ZBIRNIK ODGOVOROV NA VPRAŠANJA MINISTRSTVA ZA KULTURO

1. V okviru javnega razpisa se pričakuje, da bo konzorcij pri razvoju jezikovnotehnoloških izdelkov uporabil že izdelane odprtokodne jezikovne vire in tehnologije. Na spletni strani MK (http://www.mk.gov.si/si/delovna_podrocja/sluzba_za_slovenski_jezik/predstavitev_podrocja/dogodki_javne_razprave/) je objavljen zbirnik dosežkov s področja digitalnih jezikovnih virov, tehnologij (JVT) in storitev, zanima pa nas (najprej splošno vprašanje), koliko so navedeni dosežki uporabni za nadaljnji razvoj in gradnjo JVT?

Inštitut za slovenski jezik Frana Ramovša ZRC SAZU:
V seznamu gre za zbirnik vse mogočih gradiv – od baz do orodij, pri čemer iz podatkov v tabeli niso razvidne njihove tehnične karakteristike, obseg ipd. Zdi se smiselno, da je MK predstavilo nabor, ki ga bo v posameznih segmentih mogoče uporabiti, vendar je treba to presojo prepustiti izvajalcem konkretnih nalog/projektov.

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Načeloma so vsi zbrani jezikovni viri uporabni za projekt, njihova uporabnost je pogojena predvsem z dostopnostjo – čim bolj je gradivo dostopno (pod odprtimi licencami), tem lažje ga je uporabiti za razvoj JVT. Enako velja za orodja - odprtokodnost pomeni boljše možnosti za uporabo. Druga omejitev je prilagojenost obstoječe rešitve sodobnim tehnologijam, tj. zastarelost programske opreme starejšega datuma. V svetu so v zadnjih treh letih nove tehnologije razvijajo na osnovi globokih nevronskih mrež. Med obstoječimi tehnologijami takšnih rešitev ni, kar pomeni, da jih bo verjetno treba razviti na novo.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
V okviru razvoja razpoznavalnika govora (ASR) je smiselno uporabiti in nadgraditi naslednje že obstoječe jezikovne vire: korpus Gos ter Gos Videolectures (kot del govorne baze za ASR), korpus Gigafida (pri čemer bi bil potreben dostop do n-gramov oz. do tekstovnih datotek za učenje jezikovnih modelov), Sloleks (pri čemer je potrebna nadgradnja zlasti s fonetičnim zapisom). Pri tem je treba poudariti, da so ti navedeni viri na voljo samo pod licenco za nekomercialno rabo.

V okviru področja strojnega prevajanja je smiselno združeno uporabiti vse na spletni strani MK objavljene vzporedne korpuse in jih nadgraditi oz. razširiti z novimi korpusi iz različnih spletnih virov. Veljalo bi razmisliti tudi o gradnji novega vzporednega korpusa, ki bi temeljil na izhodu ASR in pridruženih prevodih, v kolikor je ideja tudi prevajanje govor v govor. Pri uporabi obstoječih virov bi bilo potrebno uravnotežiti uporabo virov pisanega in govorjenega jezika, glede na to ali gradimo prevajalnik govora ali prevajalnik pisanega jezika.

V okviru semantike in razumevanja govorjenega jezika (SLU) kot podpornega mehanizma k bolj zanesljivemu ASR bo potrebno poznati ciljne domene in ključne termine (besedne zveze), ki bi lahko opredeljevali posamezno domeno. V povsem odprtih domenah je namreč ena izmed nalog SLU klasificirati domeno oz. kontekst v katerem nahaja ter skozi pomen definirati manjkajoče enote informacije ali celo zapolniti nekatere manjkajoče enote. Podobne tehnike bi bilo moč razširiti tudi na ASR in druge jezikovno-tehnološke izdelke (npr. prevajalnik, sintetizator, itn.). Načeloma viri za SLU niso nujno več-modalni in so

lahko ustrezno klasificirana besedila, ki vključujejo dovolj 'ključnih besed' in pomenskih enot s katerimi obogatimo. Možno je uporabiti obstoječe vire in transkripcije, seveda pa je zaželena anotacija novih nastalih v ciljnih domenah. Se pa v splošnem ti viri dobro povezujejo tudi s preostalimi načrtovanimi izdelki (tako prevajalnik, kot tudi terminološki slovar).

Raziskovalci iz Alpineona in Fakultete za elektrotehniko Univerze v Ljubljani (Laboratorij za strojno inteligenco):
Vprašanje je zelo splošno zastavljeno, zato odgovor podajamo izključno za sklop 2 – govorne tehnologije – razvoj razpoznavalnika govora. Zvočni posnetki najbolj obsežnega govornega korpusa GOS 1 so dostopni zgolj preko spletnega konkordančnika, pa še ti govorni posnetki niso izvorni, ampak so bili digitalno obdelani, da se zakrije identiteta govorca. Za razvoj razpoznavalnika govora so uporabni ostali govorni korpusi, ki so navedeni v omenjenem zbirniku (denimo SOFES 1.0 in SNABI), pa tudi novejši govorni korpusi, ki v zbirniku niso omenjeni (denimo GOS-Videolectures, SI TEDx-UM). Vendar večina omenjenih virov ni na voljo pod CC BY 4.0 licenco in jih ni možno uporabljati za izgradnjo komercialnih produktov, kar predstavlja enega izmed ciljev načrtovanega programa.

