Univerza v Ljubljani

Online Notes

Ljubljana, 15. december 2021

Poročilo o izvedbi pilotnih predavanj

Projekt Online Notes (ON)

V okviru projekta Online Notes (ON) je bilo v drugi polovici leta 2021 (oz. v prvem semestru študijskega leta 2021/2022) na Univerzi v Ljubljani izvedenih pet pilotnih predavanj, na katerih se je v praksi testiral sistem za strojno prevajanje slovenskih predavanj v angleški jezik. Predavanja so zajemala tako družboslovna kot tehnična področja na štirih različnih fakultetah:

1. prof. dr. Marko Bajec: **Osnove podatkovnih baz** (FRI, predavalnica P1, 18. november 2021)
2. viš. pred. dr. Robert Rozman: **Računalniška arhitektura** (FRI, predavalnica PA, 18. november 2021)
3. red. prof. dr. Roman Kuhar: **Uvod v gejevske in lezbične študije** (FF, predavalnica 4, 25. november 2021)
4. izr. prof. dr. Katja Lozar Manfreda: **Statistika** (FDV, Velika dvorana, 30. november 2021)
5. izr. prof. dr. Anton Rafael Sinigoj: **Osnove elektrotehnike** (FE, predavalnica P1, 3. december 2021)

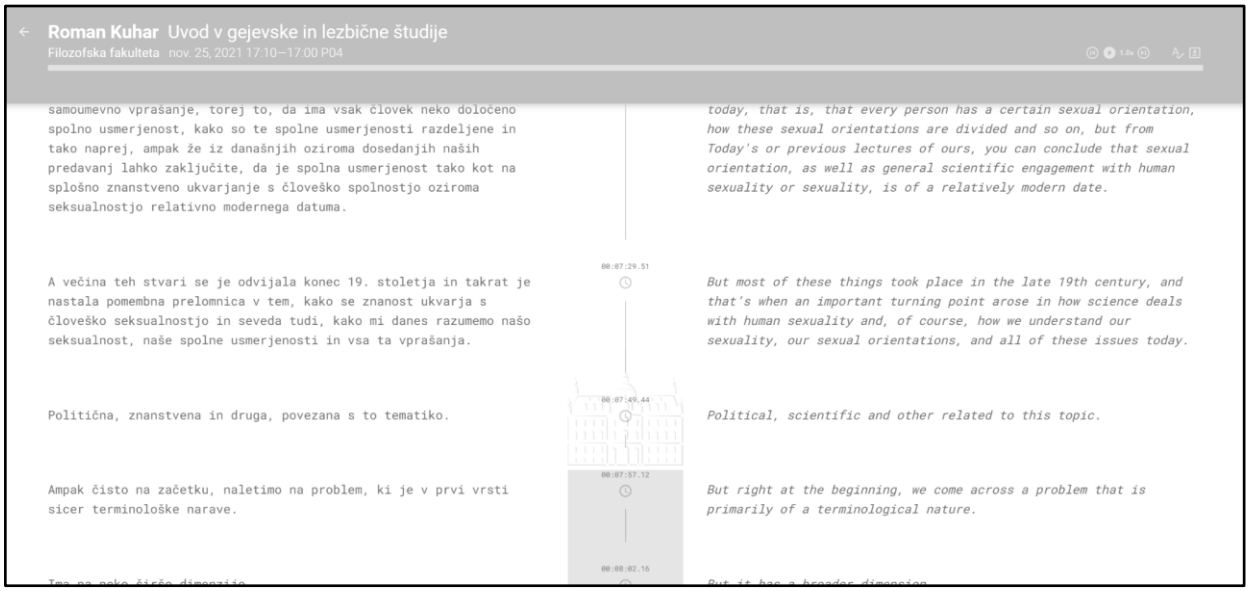
Vsebina predavanj je bila s predavatelji[1] predhodno dogovorjena, posnetki predavanj pa transkribirani. Transkripcije so bile uporabljene tako za širjenje slovarja za razpoznavo govora kot za gradnjo jezikovnega modela za razpoznavalnik govora.

Pred izvedbo predavanj se je tehnična ekipa projekta ON sestala s tehničnimi sodelavci na vseh štirih fakultetah in preverila razpoložljivost opreme ter preizkusila

[1] Izrazi, ki se v poročilu nanašajo na osebe in so zapisani v moški slovnični obliki, so uporabljeni kot nevtralni za ženski in moški spol.

zajem zvoka v prostoru. Sama predavanja so potekala običajno - predavatelji so bili opremljeni z mikrofonom, ki je zaznaval njihov govor in ga pošiljal sistemu za razpoznavo govora in prevajalniku. Študenti so imeli možnost razpoznavo govora in strojni prevod spremljati na pilotni platformi ON (izjema je predavanje izr. prof. dr. Antona Rafaela Sinigoja na FE, kjer je bila razpoznava govora s strojnim prevodom projecirana na platno).

