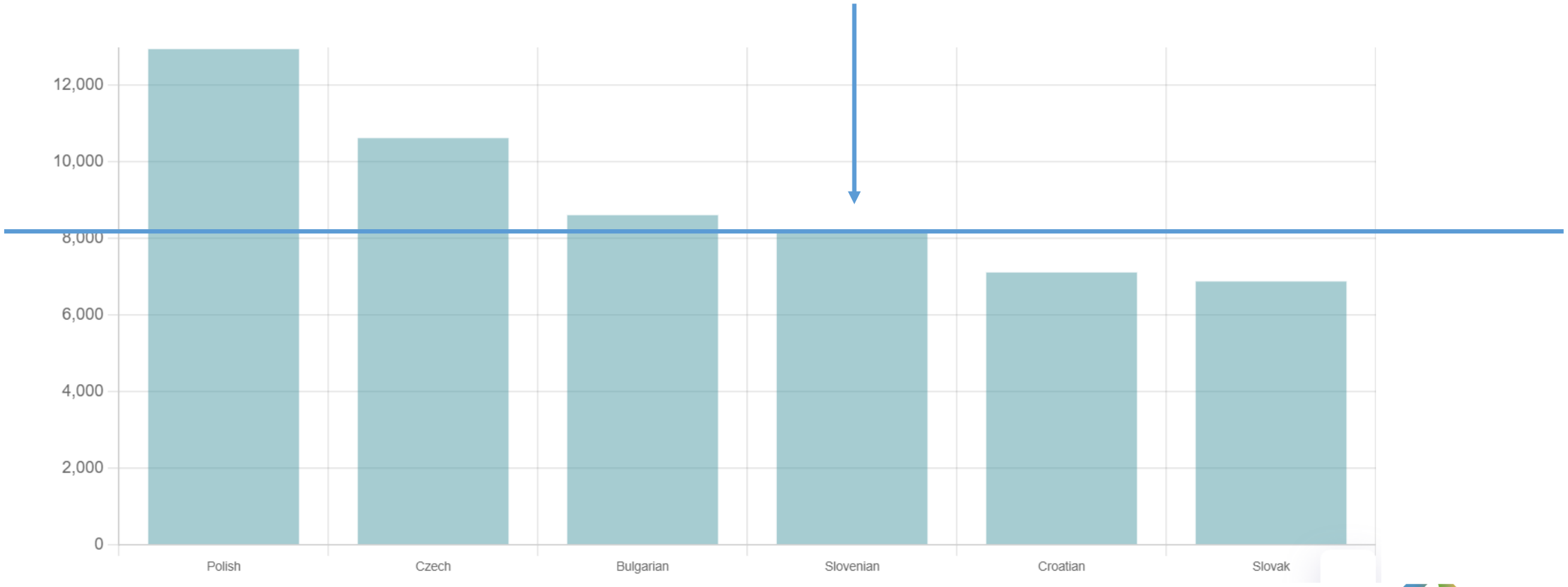
REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI STRUKTURNI IN INVESTICIJSKI SKLADI
NALOŽBA V VAŠO PRIHODNOST

