EUROPEAN LANGUAGE RESOURCE COORDINATION

TRETJA DELAVNICA ELRC V SLOVENIJI

SIMON KREK

27. maj 2022

PROGRAM DELAVNICE

Organizacija

- European Language Resources Coordination (ELRC)
  - Razvoj slovenščine v digitalnem okolju (RSDO)
  - Federated eTranslation TermBank Network (FedTerm)
- European Language Grid (ELG)

Od – do	Sekcija
09:00 – 09:30	**Uvodni nagovor** *Simon Krek, Institut Jožef Stefan in Center za jezikovne vire in tehnologije, Univerza v Ljubljani*
09:30 – 09:55	**eTranslation** *Miha Žličar, Generalni direktorat za prevajanje, Evropska komisija*
09.55 – 10:20	**Novosti ELRC** *Thierry Declerck, DFKI / Koordinacija evropskih jezikovnih virov*
10.20 – 10:30	**Razprava/vprašanja**
10:30 – 10:45	**Odmor za kavo**
10:45 – 11:00	**Predstavitev strojnega prevajanja za slovenščino** *Iztok Lebar Bajec, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani*
11:00 – 11:15	**Predstavitev govornih tehnologij za slovenščino** *Marko Bajec, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani*
11:15 – 11:30	**Predstavitev semantičnih tehnologij za slovenščino** *Slavko Žitnik, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani*

Od – do	Sekcija
11:30 – 11:45	**Slovenski terminološki portal** *Miro Romih, Amebis*
11:45 – 12:00	**Projekt FedTerm CEF** *Andraž Repar, Institut Jožef Stefan*
12:00 – 12:15	**Nacionalni program za spodbujanje razvoja in uporabe umetne inteligence do leta 2025** *Samo Zorc, Služba vlade za digitalno preobrazbo*
12:15 – 12:45	**Razprava/vprašanja.** Govorniki z dopoldanske sekcije v panelu
12:45 – 13:30	**Odmor za kosilo**
	DELAVNICA ELG
13:30 – 14:00	**European Language Grid (evropska jezikovna mreža) in projekti evropske jezikovne enakosti** *Penny Labropoulou, Inštitut za obdelavo jezika in govora/R.C. "Athena", Grčija*
14:00 – 14:20	**Panel o ELG in možnih sinergijah s slovensko jezikovnotehnološko skupnostjo**
14:20 – 14:40	**Zaključek**

SLOVENSKA JEZIKOVNOTEHNOLOŠKA KRAJINA

- Nacionalni program spodbujanja razvoja in uporabe umetne inteligence v Republiki Sloveniji do leta 2025 (NpUI): **maj 2021**
  - prioritetna področja: 3. Jezikovne tehnologije, kulturna identiteta in raziskovalna umetnost
- Resolucija o nacionalnem programu za jezikovno politiko 2021–2025 (ReNPJP21–25): **junij 2021**
  - 2.3 Jezikovna opremljenost -> 2.3.6 Jezikovne tehnologije
- Načrt razvoja raziskovalne infrastrukture 2030 (NRRI 2030): **maj 2022**
  - 2.2.1 Družbene in kulturne inovacije -> 2.2.1.2 CLARIN

RESOLUCIJA O NACIONALNEM PROGRAMU ZA JEZIKOVNO POLITIKO 2021–2025

- 2.3 Jezikovna opremljenost
  - 2.3.2 Jezikovni opis
  - 2.3.3 Standardizacija
  - 2.3.4 Terminologija
  - 2.3.5 Večjezičnost
  - **2.3.6 Jezikovne tehnologije**
  - 2.3.7 Digitalizacija
  - 2.3.8 Osebe s posebnimi potrebami in prilagojenimi načini sporazumevanja

2.3.6 JEZIKOVNE TEHNOLOGIJE

- **Cilj**: Gradnja, posodabljanje in vzdrževanje temeljnih jezikovnih tehnologij za slovenščino in druge jezike, ki sodijo v okvir slovenske jezikovne politike, ter zagotavljanje njihove čim bolj proste dostopnosti
- Ukrepi:
  - redna **analiza stanja** za prepoznavanje potreb na področju izgradnje oziroma razvoja in posodabljanja jezikovnih virov in tehnologij s preučitvijo dobrih praks razvoja in financiranja posameznih virov in tehnologij (npr. presoja, za katere vrste besedil je smiselno razvijati strojnoprevajalne sisteme – splošni jezik, strokovna in znanstvena besedila);
  - oblikovanje **seznama** jezikovnih virov in tehnologij za prednostno izgradnjo oziroma razvoj in posodobitev na podlagi raziskav potreb uporabnikov;

