POROČILO Z DELAVNICE PROJEKTA *EUROPEAN LANGUAGE RESOURCE COORDINATION* (ELRC) V LJUBLJANI (8. 12. 2015)

Katja ZUPAN
Institut »Jožef Stefan«

Zupan, K. (2016): Poročilo z delavnice projekta European Language Resource Coordination (ELRC) v Ljubljani (8. 12. 2015). Slovenščina 2.0, 4 (1): 1–9.
DOI: http://dx.doi.org/10.4312/slo2.0.2016.1.1-9.

Delavnica projekta *European Language Resource Coordination (ELRC)* je potekala 8. decembra 2015 na Institutu »Jožef Stefan« (IJS) v Ljubljani. Organizirala sta jo Center za prenos znanja na področju informacijskih tehnologij ter Laboratorij za umetno inteligenco IJS skupaj s Predstavništvom Evropske komisije v Sloveniji. Nacionalni koordinator dogodka je bil predstavnik ELRC v Sloveniji Simon Krek z IJS, konzorcij ELRC pa je zastopal Stelios Piperidis. Delavnice se je udeležilo 38 udeležencev, večinoma predstavnikov ministrstev in drugih javnih služb, pa tudi računalniški strokovnjaki in samostojni prevajalci. Videoposnetek delavnice in posamezne predstavitve si je mogoče ogledati na portalu Videolectures.

V okviru projekta ELRC strokovnjaki Evropske komisije, jezikovni tehnologi, ponudniki jezikovnih storitev in nacionalne javne uprave ter vladne službe sodelujejo pri prepoznavanju potreb glede strojnega prevajanja, pri odkrivanju relevantnih večjezičnih jezikovnih virov, obravnavajo pa tudi tehnična in pravna vprašanja, povezana z uporabo podatkov za potrebe strojnega prevajanja. Evropska komisija je v sklopu programa *Connecting Europe Facility* (CEF) vzpostavila platformo za strojno prevajanje CEF.AT. Platforma temelji na sistemu MT@EC, razvitem v Generalnem direktoratu za prevajanje Evropske komisije in ponujenem v uporabo vsem javnim upravam v državah članicah. Ključni cilj platforme je omogočiti večjezično komuniciranje in

izmenjavo dokumentov med javnimi upravami v Evropi, za dosego tega cilja pa bi bilo treba izboljšati pokritost in kakovost prevajalnega sistema, da bo ta čim bolje zadostil potrebam uporabnikov. Namen delavnice ELRC v Ljubljani je bil spodbuditi diskusijo o tem, kakšne so potrebe slovenske javne uprave pri jezikovnih virih in katere večjezične jezikovne vire lahko ta ponudi za izboljšanje delovanja MT@EC.

V pozdravnih govorih so svoj pogled na vlogo strojnega prevajanja kot sredstva za podporo večjezičnosti podali predstavniki ključnih deležnikov. Bronka Straus z ministrstva za izobraževanje, znanost in šport je poudarila pomen vključenosti področja strojnega prevajanja v dokumente, ki urejajo nacionalno jezikovno politiko za obdobje do l. 2018. Maja Pavlović s Predstavništva Evropske komisije v Sloveniji je izpostavila, da je zagotavljanje večjezičnosti postavljeno pred mnogo izzivov. Glavni izziv je dejstvo, da se potrebe po dostopnosti besedil v več jezikih povečujejo, časa za njihovo pripravo pa je vedno manj. Zaradi hitrega napredka tehnologije se spreminja tudi način dela prevajalcev. Pri svojem delu si lahko vse bolj pomagajo s strojnim prevajanjem, ob povečani storilnosti pa je ključni izziv zagotavljati ustrezno kakovost prevoda, tj. pripraviti besedilo, ki ga bralec dejansko razume. Prevajalska služba EU, tj. Generalni direktorat za prevajanje, si prizadeva združiti oba izziva, slediti razvoju in zagotavljati kakovost – l. 2013 je bilo zato uvedeno orodje za strojno prevajanje MT@EC, ki je prilagojeno prevajanju besedil EU, saj kot osnovo (učno množico) uporablja besedila, terminologijo in besedne zveze iz obsežne količine človeških prevodov institucij EU, nastalih v zadnjih desetletjih. Vključuje vseh 24 uradnih jezikov EU oz. 552 jezikovnih parov. Orodje vzdržujejo in stalno nadgrajujejo interni strokovnjaki ob tesnem sodelovanju s prevajalci iz evropskih ustanov. Vizija strojnega prevajanja je krepitev večjezičnosti vseevropskih digitalnih storitev, kot so npr. e-pravosodje, e-zdravje, e-računi, pa tudi prostega pretoka digitalnih storitev med gospodarstvi držav EU. Trenutno namreč le 15 % kupcev kupuje v spletnih trgovinah v drugi državi članici EU in le 7 % malih oz. srednjih podjetij svoje

izdelke prodaja prek spleta v druge države članice. To kaže, da ima digitalni trg v EU še zelo veliko razvojnega potenciala, tudi v smislu zagotavljanja novih delovnih mest. Ključno je torej olajšati komuniciranje med podjetji, organi posameznih držav EU in državljani samimi.

