European Language Resource Coordination
Connecting Europe Facility

Jezici i jezične tehnologije u Hrvatskoj

Prof. dr. sc. Marko Tadić
Sveučilište u Zagrebu, Filozofski fakultet

Je li Hrvatska jednojezična zemlja?

- Republika Hrvatska ima jedan službeni jezik: hrvatski
  - Ustav Republike Hrvatske (čl. 12, st. 1)
  - „U Republici Hrvatskoj u službenoj je uporabi hrvatski jezik i latinično pismo.”
- u stvarnome životu prilike su ponešto drukčije
  - turizam
  - trgovina
  - prekogranična suradnja
    - promet
    - energija i klima
    - zaštita okoliša
    - pravna pitanja (rođenja, vjenčanja, nasljedstva…)
  - nacionalne manjine

Jezici u Republici Hrvatskoj

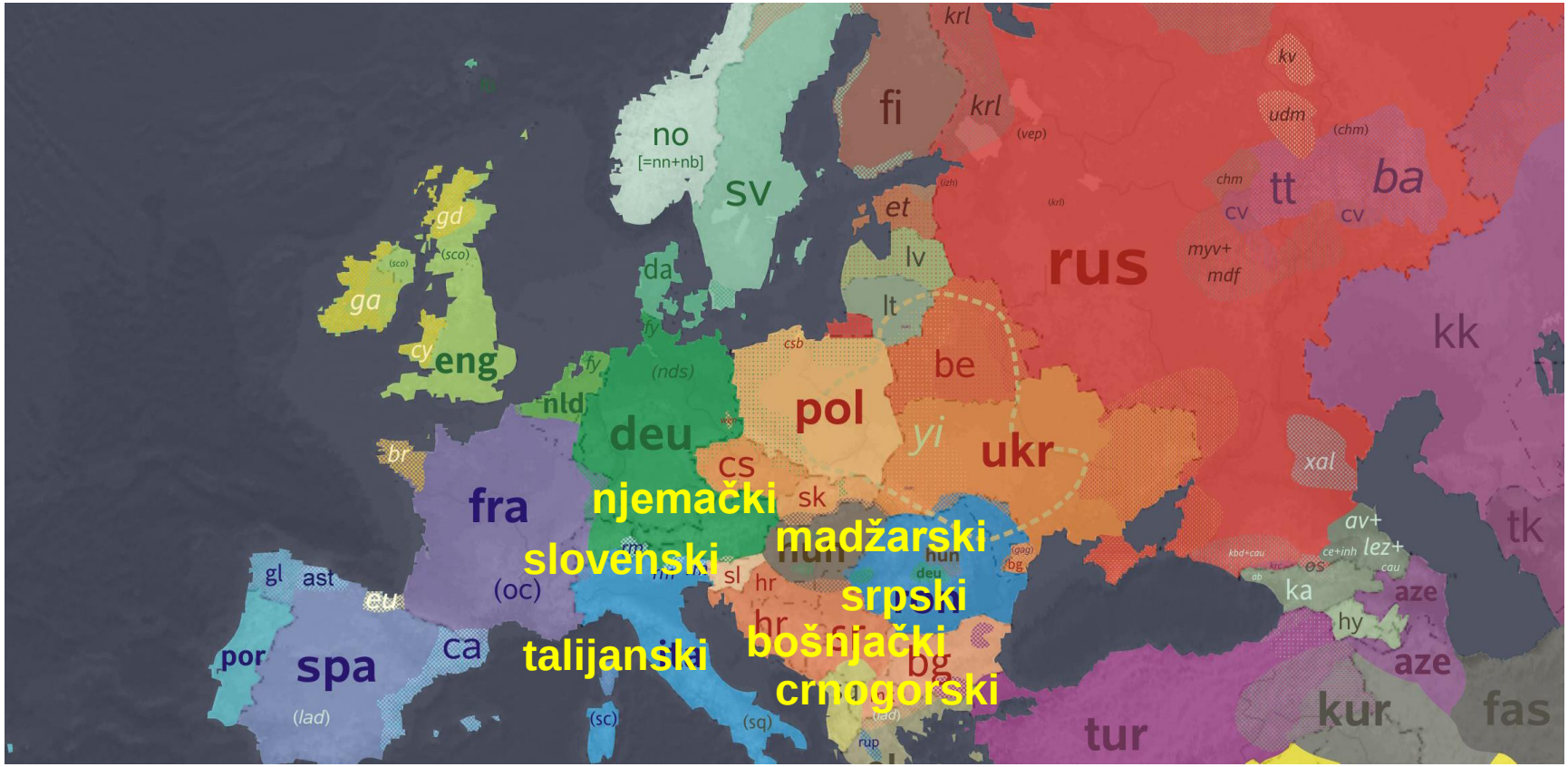
- zaštita prava nacionalnih manjina i u slučaju jezika i pisma
  - Ustav Republike Hrvatske (čl. 12, st. 2)
  - „U pojedinim lokalnim jedinicama uz hrvatski jezik i latinično pismo u službenu se uporabu može uvesti i drugi jezik te ćirilično ili koje drugo pismo pod uvjetima propisanima zakonom."
- Popis stanovništva Republike Hrvatske, DZS (2011.): materinski jezik (RH: 4.285.889 stanovnika)
  - hrvatski: 95,60%
  - srpski: 1,23%
  - talijanski: 0,43%
  - albanski: 0,40%
  - bošnjački: 0,39%
  - romski: 0,34%
  - ostali: < 0,3%

