European Language Resource Coordination
Connecting Europe Facility

Radionica ELRC-a u Hrvatskoj

Zagreb, 2019-02-12

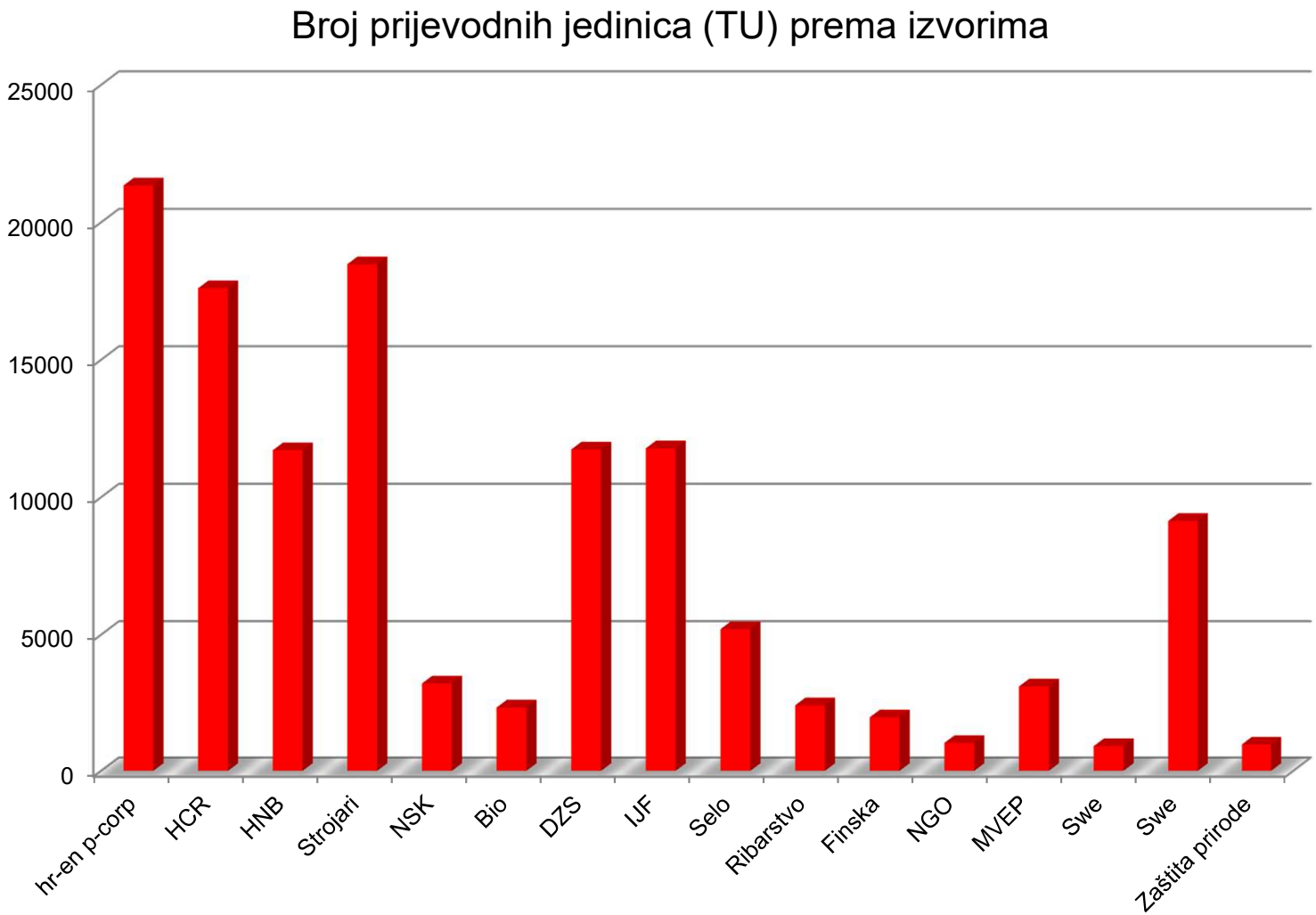
ELRC u Hrvatskoj

Marko Tadić
Sveučilište u Zagrebu, Filozofski fakultet
Mladen Stojak
Ciklopea

Prikupljeni jezični resursi za hrvatski jezik 1

- ELRC-SHARE
  - repozitorij jezičnih resursa
  - https://www.elrc-share.eu
  - 21 jezični resurs s uključenim hrvatskim jezikom
  - Vrste resursa
    - 18 dvo- ili višejezičnih korpusa
    - 1 jednojezični korpus (hr)
    - 2 terminološke zbirke
  - Ukupna okvirna statistika
    - Prijevodnih segmenata: 122,896
    - Pojavnica hr: 100.000.000 (Narodne novine)
    - Licencije: otvorena PSI: 16; javna domena: 3; razne CC inačice: 2
- Nedovoljno!

Prikupljeni jezični resursi za hrvatski jezik 2



Prikupljeni jezični resursi za hrvatski jezik 3

- Bilingual hr-en parallel corpus from Croatian Mine Action
  - Jezici: hr/en, format: TMX, veličina: 17,601 prijevodnih jedinica (TU)
- Bilingual hr-en parallel corpus from Croatian National Bank
  - Jezici: hr/en, format: TMX, veličina: 11,707 TU