Izpis slovenske transkripcije predavanja in njenega angleškega prevoda, kot je bil na voljo študentom med predavanji, je prikazan na Sliki 1.



Slika 1: Vmesnik za spremljanje slovenske razpoznave govora in angleškega prevoda.

V tem poročilu podajamo preliminarno analizo uspešnosti sistema in navajamo nekaj točk za morebitne izboljšave v prihodnosti.

1 Tehnološke omejitve

Prvo pilotno predavanje (prof. dr. Marko Bajec, FRI) je služilo tudi kot stresni test. Da bi preverili zmogljivost programske opreme in zagotovili pravilno delovanje v normalnih razmerah, smo takrat prisotne študente povabili, da se povežejo na platformo. Zaradi sočasnega dostopa večjega števila uporabnikov se je na zalednem sistemu pojavila težava pri komunikaciji s podatkovno bazo. V nastavitvah sočasnih povezav, ki jih podatkovna baza dovoljuje, je bila uporabljena privzeta vrednost, namenjena testiranju. Z ustrezno prilagoditvijo te nastavitve smo že tekom prvega pilotnega predavanja napako odpravili. V kasnejših predavanjih se ni več pojavila, prav tako nismo opazili drugih večjih težav pri delovanju sistema.

Odziv študentov, ki so se na platformo prijavili, je bil pozitiven. Komentirali so, da je včasih v prevodu »dodana kakšna čudna beseda, sicer je pa prevod dovolj dober za splošno razumevanje vsebine predavanja« (FDV). Študentje so opazili določene

težave pri segmentiranju in prevajanju, sicer pa so menili, da »se kar dobro znajde« (FF).

2 Analiza napak

Izvedli smo preliminarno analizo napak, pri kateri smo pri vsakem predavanju vzeli približno 50 segmentov (odvisno od konteksta smo zadnjemu odseku dodali še enega ali dva, da so bile izjave pomensko zaključene). Segmenti v tem primeru predstavljajo vzporedne enote slovenske razpoznave govora in angleškega strojnega prevoda (torej besedila, kot jih spremljajo udeleženci predavanja). Primer s predavanja FE (Sinigoj) je prikazan spodaj - uporabnik na levi strani vidi transkripcijo predavateljevega govora, na desni angleški prevod segmenta, vmes pa časovni žig, ki nakazuje, ob katerem času se je segment pojavil v vmesniku (ki je bil prikazan na Sliki 1).

Slovenska razpoznava govora	**Časovni žig**	**Angleški strojni prevod**
Tako torej, naslov nekako bo: C vezja; torej kratko malo, analiza C vezij, zdaj, mogoče že takoj na začetku: kaj je sploh vprašanje ali pa, kakšna so lahko vprašanja? V C vezjih, torej vezi, je dano; so, kakšni viri, kakšna stikala, kakšni Kondenzatorji, kaj po vaše so, bi rekel?	00:02:40.94	So, the address will somehow be: C circuit; so a little bit, analysis of C circuits, now, maybe right at the beginning: what is the question at all or what can the questions be? In C circuits, i.e. bonds, it is given; are, what sources, what switches, what Capacitors, what do you think they are, I would say?
Možna vprašanja: mi pomagate, kaj bi določil v takem vezju?	00:02:45.02	Possible questions: do you help me, what would you set in such a circuit?

Podatki o evalviranih posnetkih so prikazani v Tabeli 1.

Fakulteta in predavatelj	Dolžina posnetka (min)	Število besed	Število segmentov	Povprečno število besed na segment
FRI, Bajec	4,65	773	51	15,16
FRI, Rozman	5,92	1.192	52	22,92
FF, Kuhar	7,89	1.456	51	28,55

FDV, Lozar Manfreda	11,94	1.686	51	33,06
FE, Sinigoj	4,29	800	51	15,69

Tabela 1: Podatki o evalviranih posnetkih, transkripcijah in prevodih.

Evalvirani posnetki so bili dolgi od 4 do 12 minut, razlikovali pa so se tudi po povprečni dolžini segmentov. Najdaljše segmente najdemo v predavanju dr. Lozar Manfrede (33 besed na segment), najkrajše pa v predavanju dr. Bajca (15 besed na segment). Analizirani vzorec sicer ni reprezentativen, a lahko vidimo težnjo, da so družboslovna predavanja (FF, Kuhar in FDV, Lozar Manfreda) nekoliko bolj gostobesedna kot naravoslovno-tehnična.