2. Koliko ur posnetega govora bi potrebovali za izdelavo visokozanesljivega razpoznavalnika govora?

Inštitut za slovenski jezik Frana Ramovša ZRC SAZU:
Velja, da bi za razpoznavalnik govora potrebovali od 1.000 do 10.000 ur označenih in prepisanih govornih posnetkov. Trenutno jih imamo za približno 150 ur. Poleg splošnega govora je treba poskrbeti še za druge zvrsti, dialekte, strokovni jezik itd.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
Za izdelavo visokozanesljivega razpoznavalnika govora bi ob uporabi novih tehnologij potrebovali kvalitetno obdelano govorno bazo vsaj v obsegu 400 ur govora (čim več tem bolje). Potrebno pa bi bilo pridobiti tudi dodatne, nenatančno zajete posnetke in transkripcije iz virov, kot so scenariji TV oddaj, parlament, zvočne knjige ipd., v obsegu nekaj 100 ur govora, da je možna uporaba naprednih metod učenja. Komercialni razpoznavalniki govora za jezike, kot so angleščina, kitajščina, arabščina, itd. so na primer zgrajeni na govornih bazah, ki obsegajo več 1.000 do celo nekaj 10.000 ur posnetkov. Pri tem moramo računati na problem, v kolikšni meri bodo lastniki takšnih virov pripravljeni odstopiti gradiva in pod kakšnimi pogoji. Po naši oceni je težko pričakovati, da bi dobili (vse) takšne vire odprte tudi za komercialno rabo. Glede uporabe virov tudi za komercialno rabo bo potrebno preveriti pripravljenost izvornih lastnikov avtorskih pravic in možne mehanizme, kako jo doseči. Zato bi bilo smiselno znotraj projekta predvideti tudi ekipo, ki bo poiskala primerne rešitve. V okviru priprav novega govornega materiala, bi bilo smiselno izvesti tudi segmentacijo govornega materiala na nivoju stavkov. Dodatno je potrebeno opozoriti, da obseg govornih virov ne bo v celoti zagotovil primerljivosti uspešnosti slovenskega razpoznavalnika govora s komercialnimi produkti. Slovenski jezik je zaradi svojih lastnosti bistveno zahtevnejši za razpoznavanje govora, kot pa npr. angleški jezik.

Kako dobro bo deloval ASR (zanesljivost) je v veliki meri odvisno tudi od načina zajemanja govora in seveda v kakšnem okolju se bo to zajemanje izvajalo. Okolje je lahko studijsko, kjer ocenjujemo, da bi bili rezultati ASR najboljši. Kakor pa je s strani MK razumeti, bo uporaba ASR lahko tudi v mnogo bolj šumnem okolju, kjer bo lahko šum npr. glasba v ozadju ali drugi govorci; to ima velik vpliv na delovanje in zanesljivost ASR. Tudi način zajemanja signala je zelo pomemben za zanesljivost ASR. Kar nekaj znanih aplikacij, ki

podpirajo ASR, deluje na način »push-to-talk« (pritisni in govori), kar pripomore k boljšemu delovanju in večji zanesljivosti. Če takšnega načina uporabe ne želimo, potem nujno potrebujemo kvaliteten algoritem za zaznavanje aktivnosti govora v zajetem avdio signalu. Pomembno je tudi kakšen mikrofon bo uporabljen za zajemanje signala. Mikrofon ob ustih bo načeloma boljše zajemal signal, kot bi to lahko dosegli s poljem mikrofonov.

Potrebno je poudariti, da je visoka zanesljivost zelo odvisna od domene/vrste ASR. Torej tudi, če govorimo o tekočem govoru je odvisno ali je govor spontan, govorjen v odprti ali zaprti domeni, ali želimo transkripcijo govorjenega ali pa potrebujemo koncepte na osnovi katerih se odločamo dalje (npr. prepis govora na sestankih v.s. upravljanje s pametnim domom).

dr. Darinka Verdonik (FERI UM)
Za izdelavo visokozanesljivega razpoznavalnika je potrebnih vsaj 1000 ur govora s primerno širokim spektrom govorcev in primerno kvalitetno transkripcijo. Tak obseg baze je težko doseči znotraj predvidenega razpisa, saj govorimo o obsegu ročnega transkirbiranja ca. 60.000 ur (1 minuta posnetka pomeni uro transkribiranja). K temu je treba prišteti še koordiniranje številnih transkriptorjev in delo s pridobivanjem, editiranjem in predprocesiranjem posnetkov.

Znotraj razpisa je glede na število področij, obseg sredstev ter terminski plan smiselno računati na govorno bazo v obsegu 400 do največ 500 ur posnetkov. Kvalitetne govorne baze oz. korpusi (torej z natančnimi transkripcijami in segmentiranjem) v obsegu 1000 do več 1000 ur posnetkov so stvar dolgoročnejše strateške skrbi za jezikovne vire. Posledično to seveda pomeni, da smo znotraj projekta omejeni tudi s tem, kakšno kvaliteto razpoznavalnika lahko pričakujemo.