2.3.6 JEZIKOVNE TEHNOLOGIJE

- gradnja, posodabljanje in vzdrževanje gradivskih virov, zlasti **jezikovnih korpusov**;
- razvijanje, nadgradnja, posodabljanje in vzdrževanje **govornih tehnologij** za slovenščino in druge jezike, ki sodijo v okvir slovenske jezikovne politike;
- prilagajanje in uporaba **semantičnih virov in tehnologij** globokih nevronskih mrež za semantično podporo jezikovnotehnološkim nalogam s področja slovenščine in drugih jezikov, ki sodijo v okvir slovenske jezikovne politike;
- razvijanje **strojnega prevajanja** za potrebe slovenščine in drugih jezikov, ki sodijo v okvir slovenske jezikovne politike;
- razvoj **črkovalnika** in pravopisno-slovničnega **pregledovalnika**;
- vključevanje v vzpostavljeno hrambno-distribucijsko **infrastrukturo**.

2.3.6 JEZIKOVNE TEHNOLOGIJE (SREDSTVA, NOSILCI)

- Ocenjena okvirna sredstva: **4.500.000** evrov. (skupaj pribl. 15M €)
- Predvideni učinki:
  - opremljenost jezikovne skupnosti s sodobnimi jezikovnimi tehnologijami, omogočenje sistematičnega jezikovnega raziskovanja s sodobnimi tehnologijami in na podlagi kakovostnih podatkov, omogočenje pomembnih in sodobnih jezikovnih informacij široki javnosti, omogočenje gradnje novih jezikovnih virov in tehnologij, omogočenje proste dostopnosti.
- Nosilci: MJU, MK, MIZŠ (ARRS)

NAČRT RAZVOJA RAZISKOVALNE INFRASTRUKTURE 2030 (NRRI)

- 2.2 Implementirani projekti – landmarks
- 2.2.1 Družbene in kulturne inovacije
- 2.2.1.2 CLARIN
  - Common Language Resources and Technology Infrastructure / Infrastruktura za skupne jezikovne vire in tehnologijo
    - Projekt je bil uvrščen na prednostni seznam **Roadmap ESFRI 2006** in na nacionalni prednostni seznam mednarodnih projektov v **NRRI 2011**. V pripravljalni fazi CLARIN ERIC, ki je trajala do odločitve Evropske Komisije o ustanovitvi CLARIN ERIC 29. 2. 2012, je Slovenija sodelovala kot opazovalka, načrtovano pa je bilo, da se vanj vključi kot polnopravna članica čim prej po ustanovitvi. Slovenija je poslala pristopno pismo 29. 4. 2015 in v **maju 2015** postala polnopravna članica **CLARIN ERIC**. Za izvajanje nacionalnih obveznosti CLARIN je bil junija 2014 ustanovljen **konzorcij CLARIN.SI**, z 12 partnerji in s sedežem na **Institutu Jožef Stefan** (IJS).

CLARIN.SI (SREDSTVA)

- Finančni vidik
  - Sofinanciranje CLARIN.SI se je začelo leta 2013. Zadnja vrednost letnega zneska znaša **100.000 evrov**, članarina za CLARIN ERIC pa okoli **14.000 evrov** letno. CLARIN je tudi pridobil sredstva Evropske kohezijske politike v okviru RI-SI-CLARIN v višini **466.000 evrov**. Dosedanja vlaganja v nacionalno raziskovalno infrastrukturo znašajo skupaj okoli **1,1 mio evrov**.
  - Za obdobje do leta 2030 je za dolgoročno vzdržno delovanje in raziskovalno-razvojno delo CLARIN.SI ocenjen znesek v višini do **250.000 evrov letno**.