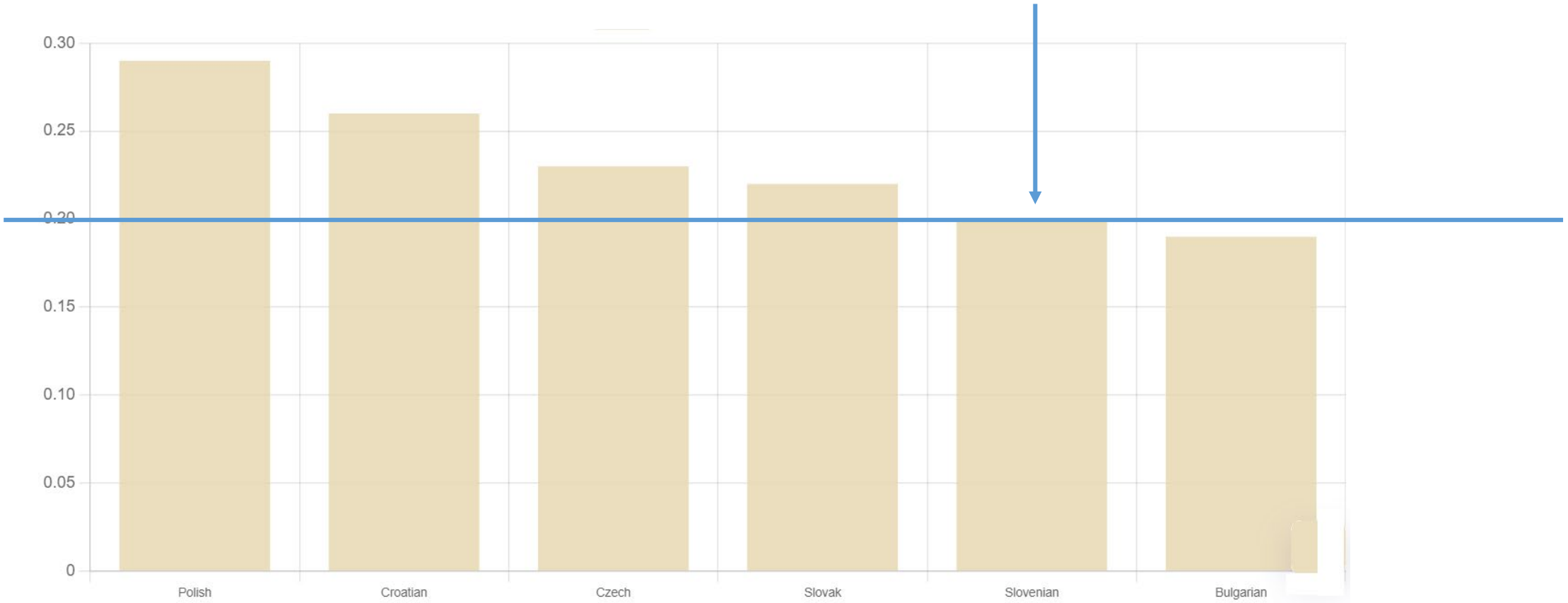
Kontekstualni dejavniki – uradni jeziki EU



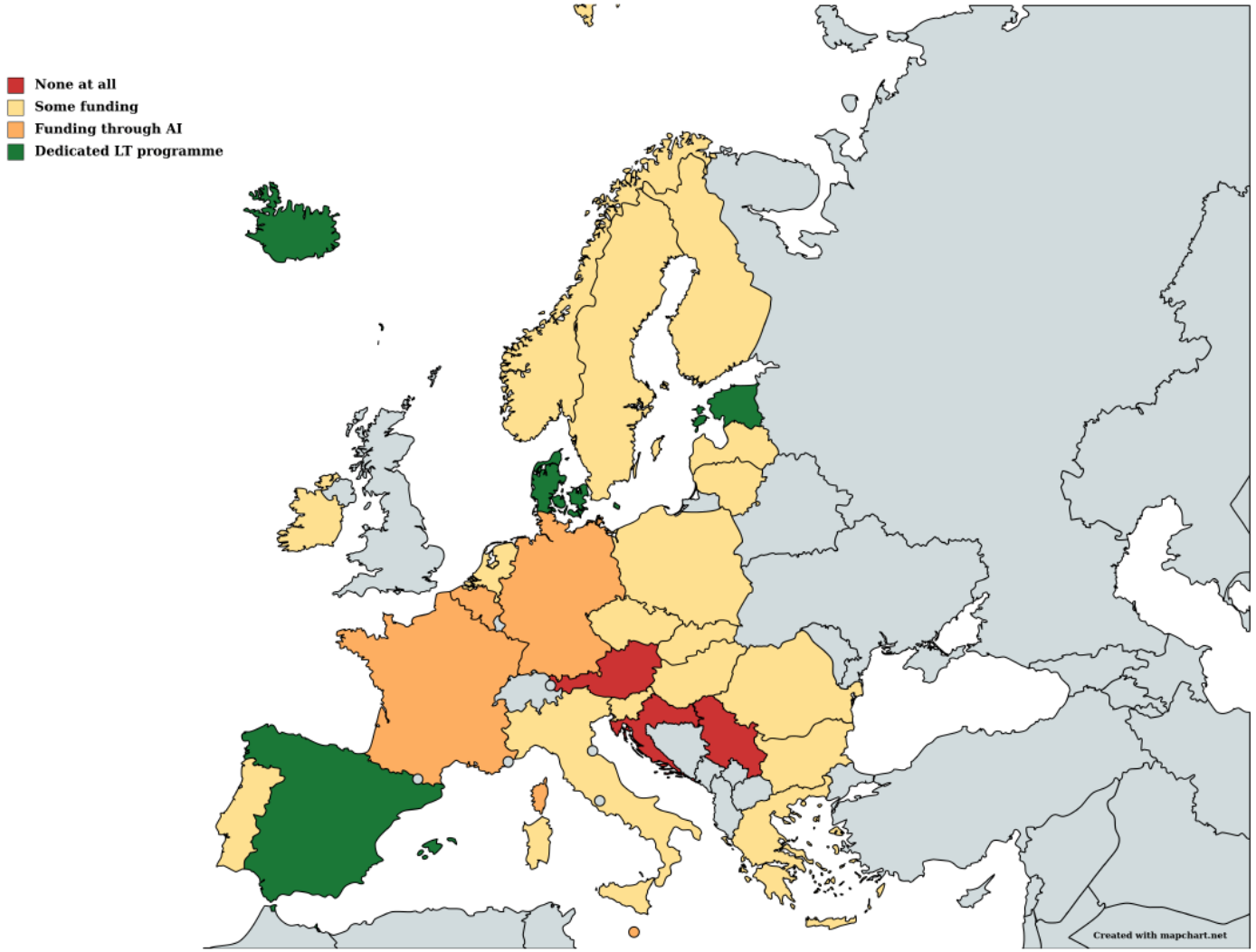
Slovanski jeziki – uradni jeziki EU (tehnološki)