- Croatian-English corpus with statistical reports and studies from the Croatian Bureau of Statistics
  - Jezici: hr/en, format: TMX, veličina: 11,737 TU
- Croatian-English parallel corpus from the Ministry of Foreign and European Affairs
  - Jezici: hr/en, format: TMX, veličina: 3,102 TU (**???**)
- Croatian-English parallel corpus from the website of the Embassy of Finland, Zagreb
  - Jezici: hr/en, format: TMX, veličina: 1,966 TU
- Parallel texts from Swedish Social Security Authority
  - Jezici: hr/en, format: TMX, veličina: 9,139 TU

European Language Resource Coordination
Connecting Europe Facility

Uključeni dionici

Nacionalne sidrišne točke

- Nacionalna sidrišna točka (NST) za javne usluge (još ju uvijek nemamo)
- Tehnološka Nacionalna sidrišna točka (Sveučilište u Zagrebu, Filozofski fakultet, Marko Tadić)

Uključeni dužnosnici

- Bernard Gršić, državni tajnik (Središnji državni ured za razvoj digitalnoga društva)
- Maja Radišič-Žuvanić (Ministarstvo gospodarstva, poduzetništva i obrta)

Dosadašnji isporučitelji podataka

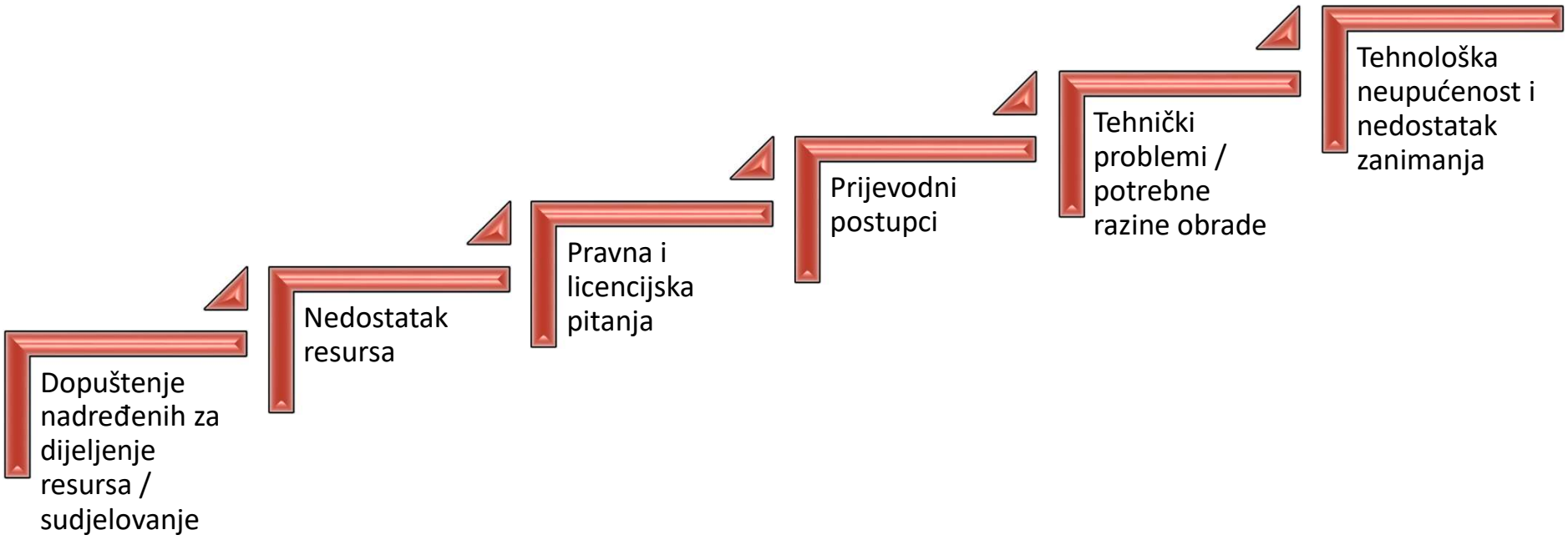
- HNB, DZS, MVEP, ...