Raziskovalci iz Alpineona in Fakultete za elektrotehniko Univerze v Ljubljani (Laboratorij za strojno inteligenco):
Količina potrebnih podatkov za učenje je zelo odvisna od predvidene vrste in namena uporabe. Se pričakuje razpoznavalnik govora za narekovanje ali transkcipcijo kvalitetno zajetih multimedijskih posnetkov, ki ga lahko tudi prilagodimo na govorca? Želimo razviti razpoznavanje govornih ukazov v prostoru za krmiljenje pametnega doma? Ali pa razpoznavanje telefonskih pogovorov s hrupnim ozadjem? Mora biti sposoben sprotne razpoznave govora? In kakšna je definicija visokokakovostnega razpoznavalnika? Kakšen uporabniški vmesnik se pričakuje, na katerih operacijskih sistemih naj bo podprt?

Minimum govornega gradiva, ki je potreben za razvoj razpoznavalnika govora solidne kakovosti, obsega vsaj 1.000 ur ustrezno posnetih in transkribiranih govornih posnetkov, ki zajemajo čim bolj širok spekter govorcev jezika. Priporočljivo je, da razvojna skupina za sklop 2 poda smernice za razvoj govornega korpusa za potrebe razpoznave govora in pri izgradnji tega jezikovnega vira tudi aktivno sodeluje. S povečevanjem obsega posnetkov na 2.000-3.000 ur se pričakuje še kakšnih 10% relativnega izboljšanja zanesljivosti.

Za razvoj jezikovnega modela razpoznavalnika govora, ki ima ključen vpliv na njegovo zanesljivost delovanja, prav tako velja načelo, da več besedilnih učnih podatkov izboljšuje njegovo zanesljivost in da je za razvoj razpoznavalnika splošnega govora potrebno uporabiti obsežne besedilne vire, ki vsebujejo od 100 in tudi do 1000 milijonov besed.

3. Prosimo, ocenite stroške izdelave razvoja razpoznavalnika, upoštevajoč vse dosedanje uporabne dosežke.

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Obstoječe tehnologije so zaradi prehoda na globoke nevronske mreže neuporabne, deloma pa so uporabni obstoječi viri. V razpisu je treba zahtevati izdelavo razpoznavalnika z globokimi nevronskimi mrežami, ki dajejo opazno boljše rezultate od obstoječih tehnologij. Ocena stroškov izdelave je sicer odvisna od odločitev glede potrebne količine zbranega gradiva, uporabljenih tehnologij in drugih dejavnikov, kar je težko določiti brez finančne konstrukcije celotne prijave. Na primer, treba je določiti nivo tehnološke zrelosti, podprte platforme, potrebnost razvoja uporabniških vmesnikov itd. Stroški lahko segajo od 100.000 EUR do približno polovične vrednosti predvidenih razpisanih sredstev.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
Kakor je razbrati iz vprašanj, govorimo o razpoznavalniku spontanega govora za slovenski jezik, ki bi bil uporabljan v različnih okoljih z različnimi možnostmi/zahtevnostmi glede zajemanja in tudi razpoznaval v različnih domenah (npr. predavanja ali v inteligentnih okoljih). Ocenjujemo, da bo v sklopu navedenih ciljev največji del stroškov vezan na izdelavo ustreznih govorjenih in pisnih jezikovnih virov. Kakšni bi bili ti stroški je težko oceniti, saj je najprej potrebna specifikacija sistema, ki bi jo morali v konzorciju ustrezno oblikovati, hkrati pa tudi določiti, katere obstoječe vire bomo imeli možnost uporabiti pri razvoju. Torej ali bo možno kaj pouporabiti, nadgraditi, ali pa bo potrebno potrebne vire na novo ustvariti. Posledično so že zaradi tega celotni stroški izdelave razpoznavalnika v tem trenutku težko določljivi. Od kakovosti razpoložljivih virov je tudi odvisno, kako kvalitetno je mogoče zajeto gradivo avtomatsko procesirati, npr. avtomatsko segmentirati, in koliko je pri tem še potrebnega ročnega dela, zlasti ker želimo visoko zanesljiv sistem. Nadalje so tu razpoložljivi človeški viri in njihova cena, ki tudi predstavljajo neznanko, dokler konzorcij ni sestavljen.

Pri tem je potrebno poudariti, da je danes velika večina sistemov avtomatskega razpoznavanja govora razvitih in uporabljanih v smislu storitev najboljših zmožnosti (best effort) in ne storitev z zagotovljeno kvaliteto (quality of service – QoS), kar pomeni, da ne obstaja prag zagotovljene kvalitete storitve v vseh pogojih delovanja. To je pomembno predvsem pri definiranju zmogljivosti avtomatskega razpoznavalnika govora za različna področja in akustična okolja uporabe.

Raziskovalci iz Alpineona in Fakultete za elektrotehniko Univerze v Ljubljani (Laboratorij za strojno inteligenco):
Ocena stroškov je zelo okvirna, ker je močno odvisna od pričakovane vrste rabe razpoznavalnika govora. Ob upoštevanju razpoložljivih jezikovnih virov ocenjujemo stroške izdelave dodatnih potrebnih govornih virov na ekvivalent 1.5 malega ARRS projekta (okoli 450.000 EUR). Stroške razvoja visokokakovostnega razpoznavalnika govora do nivoja uporabnega produkta ocenjujemo na ekvivalent do dveh velikih ARRS projektov (okoli 1.200.000 EUR).