RAZVOJ SLOVENŠČINE V DIGITALNEM OKOLJU (RSDO)

- 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: **slovenscina.eu**

CILJ IN DELOVNI 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**.
  - Vzdrževanje in nadgradnja korpusov (jezikovni viri)
  - Govorne tehnologije (razpoznava / sinteza govora)
  - Semantični viri in tehnologije (razumevanje naravnega jezika)
  - Strojno prevajanje
  - Terminološki portal
  - Vzdrževanje infrastrukturnega centra (CLARIN.SI)
  - Koordinacija in informiranje

REZULTATI PROJEKTA

- V okviru razpisa razviti jezikovni viri morajo biti v repozitoriju CLARIN.SI objavljeni najmanj tri mesece pred iztekom operacije oziroma projekta (**november 2022**), jezikovnotehnološki izdelki pa najmanj dva meseca pred iztekom (**december 2022**).
- Upravičenec mora na ustreznem **promocijskem dogodku** zainteresirani javnosti predstaviti rezultate projekta in vzpostaviti **sistem za sprejemanje pripomb**. Čas med objavo rezultatov in iztekom operacije mora upravičenec nameniti evalvaciji in izboljšavam glede na odkrite pomanjkljivosti ter smiselne pripombe zainteresirane javnosti.

EUROPEAN LANGUAGE EQUALITY PROJECT

- Poročilo o slovenščini: **februar 2022**
  - 4 Language Technology for Slovene
    - 4.1 Language Data
    - 4.2 Language Technologies and Tools
    - 4.3 Projects, Initiatives, Stakeholders
  - 5 Cross-Language Comparison
    - 5.1 Dimensions and Types of Resources
    - 5.2 Levels of Technology Support
    - 5.3 European Language Grid as Ground Truth
    - 5.4 Results and Findings


EUROPEAN LANGUAGE EQUALITY

D1.31

Report on the Slovenian Language

Author	Simon Krek
Dissemination level	Public
Date	28-02-2022

JEZIKOVNI PODATKI (POROČILO ELE)

- 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)

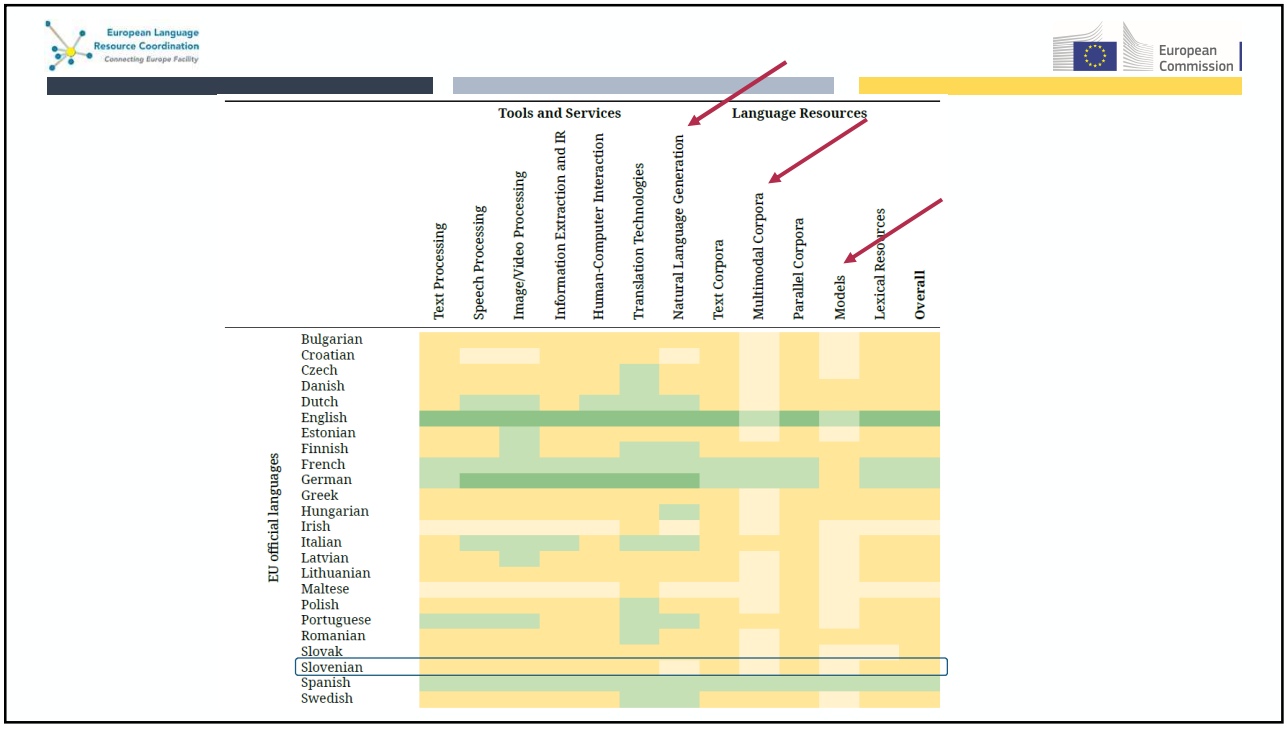
JEZIKOVNE TEHNOLOGIJE (POROČILO ELE)