Jezici u Republici Hrvatskoj

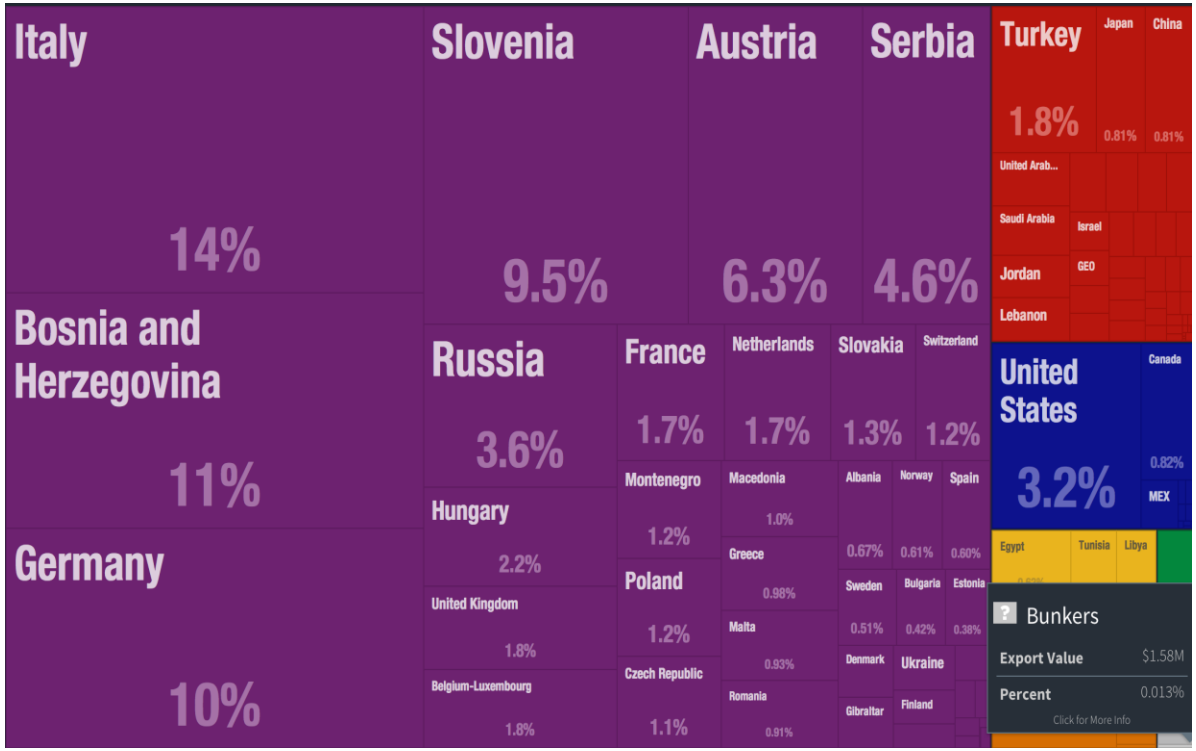
- većina državljana RH govori barem jedan strani jezik
  - prvi strani jezik – obvezatan već u osnovnoj školi
  - drugi strani jezik – dodaje se u srednjoj školi
  - najčešći strani jezici
    - engleski
    - njemački
    - talijanski
  - bitan, a često zanemaren uzrok dobroga poznavanja stranih jezika
    - u Hrvatskoj se strani filmovi ne sinkroniziraju
    - osim filmova za djecu mlađu od 6 godina
- glavna područja uporabe (aktivno i pasivno) stranih jezika u RH
  - elektronički mediji (TV)
  - turizam
  - trgovina
  - komunikacija s ostalim državama-članicama EU-a (znatan porast)