4. Prosimo za oceno delovnih ur za izdelavo razpoznavalnika.

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Odgovor je enak kot pri prejšnjem vprašanju. Predvidevamo, da bi približno polovica sredstev šla za zbiranje in transkripcijo gradiva, pol za razvoj aplikacije.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
Menimo, da je za oceno delovnih ur za izdelavo razpoznavalnika najprej potrebno določiti ključne zahteve razpoznavalnika oziroma izvesti njegovo osnovno funkcionalno specifikacijo. Glede na trenutno stanje tehnike in raznolikost možnih področij uporabe ter pogojev delovanja je to predpogoj za kakršnokoli realno oceno potrebnega vložka dela.

Raziskovalci iz Alpineona in Fakultete za elektrotehniko Univerze v Ljubljani (Laboratorij za strojno inteligenco):
Predvideni stroški podizvajalcev so razmeroma majhni v primerjavi s celotnim stroškom izdelave razpoznavalnika govora, tako da ocena potrebnega števila delovnih ur neposredno izhaja iz predvidenih stroškov, ki so ocenjeni v točki 2.

5. Prosimo, ocenite stroške izdelave izboljšanega prevajalnika za jezikovni par SL-AN in AN-SL.

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Odgovor je podoben kot v primeru razpoznavalnika govora. Potreben je prehod na tehnologijo globokih nevronskih mrež. V tej fazi stroškov realna ocena ni možna brez razmeroma kompleksnih odločitev glede količine zbranega gradiva (minimalno 10 milijonov stavkov v vzporednih korpusih) in izbranih tehnologij. Ocena stroškov je v grobem podobna kot pri razpoznavalniku in sega nekje od 100.000 EUR do četrtine vrednosti razpisa.
Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
Preden je mogoča kakršnakoli realna ocena stroškov izdelave izboljšanega prevajalnika je potrebno definirati osnovne zmogljivosti obstoječega prevajalnika in določiti ciljne vrednosti izboljšanega prevajalnika. Tako kot pri sistemu avtomatskega razpoznavanja govora so tudi v primeru prevajalnika ključni dovolj obsežni jezikovni viri, konkretno vzporedni korpusi, delno tudi obsežni pisni korpusi, kot je Gigafida. Koristijo lahko tudi dvojezični slovarji. Enako kot pri razpoznavalniku je pomembno, na katero področje/domeno je prilagojen prevajalnik oz. ali je splošni ali specifični ter razpoložljivost/izdelava ustreznih virov za področje/domeno. Zato je tudi v tem primeru najprej potrebno izvesti oceno ustreznosti obstoječih jezikovnih virov ter oceno obsega dela pri pripravi novih oziroma dopolnjenih jezikovnih virov, glede na zahteve izboljšanja prevajalnika za jezikovni par SL-AN in AN-SL.

6. Koliko delovnih ur bi potrebovali zato?

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Odgovor je enak kot pri prejšnjem vprašanju. Približno pol sredstev bi šlo za zbiranje gradiva (vzporedni korpusi prevodov itd), pol za razvoj aplikacije in drugih potrebnih orodij, med katere spada tudi čim boljši strojni anonimizator besedil.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
Ocena ur bo mogoča ob predhodni specifikaciji zahtev izboljšanja zmogljivosti prevajalnika in ocene stroškov izdelave potrebnih jezikovnih virov.

7. Je morda v razvoju prevajalnik še za kakšen drug jezikovni par? Če da, kako daleč je ta razvoj?

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Načeloma je s tehnologijami nevronskih mrež mogoče za vsak par, za katerega obstaja vzporedni korpus, razviti enak prevajalni sistem. Uporabljena tehnologija bi bila približno enaka, tako da je mogoče predvidevati, da je strojni prevajalnik mogoče razviti vsaj za vse uradne jezike EU, ker obstaja odprto dostopen korpus prevedene zakonodaje EU. Verjetno

pa bi strojni prevajalnik za druge jezikovne pare v vsakem primeru deloval slabše, ker toliko gradiva, kot ga je mogoče dobiti za par z angleščino, ne bo na voljo za noben drug jezik.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
Prehod na nove jezikovne pare je v veliki meri odvisen od razpoložljivih vzporednih korpusov.

8. Koliko delovnih ur bi potrebovali za razvoj semantičnih tehnologij (nevronskega lematizatorja, oblikoskladenjskega označevalnika in skladenjskega razčlenjevalnika)?

