META-NET niz Bijele knjige

Jezici u Europskom informacijskom društvu

Hrvatski

Pripremno predizdanje

META-FORUM 2011

27.-28. lipnja 2011.

Budimpešta, Madžarska

Ova je bijela knjiga nastala potporom Sedmoga okvirnoga programa i Potpornoga programa za ICT politiku Europske komisije u skladu s ugovorima T4ME (Ugovor br. 249119), CESAR (ugovor br. 271022), METANET4U (ugovor br. 270893) i META-NORD (ugovor br. 270899).

Ova je bijela knjiga namijenjena nastavnicima, novinarima, političarima, jezičnim zajednicama i svima drugima koji žele uspostavu istinski višejezične Europe.

Ova je bijela knjiga dio niza knjiga koji promiče znanje o jezičnim tehnologijama i njihovim mogućnostima. Dostupnost i uporaba jezičnih tehnologija varira među raznim europskim jezicima. Sljedno tome, aktivnosti potrebne za daljnje podupiranje istraživanja i razvoja jezičnih tehnologija također se razlikuju od jezika do jezika. Aktivnosti koje treba poduzeti ovise o mnogim čimbenicima kao što su strukturna složenost određenoga jezika, ali i broj govornika u toj jezičnoj zajednici.

META-NET kao mreža izvrsnosti Europske komisije obavio je analizu postojećih jezičnih resursa i tehnologija. Ta je analiza usredotočena na 23 službena jezika Europske unije kao i na ostale važne regionalne europske jezike. Rezultati ove analize pokazuju kako postoje mnogi značajni jazovi u istraživanju svakoga od spomenutih jezika. Podrobnija stručna analiza i procjena trenutačnoga stanja pomoći će učinkovitost dodatnih istraživanja ali i umanjiti moguće rizike.

META-NET se sastoji od 44 istraživačka središta iz 31 zemlje koja surađuju sa svim dionicima tih aktivnosti kao što su komercijalne tvrtke, vladine agencije, industrija, istraživačke ustanove, softverske tvrtke, isporučitelji tehnologije i europska sveučilišta. Oni svi zajedno stvaraju zajedničku tehnološku viziju kroz razvijanje strateškoga istraživačkoga plana koji će pokazati kako će se do 2020. premostiti jazovi u primjeni jezičnih tehnologija.

META-NET
DFKI Projektbüro Berlin
Alt-Moabit 91c
10559 Berlin
Germany

office@meta-net.eu
http://www.meta-net.eu

Autori

Prof. Dr. Marko Tadić, University of Zagreb, Faculty of Humanities and Social Sciences
Prof. Dr. Dunja Brozović-Rončević, Institute of Croatian Language and Linguistics
Prof. Dr. Amir Kapetanović, Institute of Croatian Language and Linguistics

Zahvale

Nakladnik zahvaljuje autorima Bijele knjige o njemačkome jeziku za dopuštenje korištenja građe iz te knjige.

Kazalo

Kazalo **2**
Sažetak **3**
Opasnost za naše jezike i izazov za jezične tehnologije **4**
Jezične granice ometaju uspostavu europskoga informacijskoga društva 5
Opasnost za naše jezike 5
Jezične su tehnologije ključne potporne tehnologije 6
Mogućnosti jezičnih tehnologija 6
Izazovi koji stoje pred jezičnim tehnologijama 7
Usvajanje jezika 8
Hrvatski u europskome informacijskome društvu **10**
Opće činjenice 10
Hrvatska narječja 12
Standardizacija hrvatskoga jezika 13
Osobine hrvatskoga jezika 14
Fonetika, fonologija, morfonologija *14*
Morfologija *15*
Rječnik, frazeologija, nazivlje *16*
Sintaksa *17*
Pravopis *18*
Onomastika *18*
Odnos hrvatskoga standardnoga jezika s ostalim jezicima štokavske osnovice 19
Skrb o jeziku u Hrvatskoj 20
Jezik u obrazovanju 21
Međunarodni odnosi 21
Hrvatski na Internetu 22
Jezičnotehnološka podrška za hrvatski **24**
Jezične tehnologije 24
Arhitekture jezičnotehnoloških aplikacija 24
Osnovna područja primjene jezičnih tehnologija 25
Jezična provjera *25*
WWW tražilice *26*
Govorna interakcija *28*
Strojno prevođenje *30*
Jezične tehnologije 'iza zavjese' 33
Jezične tehnologije u obrazovanju 34
Jezičnotehnološki programi 34
Dostupnost alata i resursa za hrvatski jezik 36
Tablica alata i resursa 38
Zaključci 39
Interpretacija tablice o hrvatskim resursima i alatima *39*
Što bi trebalo poduzeti? *39*
Literatura **41**
META-NET **43**
Što je cilj META-NET-a? 43
Prva META-NET zbivanja u 2010 43
Trenutačni sastav META tehnološkoga vijeća 45
Članstvo mreže izvrsnosti META-NET (*: osnivački članovi) 45

Sažetak

Mnogi se europski jezici nalaze u opasnosti postati žrtvama digitalnoga doba zbog svoje nedovoljne zastupljenosti on-line i nedovoljne razvijenosti svojih jezičnih resursa. Ogromne mogućnosti koje se otvaraju na regionalnim tržištima do danas još nisu iskorištene zbog jezičnih prepreka. Ukoliko se već sad ne povuku odgovarajući potezi, mnogi će se građani Europske unije naći socijalno i ekonomski zakinuti jer govore samo svoj materinski jezik.

Inovativne su jezične tehnologije (JT) posrednik koji će omogućiti europskim građanima sudjelovanje u egalitarnom, uključivom i gospodarski uspješnom informacijskome društvu i društvu znanja. Višejezične jezične tehnologije bit će kanali za trenutačnu, jeftinu i jednostavnu komunikaciju i interakciju zaobilazeći jezične prepreke.

Danas razne vrste jezičnotehnoloških usluga nude ponajprije komercijalni isporučitelji iz SAD. Besplatna usluga prevođenja Google Translate samo je jedan od primjera. Nedavni uspjeh Watsona, IBM-ova računalnoga sustava koji je pobijedio dvojicu ljudi-natjecatelja u epizodi kviza Jeopardy, dobar je primjer velikoga potencijala jezičnih tehnologija. Međutim, mi si kao Europljani moramo postaviti nekoliko gorućih pitanja:

- Treba li naša komunikacijska infrastruktura i infrastruktura znanja ovisiti o monopolističkim tvrtkama?
- Možemo li se doista oslanjati na jezičnotehnološke usluge koje bi netko u bilo kojem trenutku mogao prestati nuditi?
- Natječemo li se dovoljno aktivno na svjetskome tržištu istraživanja i razvoja jezičnih tehnologija?
- Možemo li doista očekivati da će se itko s drugih kontinenata htjeti uopće baviti našim prevoditeljskim problemima i drugim problemima vezanim uz europsku višejezičnost?
- Može li naše europsko kulturno nasljeđe pomoći u oblikovanju društva znanja nudeći bolju, sigurniju, precizniju, inovativniju i otporniju visokokvalitetnu tehnologiju?

Ova bijela knjiga o hrvatskome jeziku pokazuje kako u Hrvatskoj postoji temeljno okružje za istraživanja jezičnih tehnologija, međutim to nije rezultiralo razvojem jezične industrije. Unatoč tome što su za hrvatski izrađeni neki jezični resursi i tehnologije, znatno ih je manje nego za druge slavenske jezike, npr. češki, a još ih je manje razvijeno u odnosu na veće europske jezike kao što su engleski, njemački ili francuski. Prema procjenama podrobnije iznesenim u ovome izvješću, potrebno je poduzeti niz ciljanih mjera ne bi li se hrvatski jezični resursi i alati doveli na istu razinu razvijenosti glede njihove kakvoće i količine, kakva je razina već dosegnuta za druge europske jezike.

META-NET pridonosi izgradnji snažnoga, višejezičnoga europskoga digitalnoga informacijskoga prostora. Ostvarenjem toga cilja multikulturna će unija naroda napredovati i postati primjerom za uspješnu miroljubivu i jednakopravnu međunarodnu suradnju. Kad se taj cilj ne bi postigao, Europa bi se mogla naći u pred teškim izborom između žrtvovanja svojih kulturnih identiteta ili gospodarskoga neuspjeha.

Opasnost za naše jezike i izazov za jezične tehnologije

Upravo svjedočimo digitalnom veleprevratu koji je usporediv s Gutenbergovim izumom tiska pomičnim slovima.

U ovome trenutku svjedočimo digitalnoj revoluciji koja korjenito utječe na našu komunikaciju i naše društvo. Najnoviji razvoj digitalnih i mrežnih komunikacijskih tehnologija ponekad se uspoređuju s Gutenbergovim izumom tiska pomičnim slovima. Što nam ta analogija može reći o budućnosti europskoga informacijskoga društva i o našim vlastitim jezicima?

Nakon Gutenbergova izuma pravi su proboji u komunikaciji i razmjeni znanja postignuti pothvatima kao što je Lutherov prijevod Biblije na narodni jezik. U stoljećima koja su slijedila razvijeni su različiti kulturni postupci koji su omogućili obradbu jezika i razmjenu znanja:

- pravopisno i gramatičko normiranje većih jezika omogućilo je brzu razmjenu novih znanstvenih ideja;
- uspostavljanje službenih jezika omogućilo je građanima komunikaciju unutar određenih (često politički) granica;
- poučavanje jezika i prevođenje omogućilo je razmjenu preko jezičnih granica;
- stvaranje novinarskih i bibliografskih normi osiguralo je kakvoću i dostupnost tiskovina;
- stvaranjem različitih medija kao što su knjige, novine, radio, televizija i drugi mediji zadovoljavaju se komunikacijske potrebe pučanstva.

U zadnjih dvadeset godina informacijska tehnologija omogućila je olakšavanje i automatizaciju mnogih procesa:
In the past twenty years, information technology helped to automate and facilitate many of the processes:

- računalna potpora pripremi za tisak zamjenjuje tipkanje i prijelom;
- Microsoft PowerPoint zamjenjuje projiciranje s prozirnica;
- e-pošta odašilje i prima dokumente brže od telefaks uređaja;
- Skype omogućuje internetske pozive i održavanje virtualnih sastanaka;
- zajednički formati zapisa zvučnih i vizualnih podataka omogućuje jednostavnu razmjenu multimedijskoga sadržaja;
- tražilice omogućuju pristup www-stranicama na temelju pretrage ključnim riječima;
- on-line usluge poput Google Translate nude brze ali približne prijevode;
- društvene mreže omogućuju suradnju i razmjenu obavijesti.

Premda su takve aplikacije i usluge višestruko korisne, ipak se na njima ne može uspostaviti održivo, višejezično europsko informacijsko društvo kao suvremeno i uključivo društvo gdje je protočnost informacija i roba slobodna.

Jezične granice ometaju uspostavu europskoga informacijskoga društva

Ne možemo točno znati kako će u budućnosti izgledati informacijsko društvo. U raspravama o zajedničkoj europskoj energetskoj strategiji ili vanjskoj politici željeli bismo slušati europske ministre vanjskih poslova kako govore na njihovom vlastitom jeziku. Također bismo htjeli vidjeti platformu s pomoću koje će ljudi, koji se služe s više različitih jezika ne znajući ih sve podjednako dobro, moći raspravljati o određenoj temi dok će tehnološki uređaji prikupljati njihove stavove i generirati njihove sažetke. Također, jednoga ćemo dana možda imati potrebu razgovarati s pozivnim središtem za zdravstveno osiguranje koje je smješteno u stranoj zemlji.

Globalizacija gospodarstva i informacijskoga prostora suočava nas sa sve više različitih jezika, govornika i sadržaja.

U usporedbi sa stanjem od prije nekoliko godina, današnje komunikacijske potrebe izgledaju sasvim drukčije. U globalnome gospodarstvu i informacijskome prostoru dostupno je sve više sadržaja koje generira sve više govornika na sve više jezika, a od nas se traži da budemo u stanju služiti se sa sve novijim medijima. Trenutačna popularnost društvenih mreža (Wikipedia, Facebook, Twitter i YouTube) predstavlja samo vršak ledenoga brijega.

Danas bez ikakvih prepreka možemo u nekoliko sekunda na drugu stranu svijeta prebaciti gigabajte teksta prije nego što uopće shvatimo kako je on na jeziku koji uopće ne razumijemo. Prema nedavnome izvješću, koje je zatražila Europska komisija, 57% internetskih korisnika u Europi kupuje *on-line* robu i usluge na jeziku koji nije njihov vlastiti (engleski je najčešći strani jezik, a slijede ga francuski, njemački i španjolski). 55% korisnika čita sadržaje na stranome jeziku dok ih samo 35% koristi strani jezik za pisanje poruka e-pošte ili ostavljanje komentara na www-u. [1] Prije nekoliko godina engleski je možda bio lingua franca www-a jer je većina sadržaja na www-u tada bila na engleskome, međutim, danas su se prilike u mnogome promijenile. Moglo bi se reći kako je količina sadržaja na drugim jezicima (osobito azijskim i na arapskome) upravo eksplodirala.

Neobično je kako sveprisutan digitalni jaz prouzrokovan jezičnim preprekama nije privukao dovoljno pozornosti u javnim raspravama; pa ipak, upravo nas on navodi na goruće pitanje: „Koji će europski jezici napredovati i održati se u umreženome informacijskome društvu i društvu znanja?“

Opasnost za naše jezike

Otkriće tiska neizmjerno je pridonijelo razmjeni informacija u Europi, no istodobno je dovelo do izumiranja mnogih europskih jezika. Kako se na regionalnim i manjinskim jezicima tiskalo rijetko, mnogi su jezici, npr. cornwallski ili dalmatski, bili ograničeni samo na govorni oblik komunikacije što je ograničilo njihovo usvajanje, širenje i uporabu.

Šezdesetak europskih jezika jedno su od njezinih najbogatijih i najvažnijih kulturnih nasljeđa. Europsko mnoštvo jezika čini i sastavni dio njezina društvena uspjeha. [2] Dok će prošireni jezici kao engleski ili španjolski zacijelo održati svoju prisutnost na rastućem tržištu

digitalnoga društva, mnogi će europski jezici biti isključeni iz digitalnih komunikacijskih kanal i postati nevažni za internetizirano društvo. Takav razvoj se nipošto ne može podraviti. S jedne bi se strane izgubila strateška pozicija i oslabila bi se ukupna pozicija Europe, a s druge strane, takav bi razvoj bio u suprotnosti s ciljem jednakoga sudjelovanja svakoga građanina EU bez obzira na njegov jezik. Prema izvješću UNESCO-a o višejezičnosti jezici su ključni medij za ostvarivanje temeljnih ljudskih prava kao što su iskazivanje političkoga stava, obrazovanje i sudjelovanje u društvu.[3]

Znatna raznolikost jezika u Europi jedno je od najvažnijih kulturnih dobara i bitan je dio europskoga uspjeha.

Jezične su tehnologije ključne potporne tehnologije

U prethodnim razdobljima investiranje se usredotočivalo na učenje jezika i prevođenje. Na primjer, prema nekim procjenama europsko tržište prevođenja, tumačenja, lokalizacije programske podrške i prevođenja www-stranica vrijedilo je 8,4 milijarde eura, a očekivao se njegov rast od 10% godišnje.[4] No čak i uz takve prognoze rasta postojeći kapaciteti nisu dovoljni za zadovoljenje potreba niti sadašnjih, a kamoli budućih potreba.

Jezične su tehnologije ključna potporna tehnologije koje mogu zaštititi i očuvati europske jezike. Jezične tehnologije pomažu ljudima da surađuju, obavljaju poslove, dijele znanje i sudjeluju u društvenim i političkim raspravama bez obzira na njihovu jezičnu pripadnost ili računalnu pismenost. Jezične nam tehnologije već pomažu u svakodnevnim zadatcima, kao što su pisanje e-pošte, obavljanje on-line pretrage ili rezervacije zrakoplovne karte. Izravne koristi od jezičnih tehnologija imamo kad:

Jezične tehnologije pomažu ljudima u suradnji, obavljanju poslova, razmjeni znanja i sudjelovanju u društvenim i političkim raspravama na različitim jezicima.

- tražimo obavijesti korištenjem internetskih tražilica;
- provjeravamo pravopis ili gramatiku u obradniku teksta;
- gledamo preporuke za proizvode u on-line dućanima;
- slušamo glasovne upute navigacijskoga sustava;
- prevodimo www-stranice uporabom on-line usluge.

Jezične tehnologije o kojima se govori u ovoj bijeloj knjizi čine srž budućih inovativnih aplikacija. Jezične su tehnologije tipično potporne tehnologije unutar veće aplikacije kao što su navigacijski sustav ili tražilica. Ove bijele knjige prikazuju stanje osnovnih postignuća u jezičnim tehnologijama za svaki pojedini jezik.

Uskoro će nam za sve europske jezike trebati jezične tehnologije koje će biti dostupne, prihvatljive i usko povezane u širim okružjima programske podrške. Interaktivno, multimedijsko i višejezično korisničko iskustvo uporabe uređaja ili usluga nije moguće bez razvijenih jezičnih tehnologija.