Po uvodnih govorih je Stelios Piperidis predstavil potek in cilje delavnice. Program delavnice je bil sestavljen iz predavanj predstavnikov ELRC, Evropske komisije, Filozofske fakultete UL in Inštituta za intelektualno lastnino, med katera pa sta bili umeščeni tudi dve panelni diskusiji, ki so ju pripravili predstavniki Sektorja za prevajanje Generalnega sekretariata Vlade RS.

Piperidis je nato kot prvi predavatelj opisal prednosti nadgradnje dela prevajalcev z (delno) avtomatiziranimi (AT) ali strojnimi (MT) prevajalskimi sistemi. Pri prevajanju, katerega cilj so visoko kakovostni prevodi, je smiselna uporaba avtomatiziranega sistema. V EU so tako že dalj časa v uporabi avtomatizirani sistemi (pomnilniki prevodov, terminološke baze in druga orodja) kot tehnološka podpora prevajalcem v institucijah EU, novi CEF.AT pa bo brezplačno na voljo tudi javnim upravam držav članic EU. Za čim boljše delovanje pa tovrstni delno ali povsem avtomatizirani sistemi potrebujejo čim večjo količino podatkov, tj. učno množico prevodov. Ključni deležniki, ki lahko priskrbijo primerne podatkovne vire, so po besedah Piperidisa javni sektorji (predvsem državna uprava in, širše, javna uprava) držav članic, nevladne organizacije in institucije EU. Proces pridobivanja virov kot posrednik usmerja ELRC s svojimi nacionalnimi točkami. V Sloveniji sta predstavnika tehnične oz. znanstvene javnosti Simon Krek in Marko Grobelnik, nacionalni organi oz. javni sektor bodo imenovali svojega zastopnika, prevajalce pa zastopa Peter Jakša s Predstavništva Evropske komisije v Sloveniji.

V telekonferenčnem predavanju je Spyros Pilos z Generalnega direktorata za prevajanje opisal, kako lahko javne ustanove uporabljajo platformo CEF.AT. Podrobno je opisal sistem MT@EC (ki je tudi gradnik te platforme), in sicer z

vidika podprtih jezikov, tehnologij, uporabniškega vmesnika in drugih tehničnih lastnosti, kot so vhodni format, prikaz rezultatov in varnost dokumentov. Predstavil je tudi nekaj statističnih podatkov o uporabi sistema in povabil predstavnike javnih ustanov, naj ga preizkusijo. Izpostavil je tudi koristi, ki jih uporaba CEF.AT prinaša uporabnikom tako glede kakovosti prevoda, varnosti kot tudi prilagojenosti specifičnim potrebam uporabnika in posameznim področjem.

Peter Jakša je predstavil politiko večjezičnosti na ravni EU in vlogo – tudi samodejnega – prevajanja. Najpogostejši način dela prevajalskih služb v Evropski komisiji je prevajanje iz angleščine v ciljni jezik. Slovenska služba tako kar v 80 % primerov prevaja besedila iz angleščine v slovenščino, skupno pa prevede približno 80.000 strani letno. V splošnem se količina prevodov povečuje, čas za prevode pa krajša. Kot je opozoril Jakša, je enoten digitalni trg še en korak k enotnemu trgu blaga in storitev, hkrati pa tudi evropske javne uprave nimajo enotnega jezika sporazumevanja, zato je cilj CEF.AT podpora večjezičnosti evropskih javnih spletnih storitev s poudarkom na tistih, ki so tipično vseevropske, kot so npr. javna uprava, javna naročila oz. razpisi in zdravstvene storitve. Na praktični ravni je za dosego tega cilja potrebna uvedba izpopolnjenih jezikovnih tehnologij (kar je tudi cilj npr. MT@EC) na varni platformi. Tudi Jakša je predstavil statistične podatke, ki kažejo potrebo po večjezičnosti enotnega digitalnega trga: okrog 90 % evropskih potrošnikov najraje išče po spletu v lastnem jeziku, več kot 80 % spletnih trgovin v EU pa je enojezičnih.