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Dejansko lematizator, oblikoskladenjski označevalnik in skladenjski razčlenjevalnik ne spadajo k semantičnim tehnologijam, ampak predstavljajo osnovna morfološko-sintaktična orodja za obdelavo naravnega jezika. Semantične tehnologije se začno z srednjim nivojem tehnologij, za naloge kot so prepoznavanje imenskih entitet, povezav, anafore in koreferenčnosti ter nadaljujejo z nevronskimi jezikovnimi modeli, vektorskimi vložitvami, označevanjem udeleženskih vlog, semantičnimi okvirji in razdvoumljanjem pomena besed v sobesedilu. Vrhnji nivo semantičnih tehnologij so naloge razumevanja jezika: strojno prevajanje, avtomatsko povzemanje, odgovarjanje na vprašanja, generiranje naravnega jezika, napredni pripomočki za pisanje itd. Za vse te tehnologije je smiselno uporabljati tehnologije globokih nevronskih mrež, ki dajejo bistveno boljše rezultate od obstoječih tehnologij. Tehnologijo globokih nevronskih mrež je treba tudi nekoliko prilagoditi slovenščini. Na kratko so posamezni nivoji zanje na kratko opisani tu: https://nlpforhackers.io/intro-natural-language-processing/ (The NLP Pyramid). Ocena delovnih ur za razvoj semantičnih tehnologij je ponovno odvisna od zahtevanega nivoja tehnološke zrelosti. Za res uspešna orodja je potrebno povečati učne množice in jih razširiti predvsem z novimi učnimi primeri za jezikovne kategorije, kjer so analize obstoječih rešitev pokazale nezadovoljive rezultate. Za vsako od osnovnih orodij je za tehnološki razvoj tipično treba računati vsaj 100.000 EUR (polovica za razvoj orodja, polovica za širitev učne množice).

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
Navedeni primeri semantičnih tehnologij predstavljajo zelo nizko raven in praviloma niti niso razumljeni kot semantični, pač pa kot oblikoslovni in skladenjski viri. V okviru semantičnih virov in tehnologij je smiselno razmišljati o virih in tehnologijah, kot so označevanje in klasifikacija konceptov (oz. model bag of concepts), razreševanje anafor, označevanje/analiza imenskih entitet, Wordnet, Framenet ipd.

9. Ali je treba infrastrukturo, torej repozitorij CLARIN.SI, za namen razpisa dograditi s strojno ali programsko opremo?

Institut »Jožef Stefan«:
Če bo CLARIN.SI uspešen pri neposrednem pozivu MIZŠ za nakup raziskovalne opreme raziskovalnih infrastruktur, potem ne.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
UM FERI, Laboratorij za digitalno procesiranje signalov, že razpolaga z najnovešjo strojno in programsko opremo, ki jo lahko zelo učinkovito uporabimo za namen razvoja razpoznavalnika govora in sistem strojnega prevajanja.

dr. Darinka Verdonik (FERI UM)
CLARIN.SI se res poteguje na razpisu za novo infrastrukturo, ki bo – ob pogoju, da bo prijava uspešna, česar še ne vemo – zadoščala za pokritje potreb repozitorija. Vendar bo sam projekt zahteval tudi močno strojno opremo za razvoj samih aplikacij. Koliko te opreme je na voljo, pa je odvisno od posameznih konzorcijskih partnerjev. Vsekakor pa je za pričakovati, da bo za uspešno izvedbo projekta potrebnih nekaj vlaganj tudi v strojno opremo.

Raziskovalci iz Alpineona in Fakultete za elektrotehniko Univerze v Ljubljani (Laboratorij za strojno inteligenco):
Infrastruktura CLARIN.SI se poteguje za sredstva za nadgradnjo strojne opreme, kar bo predvidoma zadoščalo za realizacijo sklopa 2 za razvojne potrebe. V fazi produkcije in široke uporabe bo potrebno storitev preseliti na bolj zmogljivo produkcijsko infrastrukturo, če se pričakuje strežniška storitev razpoznave govora.

10. Ali je v okviru razpisanih sredstev mogoče podpreti vse navedene pričakovane rezultate?

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Načeloma je odgovor pozitiven za naštete tehnologije, medtem ko sredstev in časa ni dovolj za razvoj višjenivojskih semantičnih orodij za razumevanje jezika (avtomatsko povzemanje, odgovarjanje na vprašanja, generiranje naravnega jezika) in pripadajočih virov.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
V primeru razvoja razpoznavalnika spontanega govora za slovenščino je z razvojem potrebnih govornih virov možno zagotoviti pričakovane rezultate. Vendar je znotraj danega časovnega in finančnega okvira na ravni jezikovnih virov vprašanje ali bomo zmogli premostiti zaostanek za po velikosti in kompleksnosti obdelave sicer primerljivimi evropskimi jeziki, kot so nizozemski, češki, slovaški, estonski. Prav tako tudi za sistem strojnega prevajanja. Potrebno je opozoriti, da glede na znanstvene rezultate, obseg govornih virov ne bo v celoti zagotovil primerljivosti uspešnosti slovenskega razpoznavalnika govora s komercialnimi produkti. Slovenski jezik je zaradi svojih lastnosti bistveno zahtevnejši za razpoznavanje govora, kot pa npr. angleški jezik. Na primeru razpoznavalnika: jeziki z dobro razvitimi govornimi viri imajo tipično na voljo za osnovni razvoj nekaj 100 ur govorne baze, velike multinacionalke pa za svoje komercialne produkte uporabljajo govorne baze v obsegu nekaj 1000 ur posnetkov. Pri tem gre pogosto za jezike, ki so glede na svoje lastnosti manj kompleksni za razpoznavanje govora, kot je slovenski jezik (tukaj je pričakovano zaostajanje na račun kompleksnosti približno 10 % do 20 %).