Mogućnosti jezičnih tehnologija

Jezične nam tehnologije omogućuju strojno prevođenje, stvaranje sadržaja, obradbu obavijesti i upravljanje znanjem na svim europskim jezicima. Jezične tehnologije također mogu unaprijediti daljnji razvoj intuitivnih jezično utemeljenih sučelja za kućnu elektroniku, strojeve, vozila, računala i robote. Premda već postoje mnogi prototipovi, komercijalne i industrijske primjene su još uvijek u ranim stupnjevima

razvoja. Nedavna su postignuća u istraživanjima i razvoju stvorila jedinstvene mogućnosti. Na primjer, strojnim prevođenjem (MT) već se mogu dobiti prijevodi prihvatljive točnosti unutar posebnih područja, dok istodobno neke eksperimentalne aplikacije već omogućuju dohvat višejezičnih obavijesti i upravljanje znanjem, kao i proizvodnju sadržaja na mnogim europskim jezicima.

Jezične aplikacije, govorno temeljena korisnička sučelja i razgovorni sustavi tradicionalno se mogu naći u visokospecijaliziranim područjima uporabe, ali često su ograničene kakvoće. Jedno od aktivnijih područja istraživanja jest uporaba jezičnih tehnologija u spasilačkim operacijama u unesrećenim područjima. U takvim okružjima visoke opasnosti točnost prijevoda može značiti razliku između života i smrti. Ista se načela mogu primijeniti i pri uporabi jezičnih tehnologija u industriji zdravstvene skrbi. Inteligentni roboti s višejezičnim sposobnostima mogli bi spašavati živote ljudi.

Velike su tržišne mogućnosti za uključivanje jezičnih tehnologija već otvorene u obrazovanju i industriji zabave, kao što su računalne igre, obrazovanje kroz zabavu, simulatorska okruženja ili programi obuke. Mobilne obavijesne usluge, strojno potpomognuti programi učenja jezika, okružja za e-učenje, alati za samoprocjenu i sustavi za pronalaženje plagijata samo su još neki od primjera gdje jezične tehnologije igraju značajnu ulogu. Popularnost društvenih mreža kao što su Twitter ili Facebook nagoviještaju dodatne potrebe za razrađenim jezičnih tehnologijama koje bi mogle nadgledati poruke, sažimati rasprave, predlagati opća kretanja u stavovima i mišljenima sudionika, otkrivati emocionalne afinitete, uočavati kršenje autorskih prava ili pratiti zloporabu.

Jezične tehnologije Europskoj uniji pružaju upravo nesagledive mogućnosti koje su i ekonomski i kulturno značajne. U Europskoj je uniji višejezičnost postala pravilom. Europsko poslovanje, ustanove i škole također su višenacionalne i raznolike. Građani žele komunicirati onkraj jezičnih granica koje još uvijek postoje na europskome zajedničkome tržištu. Jezične bi tehnologije mogle pomoći u nadilaženju tih preostalih prepreka uz potpomaganje slobodne i otvorene uporabe bilo kojega jezika. Nadalje, inovativne, višejezične jezične tehnologije nama bi Europljanima također pomogle u komunikaciji s našim globalnim partnerima i njihovim višejezičnim zajednicama. Jezične tehnologije postaju potpora u mnogim međunarodnim gospodarskim prigodama.

Višejezičnost je pravilo, a ne iznimka.

Izazovi koji stoje pred jezičnim tehnologijama

Premda su u nekoliko godina jezične tehnologije napravile znatan napredak, trenutačan je tempo tehnološkoga napretka i stvaranja novih proizvoda prespor. Ne možemo čekati deset ili dvadeset godina na značajna poboljšanja koja će olakšati komunikaciju ili produktivnost u našem višejezičnome okružju.

Trenutačna brzina tehnološkoga napretka preniska je za dosezanje bitno naprednijih programskih proizvoda u sljedećih deset do dvadeset godina.

Jezične tehnologije, koje su već u širokoj uporabi kao što su provjernici pravopisa ili gramatike u obradnicima teksta, tipično su jednojezične i dostupne su samo za ograničen broj jezika. Aplikacije za višejezičnu komunikaciju zahtijevaju stanovitu razinu složenosti. Strojno

prevođenje i on-line usluge kao Google Translate ili Bing Translator izvrsni su za stvaranje općega dojma o čemu je u nekom dokumentu riječ. Međutim, takve on-line usluge, ali i profesionalne programi za strojno prevođenje, bore se s mnogim poteškoćama kad su nam potrebni visokokvalitetni i potpuni prijevodi. Mnogo je dobro poznatih primjera smiješnih pogrješnih prijevoda, npr. doslovni prijevodu imena kao što su *Bush* ('grm') ili *Kohl* ('ugljen'), koji dobro ilustriraju izazove s kojima se jezične tehnologije moraju sučeliti.

Usvajanje jezika

Kako bismo prikazali kako se računala nose s prirodnim jezikom i zašto je usvajanje jezika za njih iznimno težak zadatak, pogledajmo na kratko kako ljudi usvajaju svoj prvi i ostale jezike, a potom ćemo u grubo objasniti ako djeluju sustavi za strojno prevođenje. Postoje jasni razlozi zašto je područje jezičnih tehnologija usko vezano uz područje umjetne inteligencije.

Ljudi usvajaju jezične sposobnosti na dva različita načina. Prvo, mala djeca usvajaju jezik slušanjem i praćenjem interakcija između govornika nekoga jezika. Njihova izloženost stvarnim jezičnim uporabama koje su izrekli njihovi roditelji, rođaci ili drugi, pomaže maloj djeci da u otprilike dvogodišnjoj dobi sami počinju proizvoditi riječi i kratke fraze. To je moguće samo zato jer ljudski rod već ima genetsku predispoziciju za usvajanje jezikā.

Ljudi usvajaju jezičnu sposobnost na dva različita načina: učeći na primjerima i učeći temeljna jezična pravila.

Učenje drugoga jezika obično traži više napora ukoliko dijete nije uronjeno u jezičnu zajednicu izvornih govornika. U školskoj se dobi strani jezici obično usvajaju učenjem njihove gramatičke strukture, rječnika i pravopisa iz knjiga i obrazovnih materijala koji opisuju jezično znanje u obliku apstraktnih pravila, tablica i primjera. Učenje stranoga jezika iziskuje mnogo vremena i napora, a sa starenjem postaje sve teže.

Dvije osnovne vrste jezičnotehnoloških sustava usvajaju jezične sposobnosti na sličan način kao i ljudi. Statistički pristupi stječu jezično znanje iz golemih zbirki pojedinačnih tekstnih primjera na jednome jeziku ili iz tzv. paralelnih tekstova koji su dostupni na dvama ili više jezika. Algoritmima strojnoga učenja modeliraju se određene jezične aktivnosti na temelju kojih se potom mogu prepoznati obrasci ispravne uporabe riječi, kratkih fraza ili čitavih rečenica na jednome jeziku ili u prijevodima s jednoga jezika na drugi. Međutim, za takve statističke pristupe potreban je ogroman broj rečenica. Kakvoća izvedbe takvih sustava raste s povećanjem broja analiziranih tekstova. Nije neuobičajeno trenirati takve sustave na tekstovima koji obuhvaćaju milijune rečenica, te je to jedan od razloga zašto sastavljači tražilica teže skupiti što je više moguće pisanoga teksta. Provjera pravopisa u obradnicima teksta, dostupne on-line obavijesti i prijevodne usluge kao što su Google Search ili Google Translate počivaju na statističkim (podatkovno orijentiranim) pristupima.

Sustavi temeljeni na pravilima predstavljaju drugu osnovnu vrstu jezičnih tehnologija. Stručnjaci s područja jezikoslovlja, računalnoga jezikoslovlja i računarstva kodiraju gramatičke analize (prijevodna pravila) i sastavljaju popise riječi (leksikone). Izgradnja sustava

temeljenoga na pravilima iznimno je vremenski i poslovno zahtjevna, a ne može se provesti bez visokospecijaliziranih stručnjaka. Neki od vodećih strojnoprevoditeljskih sustava temeljenih na pravilima u stalnom su razvoju već dvadesetak godina. Prednost sustava temeljenih na pravilima je u tome što stručnjaci imaju istančanije upravljanje obradbom jezika. Zbog toga je moguće sustavno ispravljati pogrješke u programskoj podršci i pružati korisniku podrobne povratne obavijesti, osobito kad se na pravilima temeljeni sustavi koriste za učenje jezika. Na žalost, zbog financijskih ograničenja jezične tehnologije temeljene na pravilima isplative su samo za jezike s velikim brojem govornika.

Hrvatski u europskome informacijskome društvu

Opće činjenice

Hrvatski jezik pripada zapadnoj južnoslavenskoj podskupini slavenske grane inoeuropske jezične porodice. U ovome trenutku hrvatski jezik broji preko 5,5 milijuna izvornih govornika. Hrvatski se jezik sastoji od narječja i nacionalnoga standardnoga jezika Hrvata, koji je službeni jezik više od 4 milijuna stanovnika Republike Hrvatske, a uz bošnjački i srpski također je jedan od tri službena jezika Bosne i Hercegovine gdje ga govori oko 700.000 govornika. Također, hrvatskim jezikom govore i mnogi pripadnici nacionalnih manjina u Hrvatskoj kao i autohtone hrvatske etničke i jezične manjine u Srbiji, Crnoj Gori, Sloveniji, Madžarskoj, Austriji, Slovačkoj i Italiji, koje ili obitavaju na teritorijima nekadašnjih hrvatskih zemalja ili su iselili tijekom stoljeća u povijesno uvjetovanim selidbama. Zbog znatne gospodarski i politički uvjetovane emigracije u 20. stoljeću i nakon Drugoga svjetskoga rata, hrvatski se također govori u mnogobrojnim hrvatskim zajednicama u cijelome nizu europskih i prekomorskih zemalja. Najveće hrvatsko gospodarsko iseljeništvo smješteno je u Njemačkoj, potom u SAD, Kanadi i Australiji. Aktivna uporaba hrvatskoga uglavnom ovisi o iseljeničkome naraštaju kojem govornici pripadaju. Pa ipak, u mnogim zemljama, osobito europskima, postoje dodatni školski programi programi na hrvatskome koje organizira i podupire hrvatska Vlada.

Službeni je status hrvatskoga jezika u Hrvatskoj određen Ustavom Republike Hrvatske. Prema Članku 12 Ustava: „U Republici Hrvatskoj u službenoj je uporabi hrvatski jezik i latinično pismo. 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.“ Kako se 2013. očekuje pristup Hrvatske Europskoj uniji, hrvatski će jezik postati 24. službeni jezik EU.

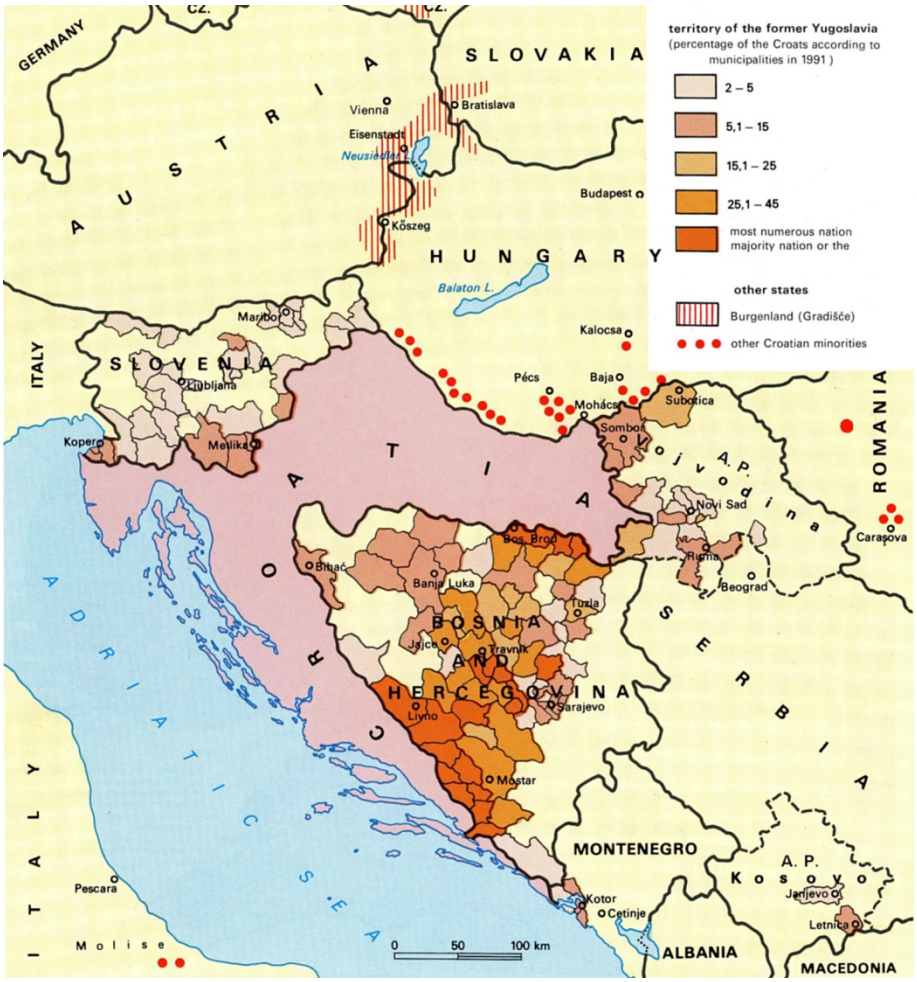
U Hrvatskoj još uvijek ne postoji jedinstven „jezični zakon“ koji bi regulirao službenu uporabu jezika u javnosti. Uvođenje zakona o jeziku pokušano je u nekoliko navrata od stjecanja hrvatske neovisnosti, ali niti u jednom slučaju nije dobivena dovoljna potpora hrvatske Vlade te niti jedan prijedlog nije ušao u saborsku proceduru. Zadnji se takav pokušaj dogodio u travnju 2010. Međutim, u zakonima o obrazovanju, sudskim postupcima itd. postoje članci koji reguliraju uporabu hrvatskoga kao službenoga državnoga jezika. Do sada, zakonodavstvo ne zahtijeva obvezatnu provjeru ili ispitivanje znanja hrvatskoga jezika kao uvjet za naturalizaciju. Zakon o hrvatskom državljanstvu[5] prepostavlja da stranac, koji traži stjecanje hrvatskoga državljanstva, poznaje hrvatski jezik i pismo.

Prema popisu stanovništva iz 2001. Hrvatska je imala 4.437.460 stanovnika od kojih su 89,63% Hrvati. Srbi su najzastupljenija nacionalna manjina s 4,54% stanovništva dok svaka od preostalih nacionalnih manjina zauzima manje od 0,5% stanovništva: Bošnjaci (0,47%), Albanci (0,34%), Slovenci (0,30%), Crnogorci (0,11%) i ostali u još manjim postotcima. Hrvatski je jezik materinski jezik za 96% stanovnika. Nacionalne manjine izjasnile su se kako govore sljedeće

jezike: albanski, bošnjački, bugarski, češki, hebrejski, madžarski, njemački, istrorumunjski, talijanski, makedonski, crnogorski, poljski, romski, rumunjski, ruski, rusinski, slovački, srpski, turski i ukrajinski. Jezici četiri manjine, srpski, madžarski, talijanski i češki, stekli su status jezika i pisma u službenoj uporabi u određenim područjima prema udjelu njihovih govornika u ukupnome stanovništvu koji mora iznositi barem 1/3 svih stanovnika na području lokalne samoupravi. Od 2009. u Hrvatskoj postoji 27 područja gdje nacionalne manjine imaju pravo službene uporabe vlastitoga jezika u lokalnoj administraciji. To se pravo u visokome omjeru primjenjuje u Istarskoj županiji gdje je talijanski materinski jezik 20.521 stanovnika, ali su dvojezični cestovni natpisi prisutni i u dijelovima gdje nema talijanske manjine. Republika je Hrvatska ratificirala Europsku povelju o regionalnim i manjinskim jezicima 1997. [6]

Još nije objavljena službena statistika o jezičnoj uporabi prikupljena nedavno provedenim popisom iz 2011. usklađenim s međunarodnim statističkim normama, pa je tako obuhvatio sve državljane Republike Hrvatske, strane državljane i apatride koji borave u Republici Hrvatskoj.

Hrvatska ima brojno iseljeništvo koje često još uvijek govori hrvatski jezik. Hrvatske etničke i jezične manjine žive u mnogim europskim zemljama kao posljedica povijesnih selidaba, započetih još u 16. stoljeću, kao nedavnih, mahom gospodarski i politički uvjetovanih. Najbrojnije skupine su tzv. Gradišćanski Hrvati u Austriji (pretpostavlja se oko 50.000), a otprilike sličan broj Hrvata živi u Madžarskoj. Gradišćanski se Hrvati u Austriji aktivno služe gradišćanskim hrvatskim. Ova je varijanta hrvatskoga jezika, standardizirana u skladu s ponešto drukčijim načelima od standardnoga hrvatskoga, jedan od austrijskih službenih manjinskih jezika. Čitav je niz dječjih vrtića i škola u Gradišću u kojima se rabi gradišćanski hrvatski. S druge strane, hrvatski standardni jezik službeni je manjinski jezik u Madžarskoj. U Italiji trenutačno živi oko 3.000 Hrvata koji se služe varijantom hrvatskoga zvanom moliški hrvatski i on se također uči u školama u tri općine nastanjene Hrvatima u Moliseu. Broj Hrvata u Srbiji, posebice u pokrajini Vojvodini gdje su Hrvati priznata nacionalna manjina, teško je točno utvrditi jer se dio etničkih Hrvata izjašnjava kao tzv. „Bunjevci“ uglavnom zbog političkih razloga. Premda je mnogo Hrvata izgnano iz Srbija nakon što je Hrvatska stekla svoju neovisnost od Jugoslavije, pretpostavlja se kako u Srbiji još uvijek živi više od 100.000 Hrvata. U ostalim europskim zemljama hrvatska autohtona manjina živi u Crnoj Gori (7.000 do 10.000), Češkoj (manje od 1.000), Slovačkoj (4.000) i Rumunjskoj (7.500). Broj Hrvata u Sloveniji je oko 50.000, ali samo je malen broj njih stvarna autohtona manjina, ponajprije u naseljima uz granicu, a većina ih predstavlja nedavno gospodarsko iseljeništvo. Hrvatski ima status manjinskoga jezika u Srbiji (kao jedan od sedam službenih jezika pokrajine Vojvodine), Crnoj Gori, Austriji, Madžarskoj i Italiji.