Slovanski jeziki – uradni jeziki EU (kontekstualni)



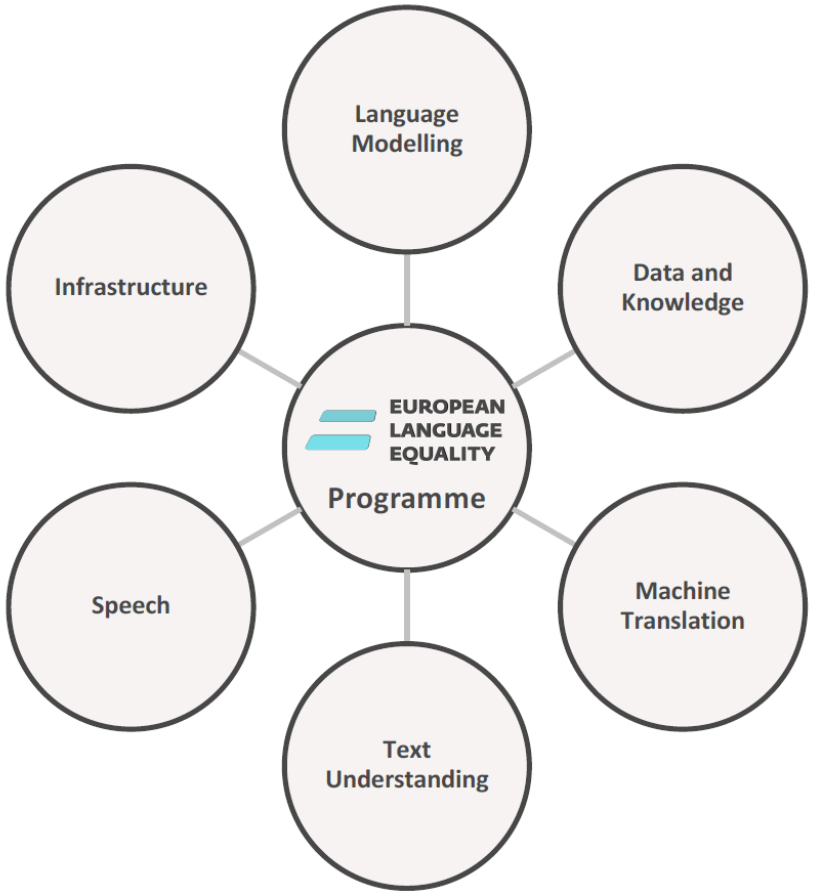
Pregled financiranja jezikovnih tehnologij v Evropi (strategije UI)



- Zelena: države z rednim programom financiranja JT
- Oranžna: države, ki zagotavljajo financiranje JT (preko UI)
- Rumena: države, pri katerih so v strategiji UI omenjene JT
- Rdeča – države brez strategije UI ali s strategijo brez omembe JT

In to je pomembno, ker…? (EU do 2030)

