PRINCIPLE

Izrada resursa za irski, norveški, hrvatski i islandski u svrhu jezičnog inženjeringa (PRINCIPLE)

Co-financed by the Connecting Europe Facility of

PRINCIPLE

- Originalni naziv: Providing Resources in Irish, Norwegian, Croatian and Icelandic for Purposes of Language Engineering (PRINCIPLE)
- Financiran od strane Instrumenta za povezivanje Europe (*Connecting Europe Facility*, CEF)
- Akcija 2018-EU-IA-0050, Financijska potpora br. INEA/CEF/ICT/A2018/1761837
- Trajanje projekta: rujan 2019. - kolovoz 2021.
- Cilj: identificirati, prikupiti i obraditi kvalitetne jezične resurse za četiri nedovoljno razvijena europska jezika: hrvatski, islandski, irski te norveški (bokmål i nynorsk)

PRINCIPLE

Partneri

- Sveučilište Dublin City (Dublin City University) (koordinator)
- Sveučilište na Islandu (University of Iceland)
- Filozofski fakultet Sveučilišta u Zagrebu
- Nacionalna knjižnica Norveške (National Library of Norway)
- Iconic Translation Machines d.o.o.

DCU Ollscoil Chathair Bhaile Atha Cliath Dublin City University

UNIVERSITY OF ICELAND
SCHOOL OF HUMANITIES

Filozofski fakultet Sveučilišta u Zagrebu

LANGUAGE BANK
National Library of Norway

Iconic
Part of the RWS Group

Domene jezičnih resursa

- Razvijamo visokokvalitetne jezične resurse radi unaprjeđenja kvalitete prijevoda u infrastrukturi digitalnih usluga (Digital Service Infrastructures, DSIs) **ePravosuđa** (eJustice) i **eNabave** (eProcurement) putem sustava za strojno prevođenje razvijenih posebno za navedene domene.
- S obzirom na izbijanje pandemije bolesti COVID-19, dodatno smo razvili resurse za **eZdravstvo** (eHealth).

Postupak identifikacije izvora jezičnih resursa

- Nekolicina nacionalnih tijela i lokalnih dionika diljem Hrvatske, Islanda, Irske i Norveške dostavili su jezične resurse konzorciju PRINCIPLE-a te postali „rani prisvajatelji”.
- Iconic Translation Machines razvili su sustave za neuronsko strojno prevođenje iz doniranih jezičnih resursa radi ovjeravanja kvalitete resursa.
- „Rani prisvajatelji” imaju pristup sustavima za strojno prevođenje za vrijeme trajanja projekta kako bi potvrdili kvalitetu u stvarnim poslovnim procedurama te kako bi pružili povratne informacije.
- Jezični resursi su postavljeni na portal ELRC-SHARE (https://www.elrc-share.eu/) za razvoj CEF-ovog sustava eTranslation (većinom javno dostupni).

Rani prisvajatelji u Hrvatskoj

- Ciklopea d.o.o.
  - Razvijena dva sustava za strojno prevođenje: područje eNabave i područje eZdravstva
- Ministarstvo vanjskih i europskih poslova
  - Razvijen sustav strojnog prevođenja iz područja ePravosuđa
- Prednosti razvijenih sustava strojnog prevođenja:
  - Razvijeni upravo za potrebe ranih prisvajatelja
  - Rade bolje od javno dostupnih sustava strojnog prevođenja (Google Translate & Bing Microsoft Translator)
  - Sigurni podaci

Ostali donatori izvora jezičnih resursa

- Središnji državni ured za razvoj digitalnog društva
- Središnji državni ured za središnju javnu nabavu
- Državna komisija za kontrolu postupaka javne nabave
- Filozofski fakultet
- Budući da se tijekom projekta dogodila pandemija i potres u Zagrebu, s dijelom donatora nismo uspjeli ponovno stupiti u kontakt.

Jezični resursi za hrvatski jezik - obrada

- Donatori su dokumente većinom slali direktno preko e-maila ili poveznica
- Dio dokumenata je bio uparen (hrvatski i engleski dokument), a dio nije
- Dokumenti su dolazili u raznim formatima: XML/TMX, PDF, DOC(X), HTML
- Cilj: konvertirati sve podatke u čisti i poravnati tekst
  - Pristupi: ručno kopiranje iz PDF i DOC datoteka, automatska ekstrakcija iz HTML-a, OCR

Jezični resursi za hrvatski jezik - obrada

- Neki od problema na koje se nailazilo prilikom obrade podataka:
  - Nekvalitetan OCR pojedinih dokumenata ili dijelova dokumenata
  - Pojavljivanje ligatura koje su se morale (polu)automatski ili ručno ispravljati
  - Pojava znakova koji zamjenjuju pojedina slova u riječi ili su nepotrebno dodani
  - Neki dokumenti sadrže i original i prijevod, bez eksplicitne granice – potrebno ih razdvojiti u dva dokumenta

PRINCIPLE

Jezični resursi za hrvatski jezik – trenutno stanje

Prijevodne jedinice							
Donator / Domena	**Središnja nabava**	**SDURDD**	**MVEP**	**Ciklopea**	**DKOM**	**FFZG**	**Ukupno**
eNabava	7,281	3,911	0	36,635	11,511	0	**59,338**
ePravosuđe	0	495,546	124,284	0	492	0	**620,336**
eZdravstvo	0	0	563	76,108	0	0	**76,671**
Opća	0	0	0	0	0	1,760,822	**1,760,822**
Ukupno	**7,281**	**499,457**	**124,847**	**112,743**	**12,003**	**1,760,822**	**2,517,167**

PRINCIPLE

Jezični resursi za hrvatski jezik – trenutno stanje

Pojavnice							
Donator / Domena	**Središnja nabava**	**SDURDD**	**MVEP**	**Ciklopea**	**DKOM**	**FFZG**	**Ukupno**
eNabava	283,555	161,737	0	1,315,683	404,626	0	**2,165,601**
ePravosuđe	0	19,355,551	8,070,074	0	24,260	0	**27,449,885**
eZdravstvo	0	0	18,044	2,074,341	0	0	**2,092,385**
Opća	0	0	0	0	0		**65,378,746**
Ukupno	**283,555**	**19,517,288**	**8,088,118**	**3,390,024**	**428,886**	**65,378,746**	**97,086,617**

PRINCIPLE

Hvala na pažnji!