Slika 1: Hrvati u susjednim državama[7]

Hrvatska narječja

Slika hrvatskih narječja sastavljena je od tri narječne skupine: čakavske, kajkavske i štokavske. Mjesni govori koji pripadaju nekom od triju narječja govore se po cijeloj Republici Hrvatskoj. Sva hrvatska narječja pripadaju srednjo-južnoslavenskome dijasistemu slavenske jezične grane i u južnoslavenskom prostoru obuhvaća dio dijalekatnoga kontinuuma između slovenskoga tipa na sjeverozapadu i makedonsko-bugarskoga tipa na jugoistoku. Imena triju narječja izvedena su iz oblika upitne zamjenice *ča, kaj* i *što* (lat. *quid*). Međutim, na južnoslavenskome prostoru ta je klasifikacija relevantna samo za hrvatske dijalekte i rezultat je potreba hrvatske jezične zajednice. Slovenci koriste zamjenicu *kaj*, ali slovenski jezik ne pripada u *kajkavsko narječje*. Bošnjaci, Crnogorci, Srbi kao i Bugari, Makedonci i svi istočni Slaveni koriste *što*, ali njihovi jezici ne pripadaju u *štokavsko narječje* u istome smislu u kojem je štokavski hrvatsko narječje. Srbi, Crnogorci i Bošnjaci nemaju oblik te upitne zamjenice kao kriterij za razlikovanje svojih narječja. Kad je riječ o štokavskome, arhaični šćakavski (tzv. slavonski) govore samo Hrvati, novoštokavski ikavski i ijekavsko-šćakavski govore Hrvati i Bošnjaci, a novoštokavski ijekavski govore Hrvati iz širega dubrovačkoga područja, ali također i ostali južni Slaveni. Hrvati u Gradišću (Austrija, Madžarska, Slovačka) mahom govore čakavski, a rijetko štokavski ili kajkavski. Hrvati u talijanskoj po-

krajini Molise govore arhaičnim štokavskim dok Hrvati u Karaševu, Rumunjska govore torlačkim narječjem.

Uslijed mnogobrojnih, često prisilnih iseljavanja, prostorna se raspodjela pojedinih hrvatskih narječja stubokom promijenila od srednjega vijeka. Čakavski i kajkavski su u prošlosti bili raspoređeni na znatno širem području, ali danas prevladava štokavsko narječje. Prije iseljavanja čakavsko se narječje rabilo na sjeveru do rijeka Kupe i Save a na istoku do crte Una-Dinara-Cetina. Nakon migracija čakavsko je narječje ograničeno na obalno područje i otoke, dok su se čakavski govori u unutrašnjosti počeli razlikovati prema količini štokavskoga utjecaja. Kajkavsko se narječje također nekad prostiralo istočnije gdje danas prevladava štokavsko.

Čakavsko, kajkavsko i štokavsko narječje razlikuju se na svim jezičnim razinama: fonološkoj, morfološkoj, sintaktičkoj i leksičkoj i svaka od tih razina uključuje brojne arhaizme, ali i inovacije karakteristične za određeno narječje.

Slika 2: Zemljovid hrvatskih narječja u Republici Hrvatskoj

Standardizacija hrvatskoga jezika

Tisućljetna povijest hrvatskoga jezika potvrđena je tekstovima pisanim još krajem 10. stoljeća ili početkom 11. stoljeća, u vrijeme kad su se tri hrvatska narječja (čakavski, štokavski i kajkavski) počela oblikovati. Sva su tri hrvatska narječja odigrala važnu ulogu u stvaranju hrvatskoga književnoga jezika (različitih narječnih osnovica) i oblikovanju hr-

vatske jezične kulture koja je dovela do standardnoga hrvatskoga jezika izgrađenoga na štokavskoj osnovici.

Prvi jasni pokušaj oblikovanja hrvatskoga standardnoga jezika pojavio se u 17. stoljeću kad je većina hrvatske etničke zajednice — osobito nakon gramatike i drugih djela Bartola Kašića (1575-1650) i rascvjetale renesansne i barokne književnosti štokavskoga Dubrovnika — prepoznala jezičnu strukturu štokavskoga (isprva s ikavskim refleksom *jata*, ali kasnije s *jekavskim*) kao najbolje polazište za sastavljanje nadregionalnoga hrvatskoga književnoga jezika. Unatoč odabiru jedne jezične osnovice za sastavljanje svoga standardnoga jezika, Hrvati nisu odbacili postignuća višestoljetne jezične kulture različitih narječnih osnovica unutar hrvatskoga književnoga jezika (kajkavsko-štokavsko-čakavski hibrid) koji je obilježio i povijest hrvatske etničke zajednice. Premda je standardizacija jezika Hrvata temeljena na štokavskome narječju započela vrlo rano, narodno se jezično jedinstvo postiglo tek u vrijeme Ilirskoga narodnoga preporoda (počevši od 1835.) kad je mala skupina Hrvata, koji su se do tada služili kajkavskim idiomom, također prihvatili štokvski hrvatski standardni jezik. Tijekom većine 20. stoljeća hrvatski se standardni jezik rezvijao u različitim južnoslavenskim državnim jedinicama pod različitim imenima, a bio je predstavljan kao varijanta tzv. hrvatsko-srpskoga (srpsko-hrvatskoga) jezika. To je napušteno s demokratskim društveno-političkim promjenama 1990.

Različite stilizacije hrvatskoga jezika oblikovane su još davno u iseljeništvu (npr. gradišćanski hrvatski, moliški hrvatski).

Hrvatska je pisana kultura obilježena uporabom triju pisama (glagoljica, ćirilica, latinica) među kojima je latinica među Hrvatima prevladava od 16. stoljeća. Njezina uporaba nije bila normirana niti ususstavljena sve do 1835. kad je Ljudevit Gaj dao hrvatskoj latinici današnji oblik.

Osobine hrvatskoga jezika

Fonetika, fonologija, morfonologija

Fonemski inventar hrvatskoga standardnoga jezika sastoji se od 5 samoglasnika (*a, e, i, o, u*) i 25 suglasnika (*m, v, n, l, r, j, nj, lj, p, b, f, s, z, c, t, d, ć, đ, š, ž, č, dž, h, k, g*). Akustične i artikulacijske osobine samoglasnika ne mijenjaju se s obzirom na mjesto izgovora (bez obzira nalazi li se u kratkom, dugom, naglašenom ili nenaglašenom slogu). Uz tih 5 samoglasnika postoji i samoglasničko *r* (*crn* 'niger') i dvoglas *ie*, koji se u pismu bilježi kao *je/ije* (*djelo, odijelo*).

Naglasni sustav sastoji se od 4 naglaska (dva duga naglaska s uzlaznim i silaznim tonom i dva kratka naglaska s uzlaznim i silaznim tonom) i zanaglasne dužine. Naglasni je sustav standardnoga hrvatskoga jezika novoštokavski premda danas postoje mnoga odstupanja od naglasnih modela kodificiranih u drugoj polovici 19. stoljeća. Mjesto naglaska nije vezano uz pojedini slog, nego raspodjela naglasaka podliježe stanovitim ograničenjima (npr. zadnji slog višesložne riječi u načelu ne može biti naglašen, silazni naglasci ostvaruju se samo na prvome slogu riječi koje nisu složenice, itd.) Ova se pravila krše u svakodnevnome govoru, osobito u velikim gradskim središtima koja su smještena izvan novoštokavskoga područja (npr. *kontinuitêt* / *kontinuìtēt*). Naglasak i

dužina mogu se povremeno rabiti za razlikovanje značenja između leksičkih jedinica ili njihovih oblika, npr. *gr'ad* : *grâd*, *žènē* (gen. jd.) : *žène* (nom. mn.).

U hrvatskome neke riječi nemaju vlastiti naglasak (naslonjenice) već u naglasnoj riječi prednaglasnice mogu preuzeti naglasak prenesen s naglašene riječi ukoliko je naglasak silazni i na prvom je slogu (*grâd* : *'u grād*). Kod zanaglasnica to nije moguće. Prenošenje naglaska na prednaglasnicu postaje sve rjeđe, osobito u gradskim središtima izvan neoštokavskoga područja.

U hrvatskome se standardnome jeziku nalaze mnoge fonološki (nom. jd. *sla**d**ak* : gen. jd. *sla**t**koga*, nom. jd. *dio* : gen. jd. *dijela*) i morfonološki uvjetovane promjene (nom. jd. *maj**k**a* : dat. jd. *maj**c**i*, nom. jd. *juna**k*** : vok. jd. *juna**če***).

Regionalna primjena hrvatskoga standardnoga jezika često je u govoru pod utjecajem lokalnoga narječja, npr. na čakavskome Kvarneru prevladava zatvorno *t'* umjesto bezvučnoga poluzatvornoga *ć*, ili u sjeverozapadnome kajkavskome znakovito je nerazlikovanje između *č* – *ć* ili *đ* – *dž*.

Morfologija

Hrvatski standardni jezik razlikuje 10 vrsta riječi od kojih su pet promjenljive (imenice, pridjevi, zamjenice, brojevi, glagoli), četiri nepromjenljive (prijedlozi, veznici, uzvici, čestice), a prilozi su promjenljivi samo u komparaciji.

Gramatičke kategorije koje se nalaze u većine promjenljivih riječi jesu rod (tri vrijednosti: muški, ženski, srednji), broj (dvije vrijednosti: jednina, množina), padež (sedam vrijednosti: nominativ, genitiv, dativ, akuzativ, vokativ, lokativ, instrumental). Neke sklonjive riječi imaju i neke posebne kategorije (npr. određenost se obilježava na pridjevima zasebnim nizom flektivnih nastavaka; živo/neživo se obilježava odabirom nastavka za akuzativ jednine imenica muškoga roda; imenice mogu biti konkretne, tvarne, kategorijalne ili zbirne; itd.). Konjugirane riječi (glagoli) obilježene su kategorijama: načina (četiri vrijednosti: indikativ, imperativ, kondicional, optativ), lica (tri vrijednosti: prvo, drugo, treće), broja (dvije vrijednosti: jednina, množina), stanja (dvije vrijednosti: aktiv, pasiv) i vremena (sedam vrijednosti: prezent, imperfekt, aorist, perfekt, pluskvamperfekt, futur I., futur II.). Glagoli *biti* ('esse') and *htjeti* ('volere') u hrvatskome su pomoćni glagoli. Glagoli također posjeduju složen sustav glagolskih vidova (svršeni i nesvršeni s dodatnim podvrijednostima kao što su početni, učestali itd.), a mogu uključivati i osobinu prijelaznosti. Pridjevi i prilozi mogu se pojaviti i u kompariranim oblicima (tri vrijednosti: pozitiv, komparativ, superlativ).

U hrvatskome postoje dvije osnovne vrste sklonidbe: imenična sklonidba (imenice i neodređeni oblici pridjeva) i zamjenično-pridjevska sklonidba (zamjenice, određeni oblici pridjeva, brojevi). Svaki imenični rod ima svoju sklonidbu (a-vrsta za muški i srednji rod, e-vrsta za ženski), a postoji i posebna i-vrsta (imenice ženskoga roda).

imenična sklonidba	nom. i gen. jd.	nom. mn.

a-vrsta muški rod	*opis, opis**a***	*opis**i***
a-vrsta srednji rod	*sunce, sunc**a***	*sunc**a***
e-vrsta ženski rod	*žen**a**, žen**e***	*žen**e***
i-vrsta ženski rod	*noć, noć**i***	*noć**i***

Nastavci za zamjenično-pridjevsku sklonidbu prikazani su u ovoj tablici:

padež	**muški rod**	**srednji rod**	**ženski rod**
jednina			
N	-i	-o -e	-a
G	-og(a) -eg(a)		-e
D	-om(u/e) -em(u/e)		-oj
A	= N / = G	= N	-u
V	= N	= N	= N
L	-om(u/e) -em(u/e)		= D
I	-im		-om
množina			
N	-i	-a	-e
G	-ih		
D	-im(a)		
A	-e	= N	= N
V	= N	= N	= N
L	= D		
I	= D		

Riječi se u hrvatskome tvore derivacijom i slaganjem. Postoji nekoliko različitih načina tvorbe riječi: sufiksalna (*star-ac*), prefiksalno-sufiksalna (*do-život-an*), nesufiksalno slaganje (*plačidrug*), sufiksalno slaganje (*vanjskopolitički*), srastanje (*uz-brdo*), slaganje pokrata (*Varteks*) i pretvorba (*mlada*). Najčešća je sufiksalna tvorba.

Rječnik, frazeologija, nazivlje

Temeljni se leksički sloj hrvatskoga standardnoga jezika, osim praslavenskoga leksičkoga nasljeđa, sastoji od štokavskoga vokabulara uz primjese vokabulara drugih hrvatskih narječja i vokabulara naslijeđenoga iz književnoga jezika raznih dijalekatnih stilizacija starijega podrijetla (npr. iz kajkavskoga *kukac, hlače, rječnik*, ili iz čakavskoga *spužva*). Pored toga, cjelina hrvatskoga jezika bila je stalno izložena izravnim ili neizravnim dodirima s drugim jezicima i kulturama. Hrvatski se jezik ističe između ostalih južnoslavenskih jezika znatnim leksičkim utjecajima pristiglim iz romanskih jezika (supstratni tragovi dalmatskoga jezika jesu npr. *jarbol, tunj*). Talijanski je jezik bio utjecajan u priobalju (osobito u dijelovima pod negdašnjom mletačkom dominacijom), a njemački i do neke mjere madžarski, u kontinentalnoj Hrvatskoj.

Crkvenoslavenski je književni jezik ostavio tragove u starijim razdobljima hrvatskoga jezika, ali nije imao značajnijega utjecaja tijekom razdoblja u kojem se oblikovao standardni jezik. Ruski jezik nije ostavio tako dubokoga traga u hrvatskome kao što je to učinio u susjednome srpskome standardnome jeziku. Utjecaj vokabulara klasičnih jezika (latinskoga i grčkoga) sveprisutan je u hrvatskoj kulturi, a ponajprije u intelektualnom vokabularu i znanstvenome nazivlju. Tijekom razdoblja srednjohrvatskoga jezika (16.-18. stoljeće) intenzivno su u hrvatski ulazile posuđenice iz turskoga, osobito riječi za predmete iz svakodnevnoga života. Važno je napomenuti kako zbog ranoga iseljavanja, u gradišćanskome hrvatskome nema turskih posuđenica, pa čak niti onih koje se u standardnome hrvatskome više niti ne osjećaju stranim riječima (npr. b*ubreg, čizma, jastuk* itd.). Umjesto tih riječi u gradišćanskome hrvatskome rabe se starije hrvatske riječi zajedničkoga slavenskoga podrijetla, te je stoga on vrlo bitan za uvid u povijest hrvatskoga leksičkoga inventara. Njemački i francuski nekad su također utjecali na hrvatski vokabular, a od druge polovice 20. stoljeća utjecaj engleskoga jača. Češki, premda ne u izravnome kontaktu, imao je značajan utjecaj na hrvatski vokabular u nekoliko navrata, osobito tijekom 19. stoljeća za vrijeme izgradnje stručnoga nazivlja koju je bio izveo Bogoslav Šulek (npr. *časopis*, *kisik, dušik, vodik*). Za vrijeme Jugoslavija, na hrvatski je utjecao i srpski, a osobitu je za to zaslugu imala federalna administracija. Purističke težnje u vokabularu pojavljivale su se od vremena do vremena od 16. do 20. stoljeća (npr. Zoranić, Ritter Vitezović, Reljković, razdoblje 1941.-1945.).

Kontinuitet od davnih vremena do suvremenoga hrvatskoga standardnoga jezika i sudjelovanje triju narječja u izgradnji hrvatskoga standardnoga jezika može se uočiti u njegovoj razvijenoj i bogatoj frazeologiji (npr. u svojim umjetničkim tekstovima iz 16. stoljeća Marulić rabi frazem *zgubiti glas* = 'biti postiđen, izgubiti lice', dok Zoranić rabi frazem *u magnutje oka* = 'odmah', koji su gotovo isti kao frazemi *izgubiti glas* i *u trenu oka* u današnjemu hrvatskome standardnome jeziku temeljenome na štokavskoj osnovici).

Nazivlje u pojedinim stručnim područjima započela se razvijati već u 16. stoljeću, a to je potvrđeno mnogobrojnim hrvatskim (ponajviše višejezičnim) rječnicima sastavljenim od 16. do 20. stoljeća. U 19. stoljeću njemački i češki imali su iznimno jak utjecaj na hrvatsko nazivlje, a engleski je danas preuzeo tu ulogu.