European Language Resource Coordination
Connecting Europe Facility

Susjedni jezici



Hrvatska izvozna tržišta (2013)



Izvor: OEC Observatory of Economic Activity

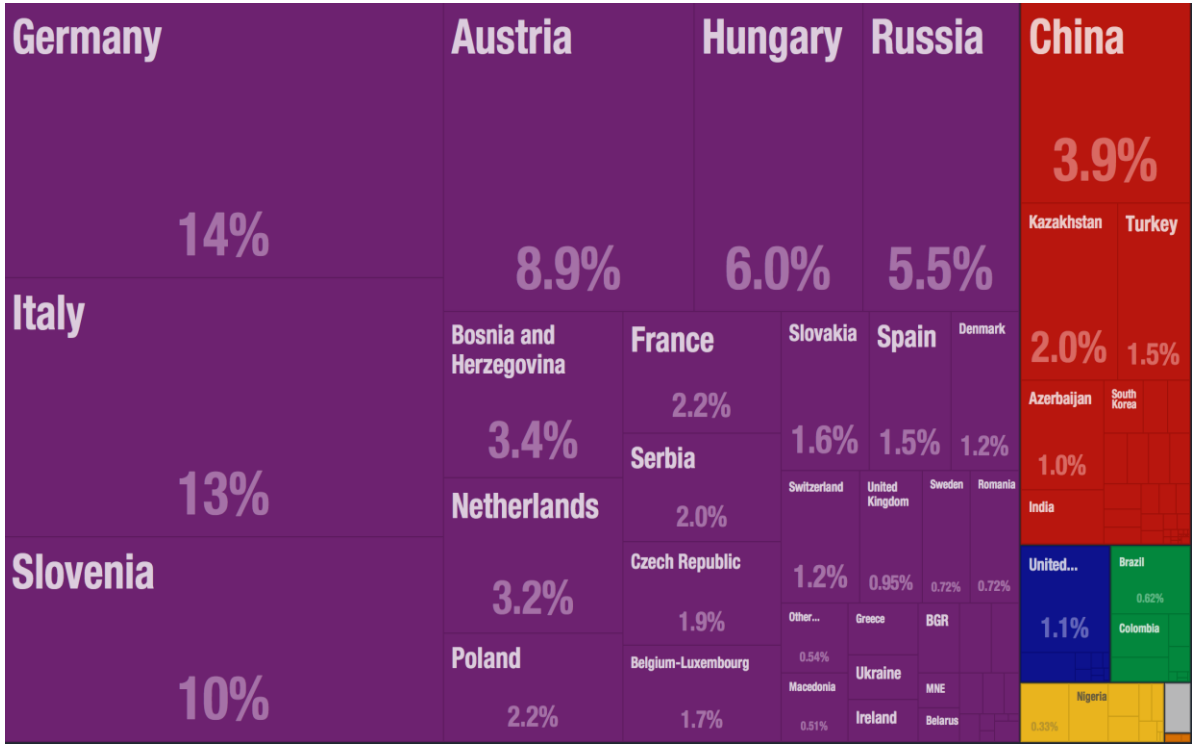
Kontinenti:

- Europa (83%)
- Azija (7,7%)
- S. Amerika (4,3%)
- Afrika (3,1%)

Jezici:

- njemački (16,3%)
- talijanski (14%)
- bošnjački (11%)
- slovenski (9,5%)
- engleski (5%)
- srpski (4,6%)
- ruski (3,6%)
- madžarski (2,2%)
- francuski (1,7%)

Hrvatska uvozna tržišta (2013)



Izvor: OEC Observatory of Economic Activity

Kontinenti:

- Europa (86%)
- Azija (11%)
- S. Amerika (1,4%)
- J. Amerika (1,2%)
- Afrika (0,99%)

Jezici:

- njemački (22,9%)
- talijanski (13%)
- slovenski (10%)
- madžarski (6%)
- ruski (5,5%)
- kineski (3,9%)
- bošnjački (3,4%)
- nizozemski (3,2%)
- francuski (2,2%)
- poljski (2,2%)

Jezične tehnologije za hrvatski jezik

- svega nekoliko središta/ustanova u kojima se razvijaju jezične tehnologije za hrvatski jezik
  - Sveučilište u Zagrebu, Filozofski fakultet
    - Zavod za lingvistiku
    - Odsjek za lingvistiku (Katedra za algebarsku i računalnu lingvistiku)
    - Odsjek za informacijske i komunikacijske znanosti (NLP Lab)
  - Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
    - Zavod za elektroniku, mikroelektroniku, računalne i inteligentne sustave (KTLab)
    - Zavod za osnove elektrotehnike i električka mjerenja
  - Institut za hrvatski jezik i jezikoslovlje
    - Odjel za opće jezikoslovlje
  - Sveučilište u Rijeci
    - Odjel za informatiku
- Hrvatsko društvo za jezične tehnologije (hdjt.hr, od 2004.)
- Portal Jezične tehnologije za hrvatski jezik (jthj.ffzg.hr, od 2000.)
- hrvatski META-SHARE čvor (meta-share.ffzg.hr, od 2013.)