Segmente smo nato pregledali in jih ocenili glede na uspešnost razpoznave govora in strojnega prevoda. Pri razpoznavi govora smo označevali **manjše napake** (npr. če je razpoznavalnik izpisal napačen predlog ali zaimek, a to ni vplivalo na pomen odstavka; če je bil izpisan samostalnik namesto pridevnika, a je pomen ostal jasen), **srednje napake** (npr. problemi s segmentacijo, ki so vplivali na delitev izjav in na pomen) in **večje napake** (npr. povsem napačno razpoznane besede ali številke). Rezultati so prikazani v Tabeli 2 (število napak je normalizirano glede na število besed v evalviranem odseku; vrednosti torej prikazujejo število napak na vsakih 100 besed).

Fakulteta in predavatelj	Manjše napake	Srednje napake	Večje napake	Skupaj
FRI, Bajec	1,68	2,33	1,68	5,69
FRI, Rozman	0,42	2,10	3,1	5,62
FF, Kuhar	0,55	2,13	0,55	3,23
FDV, Lozar Manfreda	1,07	1,54	2,31	4,92
FE, Sinigoj	1,12	2,75	3,88	7,75

Tabela 2: Pojavnost napak razpoznavalnika govora v evalviranih transkripcijah.

Pri vseh evalviranih transkripcijah se najredkeje pojavljajo manjše napake, pogostejše pa so srednje in večje napake. Zanimivo je, da sta tu nekoliko bolj v prednosti predavanji s FF (Kuhar) in s FDV (Lozar Manfreda), ki izkazujeta nekoliko manjše skupno število napak (3,23 na 100 besed oz. 4,92 na 100 besed), čeprav sta bila prav ta posnetka glede na evalvirano število besed najdaljša. Največ napak najdemo v transkripcijah predavanja s FE (Sinigoj), a gre to pripisati predvsem vsebini predavanja (veliko narekovanja enačb, zaradi česar so stavki in segmenti razsekani, same besede pa kratke in zato pogosto napačno razpoznane).

Tabela 3 prikazuje število segmentov, v katerih se je pojavila vsaj ena manjša, srednja ali večja napaka (v enem segmentu lahko najdemo tudi več napak iz različnih kategorij).

Fakulteta in predavatelj	Število segmentov z vsaj eno manjšo napako	Število segmentov z vsaj eno srednjo napako	Število segmentov z vsaj eno večjo napako	Število segmentov z vsaj eno napako
FRI, Bajec	12 23,53 %	16 31,37 %	11 21,57 %	30 59,00 %
FRI, Rozman	5 9,62 %	23 44,23 %	22 42,31 %	37 71,00 %
FF, Kuhar	7 13,73 %	24 47,06 %	8 15,69 %	35 69,00 %
FDV, Lozar Manfreda	15 29,41 %	20 39,22 %	25 49,02 %	41 80,00 %
FE, Sinigoj	7 13,73 %	15 29,41 %	21 41,18 %	34 67,00 %

Tabela 3: Število segmentov z napakami v evalviranih transkripcijah.

V nadaljevanju opozarjamo na nekaj pogostih napak, ki so se pojavljale na vseh opazovanih pilotnih predavanjih.

2.1 Napake na nivoju razpoznave govora

V Tabeli 4 so navedene najpogostejše napake razpoznavalnika govora. Navedene količine predstavljajo število napak določene kategorije, ki jih razpoznavalnik zagreši na vsakih sto besed.

Tip napake	FRI, Bajec	FRI, Rozman	FF, Kuhar	FDV, Lozar Manfreda	FE, Sinigoj

Izpust besede	0,13	0,08	0,14	0,24	0,12
Napačna beseda	2,46	3,78	1,17	3,08	4,61
Napačna delitev	1,03	0,34	0,27	0,06	0,62
Segmentacija	1,81	1,26	1,37	0,53	1,62
Diskurzni označevalec vpliva na pomen	-	0,59	0,27	0,59	0,62
Napačna postavitev vejice (vpliv na pomen)	0,13	-	-	0,12	0,25
Napačna raba vprašaja	0,13	-	-	0,24	-

Tabela 4: Število napak razpoznavalnika govora na sto besed v transkripciji.