Sintaksa

Hrvatski jezik pripada skupini jezika obilježenih SVO sintaktičkom strukturom (*Marija voli Ivana*) i relativno slobodnim redom riječi (mnogobrojne permutacije sastavnica moguće su uz neka ograničenja kao što je smještaj nenaglasnica). Glede informacijske strukture rečenica, temeljno je pravilo u stilistički neutralnome diskursu da se na prvo mjesto smješta *tema* (stara obavijest), a slijedi u *rema* (nova obavijest, primjedba).

Subjekt u rečenici ne mora biti izrijekom naveden, a njegovo je ispuštanje poželjno ukoliko bi ga se trebalo ponavljati više puta unutar neposredne okoline. Obvezatna je dvostruka negacija (*Nitko ga nije*

volio). Sročnost sastavnica u rodu, broju i padežu je tipična za strukturu hrvatskih rečenica.

U hrvatskome standardnome jeziku postoji sedam padeža, a oblici se mogu kombinirati s prijedlozima (obvezatni uz lokativ). Bitna odrednica hrvatskih glagola jest vid, a glagolski oblici također izražavaju glagolsko vrijeme i modalna značenja. Organizacija složenih rečenica može biti nezavisna ili zavisna (uz prisutnost veznika ili bez njih). Novija je pojava u suvremenome jeziku ograničenje uporabe zajedničkoga slavenskoga genitiva (*Nije volio vina*), posvojne genitivne konstrukcije izbjegavaju se u korist posvojnih pridjeva (*majčina kuća* umjesto *kuća majke*), a uporaba prošlih vremena (imperfekt, aorist i pluskvamperfekt) je sve ograničenija. U suvremenome su hrvatskome pasivne konstrukcije znatno rjeđe nego u starijem hrvatskome.

Pravopis

Premda je povijest hrvatske kulture obilježena uporabom triju pisama (glagoljica, ćirilica, latinica), latinica u Hrvata prevladava od 16. stoljeća. Hrvatska latinična abeceda nije bila u cijelosti standardizirana do 1835. kad joj Ljudevit Gaj daje današnji oblik. Sastoji se od 30 slova, od koji su tri dvoslovi (*dž*, *lj*, *nj*), a ostala su jednoslovi od čega pet s dijakritičkim znacima (*č*, *ć*, *đ*, *š*, *ž*). U akademskim krugovima, osobito pri tiskanju tekstova hrvatske pismene baštine, dvoslovi *dž*, *lj* i *nj* se mogu zamijeniti s *ǵ*, *ļ* and *ń*. Slova *q*, *x*, *y*, *w* ne postoje izvorno u hrvatskoj abecedi premda se rabe za pisanje stranih imena.

velika slova														
A	B	C	Č	Ć	D	Dž	Đ	E	F	G	H	I	J	K
L	Lj	M	N	Nj	O	P	R	S	Š	T	U	V	Z	Ž
mala slova														
a	b	c	č	ć	d	dž	đ	e	f	g	h	i	j	k
l	lj	m	n	nj	o	p	r	s	š	t	u	v	z	ž

Hrvatski je pravopis fonološko-morfonološki jer predstavlja stapanje dvaju pravopisnih načela: nadređenoga fonološkoga (npr. bilježenje asimilacije) i podređenoga morfonološkoga (npr. *podcrtati*). Razmak između riječi je logički, a ne gramatički (kakav je bio nekada). Za hrvatski je pravopis tipično da se pisanje stranih imena ne prilagođuje izgovoru ili grafemskom sastavu hrvatske abecede, a i oblični se nastavci uklapaju u čitavu riječ (npr. *John*, a ne *Džon*; *Washington*, a ne *Vašington*; *Johna*, a ne *John-a*).

Onomastika

Hrvatska imena predstavljaju važne spomenike jezičnoga, kulturnoga i društvenoga nasljeđa ljudi koji su ih napravili. Stoga i osobna imena (antroponimi) i imena mjesta (toponimi) čine važan dio hrvatske jezične kulture. Ozemlje današnje Hrvatske, u grubo ograničeno rijekom Dravom na sjeveru, rijekom Dunavom na istoku i Jadranskim morem na jugu, vrlo se ilustrativno reflektira u bogatom raslojavanu zemljopisnih imena. To obilno raslojavanje u hrvatskoj toponimiji odraz je višestoljetnoga suživota različitih etničkih skupina koje su nastanjivale istočnu obalu Jadrana i njegovo zaleđe u povijesti. Stoljeća

jezičnoga prožimanja i stapanja različitih kulturnih tradicija ostavila su neizbrisiv trag u hrvatskoj toponimiji. Štoviše, potvrđena imena mjesta često su najstarijim svjedocima najstarijih promjena u samome hrvatskome jeziku. Kako se hrvatski jezik razvijao preko vjerskih (pretkršćanstvo i kršćanstvo), kulturnih i civilizacijskih granica, tragovi Istoka i Zapada mogu se uočiti u hrvatskim imenima. Kad je riječ o imenima osoba, Hrvati su prvi slavenski narod koji je uveo prezimena (od 12. stoljeća) uzduž jadranske obale uslijed izravnoga romanskoga kulturnoga utjecaja. Najstariji sloj hrvatskih imena oblikovan je u skladu s praslavenskim imenskim obrascima koji su pak slijedili zajedničke indoeuropske obrasce oblikovanja imena. Patronimici još uvijek čine najveći dio inventara prezimena, ali za razliku od ruskoga, danas više nisu produktivni i ostaju neizmijenjeni kao zamrznuta prezimena koja su uklopljena u flektivni sustav kao imenice. U suprotnosti s hrvatskim toponomastičkim sustavom gdje gotovo da i nema turskoga utjecaja, mnoga su hrvatska prezimena oblikovana iz turskih posuđenica hrvatskim tvorbenim nastavcima. Tome je razlog činjenica kako je većina prezimena u Hrvatskoj stvorena nakon tridentinskoga koncila u 16. stoljeću, u vrijeme kad je velik dio hrvatskih zemalja bio pod turskom vlašću.

Odnos hrvatskoga standardnoga jezika s ostalim jezicima štokavske osnovice

Četiri nacionalna jezika, hrvatski, srpski i od nedavna, bošnjački i crnogorski, svi dijele štokavsku strukturnu osnovicu, međutim, tradicije i nadstrukture ovih jezika su poprilično različite. Što je razlikuje hrvatsku jezičnu povijest i kulturu od ostalih južnoslavenskih jezika jest odnos između svih triju narječja (kajkavsko, čakavsko, štokavsko) koji odnos postojano obogaćuje hrvatski standardni jezik štokavske osnovice. Zbog različitih polaznih uvjeta (nepostojanje osnovnoga, zajedničkoga standarda) i različitih tradicija u jezičnome kultiviranju i standardizaciji, zbog razjedinjenja neoštokavskih struktura i razlika u jezičnim nadstrukturama jedan zajednički monolitni standarni jezik nikad nije bio uspio biti oblikovan tijekom postojanja jugoslavenskih država, premda je postojalo nekoliko pokušaja političkoga nametanja zajedničkoga imena jezika (*srpsko-hrvatsko-slovenački* za Kraljevine Jugoslavije; *srpsko-hrvatski* ili *hrvatsko-srpski* za komunističke Jugoslavije). Za vrijeme Drugoga svjetskoga rata i nekoliko godina nakon njega svi su službeni dokumenti u Jugoslaviji objavljivani na četiri službena jezika (hrvatskome, makedonskome, slovenskome, srpskome), no ubrzo je mnogo političkoga napora uporabljeno za ponvonu konvergenciju hrvatskoga i srpskoga. Unatoč svim pokušajima da se službeno prizna postojanje hrvatskoga kao zasebnoga jezika, nametanje zajedničkoga nazivlja, vokabulara, pravopisa i drugih jezičnih normi u Jugoslaviji, dovelo je jedino do službenoga prihvaćanja jednoga zajedničkoga standardnoga jezika (*srpsko-hrvatskoga*) s dvije varijante (*istočnom* ili *srpskom* i *zapadnom* ili *hrvatskom*). Reakcija iz Hrvatske došla je ubrzo u obliku *Deklaracije o nazivu i položaju hrvatskog književnog jezika* koja se otvoreno zalagala za priznavanje samostalnoga hrvatskoga jezika i koju su jednoglasno 1967. potpisale vodeće znanstvene, kulturne i obrazovne ustanove, kao i vodeći intelektualci diljem Hrvatske, a koji

su se tako otvorenim političkim potezom nesumnjivo doveli u opasnost u komunističkim vremenima.

Tijekom zadnjih 20 godina, četiri štokavski temeljena standardna jezika razvijaju se samostalno kao nacionalni standardni jezici u prirodno divergentnim smjerovima budući da ne postoji nikakav sporazum ili koordinacija glede njihovoga zajedničkoga normiranja, pa su se time među njima razlike uvećale.

Skrb o jeziku u Hrvatskoj

Vijeće za normu hrvatskoga standardnoga jezika ustanovljeno je odlukom Ministarstva znanosti, obrazovanja i športa, 14. travnja 2005. Njegova je temeljna zadaća sustavna i znanstveno utemeljena skrb o hrvatskome standardnome jeziku. Posebni zadatci Vijeća su:

- skrb o hrvatskome standardnome jeziku;
- raspravljati o aktualnim nedoumicama i otvorenim pitanjima hrvatskoga standardnog jezika;
- upozoravati na primjere nepoštivanja ustavne odredbe o hrvatskome kao službenome jeziku u Republici Hrvatskoj;
- promicati kulturu hrvatskoga standardnog jezika u pisanoj i govornoj komunikaciji;
- skrbiti o statusu i ulozi hrvatskoga standardnoga jezika u svjetlu integracije Hrvatske u Europsku uniju;
- donositi odluke u daljnjem procesu standardizacije hrvatskoga standardnoga jezika;
- brinuti o jezičnim pitanjima i postavljati načela za pravopisnu standardizaciju.

Vijeće za normu hrvatskoga standardnoga jezika sastaje se redovite i kroz temeljite rasprave dolazi do zaključaka. Institut za hrvatski jezik i jezikoslovlje udomljuje Vijeće, pruža mu tehničku i administrativnu podršku kao i jezikoslovne savjete kad je to potrebno.

Institut za hrvatski jezik i jezikoslovlje[8] središnja je hrvatska ustanova za istraživanje hrvatskoga jezika, a jedan je od njegovih odjela (Odjel za hrvatski standardni jezik) posvećen opisu hrvatskoga standardnoga jezika s osobitom pozornošću na jezičnu kulturu (npr. pružanje javnosti jezičnih savjeta ili pisanje jezičnih priručnika). Savjeti o ispravnoj jezičnoj uporabi i jezikoslovna ekspertiza stalne su dužnosti Instituta. Savjeti se daju telefonski, e-poštom ili u pisanome obliku. Uz to, odgovori na najčešće postavljana pitanja dostupni su na portalu Jezični savjeti[9] u sastavu institutova www-sjedišta.

Institutov projekt STRUNA[10], unutar kojega se razvija hrvatsko stručno nazivlje zaslužuje posebno spominjanje. Cilje je ovoga projekta uspostava sustava koordinacije terminoloških poslova u svim stručnim područjima u Hrvatskoj i time pripomoći poboljšanju kakvoće i učinkovitosti višega obrazovanja i znanstvenih istraživanja izgradnjom jedinstvenoga provjerenoga nazivlja koje mogu rabiti stručnjaci svih polja, a i zainteresirani pojedinci i opće javnosti. Također se planira uspostava mreže istraživačkoga nazivlja kao i znanstvena suradnja između ustanova koje se bave različitim vidovima terminološkoga rada.

Pored toga, ostale hrvatske znanstvene ustanove (nekoliko sveučilišta s njihovim odsjecima za hrvatski jezik i književnost) i kulturne ustanove

(kao što je *Matica hrvatska*) također sudjeluju u skrbi o hrvatskome jeziku. Javna glasila, kao što su državna radio-televizija i neki novinski nakladnici imaju dobro razvijene korektorske i lektorske službe za hrvatski standardni jezik, te obraćaju posebnu pozornost na kakvoću jezika koji rabe u svojoj proizvodnji javno dostupnih tekstova.

Jezik u obrazovanju

Hrvatski je jezik službeni u svim osnovnim i srednjim školama osimu u područjima sa pučanstvom nacionalnih manjina. Međutim, nije određen kao obvezatan na sveučilištima. Premda u Hrvatskoj postoje težnje, posebice u prirodnim znanostima, da se predavanja održavaju na engleskome jeziku koje se težnje opravdavaju tvrdnjama kako je to svrhovito i korisno, sasvim je jasno da bi bilo iznimno štetno i neprihvatljivo ne poučavati na hrvatskome na sveučilištima. To bi imalo razarajući učinak na razvoj hrvatskoga znanstvenoga nazivlja i stručne frazeologije. Stoga je Vijeće za normu hrvatskoga standardnoga jezika preporučilo Ministarstvu da službeno odredi uporabu jezika u visokome obrazovanju.

U osnovnim i srednjim školama Hrvatski jezik i književnost poučava se kao predmet koji zauzimlje značajan dio školskih sati. Kao dio toga predmeta proučava se hrvatska gramatika, rječnik i književnost, a razvijaju se također pismeno i govorno izražavanje na hrvatskome jeziku. PISA testiranje, koje provjerava vještine učenika na svjetskoj razini, provodi se u Hrvatskoj od 2006, a prvi rezultati provjera pokazuju kako hrvatski petnaestogodišnjaci zauzimaju 26. mjesto na ljestvici svih zemalja svijeta i smješteni su ispred učenika iz deset zemalja-članica EU i SAD.

U osnovnim i srednjim školama uz hrvatski obvezatno je učenje barem jednoga stranoga jezika od četvrtoga razreda. Međutim, engleski se jezik (rijetko francuski ili njemački) nerijetko uče već u dječjem vrtiću. Engleski je uobičajeno prvi strani jezik u osnovnoj školi. Najrašireniji drugi strani jezik je njemački, potom talijanski i francuski. U srednjim školama ponekad se uče ruski i španjolski kao drugi ili treći strani jezik. Latinski i starogrčki uče se u klasičim programima koji počinju u petome razredu osnovne škole. K tome je latinski obvezatan u svim humanističkim srednjim školama. U školi židovske manjine (koja ima pravo javnosti), moguće je učiti i hebrejski. Obrazovanje na manjinskim jezicima dostupno je od dječjega vrtića do srednje škole i hrvatska ga Vlada financira za srpsku, češku, madžarsku i talijansku manjinu.

Međunarodni odnosi

Uporaba hrvatskoga standardnoga jezika u zemljama regije regulirana je zakonima tih zemalja. Status hrvatskoga standardnoga jezika kao jednoga od službenih jezika susjedne Bosne i Hercegovine od osobite je važnosti, pa hrvatske ustanove posvećuju osobitu pozornost suradnji sa znanstvenim i kulturnim ustanovama hrvatskoga naroda u Bosni i Hercegovini. Također, kulturne ustanove iz Republike Hrvatske uspostavljaju suradnju s mnogim hrvatskim iseljeničkim ustanovama diljem svijeta.

Poučavanje hrvatskoga jezika organizirano je u inozemnim školama za djecu hrvatskih državljana koji privremeno ili trajno žive u drugim zemljama. Hrvatski se jezik poučava na mnogim inozemnim ustanovama i u središtima za slavenske jezike (tako postoji 36 službenih razmjenskih lektorata za hrvatski jezik i književnosti kao i dva središta za Hrvatske studije u Australiji i Kanadi koje sve podupire Ministarstvo znanosti, obrazovanja i športa Republike Hrvatske). Velik broj središta za učenje hrvatskoga kao drugoga ili stranoga jezika postoji u Hrvatskoj, a najpoznatiji je *Croaticum*[11].

Hrvatski na Internetu

Statistika Državnoga zavoda za statistku o uporabi informacijskih i komunikacijskih tehnologija u poduzećima i kućanstvima dana je u sljedećoj tablici:

Uporaba informacijskih i komunikacijskih tehnologije (ICT) u poduzećima (%)			
	2008	2009	2010
uporaba računala	98	98	97
pristup Internetu	97	95	95
www-sjedište	64	57	61
uporaba financijskih i bankovnih usluga	84	84	85
uporaba usluga e-uprave	56	61	63

Kućanstva opremljena informacijskim i komunikacijskim tehnologijama (ICT) (%)			
	2008	2009	2010
osobno računalo	53	55	60
pristup Internetu	45	50	57
mobilni telefon	81	82	–

Najposjećenija hrvatska www-sjedišta su: *net.hr* (portal za vijesti, sport, zabavu i zbivanja), *index.hr* (opći www-portal, informacije, usluge, vijesti, sport, zabava, vozila, gastronomija), *jutarnji.hr* (www-sjedište dnevnih novina"Jutarnji list"), *24sata.hr* (www-sjedište dnevnih novina "24 sata"), *tportal.hr* (portal HT-a, Hrvatskih telekomunikacija), *njuskalo.hr* ("Njuškalo" portal s oglasima), *vecernji.hr* (www-sjedište dnevnih novina "Večernji list"), *forum.hr* (najveći hrvatski www-forum na kojem se raspravlja o temama iz društva, kulture, zabave itd.). Sedam dnevnih novina svakodnevno objavljuje svoje članke i na vlastitim www-sjedištima pored papirnatih izdanja.

Institut za hrvatski jezik i jezikoslovlje održava www-stranicu o hrvatskome jeziku koja donosi iscrpan popis hrvatskih jedno- i višejezičnih rječnika, gramatika i pravopisnih priručnika. Na

Filozofskome fakultetu Sveučilišta u Zagrebu održava se slična www-stranica[12].