Raziskovalci iz Alpineona in Fakultete za elektrotehniko Univerze v Ljubljani (Laboratorij za strojno inteligenco):
Pričakovani rezultati po našem mnenju presegajo razpoložljiva sredstva. Že razvoj jezikovnih virov za razvoj visokokakovostnega razpoznavalnika govora ter sam razvoj takega razpoznavalnika govora po grobi oceni zahtevata vsaj tretjino razpoložljivih virov, gl. odgovor na vprašanje št. 3.

11. Ali lahko pričakujemo, da bomo dobili uporaben končni produkt, npr. strojni prevajalnik, ki ga bo mogoče uporabiti za spontano prevajanje predavanj v angleškem jeziku (in obratno – iz SL v AN)?

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Da, če se to postavi kot želeni cilj, je to mogoče upoštevati in zagotoviti (podobne rešitve že obstajajo, npr. https://lecture-translator.kit.edu/#/).

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
V sklopu 2 je govora samo o razpoznavanju govora za slovenski jezik. Da bi takšen končni produkt deloval v smeri prevajanja predavanj v angleškem jeziku, je potreben še razpoznavalnik govora za angleški jezik, ki ni del sklopa 2. **(Alpineon:** in tudi (odprtokodna) sintetizatorja govora za oba jezika).

V osnovi lahko pričakujemo, da lahko dobimo v okviru konzorcija uporaben končen produkt oz. sistem strojnega prevajalnika za prevajanje predavanj v slovenskem jeziku (SL-AN). Predstavlja pa takšen sistem zelo visoko stopnjo zahtevnosti razvoja (razpoznava spontanega govora, prevajanje takšnega izhoda). Potrebovali bi na primer dodatne specifične vire (govorno bazo predavanj, semantične vire, nadgrajen Sloleks in terminologijo). Za učenje strojnega prevajalnika, prilagojenega takšnemu spontanemu govoru, bi potrebovali korpus predavanj, ki vključuje takšen spontani govor, in morali tudi zagotoviti pridružene prevode. Takšnega korpusa po naših informacijah še ni. Če namesto njega uporabimo druge vire, je uspešnost prevajanja slabša. Problem lahko predstavlja tudi terminologija, če so predavanja strokovna/znanstvena in manj poljudna. Na uporabnost sistema v tem primeru vpliva tudi prenos napak sistema avtomatskega razpoznavanja govora v sam proces prevajanja. Če bi šlo za nesprotno (off-line) prevajanje, ga je mogoče kombinirati s post-procesiranjem in v tem kontekstu zagotavljati »pohitreno prevajanje« v primerjavi z ročnim, kar pa v omenjenem sistemu ocenjujemo, da ni preveč uporabna opcija. Če bi želeli, da sistem prevajanja prevode tudi izgovarja, je potrebno vključiti tudi sintezo govora (angleščina). Takšen sistem bi bil uporaben v smislu, da bi lahko razumeli večino predavanja, je pa potrebno računati tudi na večje napake v prevodih zaradi spontanega govora.

Če bi bila ideja strojnega prevajanja tudi AN-SL, je potrebno ali razviti razpoznavalnik spontanega angleškega govora (z uporabo javno dostopnih baz) ali uporabiti katerega izmed obstoječih komercialnih razpoznavalnikov govora različnih ponudnikov, kjer pa lahko predstavlja problem večja zaprtost ali nedostopnost dane platforme. Izvedba sistema strojnega prevajanja v smeri AN-SL prav tako zahteva razpoložljivost ustreznih virov. Še najmanjša težava bi bila v tej smeri izvedba sinteze govora (slovenščina).

Uspešnost delovanja sistema je zelo odvisna od vseh zgoraj izpostavljenih odločitev, ki so: ali deluje sprotno (online) ali nesprotno (offline), ali je način pritisni in govori ali kako drugače, kakšen je mikrofon, kako šumno je okolje, koliko je govor spontan in koliko vnaprej pripravljen, ali velja kar za vsa področja ali samo za neko specifično domeno. Problem je tudi zagotavljanje ustreznih virov: enojezični pisni korpus v obsegu nekaj 100 mio. besed za področje predavanj (v čim večji meri kot transkribiran govor), govorna baza v obsegu najmanj nekaj sto ur za domeno predavanj, obsežni (več sto milijonov besed) vzporedni korpus prevodov za področje predavanj, najbolj idealno seveda transkribiranih (česar realno ni mogoče izvesti v okviru predvidenega projekta), ustrezno obsežen slovar terminologije. Zaradi velike težavnosti izdelave in dostopnosti virov se običajno uporabljajo drugi razpoložljivi viri, ki ne izvirajo iz govorjene rabe, vendar to bistveno poslabša kvaliteto prevodov. Dodatno tak sistem zahteva še rešitve, ki zdaj v razpisu niso predvidene, recimo

v smeri AN-SL potrebujemo razpoznavalnik za angleščino, vprašanje je tudi, ali so prevodi v obliki podnapisa ali govorjeni (v tem primeru potrebujemo še sintezo za angleščino in slovenščino).