Število napak, ki nastanejo pri razpoznavi govora, je pričakovano. Najpogostejša napaka je izpis napačne besede. Pogost razlog je bližnja homonimija (program izpiše *»ceni jelki«* namesto *»cenilki«*), včasih pa je razlog za napačen zapis ta, da besede še ni v slovarju, program zapiše tisto, ki je po zvočni podobi najbližje. Program včasih izpusti določeno besedo *(“In, če **(je)** ping count ena pomeni trenutno”)*, zgodi pa se tudi, da je dodana odvečna beseda (*“je Richard von Krafft Ebing, **ki** Rihard von Krafft Ebing, ki je deloval v Avstriji”*).

Pogosta je napačna segmentacija, ki besedilo prelomi na nepričakovanem mestu in ustvari dve ločeni povedi. Napaka včasih vpliva tudi na pomen (*“torej s 95 odstotno gotovostjo. // Ocenjujemo, da ta interval vsebuje populacijski parameter”*). Zaradi diskurznih označevalcev prihaja do negacij (*“Zdaj, ko ocenjujemo populacijski parameter **a ne** interval zaupanja*”). Pojavlja se tudi napačna delitev (*“Potencial**.** Te točke poznamo”),* ki vpliva na razumevanje besedila. Postavljanje nekončnih ločil v poročilu ni analizirano, opazovana je bila le raba končnih, in sicer vprašaja, ki se pojavlja tudi pri nevprašalnih povedih (“Torej je to v resnici neka spremenljivka**?**”).

Iz tabele je razvidno, da je pri določenih predavateljih prisotnih manj napak. Do odstopanj prihaja zaradi razlik v načinu govora posameznikov (diskurzni označevalci, mašila, vpliv narečja ipd.). Najmanj napak se pojavi pri red. prof. dr. Romanu Kuharju, iz posnetkov je razvidno, da je govor tekoč, jasen in primerno hiter. Viš. pred. dr. Robert Rozman v posnetkih govori nestandardno, zato je napak več. Največje število

napačno zapisanih besed se pojavi pri izr. prof. dr. Antonu Rafaelu Sinigoju, razpoznavalnik ima težave pri razpoznavanju simbolov in izpisovanju enačb (*“4, cilj* **[cele]** *osem; minus, 4, cilje* **[cele]** *osem; se pravi, minus 4, cele osem, Milji! Kolona* ***[milicoulomba]****”*).

2.1.1 Segmentacija

Razpoznavalnik občasno ustvarja preveč obsežne segmente slovenskega besedila. Ker prevod nastane šele takrat, ko se slovenski segment zaključi, se na nov segment čaka tudi po več kot pol minute. To predstavlja težavo pri poslušalcih, ki profesorja delno razumejo, saj prevoda besed, ki so jim nepoznane, ne prejmejo dovolj hitro. Tabela 5 prikazuje razlike v dolžinah segmentov pri posameznem govorcu (v sekundah). Čas izpisovanja je enak trajanju čakanja na izpis angleškega prevoda.

Fakulteta in predavatelj	Dolžina najkrajšega segmenta	Dolžina najdaljšega segmenta	Dolžina povprečnega segmenta	Mediana
FRI, Bajec	0,63	13,62	5,47	4,17
FRI, Rozman	0,15	26,79	6,83	4,33
FF, Kuhar	0,84	23,91	9,28	7,65
FDV, Lozar Manfreda	0,99	47,04	14,04	10,68
FE, Sinigoj	0,15	15,00	5,05	4,20

Tabela 5: Razlike v dolžinah segmentov pri posameznem govorcu.

Dolžina najdaljšega segmenta pri izr. prof. dr. Katji Lozar Manfreda (spodnji primer) dosega najvišjo vrednost, in sicer 47,04 sekunde:

Skripti lahko preberete še doslednost pa še 14 so: ampak recimo da sta te dve najbolj pomembni. Učinkovitost, učinkovita je, če ima majhno variabilnost v primerjavi z ostalimi cenilka. Mi bi namreč želimo, da je vzorčna dobra, vzor si! Želimo si, da bi bila vzorčna porazdelitev čim manj variabilna, ne mi želimo imeti ocene čim bolj blizu sredine porazdelitve, ne čim manjšo variabilnost in dobre cenilka. Seveda je tista, ki ima	You can read the scripts for consistency and there are 14 more: but let's say these two are the most important. Efficiency is effective if it has little variability compared to other appraisers. Namely, we would like the model to be good, you are a model! We want the sample distribution to be as variable as possible, we do not want to have the estimate as close as possible to the middle of the distribution, not the

manjšo variabilnost, torej če bi primerjali recimo, porazdelitev vzorčnih aritmetično svinj ali pa porazdelitev modusov, bi videli, da modus ima bistveno večjo variabilnost od aritmetična sredina.	smallest variability and a good estimator. Of course, it is the one that has less variability, so if we were to compare, say, the distribution of sample arithmetic sows or the distribution of modes, we would see that the modus has significantly greater variability than the arithmetic mean.