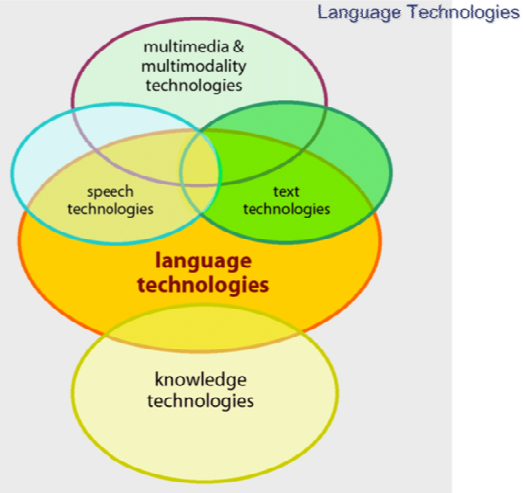
Wikipedija na hrvatskome jeziku osnovana je 2003. i trenutačno broji 100.708 članaka, te je 30. Wikipedija po službenome broju članaka.

Pristup jezičnim resursima za hrvatski jezik u zadnje je vrijeme olakšan zbog broja hrvatskih ustanova i organizacije koje provode postupke digitalizacije (uključujući značajne projekte koje podupire Ministarstvo znanosti, obrazovanja i športa i Ministarstvo kulture u digitalizaciji hrvatske kulturne baštine), a koja je uvećala vidljivost hrvatskoga jezika među ostalim internetskim izvorima.

Jezičnotehnološka podrška za hrvatski

Jezične tehnologije

Jezične su tehnologije specijalizirane tehnologije za obradbu prirodnoga jezika na razmeđi humanističkih, društvenih i tehničkih znanosti. Te se tehnologije mogu pojaviti i pod nazivom Prirodnojezične tehnologije. Prirodni se jezik pojavljuje u govorenom i pisanom obliku. Dok je govor najstariji i najprirodnij oblik jezične komunikacije, složene obavijesti i većina ljudskoga znanja sadržana je i prenosi se putem teksta. Govorne tehnologije i tehnologije obradbe teksta obrađuju jezik u ovim dvama načinima njegova ostvaraja. Unutar jezika, međutim, postoje elementi koji se dijele između govora i teksta kao što su rječnici, većina gramatike i rečenična značenja. Stoga se velik dio jezičnih tehnologija ne može se podvesti samo pod govorne tehnologije ili tehnologije obradbe teksta. Među njima su i tehnologije koje povezuju jezik sa znanjem. Slika 3 (desno) prikazuje okružje jezičnih tehnologija. U našem priopćavanju miješamo jezik s ostalim oblicima priopćavanja i drugih medija. Kombiniramo govor s gestikulacijom i mimikom. Digitalni tekst kombiniramo sa slikama i zvukovima. Filmovi mogu sadržavati jezik u govorenom ili pisanom obliku. Stoga se govorne tehnologije i tehnologije obradbe teksta preklapaju i prožimaju se s mnogim drugim tehnologijama koje pospješuju obradbu višekanalnoga priopćavanja i multimedijskih dokumenata.

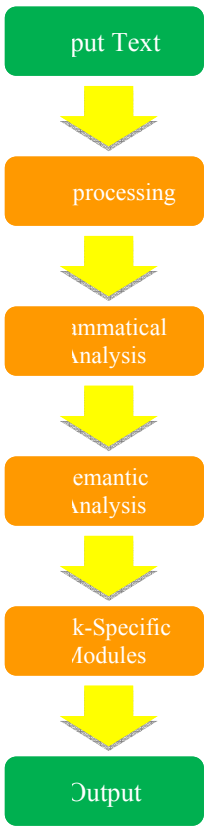


Arhitekture jezičnotehnoloških aplikacija

Tipična programska aplikacija za obradbu jezika sastoji se od nekoliko dijelova koji se bave različitim jezičnim slojevima i od osnovne zadaće za koju je ta aplikacija razvijena. Slika 4 (desno) pokazuje znatno pojednostavnjenu arhitekturu kakva se može naći u sustavima za obradbu teksta. Prva tri modula bave se strukturom i značenjem ulaznoga teksta:

- predobradba: čišćenje podataka, uklanjanje oblikovanja, prepoznavanje ulaznoga jezika, ponekad umetanje nedostajućih dijakritičkih znakova u hrvatskome, itd.
- gramatička raščlamba: pronalaženje glagola i njegovih objekata, subjekata, njihovih atributa itd.; prepoznavanje rečenične strukture.
- semantička raščlamba: razobličenje (Koje značenje riječi „glava“ je ispravno u danom kontekstu?), razrješenje anafore i odnosnih izraza kao što su „ona“, „kojemu“, itd.; predstavljanje značenje rečenice u strojno čitljivome obliku.

Zadatkovno-orijentirani moduli tada izvode mnoge specifične postupke kao što su automatsko sažimanje ulaznoga teksta, pretraga baza podataka i mnoge druge. U nastavku ćemo prikazati **osnovna područja primjene jezičnih tehnologija** i istaknuti neke module u različitim arhitekturama u svakome odjeljku. Prikazane su arhitekture znatno pojednostavnjene i idealizirane, te služe kao ilustracija složenosti jezičnotehnoloških aplikacija na način razumljiv i nestručnjacima.



Nakon uvodnoga dijela o osnovim područjima primjene jezičnih tehnologija, dat će se kratak pregled stanja jezičnih tehnologija u istraživanjima i obrazovanju koji će se zaključiti pregledom prošlih, sadašnjih i budućih istraživačkih programa razvoja jezičnih tehnologija za hrvatski jezik[13]. Na kraju ovoga odjeljka prikazat će se stručna procjena stanja osnovnih jezičnih resursa i alata za hrvatski jezik sagledanoga kroz niz kategorija kao što su dostupnost, zrelost ili kakvoća. Tablica s rezultatima daje dobru sliku stanja jezičnih tehnologija za hrvatski jezik.

Najvažniji resursi i alati koji se opisuju u tekstu su podcrtani, a može ih se također naći u tablici na kraju poglavlja.

Osnovna područja primjene jezičnih tehnologija

Jezična provjera

Svatko tko koristi obradnike teksta kao što je npr. Microsoft Word, naišao je pravopisni provjernik koji obilježava pogrješke u tipkanju i predlaže ispravke. Četrdeset godina nakon što je Ralph Gorin napravio prvi pravopisni provjernik, jezični provjernici ne uspređuju samo riječi iz teksta s rječnikom ispravno napisanih riječi, već su postali znatno razrađeniji. Uz dodatak jezično ovisnih algoritama za obradbu morfologije (npr. prepoznavanje različitih padeža) neki već mogu prepoznati sintaktičke pogrješke kao što je ispuštanje glagola ili nesročnost između subjekta i predikata u broju i rodu, npr. „Ona je *pisao pismo.“ Pa ipak, i najnapredniji provjernici (uključujući i Microsoft Word) ne mogu pronaći pogrješke u prvoj strofi pjesme Jerrolda H. Zara (1992):

> *Eye have a spelling chequer,*
> *It came with my Pea Sea.*
> *It plane lee marks four my revue*
> *Miss Steaks I can knot sea.*

Za razrješivanje ovakvih pogrješaka u mnogim je slučajevima potrebna raščlamba konteksta, npr. treba li u hrvatskome imenicu pisati velikim (žensko osobno ime) ili malim (opća imenica) početnim slovom kao u slučaju:

> *Slatka je ova višnja.*
> *[This cherry is sweet.]*
>
> *Slatka je ova Višnja.*
> *[This Cherry is sweet.]*

Takva raščlamba zahtijeva bilo oblikovanje jezično posebnih gramatičkih pravila, izradba kojih uključuje visoku stručnost i mnogo radnih sati, ili uporabu tzv. statističkih jezičnih modela. Takvim je modelima točno izračunana vjerojatnost pojvljivanja neke riječi u određenome kontekstu (tj. s obzirom na prethodeću ili slijedeću riječ). Na primjer, u hrvatskome je *jaz između* mnogo češći niz dviju riječi nego *jaz generacija*. Statistički se jezični model može proizvesti automatski iz velike količine (ispravnih) jezičnih podataka (tj. iz korpusa). Do sada su se ovakvi modeli razvili i provjerili na jezičnim podatcima za engleski jezik. Međutim, na hrvatski jezik takva rješenja nisu izravno primjenljiva zbog njegove bogate fleksije i slobodnijega reda riječi u rečenici koji

u mnogome pridonose takvim sustavima problematičnoj raspršenosti podataka.

Uporaba jezičnih provjernika nije ograničena samo na obradnike teksta, već se koristi i u potpornim alatima za stvaranje teksta. Prateći porast različitih tehničkih proizvoda, količina tehničke dokumentacije značajno je narasla tijekom nekoliko proteklih desetljeća. Bojeći se korisničkih pritužaba o pogrješnoj uporabi ili odštetnih zahtjeva zbog nepreciznih ili loše shvaćenih korisničkih uputa, tvrtke se sve više okreću prema stvaranju što kvalitetnijih korisničkih uputa dok se istodobno pokušavaju širiti na međunarodnome tržištu. Napredak u računalnoj obradbi prirodnoga jezika doveo je do razvoja potpornih programa za pisanje teksta koji pomažu piscima tehničke dokumentacije pri uporabi kontroliranoga jezika u kojem je u skladu s (korporativnim) pravilima ograničene uporabe leksičkih jedinica, stručnoga nazivlja ili jednostavijih sintaktičkih struktura. Za hrvatski takvi alati još nisu na raspolaganju.

Premda su istraživanja računalnih modela hrvatske flektivne morfologije postojala još u 1980-ima, prvi je komercijalni prvopisni provjernik *Hrvatski računalni pravopis* objavljen tek 1996.[14] Ubrzo ga je preuzeo Microsoft i danas je sastavni dio Microsoft Officea te je u najširoj uporabi. Nekoliko je privatnih tvrtki također razvilo pravopisne provjernike, ali niti jedan nije bio toliko uspješan. Za hrvatski također postoji i besplatni prvopisni provjernik temeljem na ispell/aspell aplikaciji, a uporabiv je na svm platformama na kojima je dostupan OpenOffice. Svi se ovi programi temelje na vrlo velikim leksikonima ispravno napisanih riječi, a taj pristup ima dvije osnovne manjkavosti: 1) nizovi pismena koji predstavljaju ispravno napisane riječi mogu se pojaviti u pogrješnome kontekstu; 2) nemogućnost prepoznavanja ispravno napisanih riječi koje su nepoznate leksikonu.

Osim provjernika pravopisa i potpornih programa za stvaranje teksta, jezična je provjera važna i na području strojno potpomognutoga učenja jezika, a primjenjuje se i kod automatskoga ispravljanja upita poslanih www-tražilicama, npr. Googleove „Jeste li mislili…“ preporuke.

WWW tražilice

Pretraga na mreži, na intranetima ili u digitalnim knjižnicama danas je vjerojatno najšire korištena, a ipak još nedovoljno razvijena, jezična tehnologija. Tražilica Google, koja je započela 1998., danas se rabi za otprilike 80% svih pretraga u svijetu. [15] Od 2004. u hrvatskome se koristi glagol *guglati/googlati* i njegove tvorenice *(iz-/na-/pre-/pro-/u-)guglati/(iz-/na-/pre-/pro-/u-)googlati* premda si još nije izborio mjesto u tiskanim rječnicima (zabilježeni su čak i složenije tvorenice kao npr. *ugugljiv*). Ni sučelje za pretragu, niti prikaz dohvaćenih rezultata nisu se značajno promijenili od prve inačice. U trenutačnoj inačici Google nudi pravopisnu provjeru za pogrješno napisane riječi, a od 2009. uključuje u pretragu osnovne semantičke elemente unutar svoje mješavine algoritama[16] kojima je moguće poboljšati točnost pretrage analizom značenja upita u danome kontekstu. Uz pomoć ovoga algoritma Google je počeo pokrivati hrvatske riječi u nekima od oblika u kojima se pojavljuju u tekstu. Za razliku od npr. engleskih imenica gdje postoje samo četiri moguća oblika (*hand, hand's, hands, hands'*), hrvatske se teoretski mogu pojaviti u 14 različitih oblika, ali su

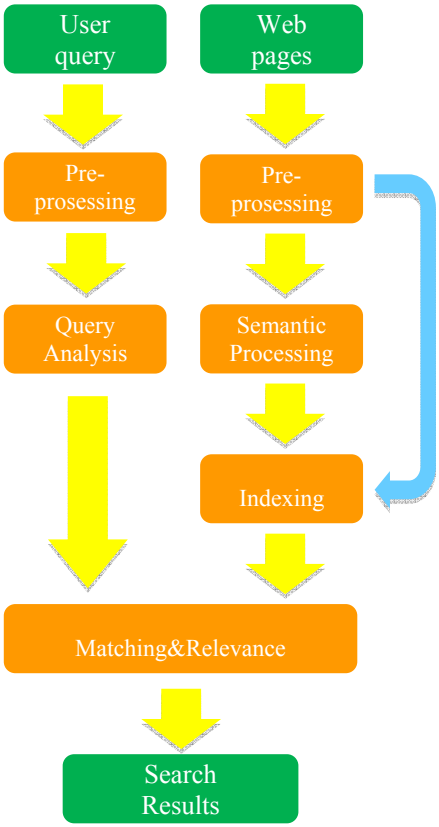
prosječno predstavljeni s 10 različitih nizova (*ruka, ruke, ruci, ruku, rukom, rukama*...). Googleova tražilica može pronaći oblike kao što su *ruka* ili *ruke*, ali *ruci* nije povezano uz imenicu *ruka*. Ima još dosta prostora za poboljšanja u Googleovoj tražilici kod flektivno bogatih jezika kod kojih se mora nositi s činjenicom da se pojedine riječi mogu pojaviti u velikome broju oblika. Međutim, uspjeh Googlea pokazuje kako se s golemim količinama podataka i učinkovitim tehnikama njihova indeksiranja, pretežito statistički utemeljeni pristup može dovesti do zadovoljavajućih rezultata, no njihova kakvoća također ovisi o samoj strukturi jezika.

Pa ipak, za razrađenije pretrage za obavijestima, uključivanje dubljega jezičnoga znanja je ključno. U istraživačkim laboratorijima provode se eksperimenti u kojim se rabe strojno čitljivi tezaurusi i ontološki organizirani jezični resursi, kao što je WordNet (ili njegova hrvatska inačica – CroWN). Ti eksperimenti pokazuju ozbiljan napredak omogućujući pronalaženje www-stranica na temelju sinonima upitnih riječi, npr. *nuklearna energija* i *atomska energija* ili na temelju još riječi još nepovezanijih s upitnim riječima.

Sljedeći će naraštaj tražilica morati uključitvati razrađenije jezične tehnologije. Naime, ukoliko upit bude sadržavao pitanje ili kakvu drugu rečenicu umjesto popisa upitnih riječi, pronalaženje relevantnih odgovora na taj upit zahtijevat će raščlambu te rečenice na sintaktičkoj i semantičkoj razini kao i tako indeksirane dokumente ne bi li odgovori bili zadovoljavajuće brzo dostupni. Na primjer, zamislite korisnički upit „Hoću popis svih tvrtki koje su neke druge tvrtke preuzele u zadnjih pet godina!“ Za zadovoljavajući odgovor na ovo pitanje potrebna je primjena parsanja kako bi se raščlanila sintaktička struktura rečenice i odredilo kako korisnik traži tvrtke koje je neka tvrtka preuzela, a ne tvrtke koje su preuzele neku tvrtku. Također, izraz *u zadnjih pet godina*, treba obraditi kako bi se odredilo o kojih se točno pet godina riječ.

Konačno, obrađeni upit se mora sraziti s golemom količinom nestrukturiranih podataka kako bi se pronašao komadić ili komadići obavijesti koje korisnik traži. Taj se zadatak obično naziva pretraživanje obavijesti i uključuje pretragu i rangiranje relevantnih dokumenata. K tome pri sastavljanju zahtjevanoga popisa tvrtki, sustav mora moći prepoznati u pretraživanim dokumentima da određen niz pismena doista predstavlja ime tvrtke. Takav se zadatak zove prepoznavanje imena i obavlja ga specijalizirana aplikacija za tu namjenu.

Još su zahtjevniji pokušaji da se na temelju upita pronađu relevantni dokumenti pisani na drugim jezicima. Za takvo višejezično pretraživanje obavijesti moramo strojno prevesti upit na se moguće jezike i pronađene dokumente prevesti na jezik upita. Rastuća količina podataka dostupnih u netekstnim oblicima potiče stvaranje usluga koje bi omogućile multimedijsko pretraživanje obavijesti, npr. pretragu u slikovnim, audio- ili video-zapisima. Za audio- i video-zapise još je potreban i modul za raspoznavanje govora koji bi omogućio pretvorbu govora u tekst ili njegov fonetski zapis u kojem se onda može obavljati pretraga.



Slika 5: Arhitektura www-tražilice

Za flektivno bogate jezike kao što je hrvatski, tražilice moraju omogućiti pretragu odjednom po svim oblicima u kojima se neka riječ može pojaviti, umjesto da se svaki oblik mora unositi pojedinačno. Takav oblik pretrage moguće je izvesti s pomoću Hrvatskoga lematizacijskoga poslužitelja koji je razvijen na Odsjeku za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu i slobodno je dostupan preko Interneta[17] omogućujući pristup Hrvatskome morfološkome leksikonu, opsežnoj bazi podataka hrvatskih riječi i svih njihovih oblika. Ta baza sadrži 110.000 natuknica iz kojih je generirano preko 4 milijuna oblika tako da svaki zapis u bazi sadrži natuknicu, oblik i MSD-oznaku tj. popis svih gramatičkih kategorija koje su se ostvarile tim oblikom. Taj je zapis usklađen s MulText East[18] preporukama.