- Jezikovna infrastruktura
- Jezikovni modeli
- Jezikovni podatki in znanje
- Strojno prevajanje
- (Strojno) razumevanje besedila
- Govor (razpoznava in sinteza)

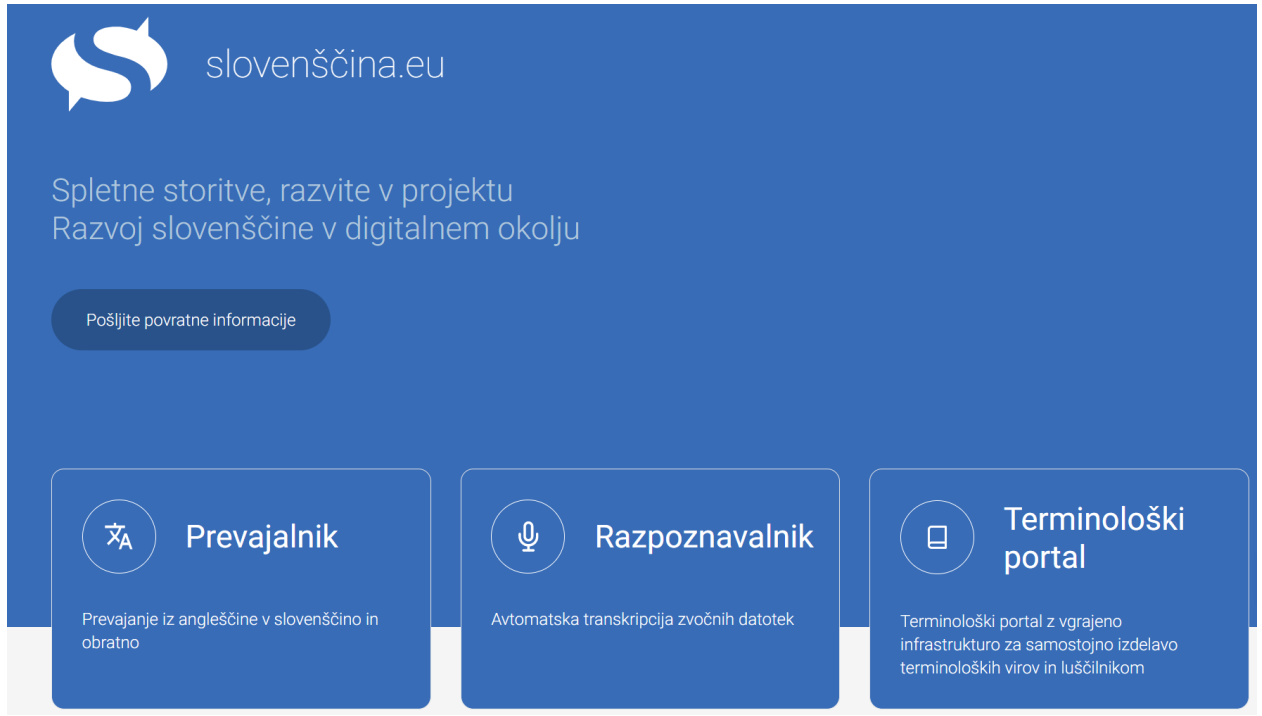


Kaj storiti (po RSDO)?

- **Vzdrževanje**: Načrt razvoja raziskovalne infrastrukture 2030 (NRRI 2030)
  - za CLARIN.SI za obdobje do leta 2030 namenja znesek v višini do 250.000 EUR letno letno
  - za **osnovno vzdrževanje** razvitih virov in orodij RSDO, torej brez nadaljnjega razvoja, bilo potrebno zagotoviti financiranje najmanj v višini teh sredstev
- **Načrtovanje**: Resolucija o nacionalnem programu za jezikovno politiko 2021–2025 (ReNPJP21–25)
  - Ukrepi: **ustanovitev vladnega delovnega telesa** oziroma sveta za usmerjanje in spremljanje razvoja jezikovnih virov in tehnologij, za podpiranje celovitih rešitev na področju digitalizacije slovenskega jezika ter za skrb za slovenski jezik;
- **Razvoj**: Nacionalni program spodbujanja razvoja in uporabe UI v Republiki Sloveniji do leta 2025 (NpUI)
  - **nadaljevanje** projekta Razvoj slovenščine v digitalnem okolju **kot program**
  - področje 3: Jezikovne tehnologije, kulturna identiteta in raziskovalna umetnost