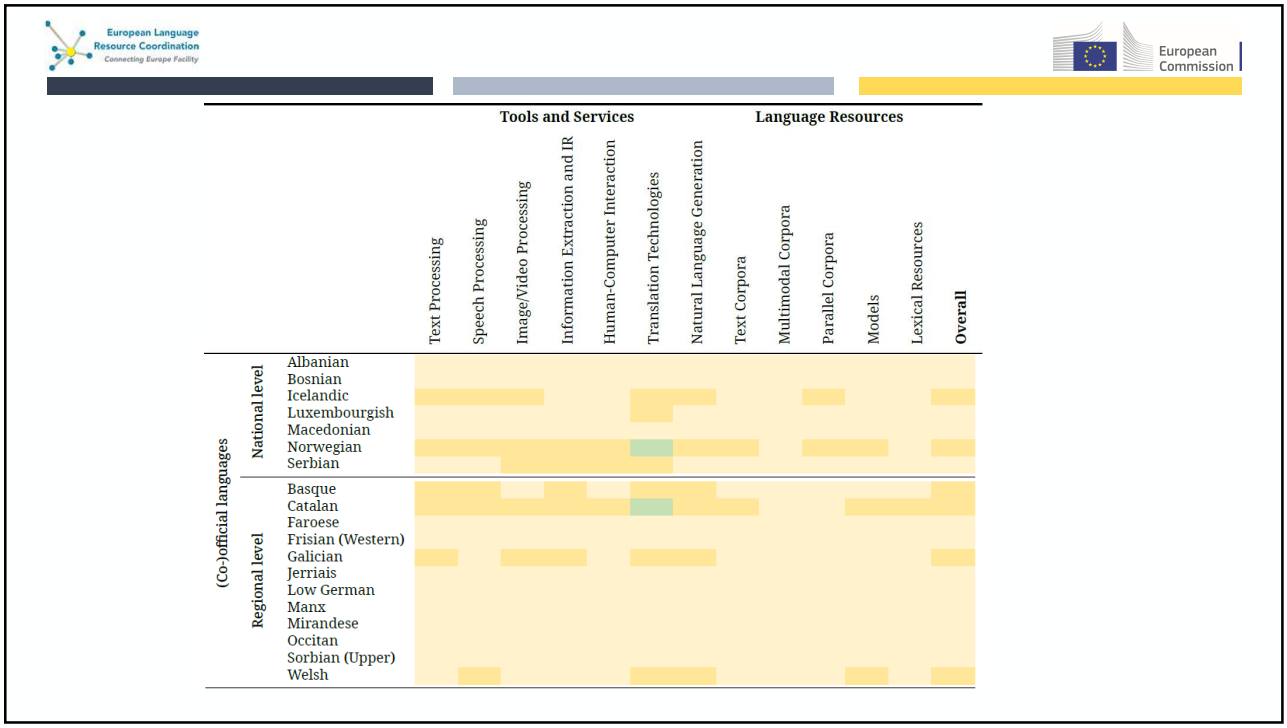
- Analiza besedil (Text Analysis)
  - Prepoznavanje in označevanje jeziko(slo)vnih informacij, ki jih vsebujejo besedila v naravnem jeziku, npr. prepoznavanje besed, zvez, stavkov, oblikoslovnih ali besedotvornih lastnosti, skladenjskih ali semantičnih vlog itd.
- Govorne tehnologije (Speech processing)
  - 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.