Potencijalni isporučitelji podataka

- Mnoge druge ustanove/tijela iz javnoga sektora

Prepreke i postignuća

Glavne prepreke



Prepreke i postignuća

Glavna postignuća

- Prikupljeno i obrađeno 20-ak hr resursa
- Veći broj isporučitelja resursa
- Nadilaženje manjka resursa za hrvatski: CEF-projekti MARCELL i NEC TM Data

MARCELL

- Multilingual Resources for CEF-AT in the Legal Domain
- Projekt iz područja CEF Telekomunikacije
- Potpora
  - Ukupno: 1.883.714,67 EUR
  - Procijenjen udio CEF-potpore: 1.412.786,00 EUR
- Trajanje
  - 2018-10-01 – 2020-09-30 (24 mjeseca)
- Mrežna stranica
  - http://marcell-project.eu

Partneri na projektu

- Istraživački institut za lingvistiku Madžarske akademije znanosti (RILMTA), Budimpešta (koordinator)
- Institut za Bugarski jezik "Prof. Ljubomir Andrejčin" (IBL), Sofija
- Sveučilište u Zagrebu, Filozofski fakultet (FFZG), Zagreb
- Institut za računalne znanosti, Poljska akademija znanosti (IPI-PAN), Varšava
- Institut za istraživanje umjetne inteligencije, Rumunjska akademija (RACAI), Bukurešt
- Jezikoslovni institut Ljudevita Štura Slovačke akademije znanosti (LSIL), Bratislava
- Institut Jožef Stefan (JSI), Ljubljana

Ciljevi projekta

- Opći cilj projekta MARCELL:
  - Unaprijediti AT nacionalnoga zakonodavstva
    - Ti tekstovi još nisu dostupni CEF-AT-u za uporabu
  - Iz 7 sredno- i istočnoeuropskih zemalja
    - Bugarska, Hrvatska, Madžarska, Poljska, Rumunjska, Slovačka, Slovenija
- Očekivani rezultati
  - 7 velikih jednojezičnih korpusa
    - obrađeni morfološki i s prepoznatim imenima
    - Klasificiranih u domene prema vršnim EUROVOC temama
    - Obogaćeni EUROVOC i IATE terminima u metapodatcima
  - Usporedivi 7-jezični korpus
    - Usklađen prema domenama identificiranim EUROVOC deskriptorima
  - Hrvatsko-engleski usporedni korpus 1800 hr dokumenata

Očekivani utjecaj

- Sveukupno poboljšanje CEF-AT sustava za 7 dotičnih jezika
- Rezultati MARCELL-a imat će izravna utjecaja na
  - ePravosuđe
  - *Online* rješavanje sporova

  jer se resursi usredotočuju na nacionalno zakonodavstvo, koje je izravno relevantno u oba DSI-ja
- Posredan utjecaj moguć je i na ostale digitalne usluge
  - Europeana
  - ...