Godine 2009., kao rezultat zajeničkoga flamansko-hrvatskoga projekta CADIAL[19], vladina agencija HIDRA omogućila je javni www-pristup svim hrvatskim zakonskim i podzakonskim dokumentima putem flektivno osjetljive tražilice[20]. Ta tražilica također omogućuje i višejezičnu pretragu dokumenata s obzirom na to da su svi dokumenti indeksirani deskriptorima iz EUROVOC-a što omogućuje uporabu i engleskih deskriptora u upitu.

Govorna interakcija

Tehnologija za govornu interakciju temelji se na stvaranju sučelja koja omogućuju komunikaciju sa strojevima korištenjem govorenoga jezika umjesto, npr., grafičkoga sučelja, tipkovnice ili miša. Danas su takva govorna korisnička sučelja (*voice user interfaces, VUI*) uključena u djelomično ili potpuno automatizirane usluge koje razne tvrtke nude svojim korisnicima, zaposlenicima ili partnerima putem telefona. Područja koja danas u mnogome ovise o uporabi VUI-ja su bankarstvo, logistika, javni prijevoz i telekomunikacije. Drugi oblici uporabe tehnologija za govornu interakciju su sučelja prema pojedinim uređajima, npr. sustavi za cestovnu navigaciju ili sustavi gdje je govorna interakcija zamjena za ulazno/izlazne podatke grafičkih sučelja, npr. kod pametnih telefona.

Sustavi koji koriste tehnologiju za govornu interakciju sastoje se od četiri različita podsustava s pripadajućim tehnologijama:

- Automatsko prepoznavanje govora (*automatic speech recognition, ASR*) određuje koje su riječi zaista izgovorene u nizu glasova koje je korisnik izrekao.
- Sintaktička raščlamba i semantička interpretacija bave se sintaktičkom strukturom korisnikova iskaza i njegovom interpretacijom u skladu s namjenom određenoga sustava.
- Upravljanje razgovorom potrebno je za određivanje kakvu akciju sustav mora poduzeti na temelju korisničkove upute i na temelju ukupne funkcionalnosti toga sustava.
- Sinteza govora (*text-to-speech, TTS*) tehnologija se rabi za pretvaranje pojedinih riječi nekoga iskaza u niz glasova koji će biti odaslani korisniku.

Jedan od glavnih izazova jest kako da ASR-sustav što točnije prepozna riječi koje je korisnik uporabio. To zahtijeva ili ograničenje broja mogućih korisničkih iskaza na ograničen skup riječi, ili ručno stvaranje jezičnoga modela koji pokriva širok raspon mogućih iskaza na prirodnome jeziku. Dok prvi pristup rezultira prilično krutim i nefleksibilnim

VUI-jem i dovodi do njegovoga moguće slabe prihvatljivosti, sastavljanje, podešavanje i održavanje jezičnih modela može u mnogome podići cijenu čitava sustava. Međutim, sustavi koji rabe jezične modele, te već u polasku dopuštaju korisnika da se izražava slobodnije, npr. započinjanjem razgovora rečenicom „Kako vam mogu pomoći?“, pokazuju i viši stupanj automatizacije i viši stupan korisničkoga prihvaćanja, te se stoga mogu smatrati naprednijima od pristupa koje koriste izravno kodirani razgovorni sustavi.

Za određene dijelove VUI-ja, tvrtke nerijetko rabe snimljene iskaze profesionalnih spikera, nerijetko vlastitih glasnogovornika. Tako snimljeni statični iskazi, u kojima se riječi ne mijenjaju ovisno o kontekstu uporabe ili o osobnim podatcima pojedinoga korisnika, korisniku pružaju kvalitetu govora koju očekuje. Međutim, što je sadržaj dinamičniji, takva mora biti i njegova govorna izvedba. Tada je i vjernost govora manja jer je potrebno umjetno povezivati velik broj malih snimki, pa to pridonosi stradavanju rečeničnoga i rječnoga naglaska. Nasuprot ovim sustavima, današnji sustavi za TTS dokazano daju vjerniji govor budući da ih je moguće podešavati do željene kakvoće s obzirom na prirodnost naglaska dinamično organiziranih iskaza.

U proteklom je desetljeću na tržištu tehnologija za govornu interakciju došlo do uznapredovale standardizacije sučeljavanja različitih tehnoloških komponenta kao i izgradnje standarda za stvaranja pojedinih programskih modula unutar pojedinih aplikacija. Također je u proteklome desetljeću došlo do značajnoga okrupnjivanja na tržištu, osobito kad je riječ o ASR i TTS sustavima. Na tržištima u zemljama G20 tj. gospodarski jakih zemalja značajne populacije, prevladava oko od pet svjetski relevantnih tvtrki, s tim što su Nuance i Loquendo najzastupljenije u Europi.

Premda je baza hrvatskih difona razvijena još 1998. unutar projekta MBROLA[21] u kojem je sudjelovao Odsjek za fonetiku Filozofskoga fakulteta Sveučilišta u Zagrebu, do danas još uvijek ne postoji niti jedan komercijalni sustav za hrvatski ATS ili TTS razvijen u Hrvatskoj. Istraživanja na ovome području provode se i na Fakultetu elektrotehnike i računarstva istoga sveučilišta, ali i na Sveučilištu u Rijeci jaka skupina istraživača radi na razvoju resursa i alata za obradbu hrvatskoga govora.

Ako bi se pokušao baciti pogled onkraj sadašnjega stanja ove tehnologije, mogle bi se očekivati značajne promjene s obzirom na ubrzano širenje pametnih telefona kao nove platforme za odnose s korisnicima uz već postojeće kao što su telefon, Internet i e-pošta. Ovakav se razvoj prilika može očekivati i u slučaju primjene tehnologije za govornu interakciju. S jedne će strane i dugoročno gledano potrebe za klasičnim telefonskim VUI zacijelo opadati, a s druge će strane uporaba govora kao izvora ulaznih podataka za pametne telefone zacijelo biti u porastu. Taj smjer razvoja također se može prepoznati s obzirom na vidan napredak točnosti prepoznavanja govora neovisnoga o govorniku u sustavima za diktiranje koji se već nude kao usluga korisnicima pametnih telefona. S obzirom da se zadatak prepoznavanja govora prebacuje na aplikacijsku infrastrukturu, primjena temeljnih jezičnih tehnologija samo će postajati sve važnijom.

Strojno prevođenje

Zamisao uporabe računala za prevođenje s jednoga prirodnoga jezika na drugi prvi je donio A. D. Booth 1946., a uslijedila joj je značajna potpora za istraživanja u tom području tijekom 1950-ih i ponovno u 1980-ima. Pa ipak, strojno prevođenje (*machine translation, MT*) još uvijek ne uspijeva ispuniti visoka očekivanja koja su se oblikovala još u njegovim najranijim godinama.

U svom najjednostavnijem obliku MT samo zamjenjuje riječi jednoga prirodnoga jezika riječima iz drugoga. To može biti uporabivo u uskim područjima s izrazito ograničenim, formulaičnim izrazima, npr. kod vremenskih izvješća. Međutim, za dobar prijevod manje ograničenih tekstova, veći tekstni odsječci (fraze, rečenice ili čitavi odlomci) moraju se što više u prijevodu približiti svojim adekvatima u polaznome jeziku. Najveći problem u tome što je ljudski jezik višeznačan, a taj se problem mora rješavati na više razina, npr. na razini razobličenja značenja riječi (*word sense disambiguation, WSD*) na leksičkoj razini („Jaguar“ može značiti životinju ili automobilsku marku) ili na razini smještanja prijedložnoga izraza u sintaktičkoj strukturi kao u primjeru:

> *Policajac je uočio čovjeka bez teleskopa.*
> [The policeman spotted a man without a telescope.]
>
> *Policajac je uočio čovjeka bez pištolja.*
> [The policeman spotted a man without a pistol.]

Jedan od mogućih pristupa ovome zadatku je temeljen na jezičnim pravilima. Za prevođenje između bliskosrodnih jezika, izravno prevođenje bi moglo biti izvedivo, no često sustavi temeljeni na pravilima (ili na znanju) raščlanjuju ulazni tekst i od njega stvaraju posrednu simboličku prezentaciju iz koje se onda generira tekst na ciljnome jeziku. Uspjeh ovakvih pristupa u mnogome ovisi o dostupnim opsežnim leksikonima s morfološkim, sintaktičkim i semantičkim podatcima, ali i o velikom broju gramatičkih pravila koje su brižljivo izradili visokostručni jezikoslovci.

Počevši s krajem 1980-ih, kad je porasla snaga računala i kad su ona postala manje skupa, više se zanimanja počelo posvećivati statističkim modlima u MT-u. Parametri tih statističkih modela izvedeni su iz raščlambe dvojezičnih korpusa kao što su Europarl usporedni korpus, koji sadrži zapisnike sjednica Europskoga parlamenta na 21 europskome jeziku, ili JRC-Acquis višejezični usporedni korpus[22], u kojem se nalaze svi tekstovi zakona Europske unije koji se odnose na zemlje-članice i to na 23 službena jezika EU. Kad im se osigura dovoljno podataka, statistički sustavi za MT rade dovoljno dobro za dobivanje približnoga značenja teksta na stranome jeziku. Međutim, za razliku od sustava temeljenih na znanju (ili pravilima), MT sustavi temeljeni na statistici (ili podatcima) često generiraju tekst koji nije ovjeren tj. nije usklađen s gramatikom ciljnoga jezika. K tome trenutačno raspoložive metode ne rade jednako dobro za sve jezične parove. Kad se gledaju samo europski jezici, prihvatljivi se prijevodi mogu dobiti za engleski i romanske jezike, no kakvoća prijevoda značajno opada za ostale germanske, slavenske, ugrofinske ili baltičke jezike[23].

Kako su prednosti i nedostatci sustava za MT temeljenih na znanju i sustava temeljenih na podatcima upravo komplementarno raspoređeni, danas su istraživači mahom usmjereni na hibridne pristupe koji kombiniraju metodologije obiju vrsta ovih sustava. Hibridizacija se može obaviti na nekoliko načina. Jedan je uporaba obje vrste sustava za obavljanje prevođenja, a potom selekcijski modul odlučuje o tome koji je rezultat više kakvoće za svaku pojedinu rečenicu. Na žalost, u slučaju duljih rečenica, niti jedan sustav još uvijek ne daje dovoljno kvalitetan prijevod. Bolji je pristup kombinirati najbolje dijelove rečenica iz višestrukih mogućih prijevoda, a oni mogu biti poprilično složeni, s obzirom da odgovarajući dijelovi višestrukih prijevodnih rješenja nisu uvijek lako uočljivi, te se moraju posebno sravnjivati.

Strojno je prevođenje s i na hrvatski jezik osobito izazovan zadatak. Slobodniji red riječi u rečenici i bogata fleksija predstavljaju probleme pri generiranju ispravnih rečeničnih konstrukcija i oblika riječi koji obličnim nastavcima kodiraju gramatičke kategorije roda, padeža, broja, načina, vremena itd. Dodatne probleme često postavlja i zahtjev za sročnošću glede tih kategorija između npr. atributa i imenice u rodu, broju i padežu ili samo u rodu i broju kad je riječ o subjektu i predikatu.

Europska je komisija pokrenula i potpomaže nekoliko projekata kako bi se razvila istraživanja i razvoj strojnoga prevođenja za tzv. jezike s nedovoljno razvijenim resursima, a među njima je i hrvatski. Tako CIP ICT PSP projekt LetsMT![24] i FP7 projekt ACCURAT[25] razvijaju nove metode za što jednostavnije prikupljanje podataka potrebnih za strojno prevođenje i izgradnju takvih sustava prilagođenih različitim domenama i uporabama. U oba ova projekta kao hrvatski partner sudjeluje skupina istraživača s Filozofskoga fakulteta Sveučilišta u Zagrebu.

Projekt ACCURAT[26] istražuje nove metode uporabe usporedivih korpusa ne bi li se nadoknadila nestašica jezičnih resursa i posredno poboljšalo strojno prevođenje za jezike s nedovoljnim resursima i za uske domene[27]. Cilj je projekta ACCURAT postići značajan napredak u kakvoći strojnoga prijevoda za čitav niz novih službenih jezika EU i jezika zemalja-pristupnica (estonski, grčki, hrvatski, letonski, litavski i rumunjski), kao i predložiti nove pristupe za prilagodbu tehnologija za strojno prevođenje u pojedinim uskim domenama i time značajno povećati pokrivenost različitih jezika i domena strojnim prevođenjem.

Projekt LetsMT! [28] izgrađuje novi vrstu *on-line* suradne platforme za dijeljenje usporednih tekstova i automatsko stvaranje vlastitih sustava za strojno prevođenje. Ova platforma smještena u računalnome oblaku omogućit će svim vrstama korisnika slanje u posebno zaštićen repozitorij vlastitih jezičnih resursa na temelju kojih će se potom automatski izraditi vlastiti sustav za statističko strojno prevođenje treniran upravo na temelju tih vlastitih jezičnih resursa. Takav sustav za strojno prevođenje potom se može podijeliti s ostalim korisnicima. Strojnoprevoditeljske usluge projekta LetsMT! mogu se rabiti na nekoliko načina: kroz www-portal, kroz *widget* koji se može slobodno preuzeti i uključiti na svoje www-stranice, kroz dodatak za popularne prebirnike (Mozilla Firefox i Internet Explorer) kao i kroz integraciju u

postojeće sustave za strojno potpomognuto prevođenje (*machine assisted translation, MAT*) kako *on-line*, tako i *off-line*.

Google Translate nudi prijevode na i s hrvatskoga od 2008. Kakvoća njegovih prijevoda bila je niska u početku, ali se poboljšava kako je sve više i više usporednih hrvatsko-engleskih podataka dostupno *on-line*.

Još uvijek se smatra kako upravo na području poboljšanja kakvoće prijevoda ima još mnogo prostora za napredak kod sustava za strojno prevođenje. Napredak se očekuje u prilagodbi jezičnih resursa određenom području uporabe ovisno o temi ili korisniku, kao i u integraciji s postojećim sustavima za strojno potpomognuto prevođenje u kojima se već rabe velike terminološke baze i prijevodne memorije.

Postupci vrjednovanja omogućuju uspoređivanje kakvoće prijevoda sustava za strojno prevođenje, njihove različite pristupe i stanje strojnoprevoditeljskih sustava za različite jezike. U okviru projekta Euromatrix+ sastavljena je Tablica 1 u kojoj je prikazana kakvoća za sve parove strojnih prijevoda između 22 službena jezika EU-a (irski jedino nedostaje) iskazana s pomoću BLEU mjere[29].

	Target Language																					
	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

Tablica 1: Kakvoća prijevoda za sve parove između 22 službena jezika EU-a (izvor: Euromatrix+)

Najbolji rezultati (prikazani u zelenome i plavome) postignuti za za jezike za koje već postoje sustavi znatno razrađeni unutar raznih istraživačkih programa i za koje jezike postoji mnogo usporednih korpusa (npr. engleski, francuski, nizozemski, španjolski, njemački), a najgori su rezultati (u crvenome) za jezike koji su u mnogome strukturno različiti od većine jezika (npr. madžarski, maltski, finski).

Jezične tehnologije 'iza zavjese'

Izgradnja jezičnotehnoloških aplikacija uključuje čitav niz podzadataka koji se ne vide na razini korisničkoga sučelja, ali 'ispod poklopca' osiguravaju bitnu funkcionalnost i uslužnost te aplikacije. Upravo ti podzadatci rezultat su rješavanja važnih istraživačkih problema koji su postali pojedine podgrane samoga računalnoga jezikoslovlja.

Odgovaranje na pitanja (*question answering*) postalo je aktivnim područje istraživanja za koje su izgrađeni obilježeni korpusi i organiziraju se posebna znanstvena natjecanja. Ideja je krenuti od pretrage temeljene na ključnim riječima (na koju stroj obično odgovara s čitavim skupom potencijalno relevantnih dokumenata) prema scenariju u kojem korisnik postavlja konkretno pitanje, a sustav pruža jedinstven odgovor, npr.: „Koliko je godina imao Neil Armstrong kad je stupio na Mjesec?“ – „38“. Premda se ovakva vrsta pretage nesumnjivo može povezati s već spomenutim www-pretraživanjem, danas se odgovaranje na pitanja smatra ponajprije sveobuhvatnim nazivom za proučavanja kakve sve *vrste pitanja* valja razlikovati i kako se s njima valja postupati, kako se dokumente koji potencijalno sadrže odgovor valja obrađivati i uspoređivati (daju li suprotstavljene odgovore?), te kako se određena obavijest – točan odgovor – može pouzdano crpiti iz dokumen(a)ta bez zanemarivanja konteksta.