Simon Krek z IJS in iz Centra za jezikovne vire in tehnologije Univerze v Ljubljani je predstavil stanje jezikovne politike ter razvoj jezikovnih tehnologij in virov v Sloveniji. Povzel je izsledke primerjalne analize stanja jezikovnih tehnologij v državah članicah EU. Kritično je analiziral vsebino nacionalnih dokumentov o jezikovni politiki in jezikovnem načrtovanju, predvsem Resolucije o nacionalnem programu za jezikovno politiko 2014–2018 in Akcijskega načrta za jezikovno opremljenost (2015). Kot svetle zglede

jezikovnega načrtovanja je naštel Estonijo, Litvo in Portugalsko. Izpostavil je problematiko opremljenosti slovenskega jezika za obstoj oz. komuniciranje v razvijajoči se digitalni dobi, kjer se nakazuje npr. razvoj aplikacij, interneta stvari (ang. *internet of things, IOT*) – ali bodo te na voljo v slovenščini? Po Krekovi oceni slovenščina pri razvoju jezikovnih tehnologij povprečno zaostaja 10 let za »velikimi« jeziki. Trenutno aktualna področja za slovenski jezik so tako komuniciranje, semantika in umetna inteligenca.

Sledil je panel o večjezičnih servisih v javni upravi. Svoje delo v Sektorju za prevajanje so predstavili Marčela Novljan Lovrinčić, Miran Željko, Adriana Sedej in Marja Adamič. Prevajalci se tu večinoma ukvarjajo z besedili v angleščini in nemščini, dobijo pa tudi nekaj (a veliko manj) besedil v francoščini in italijanščini. Pomagajo si z zunanjimi izvajalci, kot so rojeni govorci ciljnega jezika, imajo pa tudi lektorje slovenskega jezika. Besedila, ki jih dobijo v prevod, so namreč pogosto jezikovno in stilno neustrezna. S svojimi prevodi že od l. 1998 pomagajo graditi zbirki Evroterm in Evrokorpus. Pri primerih, ki jih ponudi strojno prevajanje, za zdaj opažajo, da vsebujejo še preveč šuma. Njihovo čiščenje zahteva dodatno delo, zato proces praktično ni hitrejši od ročnega prevajanja, čeprav je načeloma ravno to njegov cilj. Prevodi prevajalske službe morajo biti namreč izredno kakovostni, zato delovni proces, pri katerem bi samo pregledali oz. popravili strojni prevod z »uredniškimi« posegi, zanje trenutno ni sprejemljiv. Pri prispevanju dodatnih jezikovnih virov za gradnjo korpusa, ki bi predstavljal učno množico sistema za strojno prevajanje, sta dodatni težavi tudi stopnja zaupnosti in zahteva po varstvu osebnih podatkov, saj služba nima časa za ukvarjanje z anonimizacijo.

V nadaljevanju sta Stelios Piperidis in Simon Krek predstavila, kako deluje strojno prevajanje, ki ga uporablja orodje MT@EC. Gre za statistično strojno prevajanje na osnovi prevajanja besednih zvez, za katerega MT@EC uporablja uveljavljeni odprtokodni sistem Moses. Izpostavila sta, da strojno prevajanje za

kakovostno delovanje statističnih metod potrebuje čim večjo podatkovno učno množico, sestavljeno tako iz paralelnih prevodov kot enojezičnih podpornih virov za ciljni jezik.

Špela Vintar je predstavila sodobne smernice na področju prevajalske dejavnosti. Po njenih besedah pravica do prevoda postaja temeljna človekova pravica, s čimer se povečujeta povpraševanje po prevajalskih storitvah – v Sloveniji je registriranih kar 1.205 prevajalskih podjetij – in število jezikov, med katerimi se prevaja. Spreminja se tip besedil oz. vsebin, te so vse bolj dinamične in se generirajo v trenutku dostopa. Niso več nujno omejene na en medij, temveč so pogosto avdio-vizualne in uporabniki do njih vse pogosteje dostopajo z mobilnih naprav. Zahtevana kakovost prevodov postaja vse bolj gibljivo merilo in je odvisna predvsem od namena, za katerega bo prevedeno besedilo uporabljeno. Zlasti pri prevodih, ki so rezultat neke hipne potrebe in kot taki enkratne narave, kot je npr. prevod spletne strani, je kakovost lahko nižja. Uveljavljajo se tudi platforme za množičenje prevodov. Po mnenju Š. Vintar je prihodnost prevajanja strojno prevajanje po meri posameznika v realnem času (velike količine javno dostopnih podatkov). Trenutno se strojno prevajanje uporablja kot vtičnik prevajalskemu orodju, torej kot pomoč prevajalcu, vendar pa so storitve kakovostnega prevajanja, ki dejansko olajšajo delo, večinoma plačljive, podatkovne množice (pomnilniki prevodov, terminološke baze) pa večinoma niso javno dostopne.