Spletne strani: slovenščina.eu

- Demonstracijski portal
- Projekt
  - https://rsdo.slovenscina.eu
- Kazalniki
  - https://rsdo.slovenscina.eu/kazalniki
- Povratne informacije
  - https://rsdo.slovenscina.eu/povratne-informacije
  - info@slovenscina.eu



Projektni sklopi in koordinatorji

1. Vzdrževanje in nadgradnja korpusov (jezikovni viri)
   - Marko Robnik Šikonja (FRI UL)
   - Špela Arhar Holdt (FF/FRI UL)
2. Govorne tehnologije (razpoznava govora)
   - Marko Bajec (FRI UL)
   - Simon Dobrišek (FE UL)
   - Darinka Verdonik (FERI UM)
3. Semantični viri in tehnologije (razumevanje naravnega jezika)
   - Slavko Žitnik (FRI UL)
   - Simon Krek (FF/FRI UL)
4. Strojno prevajanje
   - Iztok Lebar Bajec (FRI UL)
   - Andraž Repar (Aikwit/IJS)
5. Terminološki portal
   - Miro Romih (Amebis)
   - Mateja Jemec Tomazin (ZRC SAZU)
6. Vzdrževanje infrastrukturnega centra (CLARIN.SI)
   - Tomaž Erjavec (IJS/ZRC SAZU)
7. Koordinacija in informiranje
   - Simon Krek (IJS/CJVT UL)

Odkup avtorskih pravic

- Deklaracije o **dostopu do** javno financiranih **raziskovalnih podatkov** (OECD)
- Priporočila Komisije z dne 17. julija 2012 o **dostopu do znanstvenih informacij** in njihovem arhiviranju
- Direktive 2013/37/EU Evropskega parlamenta in Sveta z dne 26. junija 2013 o spremembi Direktive 2003/98/ES o **ponovni uporabi informacij** javnega sektorja

- **ODPRTI PODATKI!**

Odkup avtorskih pravic: spoznanja

- Slabo stanje glede dokumentiranosti in zavedanja, kdo je lastnik avtorskih pravic za določeno avtorsko delo
- (Pomembnejših) avtorskih del, ki jih ni na listi, ni bilo mogoče odkupiti (predvsem SSKJ in Enciklopedija Slovenije)
- Treba bo razmisliti o alternativnih poteh, predvsem v povezavi z novo direktivo o avtorskih pravicah in enotnem digitalnem trgu (2019) ter spremembami v slovenski zakonodaji (2022)

Odkup avtorskih pravic: DZS

Naslov vira	Izdaja
Veliki splošni leksikon	1998
Veliki angleško-slovenski slovar Oxford-DZS (slovenski del)	2005 2006
Priročni angleško-slovenski slovar (slovenski del)	2010
Mali angleško-slovenski in slovensko-angleški slovar	2006
Veliki italijansko-slovenski slovar	2001
Mali srbsko-slovenski & slovensko-srbski slovar	2005

Odkup avtorskih pravic: individualni avtorji / lastniki avtorskih pravic

Slovensko-angleški slovar	1990	Anton Grad, Henry Leeming
Veliki slovensko-nemški slovar	1995	Božidar Debenjak, Primož Debenjak
Slovensko-francoski slovar	1990	Viktor Jesenik, Narcis Dembskij
Francosko-slovenski slovar	1990	Anton Grad
Španski-slovenski slovar Slovensko-španski slovar	1984 1979	Anton Grad
Slovensko latinski slovar	1973	Fran Bradač
Rusko-slovenski slovar	1986	Janko Pretnar
Slovensko-poljski slovar	1996	Tone Pretnar, Božena Ostromecka

Odkup avtorskih pravic: individualni avtorji / lastniki avtorskih pravic

Češko-slovenski in slovensko-češki slovar	1995	Ružena Škerlj
Veliki slovensko-italijanski slovar	2006	Sergij Šlenc
Večjezični evropski slovar	2021	Anton Rupnik
Srbskohrvatsko-slovenski slovar Slovensko-srbskohrvaški slovar	1986 1989	Janko Jurančič
Slovensko-slovaški slovar	1983	Viktor Smolej
Veliki angleško-slovenski slovar	1978	Anton Grad, Božena Škrlj, Nada Vitorovič
Romunsko-slovenski in slovensko-romunski slovar	2006	Irena Santoro