Izračunali smo tudi bralno hitrost, ki jo zahtevajo posamezni segmenti, tj. število znakov, ki jih je treba prebrati vsako sekundo, če želimo brez zaostankov dohajati izpisovanje. Tabela 6 predstavlja zahtevane bralne hitrosti za slovensko transkripcijo, Tabela 7 pa za angleške prevode.

Fakulteta in predavatelj	Povprečna bralna hitrost	Minimum	Maksimum	Mediana	Standardni odklon
FRI, Bajec	19,21	6,35	126,26	13,85	20,32
FRI, Rozman	24,84	8,72	163,33	16,6	27,49
FF, Kuhar	22,97	10,63	113,68	15,09	18,32
FDV, Lozar Manfreda	19,06	4,85	131,42	14,21	19,79
FE, Sinigoj	24,52	3,06	153,33	16,68	23,32

Tabela 6: Bralne hitrosti za slovenske transkripcije evalviranih posnetkov.

Fakulteta in predavatelj	Povprečna bralna hitrost	Minimum	Maksimum	Mediana	Standardni odklon
FRI, Bajec	20,24	6,35	127,76	14,8	20,46
FRI, Rozman	23,33	0,0	177,78	17,14	25,56
FF, Kuhar	23,29	8,3	106,19	15,49	18,45
FDV, Lozar Manfreda	20,71	5,65	143,43	15,18	21,96

FE, Sinigoj	26,86	3,93	186,67	18,16	27,61

Tabela 7: Bralne hitrosti za angleške prevode evalviranih posnetkov.

Če upoštevamo, da je pri podnaslavljanju priporočena bralna hitrost za odrasle bralce 10-12 znakov na sekundo, povprečna bralna hitrost v naših primerih ta kriterij presega. Upoštevati pa je treba, da gre pri podnapisih običajno za priredbe (skrajšana, strnjena in parafrazirana besedila), v našem primeru pa za transkripcije govora in njihove prevode. Iz teh podatkov je razvidno, da se lahko pojavljajo tudi zelo dolgi segmenti oz. taki z veliko količino znakov in kratko časovno dolžino: pri FE (Sinigoj) dosega zahtevana bralna hitrost tudi do 186 znakov na sekundo (18x več od norme)t, zato bi bilo v prihodnje smiselno razmisliti, kako doseči najboljši kompromis med besedilno celostjo na eni strani in pravočasnih strojnim prevajanjem na drugi.

2.1.2 Napačno ali nekonsistentno razpoznane besede

Čeprav so bile določene besede prisotne v slovarju izgovorjav, ki so ga pripravili študentje s transkribiranjem posnetkov, nekatere niso bile razpoznane pravilno oz. so bile razpoznane nekonsistentno.

Primeri:

- *Foucault > Fuko, Foko, Vuko*
- *Coulomb > Kulon*
- *Študij je šiht. > Študij je šiht. > (prevod) Study is shit.*
- *dirty bit - dr ti biti*

Ker se pri obravnavi določene snovi običajno ponavljajo ista imena, bi bilo dobro razmisliti, ali lahko razpoznavalnik iz konteksta razbere pravilni zapis in poveča verjetnost rabe te besede v celotnem predavanju.

“Ja, tako 400, voltov, pa 12, mikrofaradov, ja, kdor ne ve, ne, in tukaj si zapiše minus 1, 12, mikrofaradov krat, 400 ***Boljtov****.”* (FE, Sinigoj)

2.1.3 Napačno razpoznane besede zaradi homonimije / bližnje homonimije

Razpoznavalnik določene besede izpiše napačno zaradi homonimije oziroma bližnje homonimije. Razlog za napako je lahko odsotnost prave besede v slovarju, lahko pa le izbere napačno.

"Krafft Ebing, ki je bil tudi tak nastopač: on je imel javna predavanja, na predavanjih, je zdravil te seksualne, **manj dijake, [manijake]** takoimenovane z z hipnozo in tako naprej." (FF, Kuhar)

"Oziroma razlaga, ki jo je podal. On, je bil **empirik cist [empiricist]** in je opazoval vedenja in fizična telesa ljudi." (FF, Kuhar)
"To je možna, razlaga druga, po moje, bistveno bolj verjetna razlaga je to, da jaz seveda ne vem, kakšen je bil vzorec, verjetno je bil bistveno bolj kompleksen, vzorčni načrt bi mogla neka **oteževanje [uteževanje]** uporabiti in tako naprej." (FDV, Lozar-Manfreda)
"Običajno seveda ne vemo prave vrednosti, če bi uvedli [vedeli], potem **ne Beniti [ne bi niti]** delali raziskave na vzorcu a ne." (FDV, Lozar-Manfreda)
"Seveda moramo vedet; **Katie [kajti]**: če recimo ocenjujemo [...]" (FDV, Lozar-Manfreda)