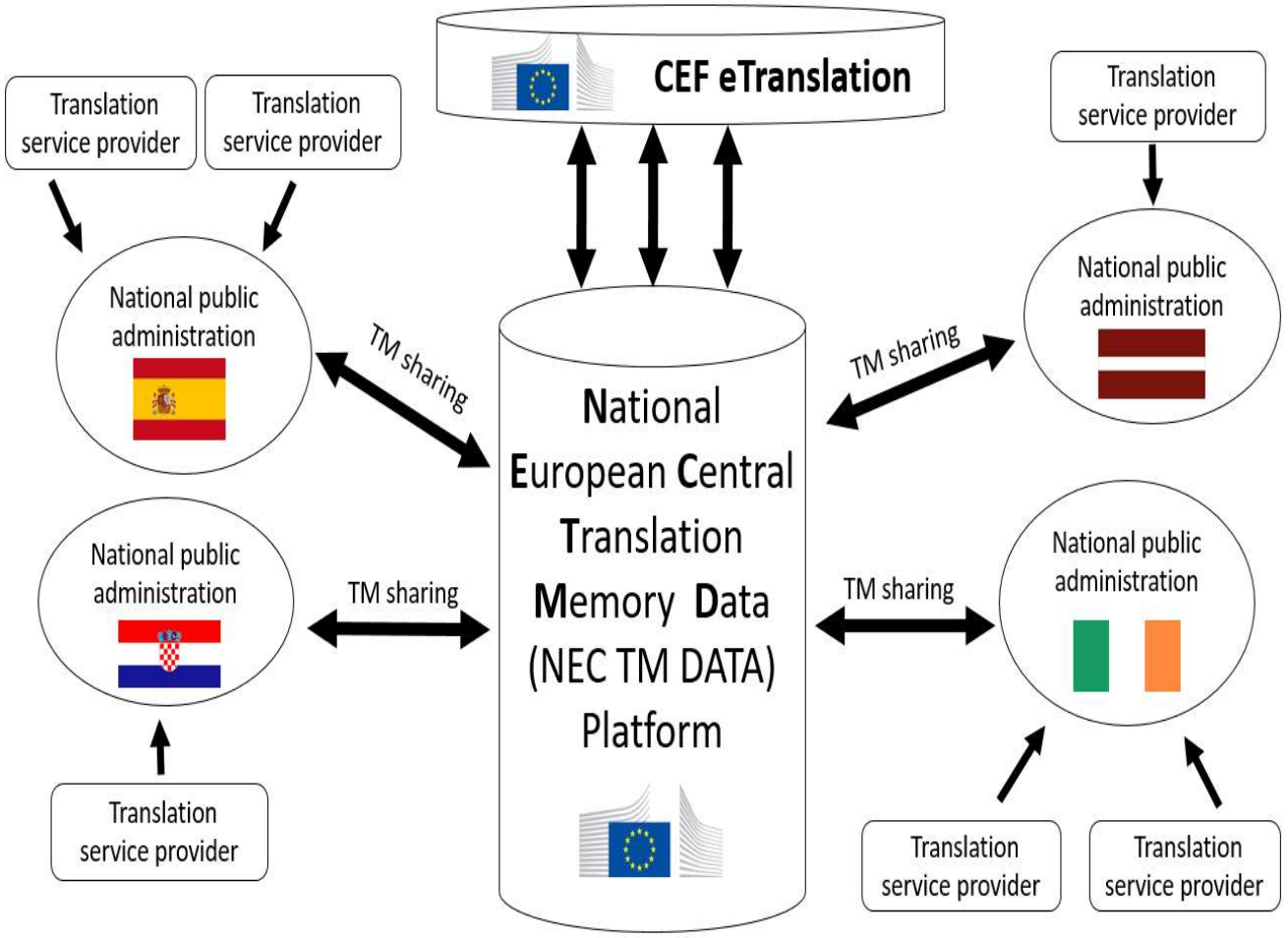
NEC TM Data

Zagreb, 2019-02-12

Ciljevi projekta

- Spoznati količinu i vrijednosti paralelnih tekstova/podataka nastalima na temelju ugovora o prevođenju na nacionalnoj razini (uloga nacionalnih sidrišnih točaka)
- Prikupljanje (paralelnih) podataka u TMX zapisu iz tvrtki s kojima su sklopljeni ugovori o prevođenju
- Povećanje opsega paralelnih podataka dostupnih EK-u te pomoć u razvoju CEF-ove platforme ePrevođenje
- Promicanje protoka TM-ova iz prijevodnih tvrtki prema tijelima javne vlasti
- Organiziranje dvojezičnih velikih zbirki podataka (*Big Data*) na nacionalnoj razini među državama-članicama
- Omogućivanje tijelima javne vlasti uporabu TM-ova ne bi li uštedjeli na ugovorima za prevođenje uporabom metoda iz profesionalne prakse (npr. približno podudaranje, *fuzzy matching*)
- Potpora radu prevoditelja omogućivanjem mrežno-organiziranoga prevođenja (suradni TM-ovi)

Arhitektura sustava (uskladiti s NST-ovima + ePrevođenjem)



Proučavanje ugovora o prevođenju i odnosa s tijelima javne vlasti

Istraživanje cilja prikupiti podatke o odnosima između svih ugovora o prevođenju, isporučitelja usluge prevođenja, količine i naručitelja kao i utvrditi prosječnu vrijednost, nacionalne troškove prevođenja itd.
TED, službeni glasnici na nacionalnoj i regionalnoj razini bit će korišteni kao izvori podataka, kao i ugledne RFP informacijske usluge i tražilice (pretraga koda, ključnih riječi itd.)
To će omogućiti uspostavljanje izvješća temeljenoga na javnim ugovorima i prosječnoj cijeni kao i projiciranim uštedama. To će također tijelima javne vlasti ponuditi osnovu za **razumijevanje koliko je u zadnjih pet godina (2015-2019) prijevoda, koje su napravile ugovorima angažirane tvrtke, dostupnio i koja je vrijednost tih prijevoda**.