Ovaj je zadatak usko povezan sa zadatkom crpljenja obavijesti (*information extraction, IE*), područjem koje je bilo iznimno popularno i utjecajno u vrijeme „statističkoga prevrata“ u računalnome jezikoslovlju, tj. u ranim 1990-im. Cilj IE-a je pronaći posebne obavijesti u posebnim klasama dokumenata, a to bi moglo biti, npr., pronalaženje u novinskim člancima ključnih osoba koje sudjeluju u preuzimanju tvrtki. Drugi mogući scenarij, razrađen iz izvješća o terorističkim napadima, tiče se postupka s pomoću kojega se iz teksta može prepoznati obrazac djelovanja, cilj, vrijeme ili mjesto napada i njegove posljedice. Popunjavanje takvih domenski ovisnih obrazaca središnja je osobina IE-a. Upravo zbog toga IE predstavlja još jedan primjer jezične tehnologije 'iza scene' koja je sama jasno odijeljena od ostalih poddisciplina (ili tehnologija), ali zbog praktičnih razloga mora biti uključena u šire uporabno okružje.

Godine 2009. Hrvatska izvještajna novinska agencija (HINA)[30] započela je s razvojem sustava za (pred)obradbu njihovih vijesti koja je uključila lematizacijsu, prepoznavanje imena, klasifikaciju vijesti prema zadanoj shemi rubrika i crpljenje ključnih riječi. Ovaj su sustav zajednički razvili Fakultet elektrotehnike i računarstva[31] i Filozofski fakultet, oba sastavnice Sveučilišta u Zagrebu.

Dva rubna područja, koja katkada igraju ulogu samostalnih, a katkada samo potpornih aplikacija, jesu sažimanje tekstova (*text summarization*) i generiranje teksta (*text generation*). Sažimanje se očito odnosi na skraćivanje dugih tekstova, a postoji kao ponuđena funkcionalnost već i u MS Wordu. Postupci sažimanja mahom su statistički utemeljeni pri čemu se prvo identificiraju „važne“ riječi u tekstu (npr. riječi koje su visoko učestale u danome tekstu, ali su izrazito nisko učestale u općoj jezičnoj uporabi), a potom se pronalaze rečenice u kojima se te riječi nalaze. Takve se rečenice u dokumentu posebno obilježavaju ili

izdvajaju ne bi li se od njih sastavio sažetak. U tom najpopularnijem scenariju sažimanje dokumenata jednako je izdvajanju rečenica: čitav se tekst reducira na podskup svojih rečenica. Svi komercijalni sustavi za sažimanje tekstova koriste isti pristup. Alternativni pristup iskušava se u nekoliko istraživačkih središta i usmjeren je na sastavljanje *novih* rečenica, tj. na izgradnju sažetka sastavljenog od rečenica koje se ne moraju pojaviti u istome obliku u sažimanome tekstu. Ovaj pristup zahtijeva stanovit oblik „dubljega razumijevanja" teksta i stoga je manje robustan, a nije uopće moguć bez modula za generiranje teksta tj. novih rečenica. Sve u svemu, generator teksta u najvećem broju slučajeva nije samostalna aplikacija već je uključen u šira programska okružja kao što su npr. klinički informacijski sustavi gdje se podatci o podinim pacijentima skupljaju, spremaju, obrađuju, a npr. generiranje izvješća o stanju pacijenta samo je jedna od mnogih funkcionalnosti takvih sustava.

Niti jedna od tehnologija iz ova dva rubna područja još ne postoje za hrvatski jezika osim nekoliko eksperimenata koji su izvedeni za sažimanje tekstova na hrvatskome jeziku[32].

Jezične tehnologije u obrazovanju

Područje jezičnih tehnologija je visoko interdisciplinarno područje koje uključuje stručnjake iz jezikoslovlja, informacijskih znanosit, računarskih znanosti, matematike, filozofije, psiholingvistike, kognitivne znanosti i neuroznanosti itd. Kako se na Odsjeku za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu neprekinuto od 1950-ih proučavaju i poučavaju algebarskolingvistički pristupi jezičnome opisu, uvođenje posebnoga smjera Računalna lingvistika u dvogodišnjem Diplomskome studiju lingvistike 2005. bilo je samo logičan nastavak te tradicije. Sličan je progam pokrenut na Sveučilištu u Zadru 2010.

Jezičnotehnološki programi

Govornika hrvatskoga jezika ima oko 5,5 milijuna i taj broj nipošto nije dovoljan da za održavanje skupoga razvoja novih jezičnohtehnoloških proizvoda isključivo iz komercijalnih izvora. Razvoj jezičnih resursa i alata za hrvatski jezik košta isto kao i za jezik s nekoliko stotina milijuna govornika. Rezultat toga jest da je broj komercijalno orijentiranih jezičnotehnoloških tvrtki za hrvatski jezik blizak nuli. Ulogu financijskoga podupiratelja jezičnotehnoloških aktivnosti djelomično preuzima država, no savsim sigurno ne u opsegu potrebnome za razvoj svih potrebnih jezičnih resursa i alata.

U Hrvatskoj su se aktivnosti oko prikupljanja jezičnih resursa, tj. računalnih korpusa, počele već 1960-ih kad je 1967. Željko Bujas sastavio prvi hrvatski računalni korpus i napravio njegovu konkordanciju[33] u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu. Od tada je ta ustanova postala središnjom ustanovom u Hrvatskoj za istraživanja s područja korpusne lingvistike. Godine 1968. u Zavodu se po prvi se puta u povijesti lingvistike uporabio usporedni računalni korpus u kontrastivnolingvističkim istraživanjima[34]. Tijekom 1970-ih i 1980-ih obavljala se računalna obradba starih hrvatskih pisaca, a sastavljanje *Jednomilijunskoga*

korpusa hrvatskoga književnoga jezika započelo je 1976. pod vodstvom Milana Moguša. Na temelju toga korpusa sastavljen je prvi hrvatski čestotni rječnik[35]. Sastavljanje Hrvatskoga nacionalnoga korpusa[36] počelo je 1998.[37], a opseg od 101 milijun riječi-pojavnica dohvatio je 2004.[38] Danas je najveći hrvatski korpus hrWaC sastavljen na istome fakultetu 2010., a obasiže 1,3 milijarde riječi-pojavnica skupljenih s .hr internetske domene. Od godine 2000. na istome se fakultetu pod vodstvom Damira Borasa odvija opsežna aktivnost digitalizacije starih hrvatskih jedno- i višejezičnih rječnika[39].

Pri Institutu za hrvatski jezik i jezikoslovlje 2004. započelo je sastavljanje opsežnoga korpusa pod nazivom *Hrvatska jezična riznica*[40] koja uključuje pisane tekstove od 11. stoljeća do današnih dana. Riznica je organizirana kao u tri glavna korpusa (starohrvatski, srednjehrvatski i suvremeni hrvatski) gdje se za prva dva rješavaju bitni problemi dijakronijskih korpusa što u hrvatskome slučaju znači, transliteracija s tri različita pisma (glagoljice, ćirilice i latinice), rješavanje nestandardnih pravopisnih rješenja, individualne varijacije u uporabi pojedinih pismena itd.

Nakon istraživačkih programa 1970-ih i 1980-ih, koji su uobičajeno bili usmjereni prema računalnoj obradbi književnih tekstova, većinu istraživačkih aktivnosti na području računalnoga jezikoslovlja, korpusnoga jezikoslovlja i jezičnih tehnologija danas podupire Ministarstvo znanosti, obrazovanja i športa kroz projekte povezane s jezičnim tehnologijama. Još je 1991. započeo prvi takav projekt pod nazivom *Računalna obradba hrvatskoga književnoga jezika*, 1996. je slijedio *Računalna obradba hrvatskoga jezika*, a 2002. *Razitak hrvatskih jezičnih resursa*. Godine 2007. iz istoga su izvora poduprta tri osnovna istraživačka programa s po nekoliko projekata usmjerenih na razvoj jezičnih tehnologija za hrvatski jezik:

- *Računalnolingvistički modeli i jezične tehnologije za hrvatski jezik*[41] gdje se sastavlja i održava čitav niz jezičnih resursa i alata (npr. Hrvatski nacionalni korpus, Hrvatsko-engleski paralelni korpus, Hrvatski morfološki leksikon, Hrvatska ovisnosna banka stabala[42], Hrvatski wordnet[43], hibridni označivač[44] i lematizator[45], ovisnosni parser za hrvatski, sustav za prepoznavanje imena, itd.);
- *Izvori za hrvatsku baštinu i hrvatski europski identitet*[46] s projektima koji se bave digitalizacijom starih hrvatskih rječnika i izradbom Hrvatskoga valencijskoga rječnika[47];
- *Hrvatska jezična riznica*[48] gdje se niz projekata bavi različitim jezikoslovnim problemima započevši od istraživanja hrvatskih narječja i etimologije, do razvoja semantičkih mreža za izgradnju leksičkih resursa. Svi ti projekti uključuju digitalizaciju skupljenih jezičnih podataka i izravno uvećavaju broj dostupnih jezičnih resursa za hrvatski jezik.

Ovi su istraživački programi otvorili mogućnost da razvoj jezičnih tehnologija za hrvatski jezik uhvati korak s ostalim europskim jezicima, a istodobno su pružili priliku za ravnopravno sudjelovanje hrvatskih istraživačkih skupina u postojećim FP7 i ICT-PSP projektima s obzirom da je zadnji takav projekt (TELRI II) u kojem su sudjelovali završio još 2002.

Jeste li znali kako je najstariji hrvatski tiskani rječnik Dictionarium quinque nobilissimarum Europae linguarum Latinae, Italicae, Germanicae, Dalmaticae et Ungaricae *Fausta Vrančića (1595) ujedno i najstariji madžarski tiskani rječnik?*

Iz Hrvatske Filozofski je fakultet Sveučilišta u Zagrebu partner na projektu CLARIN, pothvatu koji nastoji oko izgradnje istraživačke infrastrukture za istraživače s područjā humanističkih i društvenih znanosti na razini čitave Europe, a koja se infrastruktura temelji na jezičnim resursima i alatima. Hrvatska je jedna od zemalja-osnivača CLARIN ERIC-a. Isti Fakultet je partner u FP7 projektu ACCURAT i ICT-PSP projektima LetsMT! i CESAR. Sveučilište u Zadru partner je u ICT-PSP projektu ATLAS.

Dostupnost alata i resursa za hrvatski jezik

Sljedeća tablica daje pregled trenutačnoga stanja s jezičnotehnološkom potporom hrvatskome jeziku. Ocjene postojećih alata i resursa temeljene su na specijaliziranoj procjeni nekoliko vodećih stručnjaka koji su se vodili sljedećim kriterijima (u mogućem rasponu od 0 do 6).

1. **Količina**: Postoji li alat/resurs za predmetni jezik? Što više alata/resursa postoji, ocjena je viša.
    - 0: nema nikakvih alata/resursa
    - 6: postoji mnogo alata/resursa, velika raznolikost
2. **Dostupnost**: Jesu li alati/resursi dostupni, tj. u obliku otvorenoga koda (*Open Source*), slobodno ih se može koristiti na bilo kojoj platformi ili su dostupni samo uz visoku naknadu ili u ograničenim uvjetima?
    - 0: praktično su svi alati/resursi dostupni samo uz visoku naknadu
    - 6: velika je količina alata/resursa slobodno i otvoreno dostupna pod raznim oblicima Open Source ili Creative Commons licencija koje dopuštaju njihovu ponovnu uporabu, prenamjenu i/li preobliku
3. **Kakvoća**: Koliko najbolji dostupni alati/resursi zadovoljavaju odgovarajuće postavljene kriterije i indikatore kakvoće? Jesu li ti alati/resursi ažurirani i aktivno održavani?
    - 0: alat/resurs na razini prototipa ili „igračke“
    - 6: visokokvalitetan alat, resurs obilježen na razini kakvoće ljudskoga obilježavanja
4. **Pokrivenost**: Koliko najbolji dostpuni alati/resursi zadovoljavaju odgovarajuće postavljene kriterije pokrivenosti (stilova, žanrova, vrsta tekstova, jezičnih pojava, vrsta ulaznih/izlaznih podataka, broja jezika u sustavu za MT, itd.)? Koliko su resursi reprezentativni za ciljane jezične varijetete ili podvarijetete?
    - 0: alat/resurs posebne namjene, posebna primjena, iznimno malena pokrivenost, uporabivost samo za posebne i nipošto općejezične primjene
    - 6: resurs vrlo široke primjene, vrlo robustan alat, široka primjena, pokrivanje mnogo jezika
5. **Zrelost**: Može li se alat/resurs smatrati zrelim, stabilnim i spremnim za tržište? Mogu li se najbolji alati/resursi početi odmah koristiti ili ih treba prilagođavati? Je li ta tehnologija na razini za izravnu uporabu u prozvodnji ili je riječ samo o prototipu koji se još ne može uključiti u proizvodnju? Jedan je od indikatora također je li alat/resurs prihvaćen u zajednici i koristi li se uspješno u jezičnotehnološkim sustavima.

- 0: preliminarni prototip, sustav „igračka“, potvrda koncepta, primjer resursa
- 6: neposredno primjenljiva komponenta/modul/aplikacija

6. **Održivost**: Kako se dobro alat/resurs može održavati/uključivati u postojeće sustave informacijskih tehnologija? Ispunjava li alat/resurs određenu razinu održivosti glede dokumentacije/priručnika, objašnjena primjera uporabe, sučelja, grafičkih korisničkih sučelja itd.? Koristi li standardna/najbolja programirna okružja (npr. Java EE)? Postoje li industrijske norme/podnorme i ako postoje, je li alat/resurs s njima usklađen (format podataka, itd.)?
   - 0: isključivo vlastiti, *ad hoc* podatkovni formati i API-ji
   - 6: puna usklađenost s normama, potpuna dokumentacija
7. **Prilagodljivost**: Koliko se uspješno najbolji alati/resursi mogu prilagoditi/proširiti za nove zadaće/područja/žanrove/vrste tekstova/korisničke zahtjeve itd.?
   - 0: praktički je nemoguće prilagoditi alat/resurs za neki drugi zadatak, čak niti uz mnogo uloženih sredstava ili ljudskoga rada
   - 6: vrlo visoka razina prilagodljivosti; prilagodba vrlo lako i visoko učinkovito izvediva

Tablica alata i resursa

	Količina	Dostupnost	Kakvoća	Pokrivenost	Zrelost	Održivost	Prilagodljivost
Jezične tehnologije (alati, tehnologije, aplikacije)							
Opojavničenje, morfologija (opojavničenje, POS/MSD-označavnje, morfološka analiza/generiranje)	3	2	5	5	3	2	5
Parsanje (plitka ili duboka sintaktička analiza)	1	1	2	2	1	0	4
Rečenična semantika (razobličenje značenja riječi, struktura argumenata, semantičke uloge)	1	0	1	3	0	0	3
Tekstna semantika (razrješivanje koreferencije, kontekst, pragmatika, izvođenje)	0	0	0	0	0	0	0
Napredna obradba diskurza (tekstna struktura, kohrencija, retorička struktura/RST, argumentacijska područja, argumentacija, tekstni obrasci, vrste tekstova itd.)	0	0	0	0	0	0	0
Pretraživanje obavijesti (indeksiranje tekstova, multimedijsko pretraživanje obavijesti, višejezično pretraživanje obavijesti)	1	0	5	2	3	2	3
Crpljenje obavijesti (prepoznavanje imena, crpljene događaja/odnosa, prepoznavanje mišljenja/stavova, dubinska obradba teksta/tekstna analitika)	3	1	4	3	2	1	3
Generiranje jezika (generiranje rečenica, generiranje izvješća, generiranje teksta)	1	1	4	0	3	0	0
Sažimanje, odgovaranje na pitanja, napredne tehnologija pristupa obavijestima	1	0	1	0	0	0	0
Strojno prevođenje	1	0	1	3	0	0	0
Prepoznavanje govora	2	2	3	3	2	3	3
Sinteza govora	3	3	3	4	4	4	4
Upravljanje dijalogom (dijaloške sposobnosti i modeliranje korisnika)	1	2	1	1	0	2	2
Jezični resursi (resursi, podatci, baze znanja)							
Referentni korpusi	3	3	3	4	4	4	2
Sintaktički obilježeni korpusi (banke stabala, ovisnosne banke)	1	1	3	4	2	1	2
Semantički obilježeni korpusi	0	0	0	0	0	0	0
Diskursno obilježeni korpusi	0	0	0	0	0	0	0
Usporedni korpusi, prijevodne memorije	3	2	3	3	3	1	2
Govorni korpusi (sirovi govorni podatci, obilježeni govorni podatci, dijaloški govorni podatci)	3	1	3	3	4	3	4
Multimedijski i multimodalni podatci (tekstni podatci kombinirani s audio/video podatcima)	1	1	4	3	3	3	3
Jezični modeli	0	0	0	0	0	0	0
Leksikoni, terminologije	3	3	4	3	4	3	3
Gramatike	0	0	0	0	0	0	0
Tezaurusi, WordNet	2	3	3	4	3	2	2
Ontološki resursi znanja o svijetu (npr. vršne ontologije, *Linked Data*)	0	0	0	0	0	0	0

Zaključci

U ovome niz bijelih knjiga po prvi se puta za mnoge europske jezike pokušalo procijeniti njihvoo ukupno stanje s obzirom na jezičnotehnološku potporu tim jezicima i to na način koji u visokoj mjeti dopušta usporedbe i pronalaženje manjkavosti i potreba.