2.1.4 Kratice in simboli za fizikalne enote

Pogosta je težava pri razpoznavanju kratkih imen in enot. V naravoslovju se električno napetost označuje z oznako 'U', ki zvočno sovpada s predlogom 'v'. Ker je zapis predloga z 'u' napačen, tudi oznako 'U' popravi v 'v'.

2.2 Napake na nivoju prevajalnika

Tudi pri strojnih prevodih slovenskih transkripcij smo označevali napake na enak način: **manjše napake** (npr. napačni predlogi, ki niso drastično vplivali na pomen; nenavadna skladnja zaradi vrstnega reda v slovenščini), **srednje napake** (npr. nerodna segmentacija, ki razdeli izjavo na dva dela, ki sta zato okorno prevedena) in **večje napake** (npr. napaka v terminologiji, povsem napačno prevedena beseda, prazen prevod). Rezultati so predstavljeni v Tabeli 8.

Fakulteta in predavatelj	Število segmentov z vsaj eno manjšo napako	Število segmentov z vsaj eno srednjo napako	Število segmentov z vsaj eno večjo napako	Število segmentov z vsaj eno napako
FRI, Bajec	2 4 %	11 22 %	9 18 %	22 43 %
FRI, Rozman	3 6 %	9 17 %	23 44 %	34 65 %
FF, Kuhar	10 20 %	11 22 %	2 4 %	23 45 %
FDV, Lozar Manfreda	1 2 %	9 18 %	16 31 %	26 51 %
FE, Sinigoj	5 10 %	7 14 %	19 37 %	31 61 %

Tabela 8: Število segmentov z napakami v evalviranih prevodih.

Največ napak je bilo prisotnih v predavanjih FRI (Rozman) in FE (Sinigoj), pri katerih v evalviranem posnetku več kot 60 % segmentov vsebuje vsaj eno napako; 37 % oz. 44 % vsebuje večjo napako. Najmanj napak nasploh smo zabeležili pri predavanju FRI (Bajec), a je vsebovalo nekoliko več večjih napak kot npr. FF (Kuhar), kjer so bile resne napake prisotne le v 2 % evalviranih segmentov.

Tabela 9 prikazuje vrste prevajalskih napak. Količine so normalizirane na 100 besed angleškega prevoda. Največ napak se pri prevodih podeduje od napačne razpoznave govora - pri tem še nekoliko bolj izstopa FE (Sinigoj), ki ima obenem tudi največ napak zaradi napačno razpoznanih zaimkov (kar je pričakovano, saj je v evalviranem posnetku govoril o enačbah in uporabljal veliko deiktov). Kot problematična se izkaže tudi segmentacija - delitev transkribiranega posnetka na semantično manj konsistentnih delih povzroči tudi napake v prevodu.

Tip napake	FRI, Bajec	FRI, Rozman	FF, Kuhar	FDV, Lozar Manfreda	FE, Sinigoj
Prevajalska napaka zaradi napačne razpoznave govora	0,65	0,77	0,38	0,65	1,2
Prevajalska napaka zaradi napačnega zaimka	0,11	-	0,06	-	0,33
Napačna terminologija	0,22	0,46	0,13	0,05	0,43
Segmentacija	0,76	0,62	0,45	-	0,11
Diskurzni označevalec vpliva na pomen	0,11	0,54	0,32	0,55	0,87
Prevajalska napaka zaradi ločil	0,11	-	-	-	-
Prazen prevod	0,11	0,08	-	-	-
Prevajalska napaka zaradi ponovitve govorca	0,11	-	0,06	-	-
Prevajalska napaka zaradi fragmentiranih stavkov	0,22	-	-	-	-
Prevajalska napaka zaradi homonimije	-	0,08	0,19	-	-
Nerazumljiv ali slabo razumljiv prevod	-	0,24	0,06	0,05	0,22
Prevajalska napaka zaradi besednega reda	-	-	0,06	-	-
Decimalna števila	-	-	-	0,05	0,33
Elipsa	-	0,08	-	-	

Prevajalska napaka zaradi frazeologije	-	-	-	-	0,11

Tabela 9: Vrste prevajalskih napak.

V nadaljevanju navajamo primere za najpomembnejše prevajalske napake.