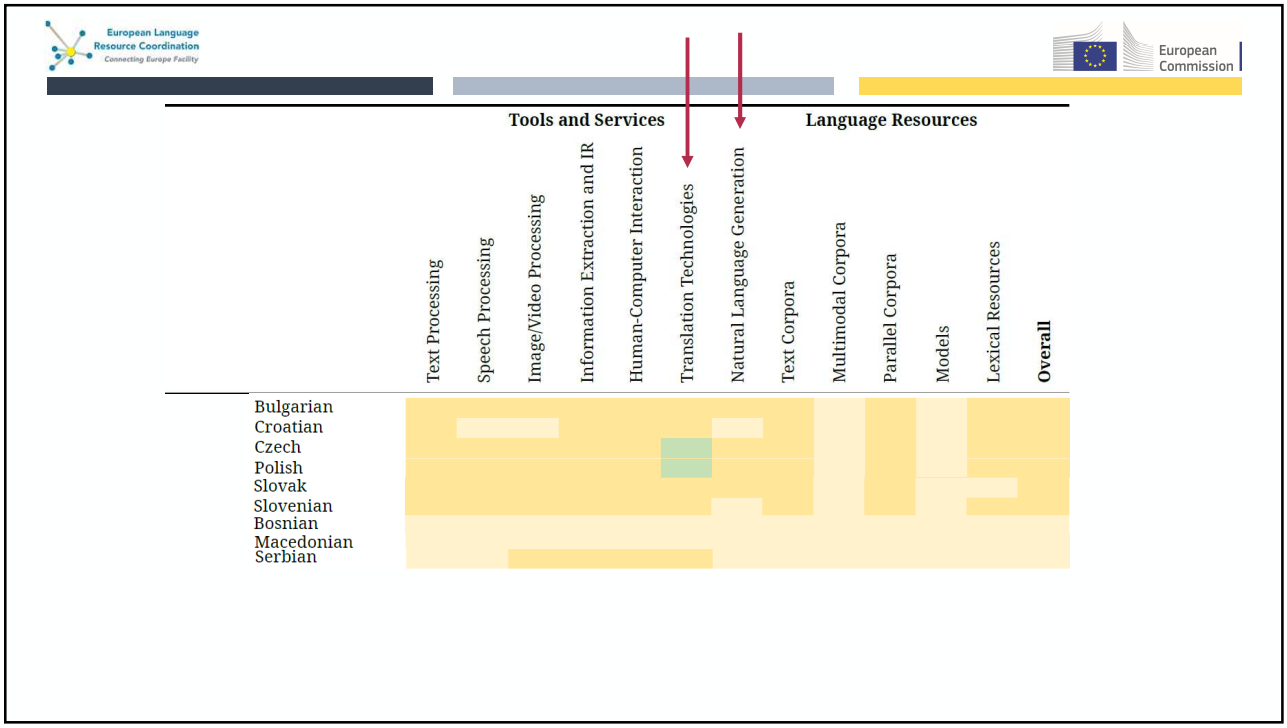
RAVEN TEHNOLOŠKE PODPORE (PRIMERJAVA)

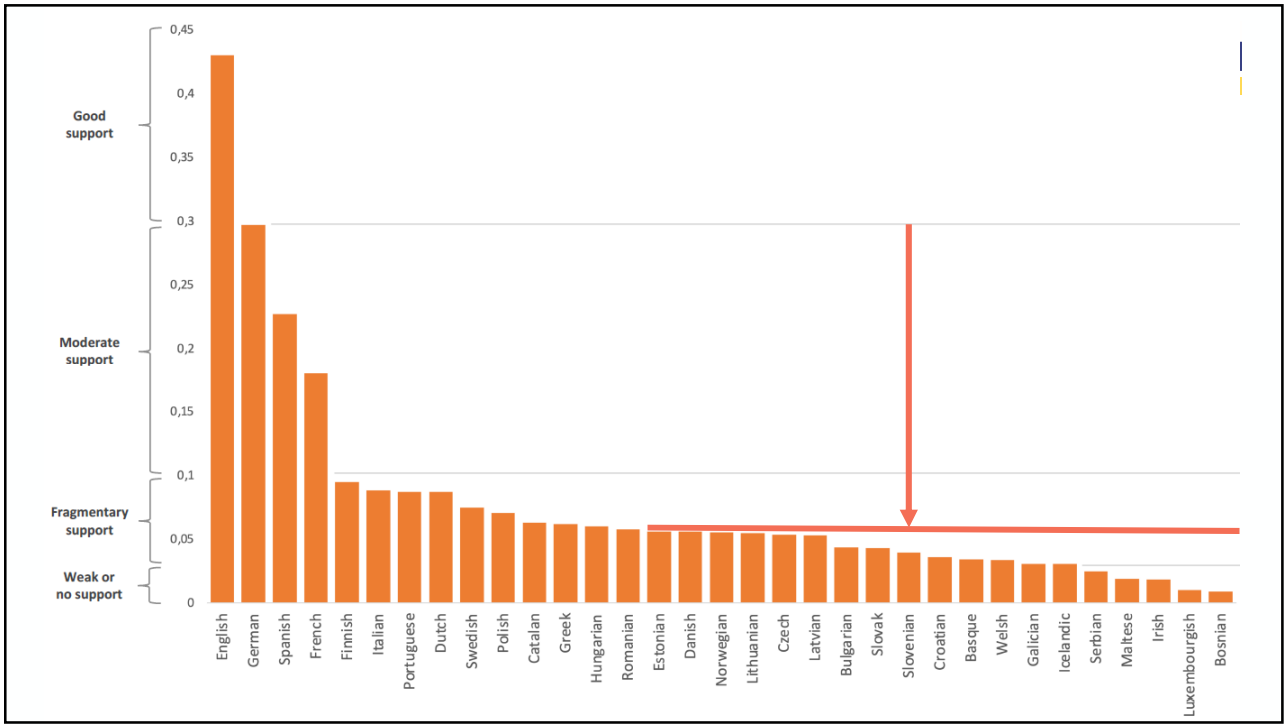
Primerjava glede na katalog European Language Grid:

1. Šibka podpora ali brez podpore
   - jezik je prisoten v manj kot <3% virov ELG enakega tipa
2. Fragmentarna podpora
   - jezik je prisoten med ≥3% in <10% virov ELG enakega tipa
3. Zmerna podpora
   - jezik je prisoten med ≥10% in <30% virov ELG enakega tipa
4. Dobra podpora
   - jezik je prisoten v več kot ≥30% virov ELG enakega tipa


European Language Resource Coordination
Connecting Europe Facility
European Commission
Tools and Services
Language Resources
Text Processing
Speech Processing
Image/Video Processing
Information Extraction and IR
Human-Computer Interaction
Translation Technologies
Natural Language Generation
Text Corpora
Multimodal Corpora
Parallel Corpora
Models
Lexical Resources
Overall
EU official languages
Bulgarian
Croatian
Czech
Danish
Dutch
English
Estonian
Finnish
French
German
Greek
Hungarian
Irish
Italian
Latvian
Lithuanian
Maltese
Polish
Portuguese
Romanian
Slovak
Slovenian
Spanish
Swedish


European Language Resource Coordination
Connecting Europe Facility
European Commission
Tools and Services
Language Resources
Text Processing
Speech Processing
Image/Video Processing
Information Extraction and IR
Human-Computer Interaction
Translation Technologies
Natural Language Generation
Text Corpora
Multimodal Corpora
Parallel Corpora
Models
Lexical Resources
Overall
(Co-)official languages
National level
Albanian
Bosnian
Icelandic
Luxembourgish
Macedonian
Norwegian
Serbian
Regional level
Basque
Catalan
Faroese
Frisian (Western)
Galician
Jerriais
Low German
Manx
Mirandese
Occitan
Sorbian (Upper)
Welsh


European Language Resource Coordination
Connecting Europe Facility
European Commission
Tools and Services
Language Resources
Text Processing
Speech Processing
Image/Video Processing
Information Extraction and IR
Human-Computer Interaction
Translation Technologies
Natural Language Generation
Text Corpora
Multimodal Corpora
Parallel Corpora
Models
Lexical Resources
Overall
Bulgarian
Croatian
Czech
Polish
Slovak
Slovenian
Bosnian
Macedonian
Serbian

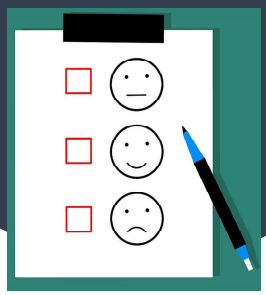

Good support
Moderate support
Fragmentary support
Weak or no support
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
English
German
Spanish
French
Finnish
Italian
Portuguese
Dutch
Swedish
Polish
Catalan
Greek
Hungarian
Romanian
Estonian
Danish
Norwegian
Lithuanian
Czech
Latvian
Bulgarian
Slovak
Slovenian
Croatian
Basque
Welsh
Galician
Icelandic
Serbian
Maltese
Irish
Luxembourgish
Bosnian

European Language Resource Coordination
Connecting Europe Facility

European Commission

PREDEN ZAČNEMO...

Povratne informacije!



Potrdilo o udeležbi?

European Language Resource Coordination
Connecting Europe Facility

European Commission

PA ZAČNIMO!