Maja Bogataj Jančič z Inštituta za intelektualno lastnino je predstavila pravni okvir pridobivanja jezikovnih podatkov. Področje na evropski ravni ureja direktiva 2013/37/EU, ki spodbuja odprti in prosti dostop do javnih podatkov, kar vključuje tudi kulturne institucije (muzeje, knjižnice). Druge relevantne zakonske podlage na nacionalni ravni so še Direktiva o ponovni uporabi informacij javnega značaja (2003), Zakon o dostopu do informacij javnega značaja (IJZ), Zakon o informacijskem pooblaščencu, Zakon o upravnem postopku in Uredba o posredovanju o ponovni uporabi IJZ. Kako ravnati ob prejetju zahtevka po IJZ, opisuje publikacija *Podatki: neusahljivi viri*

poslovnih idej!, ki jo je izdal urad Informacijskega pooblaščenca.

V drugem delu srečanja je sledil panel o testiranju okolja MT@EC. Rezultate testiranj je predstavil Milan Željko. MT@EC trenutno prevaja slabše in počasneje kot Google Translate. Dokaj dobro pozna prevode predpisov EU, če pa besedilo ni povezano s tematiko EU, prevajalnik še ni uporaben. Željko je predstavil tudi nekaj praktičnih primerov napak: orodju povzročajo težave nedeljivi presledki in ročni prelomi vrstic, pri zapisu denarnih zneskov v obliki »EUR 3.4« prevajalnik obstoječe pike ne prevede v decimalno vejico, ne pozna nekaterih tipičnih besed v neimenovalniških sklonih, in ker prevaja prek angleščine, se pojavljajo medjezikovne interference. Po besedah Željka so prevodi iz slovenščine v angleščino običajno boljši kot prevodi v nasprotni smeri.

Stelios Piperidis je nato opisal še nekaj tehničnih vidikov in dobrih praks pri pridobivanju jezikovnih virov, primernih za podatkovno množico MT@EC. Uporabni so zlasti jezikovni viri, ki imajo že privzet odprti dostop, tj. so informacije javnega značaja. Digitalni format podatkov mora biti tekstovni in ne PDF. Na praktični ravni to pomeni, da je najprej treba najti in izbrati podatke oz. primerne jezikovne vire, preveriti njihov pravni status in licence, podatkovno množico nato očistiti (pretvorba v kodiranje UTF8, odstranitev oblikovanja, konverzija v ustrezen format, npr. XML), anonimizirati s prepoznavo imenskih entitet in referenc nanje ter po preverbi vključiti v repozitorij in/ali korpus. ELRC nudi pomoč tako pri vprašanjih tehnične, pravne, kot tudi metapodatkovne narave (kako ustrezno opisati podatke), skrbi pa tudi za repozitorij META-SHARE.

V sklepnem delu je Simon Krek povzel glavna dva cilja delavnice, tj. informirati lokalno javnost o razvoju storitev in tehnologij za strojno prevajanje v okviru EU ter pozvati javno upravo v Sloveniji oz. vse institucije, ki jih tematika zadeva, k pomoči pri pridobivanju ustreznih jezikovnih virov. Ob uresničitvi teh dveh ciljev bo tudi za slovenski jezik v prihodnje mogoče doseganje ustrezne

kakovosti prevodov, izdelanih s pomočjo (delno) avtomatskega prevajanja.

Delavnica se je sklenila z vprašanji in komentarji udeležencev, ki so se osredotočili na vlogo in način dela prevajalca v prihodnosti. Izraženi so bili pomisleki o kakovosti strojnih prevodov in tudi zaskrbljenost glede prihodnosti poklica prevajalca. Prevladalo pa je stališče, da strojno prevajanje ne bo nadomestilo prevajalcev, temveč se bodo delno spremenile njihove delovne naloge. Prihranjeno jim bo bolj rutinsko, zelo repetitivno delo, hkrati pa bodo potrebne nove naloge – tehnološko usposobljeni prevajalec bo oblikoval vhodno in izhodno besedilo strojnega prevajalnika, analiziral njegove napake, pomagal pri grajenju jezikovnih virov, ki so potrebni za izboljševanje prevajalnika, luščil terminologijo ipd. Občinstvo je opozorilo tudi na potrebo po jasnejših smernicah za anonimizacijo dokumentov in na preobremenjenost prevajalcev, saj da ti nimajo časa za dodatno predelavo besedil v format, ki ga zahteva MT@EC. Udeleženci delavnice so se ob koncu strinjali, da strojno prevajanje ne bo nikoli delovalo popolno in zato tudi ne povsem samostojno, a odpira nove možnosti za učinkovito prevajanje in pomaga obvladovati vse večjo količino prevodov, kar je še posebej pomembna praktična spodbuda za pomoč pri izboljšanju sistema MT@EC.

EUROPEAN LANGUAGE RESOURCE COORDINATION (LREC) WORKSHOP IN LJUBLJANA (DECEMBER 8, 2015)