Jezične tehnologije za hrvatski jezik

- jezični resursi: korpusi
  - Hrvatski nacionalni korpus (HNK, hnk.ffzg.hr), 216 Mw
    - Zavod za lingvistiku
  - Hrvatski www-korpus (HrWaC, nlp.ffzg.hr/resources/corpora/hrwac/), 2Gw
    - Odsjek za informacijske i komunikacijske znanosti
  - Riznica hrvatskoga jezika (riznica.ihjj.hr), cca 70 Mw
    - Institut za hrvatski jezik i jezikoslovlje
  - manji korpusi
    - Hrvatski prijevodi Pravne stečevine EU (30,5 Mw)
    - SETimes korpus (desetojezični paralelni korpus), 8,8 Mw
    - Hrvatsko-engleski paralelni korpus (3,5 Mw)
    - HrEnWac (4,4 Mw)…
  - specijalizirani korpusi
    - Hrvatska ovisnosna banka stabala (HOBS, hobs.ffzg.hr), 4500 rečenica
    - SETimesTreebank (hr), 4000+2500 rečenica

Jezične tehnologije za hrvatski jezik

- jezični resursi: korpusi drugih jezika
  - SiWaC, SrWaC, BsWaC, CaWaC
- jezični resursi: leksikoni
  - Hrvatski morfološki leksikon (HML, hml.ffzg.hr), 110.000 natuknica
    - Zavod za lingvistiku
  - CroDeriV: derivacijski leksikon hr glagola (croderiv.ffzg.hr), 15.000
    - Zavod za lingvistiku
  - Hrvatski WordNet (CroWN, crown.ffzg.hr), v2.0: 23.120 sinskupova
    - Odsjek za lingvistiku
  - STRUNA (struna.ihjj.hr), preko 30.000 pojmova
    - Institut za hrvatski jezik i jezikoslovlje
    - velika terminološka baza strukovnoga nazivlja (do sada 20 struka)
  - Terminološki portal (nazivlje.hr)
    - Institut za hrvatski jezik i jezikoslovlje

Jezične tehnologije za hrvatski jezik

- jezični alati
  - Hrvatski POS/MSD-označivač (CroTag)
    - Odsjek za lingvistiku / Odsjek za informacijske i komunikacijske znanosti
  - NERC sustav (OZANA)
    - Odsjek za lingvistiku
  - CroNER
    - KTLab
  - Hrvatski ovisnosni parser
    - Odsjek za informacijske i komunikacijske znanosti
  - Hascheck (hacheck.tel.fer.hr), mrežni provjernik pravopisa
    - Zavod za osnove elektrotehnike i električka mjerenja
  - eCADIS: automatsko označavanje deskriptorima (takelab.fer.hr/ecadis)
    - KTLab
  - CADIAL tražilica po deskriptorima (takelab.fer.hr/cadial-se, cadial.org)
    - KTLab, v. isto Digitalni ured
  - TermeX: crpljenje domenskih termina (takelab.fer.hr/termex_s/)
    - KTLab

Jezične tehnologije za hrvatski jezik

- prema knjizi *Hrvatski jezik u digitalnom dobu* (2012.)
  - http://www.meta-net.eu/whitepapers/volumes/croatian
  - dio META-NET-ova niza Jezične bijele knjige (*Language Whitepapers*)

  hrvatski jezik pripada među jezike s vrlo slabo razvijenim jezičnim tehnologijama
- strojnoprevoditeljski sustavi (na i s hrvatskoga)
  - osim globalnih sustava
    - Google, Bing
  - razvijeni samo kao istraživački prototipovi
    - FP7 projekt ACCURAT (www.accurat-project.eu)
    - ICT-PSP projekt Let'sMT! (www.letsmt.org)
    - FP7 projekt XLike (www.xlike.org)
- suradnja s CEF.AT-om
  - može pospješiti razvoj jezičnih resursa i/li alata za hrvatski jezik
  - osobito za prijevod en→hr zbog specifičnosti hrvatskoga kao ciljnoga jezika

European Language Resource Coordination
Connecting Europe Facility

Zahvaljujem na pozornosti.