Raziskovalci iz Alpineona in Fakultete za elektrotehniko Univerze v Ljubljani (Laboratorij za strojno inteligenco):
Pri prevajalniku predavanj so vsaj v grobem potrebne tudi tehnološke komponente, ki ne bodo podprte v okviru načrtovanega programa. To so predvsem razpoznavalnik govora za tuji jezik ter sintetizatorja govora za slovenski jezik in za tuji jezik.

12. Realna časovnica kaže, da bo za razvoj navedenih rezultatov na voljo dobri dve leti in pol, treba je namreč računati z določbo, da morajo biti rezultati objavljeni najmanj šest mesecev pred zaključkom aktivnosti (čas za evalvacijo in morebitne izboljšave). Je ta časovni okvir realen?

Inštitut za slovenski jezik Frana Ramovša ZRC SAZU:
Razpis je s šestimi tematskimi sklopi zastavljen zelo široko, zato se sprašujemo, ali je v relativno kratkem časovnem okviru resnično mogoče doseči zastavljene rezultate.

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Glede na to, da bodo morali prijavitelji sami definirati kazalce uspešnosti, bi predlagali, da se v prijavi definira, do katere mere bodo končni kazalci doseženi 6 mesecev pred koncem aktivnosti, pri čemer bodo ob koncu projekta morali biti doseženi končni kazalci v polni meri.

Pod predpostavko, da prijavitelji torej sami določijo kazalnike uspeha, kar zagotavlja, da bodo pričakovani rezultati realno ocenjeni, je tudi časovni okvir realen. Predpostavljamo, da bodo pričakovani kazalniki prilagojeni kratkemu roku. Nekatere tehnologije so že (deloma) razvite oz. obstaja rešitev v smislu »proof-of-concept«, tako da v celoti gre predvsem za racionalno porazdelitev sredstev po posameznih ciljih.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
Pri zbiranju posnetkov za baze in gradiv za korpuse je časovna izvedba zelo odvisna od odzivnosti gradivodajalcev. Dosedanje izkušnje kažejo, da se lahko že samo za uspešno dosežen dogovor in prenos gradiv porabi več mesecev. To je seveda osnovni pogoj, da se lahko začne izdelovati baza/korpus, izdelana baza/korpus pa je dalje predpogoj, da se lahko začne razvijati razpoznavalnik govora ali strojni prevajalnik. Zato v UM FERI, Laboratoriju za digitalno procesiranje signalov predlagamo, da je rok za izdelane vire februar 2022, rok za tehnologije (prevajalnik, razpoznavalnik govora itd.) pa predlagamo ob koncu projekta. In po drugi strani, da se posebno pozornost posveti čim hitrejši in čimbolj učinkoviti izrabi že razpoložljivih virov in čim hitrejši izdelavi novih virov, da bi lažje našli kompromis z omenjeno določbo.

Raziskovalci iz Alpineona in Fakultete za elektrotehniko Univerze v Ljubljani (Laboratorij za strojno inteligenco):
Za razvoj govornih jezikovnih virov je ta rok realen, za razvoj visokokakovostnega razpoznavalnika govora pa je izvedljivost doseganja rezultatov v predlaganem časovnem možna, če bodo govorni jezikovnih viri, ki bodo razviti v okviru sklopa 1, pravočasno na voljo.

13. Ali bi za razvoj oziroma nadgradnjo katerega od jezikovnotehnoloških izdelkov potrebovali odkup licence? V javnem razpisu je sicer določba o odprtem dostopu za javno financirane izdelke, nastale v okviru tega razpisa.

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Odgovor na to vprašanje je odvisen od dogovorov konzorcijskih partnerjev ob prijavi. Vsakega od naštetih izdelkov je načeloma mogoče razviti na novo, če morda ne bi bilo mogoče uporabiti trenutnih rešitev, ki niso dostopne pod odprtimi licencami oz. odprtokodne.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
Na UM FERI, Laboratoriju za digitalno procesiranje signalov menimo, da licenca CC BY za vse, ni najboljša rešitev, saj omogoča uporabo razvitih virov in orodij (zastonj) tudi s strani multinacionalk, ki bodo prej ali slej svoje storitve na podlagi teh tehnologij nato zaračunavale oz. jih že. Na UM FERI, Laboratoriju za digitalno procesiranje signalov menimo, da je ustreznejša rešitev postopna licenca, pri kateri je brezplačna komercialna raba omogočena podjetjem z nizkim letnim prometom, manjša vsota se zaračunava podjetjem s srednjim prometom, podjetja z visokim prometom pa plačajo polno licenco. Tako bi se tudi za že razvite vire in orodja lahko pridobivalo nova sredstva za vzdrževanje in tudi nadaljni razvoj, pri tem bi bila pridobljena sredstva dodeljena tistim, ki so vire in orodja tudi razvili. V tem primeru imajo potem možnost razvoja rešitev tudi razna zagonska podjetja.

Raziskovalci iz Alpineona in Fakultete za elektrotehniko Univerze v Ljubljani (Laboratorij za strojno inteligenco):
Pri prevajalniku predavanj so vsaj v grobem potrebne tudi tehnološke komponente, ki ne bodo podprte v okviru načrtovanega programa. To so predvsem razpoznavalnik govora za tuji jezik ter sintetizatorja govora za slovenski jezik in za tuji jezik.