Interpretacija tablice o hrvatskim resursima i alatima

Rezulati iz gornje tablica mogu se sažeti u obliku nekoliko ključnih poruka koje na temelju sadašnjega stanja naglašuju ključna pitanja koja treba riješiti u daljnjem razvoju jezičnih tehnologija za hrvatski jezik:

- kad je riječ o većini temeljnih tehnoloških alata i resursa (referentni korpusi, manji usporedni korpusi, veliki flektivni rječnici, opojavničitelji, MSD-označivači, lematizatori, NERC sustavi itd.) hrvatski stoji relativno dobro.
- Međutim, veliki sintaktički obilježeni korpusi nedostaju kao i veliki usporedni korpusi (npr. hrvatski prijevod pravne stečevine EU). Mnogim postojećim resursima nedostaje standardiziran oblik, pa je potrebna ozbiljna inicijativa da se standardiziraju ti podatci kao i formati za razmjenu podataka.
- premda su u nekim područjima već započeli eksperimenti, kao što su plitko parsanje (*chunking*), sažimanje, primjena ontoloških resursa, oni se odvijaju samo u akademskim krugovima, a dosegnuti rezultati su daleko od razine razvijenosti koju pokazuju drugi europski jezici. Obradba multimedijskih i multimodalnih dokumenata dobiva na važnosti, osobito digitalizacija u kontekstu očuvanja nacionalne kulturne baštine, ali jezične tehnologije za hrvatski jezik još nisu uključene u te procese u dovoljnoj mjeri.
- Na potpodručjima kao što su npr. sinteza govora, prepoznavanje govora i crpljene obavijesti, postoje pojedini prozvodi, ali ograničene ili visokospecijalizirane funkcionalnosti.
- Alati i resursi za naprednije jezične tehnologije kao što su duboko parsanje, strojno prevođenje, tekstna semantika, obradba diskurza, generiranje jezika, upravljanje dijalogom, itd. jednostavno za hrvatski još ne postoje.

Što bi trebalo poduzeti?

Javna financijska potpora jezičnim tehnologijama u Europi je relativno niska ako ju se usporedi s troškovima prevođenja i višejezičnoga pristupa u SAD-u[49]. U Hrvatskoj je javno financiranje razvoja jezičnih tehnologija još i manje nego u mnogim usporedivim europskim zemljama, uključujući i susjedne zemlje poput Slovenije, Madžarske ili Srbije. Premda postoji goruća potreba prepoznavanja važnosti jezičnih tehnologija u osiguravanju održivoga razvoja hrvatskoga jezika u 21. stoljeću i u izazovima koje će pred njega staviti uloga jednoga od službenih jezika EU, još uvijek nije pokrenuta nikakva inicijativa koja bi skrbila o stvaranju velikih resursa, alata i servisa za hrvatski jezik, o partnerstvu između vlade, istraživanja i gospodarstva ne bi li se razvio

stručni klaster za hrvatske jezične tehnologije. Vjerujemo kako bi ta inicijativa morala imati institucionalni okvir u obliku posebnoga središta kompetencija/izvrsnosti koji bi mogao dobiti potporu iz strukturnih fondova EU s ciljem poticanja poslovno orijentiranih istraživanja, promicanja suradnje unutar sektora između tvrtki i istraživačkih ustanova na razvoju novih proizvoda i tehnologija podizanja kompetitivnost hrvatskih tvrtki na tržištu EU kojega će Hrvatska postati sastavni dio već 2013.

Literatura

Agić, Ž.; Tadić, M.; Dovedan, Z. (2008) Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis, Informatica 32, 4; pp 445 451.

Agić, Ž.; Tadić, M.; Dovedan, Z. (2009) Evaluating Full Lemmatization of Croatian Texts. In: Klopotek, M.; Przepiorkowski, A.; Wierzchon, S.; Trojanowski, K. (eds) Recent Advances in Intelligent Information Systems, Academic Publishing House EXIT, Warsaw, pp 175-184.

Batnožić, S.; Ranilović, B.; Silić, J. (1996) Hrvatski računalni pravopis (uz računalni program spelling-checker), Matica hrvatska–SYS, Zagreb.

Bekavac, B.; Tadić, M. (2007) Implementation of Croatian NERC system. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007, Special Theme: Information Extraction and Enabling Technologies, ACL, Prag, Czech Republic, pp 11-18.

Brozović Rončević D., Ćavar D. (2008) Hrvatska jezična riznica kao podloga jezičnim i jezičnopovijesnim istraživanjima hrvatskoga jezika. "Vidjeti Ohrid", Proceedings of the 14th international Slavistic Congress in Ohrid, Hrvatsko filološko društvo–Hrvatska sveučilišna naklada, Zagreb, 173-186.

Bujas, Ž. (1974) Osman, kompjutorska konkordancija, Sveučilišna naklada Liber.

Eisele, A.; Xu, J. (2010) Improving Machine Translation Performance Using Comparable Corpora. Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, European Language Resources Association (ELRA), La Valletta, Malta, pp 35-41.

Koehn, P.; Birch, A.; Steinberger, R. (2009) 462 machine translation systems for Europe. Proceedings of the Twelfth Machine Translation Summit (MT Summit XII). International Association for Machine Translation.

Mikelić Preradović, N.; Lauc, T.; Boras, D. (2007) CROXMLSUM – the System for XML Document Summarization in Croatian. International Journal of Mathematics and Computers in Simulation, 1, 1, pp 81-89.

Mikelić Preradović, N. (2010) CROVALLEX lexicon improvements: Subcategorization and semantic constraints. WSEAS transactions on computers, 9

Moguš, M.; Bratanić, M.; Tadić, M. (1999) Hrvatski čestotni rječnik, Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu – Školska knjiga.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of ACL, Philadelphia, PA, pp 311-318.

Skadiņa, I.; Vasiļjevs, A.; Skadiņš, R.; Gaizauskas, R.; Tufiş, D.; Gornostay, T. (2010) Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. Proceedings of the 3rd Workshop on Building and Using Comparable Corpora. European Language Resources Association (ELRA), La Valletta, Malta, pp 6-14.

Steinberger, R.; Pouliquen, B.; Widiger, A.; Ignat, C.; Erjavec, T.; Tufiş, D.; Varga, D. (2006) The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy.

Tadić, M. (1997) Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive. Suvremena lingvistika, 23, 43-44; pp 387-394.

Tadić, M. (2002) Building the Croatian National Corpus, Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002). Las Palmas, Spain, pp 441-446.

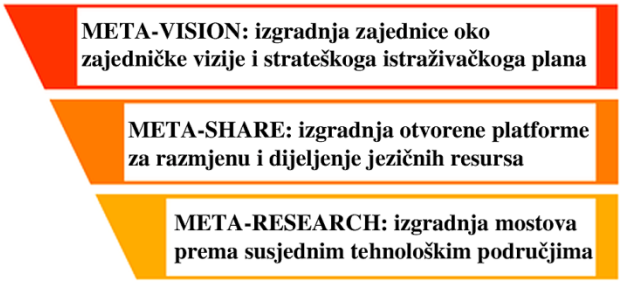
Tadić, M. (2003) Jezične tehnologije i hrvatski jezik, Exlibris, Zagreb.

Tadić, M. (2009) New version of the Croatian National Corpus. In: Hlaváčková, D.; Horák, A.; Osolsobě, K.; Rychlý, P. After Half a Century of Slavonic Natural Language Processing, Masaryk University, Brno, pp 199-205.

Vasiļjevs, A., Gornostay, T.; Skadiņš, R. (2010) LetsMT! – Online Platform for Sharing Training Data and Building User Tailored Machine Translation. Proceedings of the Fourth Baltic conference 'Human Language Technologies – the Baltic Perspective'.

META-NET

META-NET je mreža izvrnosti posvećena promicanju tehnoloških temelja višejezičnoga europskoga informacijskoga društva. Kako bi se ostvarile aplikacije koje će omogućiti strojno prevođenje, višejezično upravljanje obavijestima i znanjem, kao i proizvodnju sadržaja na svim europskim jezicima, potreban je na razini cijeloga kontinenta ozbiljan i usklađen novi korak u istraživanjima jezičnih tehnologija. U tome smjeru META-NET djeluje u tri osnova smjera aktivnosti:



Tri smjera aktivnosti META-NET-a

Što je cilj META-NET-a?

Osnovni je cilj META-NET-a izgraditi višejezični europski tehnološki savez (Multilingual European Technology Alliance, META) koji će okupiti istraživače, komercijalne isporučitelje tehnologije, privatnike ili tvrtke korisnike jezičnih tehnologija, jezične profesionalce i ostale dionike informacijskoga društva. META priprema potrebne ambiciozne zajedničke akcije usmjerene na napredak jezičnih tehnologija kao jednoga od sredstava za ostvarivanje vizije Europe ujedinjene u obliku jedinstvenoga digitalnoga tržišta i informacijskoga prostora.

META-NET štand na LREC-u (Malta, svibanj 2010.)

Prva META-NET zbivanja u 2010.

U kratkome razdoblju svoga postojanja, od veljače 2010., META-NET se već uspostavio svoju uočljivost kod većega broja uključenih zajednica. Dolje je odabir prošlih zbivanja koje je organizirala META-NET ili je na njima sudjelovala:

Translingual Europe (Berlin, lipanj 2010.)

- **Language Technology Days 2010** (ožujak 2010., Luxembourg): META-NET je predstavljen pred oko 250 ključnih sudionika u istraživačkome prostoru europskih jezičnih tehnologija; povezivanje s novopokrenutim ili nadolazećim projektima i inicijativama.
- **LREC 2010** (Language Resources and Evaluation Conference, svibanj 2010., Malta): najveća konferencija s područja računalnoga jezikoslovlja i jezičnih tehnologija usredotočena na jezične resurse (oko 1100 sudionika). META-NET je bio predstavljen s više prezentacija na glavnoj konferenciji i u radionicama uz štand na izložbi EC Projects Village.
- **theMETAnk 2010** (lipanj 2010., Berlin): poticajni sastanak s otprilike 120 ključnih istraživača iz jezičnih tehnologija ponajprije iz akademskih ustanova.

- **Translingual Europe 2010** (lipanj 2010, Berlin): pozivna konferencija s oko 150 industrijskih predstavnika u organizaciji META-NET-a.

Trenutačni sastav META tehnološkoga vijeća

Ime	Organizacija	Uloga	Zemlja
Nicoletta Calzolari	Consiglio Nazionale di Ricerche	Director of Research	Italija
Bill Dolan	Microsoft Research	Head of NLP	SAD
Josef van Genabith	Dublin City University, CNGL	Director	Irska
Yota Georgakopolou	European Captioning Institute	Managing Director	UK, Grčka
Gregory Grefenstette	Exalead	Chief Science Officer	Francuska
Jan Hajic	Karlovo sveučilište u Pragu	Professor	Češka
Theo Hoffenberg	Softissimo	CTO	Francuska
Thomas Hofmann	Google	Dir. Engineering	Švicarska
Keith Jeffrey	ERCIM	President	UK
Stefan Kreckwitz	Across	CTO	Njemačka
Claude de Loupy	Syllabs	CEO	Francuska
Elisabeth Maier	CLS Communication	CTO	Švicarska
Daniel Marcu	Language Weaver	CTO	SAD, Rumunjska
Joseph Mariani	CNRS-LIMSI, IMMI	Director	Francuska
Penny Marinou	EUATC	President	Grčka
Jaap van der Meer	TAUS	Director	Nizozemska
Roger Moore	Sveučilište u Sheffieldu	Professor	UK
Stelios Piperidis	ILSP, Research Centre “Athena”	Head of Department	Grčka
Gabor Proszeky	Morphologic	CEO	Madžarska
Georg Rehm	DFKI	Senior Consultant	Njemačka
C.M. Sperberg-McQueen	World Wide Web Consortium	Technical Staff	SAD
Daniel Tapias	Sigma Technologies	CEO	Španjolska
Alessandro Tescari	Pervoice	CEO	Italija
Hans Uszkoreit	DFKI	Scientific Director	Njemačka
AndrejsVasiljevs	Tilde	CEO	Letonija
Michel Vérel	Vecsys	CEO	Francuska
Alex Waibel	CMU, Sveučilište u Karlsruheu	Professor	SAD/Njemačka

Članstvo mreže izvrsnosti META-NET (*: osnivački članovi)

Zemlja	Član (organizacija)	Kontakt
Austrija	Universität Wien	Gerhard Budin
Belgija	Sveučilište u Antwerpenu	Walter Daelemans
	Katoličko sveučilište u Leuvenu	Dirk van Compernolle
Bugarska	Bugarska akademija znanosti	Svetla Koeva
Cipar	Sveučilište na Cipru	Jack Burston
Češka	Karlovo sveučilište u Pragu*	Jan Hajič
Danska	Sveučilište u Kopenhagenu	Bente Maegaard, Bolette Sandford Pedersen
Estonija	Sveučilište u Tartu	Tiit Roosmaa
Finska	Sveučilište Aalto*	Timo Honkela

	Sveučilište u Helsinkiju	Kimmo Koskenniemi, Krister Linden
Francuska	CNRS, LIMSI*	Joseph Mariani
	ELDA*	Khalid Choukri
Grčka	ILSP, R.C. "Athena"*	Stelios Piperidis
Hrvatska	Sveučilište u Zagrebu, Odsjek za lingvistiku	Marko Tadić
Island	Sveučilište na Islandu	Eirikur Rögnvaldsson
Irska	Dublin City University*	Josef van Genabith
Italija	Consiglio Nazionale Ricerche*	Nicoletta Calzolari
	Fondazione Bruno Kessler*	Bernardo Magnini
Letonija	Tilde	Andrejs Vasiljevs
	Sveučilište u Letoniji	Inguna Skadina
Litva	Institute za litavski jezik	Jolanta Zabarskaitė
Luxembourg	Arax Ltd.	Vartkes Goetcherian
Madžarska	Madžarska akademija znanosti	Tamás Váradi
	Tehničko sveučilište u Budimpešti	Géza Németh, Gábor Olaszy
Malta	Sveučilište na Malti	Mike Rosner
Nizozemska	Universiteit Utrecht*	Jan Odijk
Norveška	Sveučilište u Bergenu	Koenraad De Smedt
Njemačka	DFKI*	Hans Uszkoreit, Georg Rehm
	RWTH Aachen*	Hermann Ney
Poljska	Poljska akademija znanosti	Adam Przepiórkowski
	Sveučilište u Łódźu	Piotr Pezik
Portugal	Sveučilište u Lisabonu	Antonio Branco
	Institut za sistemsko inženjerstvo i računala	Isabel Trancoso
Rumunjska	Rumunjska akademija znanosti	Dan Tufis
	Sveučilište Alexandru Ioan Cuza, Iaşi	Dan Cristea
Slovačka	Slovačka akademija znanosti	Radovan Garabik
Slovenija	Institut Jozef Stefan*	Marko Grobelnik
Srbija	Sveučilište u Beogradu	Dusko Vitas, Cvetana Krstev, Ivan Obradovic
	Institut Mihajlo Pupin	Sanja Vranes
Španjolska	Barcelona Media*	Toni Badia
	Katalansko tehničko sveučilište	Asunción Moreno
	Sveučilište Pompeu Fabra	Núria Bel
Švedska	Sveučilište u Göteborgu	Lars Borin
UK	Sveučilište u Manchesteru	Sophia Ananiadou

[1] European Commission Directorate-General Information Society and Media, *User language preferences online*, Flash Eurobarometer #313, 2011 (http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf).

[2] European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).

[3] UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).

[4] European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

[5] Article 8 (NN 53/91, 28/91, 113/93, 4/94).

[6] Međunarodne novine 18/97.
[7] Atlas Republike Hrvatske, Leksikografski zavod Miroslav Krleža, Zagreb, 1993, pp 66.
[8] http://www.ihjj.hr.
[9] http://savjetnik.ihjj.hr.
[10] http://struna.ihjj.hr/o-programu.php.
[11] http://croaticum.ffzg.hr/.
[12] http://www.hrvatskijezik.eu.
[13] See also Tadić (2003).
[14] Batnožić et al. (1996).
[15] http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html.
[16] http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic _search_capabilities.html
[17] http://hml.ffzg.hr.
[18] http://nl.ijs.si/MTE.
[19] http://www.cadial.org.
[20] http://cadial.hidra.hr.
[21] http://tcts.fpms.ac.be/synthesis/mbrola.html.
[22] Steinberger et al. (2006).
[23] Koehn et al. (2009).
[24] http://www.letsmt.eu.
[25] http://www.accurat-project.eu.
[26] Skadiņa et al. (2010).
[27] Eisele & Xu (2010).
[28] Vasiļjevs et al. (2010).
[29] Što je ocjena viša, to je prijevod bolji. Ljudski prevoditelji postižu ocjenu oko 80. Papineni et al. (2002).
[30] http://www.hina.hr.
[31] http://ktlab.fer.hr.
[32] Mikelić Preradović et al. (2007).
[33] Bujas, Ž. (1974).
[34] Tadić, M. (1997).
[35] Moguš et al. (1999).
[36] http://hnk.ffzg.hr.
[37] Tadić, M. (2002).
[38] Tadić, M. (2009).
[39] http://crodip.ffzg.hr
[40] http://riznica.ihjj.hr.
[41] http://rmjt.ffzg.hr.
[42] http://hobs.ffzg.hr.
[43] http://rmjt.ffzg.hr/p3_en.html.
[44] Agić et al. (2008).
[45] Agić et al. (2009).
[46] http://zprojekti.mzos.hr/page.aspx?pid=97&lid=1.
[47] Mikelić Preradović, N. (2010).
[48] http://zprojekti.mzos.hr/page.aspx?pid=97&lid=1&progID=382&projID=1486.
[49] Gianni Lazzari: „Sprachtechnologien für Europa“, 2006: http://tcstar.org/pubblicazioni/D17_HLT_DE.pdf.