Zadatci u ovoj aktivnosti uključuju:

- Stvaranje izvještajne baze podataka s popisom ugovora za prevođenje na nacionalnoj razini
- Nacrtak za dvije bijele knjige u kojima će se analizirati ugovori za prevođenje s tijelima javne vlasti u zadnjih 5 godina
- Uspostava polaznoga dokumenta za tijela javne vlasti kako bi ona shvatila na koji je način mnoštvo podataka javno dostupno, kako bi se oni mogli prikupiti i iskoristiti u daljnjim istraživanjima, za uštede pri sastavljanju novih ugovora uz održavanje profesionalne razine prijevoda i za prilog naporima za prikupljanje podataka na europskoj razini

Ciklopejine zadaće

Dosadašnji rezultati

Opseg javne nabave:

Hrvatski trošak za ugovore o prevođenju 2015-2018: 1,3 M€;
Broj ugovora: 15
Prosječna vrijednost: 86 K€

Slovenski trošak za ugovore o prevođenju 2015-2018: 15 M€;
Broj ugovora: 40
Prosječna vrijednost: 375 K€

European Language Resource Coordination
Connecting Europe Facility

Što smo naučili?

Što nije funkcioniralo prema očekivanjima?

Kako pospješiti suradnju na nacionalnoj razini (uključivanje ključih dionika, dodatni dionici)?

Kako poboljšati pristup potencijalnim isporučiteljima tekstovnih podataka?

Kako poboljšati prijevodne postupke kako bi se olakšalo dijeljenje podataka?

Ima li mjesta za druge vrste poboljšanja?

European Language Resource Coordination
Connecting Europe Facility

Zahvaljujem na pozornosti.

e-pošta: info@lr-coordination.eu

mrežne stranice: www.lr-coordination.eu