14. Koliko okvirno stanejo licence s področja JVT?

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Vprašanje je presplošno, da bi bilo nanj mogoče smiselno odgovoriti.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
Primeri cen licenc za različne jezikovne vire so na voljo v katalogu jezikovnih virov ELRA (Evropsko združenje za jezikovne vire) za raziskovalne in komercialne namene. Cene licenc se za različne vire precej razlikujejo. Za konkretnejši odgovor na to vprašanje je zato potrebno podrobneje opredeliti jezikovne vire.

Raziskovalci iz Alpineona in Fakultete za elektrotehniko Univerze v Ljubljani (Laboratorij za strojno inteligenco):
Vprašanje je zastavljeno zelo splošno. Cena izdelka je vezana na število potencialnih uporabnikov, zato so pogosto cene licenc jezikovnotehnoloških izdelkov za jezike z manjšo baz govorcev višje od podobnih izdelkov za bolj razširjene jezike.

15. Ali je razpoložljivost visokostrokovnega kadra dovoljšnja? Prosimo za okvirno številčno oceno človeških virov (s polnim delovnim časom).

Center za jezikovne vire in tehnologije Univerze v Ljubljani:
Ocenjujemo, da je v slovenskem prostoru kadra dovolj za izvedbo projekta, ob predpostavki, da bodo v konzorcij vključene vse večje raziskovalne in druge ustanove, ki se

ukvarjajo s tem področjem (UL, UM, IJS, ZRC SAZU, Amebis, Alpineon itd). Konzorcij ob prijavi načeloma zagotavlja, da je sposoben izvesti prijavljeni projekt. Konkretne kadrovske potrebe bo mogoče oceniti šele ob sestavljanju prijave, ko bodo sprejete odločitve o konkretnih kazalnikih po posameznih sklopih.

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru:
Na UM FERI, Laboratoriju za digitalno procesiranje signalov ocenjujemo, da je za izvedbo vseh nalog v okviru konzorcija z ustreznim razporejanjem dela in vodenjem, ter z vključevanjem tudi novih sodelavcev s trga dela (za določena dela ni potrebna specifično kvalificirana delovna sila, kot je npr. veliko nalog pri izdelavi jezikovnih virov), razpoložljivost kadra v Sloveniji dovoljšnja.

Raziskovalci iz Alpineona in Fakultete za elektrotehniko Univerze v Ljubljani (Laboratorij za strojno inteligenco):
V izvajanje programa lahko vključimo do 5 strokovnjakov s področja govornih tehnologij oz. 15 FTE skupaj v predvidenem trajanju programa (podatek velja za Alpineon in UL-FE skupaj).

DODATNE PRIPOMBE O RAZPISANIH SKLOPIH:

Vzdrževanje in nadgradnja korpusov

Inštitut za slovenski jezik Frana Ramovša ZRC SAZU:
Predlagamo, naj razpisovalec natančneje opredeli, katere vrste korpusov bi bilo treba v okviru tega sklopa prioritetno nadgraditi oziroma so bolj potrebni vzdrževanja. Nujno bi bilo nadgraditi referenčni pisni korpus s stvarnimi besedili, ki niso dostopna preko spleta (slednja so bila v zadnjem času že vključena v korpusa Janes in slWAC), oziroma govorni korpus. Ustrezno bi bilo treba ob velikem številu korpusov sodobne slovenščine poskrbeti tudi za korpuse starejše slovenščine, tudi z digitalizacijo starejših pisnih in arhivskih virov ter že urejenih slovarskih kartotek.

Terminološki portal

Inštitut za slovenski jezik Frana Ramovša ZRC SAZU:
Pri tem sklopu bi bilo smiselno že v okviru razpisa kot cilj določiti izdelavo novega portala, ki naj upošteva že obstoječe rešitve (spletišča Evroterm, Termania, Terminologišče ...) ter omogoča vključitev in strukturiran prikaz raznovrstnih podatkov, ki so že na voljo.

Črkovalnik za slovenski jezik

Slovenska akademija znanosti in umetnosti:
Predlagamo, da se v načrtovani razpis namesto zdajšnje točke 5. terminološki portal (ta izdelek je namreč ARRS že financirala, prim. oznako L6-9778) vključi črkovalnik, ki je bil prepoznan kot prioriteta na 3. seji Sveta za spremljanje razvoja jezikovnih virov in tehnologij (12. 7. 2017).

Inštitut za slovenski jezik Frana Ramovša ZRC SAZU:
Predlagamo, naj se v razpis doda sklop, namenjen razvoju črkovalnika za slovenski jezik. Po presoji Sveta za spremljanje razvoja jezikovnih virov in tehnologij gre za enega izmed prioritetnih projektov s področja JVT (3. seja Sveta za spremljanje razvoja JVT, 12. 7. 2017). Trenutno aktualni črkovalnik je bil razvit še v času pred objavo Slovenskega

pravopisa 2001 in tako še ni mogel upoštevati kodifikacijskih rešitev tega priročnika; prav tako nujno ne odraža aktualne jezikovne norme in novejšega, zlasti lastnoimenskega jezikovnega gradiva.