2.2.1 Stalne besedne zveze, frazeologija in metaforika

Za prevajalnik predstavljajo težavo stalne besedne zveze in frazemi, ki so pogosto prevedene dobesedno. Napaka je pričakovana in ne predstavlja zgolj pomanjkljivosti sistema ON, temveč splošni problem pri strojnih prevajalnikih. Za izboljšanje je potrebna večja količina podatkov za učenje strojnega prevajalnika.

FF, Kuhar	"... **skočil čez plot** …"	"... **jumping over the plot** …"
FF, Kuhar	"Ulrichs s to teorijo seveda ni **prodrl**."	"Ulrichs, of course, did not **penetrate** this theory."
FE, Sinigoj	"Koliko, ja, **malo, Morgan, [malo morgen]** saj niso upori!"	"How much, yes, **little, Morgan,** for they are not rebellions!"

2.2.2 Diskurzni označevalci

Ker gre pri predavanjih za spontani govor, se v njem pogosto pojavljajo diskurzni označevalci, npr. "ne", "ane", "ali ne", "ne vem", "se pravi", "skratka". Preveliko število diskurznih označevalcev vpliva na segmentacijo ("se pravi"), nekateri pa se prevajajo napačno, ker so homonimni z neko drugo besedo ("a vidite" -> "but you see" namesto "so you see"). Precejšen problem predstavlja diskurzni označevalec ‘(a/ali) **ne**’, ki se v prevodu izrazi kot nikalnica in negira trditve.

FRI, Rozman	"Okej, smo razjasnjen, saj ni se treba tukaj nič na pamet učiti, ampak v bistvu samo toliko, da operativno. Če bo kdaj potrebno, boste te informacije lahko tudi našli, znali tudi malo razmišljat, včasih mogoče ta, zadevo, razširljiva jih dolžin zapisov, o uporabi tudi v kakšnem svojem program. Recimo če vam nekdo dal nalogo **ne vem**."	"Okay, we’re clarified, as there’s nothing to learn by heart here, but basically just enough to be operational. If necessary, you will also be able to find this information, you will also know how to think a little, sometimes maybe this, a matter of expanding the lengths of records, about the use in any of your program. Let’s say if someone gave you an assignment **I don’t know**."
FF, Kuhar	"A po drugi strani, **ne**, pa, kljub temu da pravi, da je heteroseksualnost zdrav rezultat psihoseksualnega razvoja, poudarja, da homoseksualnost,	"But on the other hand, **no**, however, despite saying that heterosexuality is a healthy result of psychosexual development, he points out that homosexuality,

	kljub tem **ne**, da, je to posledica nekih težav v psiho seksualnem razvoju, da to ni bolezen."	despite **not being yes**, is due to some problems in psychosexual development, that it is not a disease."
FF, Kuhar	"S tem je povezana ena luštna anekdota: Krafft Ebing, ki je namreč v tej knjigi popisoval seksualne fantazije svojih pacientov in pacientk: in kljub temu, da se je zavedal, da popisovanje takih **ne** seksualnih fantazij lahko pritegne zanimanje splošne javnosti. A ne, pornografija seveda ni bila na dosegu rok tako kot je danes, in je pravzaprav nekakšna taka knjiga bila edini vir nekega tovrstnega užitkom in tudi zaradi tega je bila ta knjiga tak dober bestseller."	"There is one cute anecdote associated with this: Krafft Ebing, who described the sexual fantasies of his patients in this book: and despite the fact that he was aware that listing such **non-sexual** fantasies could attract the interest of the general public. But no, pornography, of course, was not at hand as it is today, and in fact some kind of book like that was the only source of some kind of pleasure like that, and that's also why this book was such a good bestseller."

2.2.3 Ločila

Pogoste so napake pri postavljanju vejic, kar vpliva na prevod. Ker gre za spontan govor, je postavljanje vejic veliko bolj nepredvidljivo. Pojavlja se tudi napačna raba vprašaja in klicaja. Program pogosto uporablja tudi dvopičja, kar pa ne vpliva na kvaliteto prevoda. V nadaljevanju projekta bo večja količina pozornosti namenjena tudi izboljšavi modela za postavljanje ločil.

2.2.4 Elipse

Težavo predstavljajo tudi izpusti, ki so naraven del govorjene slovenščine, a jih strojni prevajalnik ne zazna in jih ne zna smiselno vključiti v prevod. Določene jezikovne prvine so v slovenskem jeziku predvidljive, njihov izpust pa se v prevodu izgubi, rezultat je manj smiselna vsebina.

FRI, Rozman	"Se pravi, a vidite ima pa tukaj nula, ena pa **[ima]** potem sedem."	"That is to say, but you see it has zero here and one then seven."

2.2.5 Homonimija

Homonimija predstavlja težavo tudi pri prevodih. Razpoznavalnik včasih slabo razbere, kateri pomen je bil tisti, ki ga je govorec v slovenskem jeziku uporabil. Tako lahko predlog (“med”) prepozna kot samostalnik (prav tako “med”) in ga prevede napačno (“honey”). Rezultat je nesmiselna poved.

FRI, Rozman	"Ne, tako, da vse je super. Z 32: bitni, mi, 64, bitni **med**, besedami, ampak tisto, kar trpi pri tem, je pa izkoriščenost pomnilnik ali pa gostota programske kode. Ne."	"No, so everything is great. With 32: bit, we, 64, bit **honey**, words, but what suffers in doing so, however, is memory utilization or program code density. No ."
FF, Kuhar	"Govori o treh: kategorijah, to je monoseksualnost, heteroseksualnost in homoseksualnost. Zdaj, ta, prvi izraz se nekako ni uveljavil. Monoseksualnost **tem** želi označiti vsa seksualna dejanja, ki jih človek počne sam s sabo."	"It speaks of three categories: monosexuality, heterosexuality, and homosexuality. Now, this, the first term somehow hasn't taken hold. Monosexuality **themes** seeks to denote all sexual acts that a man commits with himself."
FF, Kuhar	"Masturbacija in podobne: zadeve, ampak tega danes ne imenujemo tako, medtem ko heteroseksualnost in homoseksualnost, **pasta**, izraza, ki jih uporabljamo še danes, in nastaneta leta 1868, torej pred tem tudi izraza heteroseksualnost ni."	"Masturbation and the like: things, but we don't call it that today, while heterosexuality and homosexuality, **the paste**, the term we still use today, originated in 1868, so before that the term heterosexuality did not exist."
FF, Kuhar	"To je ta pravnik, ki sem ga prej omenil, s katerim **sta** bila v neki komunikaciji, skupaj z Kertbenyjem."	"It's this lawyer I mentioned earlier with whom **you** were in some communication, along with Kertbeny."

2.2.6 Deiktičnost

Težavo za prevajalnik predstavljajo tudi prevodi deiktov, saj je njihov pomen težko razbrati brez zunajjezikovnega konteksta.

FRI, Rozman	"Načine kodiranja, kjer pač najbolj pogoste podatke zakodirati z najmanjšim	"Encoding methods, where the most common data is encoded with the smallest

	številom bitov. Tako da **tam je ta podobna**."	number of bits. So **there's this one like that**."

2.2.7 Vsebinske napake pri prevajanju

Pogosta napaka pri prevajanju je bila nepovezanost besede, ki jo sistem prevaja, s sobesedilom. Naboj se lahko prevaja kot *bullet* ali (electric) *charge*. Pri predavanju iz elektrotehnike je verjetnost uporabe drugega precej višja, kljub temu pa je ON uporabil prvi prevod.

FE, Sinigoj	"Saj tisti **naboji**, ki pride semle, gre od tukaj, stran ta **naboj**! Ekipa sem dol pride gre, pa od tukaj gor a ne?"	"After all, those **bullets** that come here go from here, away from that **charge**! The team is coming down here, but from here up, right?"
FRI, Rozman	“Abeceda, ki je pa zdaj **osembitno**, ne, in sicer je, potem, bi rekel: če je bit sedem, je enako nič …”	“The alphabet, which is now **octave**, no, and that is, then, I would say: if bit seven is equal to zero …”

2.2.8 Nekonsistentna raba terminologije

Četudi je program besedo že zaznal pravilno in izbral ustrezen prevod, se je lahko ob naslednji omembi pojavila napaka. Tako je kljub temu, da je “medpomnilnik” že prevedel kot “cache”, besedo naslednjič prevedel kot “buffer”.

FRI, Bajec	“Kaj lahko naredimo: kajpa ja, ne pozabit, da nekatere strani, ki jih tukaj imamo v **medpomnilniku**, morda trenutno uporabljajo transakcije.” V **medpomnilniku**, seveda, da če je stran v medpomnilniku, je super ne, lahko jo takoj vrnemo. Naslov in povečamo.	“What we can do: well yeah, don’t forget that some of the pages we have here in the **cache** may be using transactions at the moment.” In the **buffer**, of course, if the page is in the buffer, it’s a great no, we can return it right away. Title and increase.

Podobno tudi:

- interval zaupanja: confidence interval/*trust interval
- množice:sets/*crowds
- upor: resistance/*rebellion

2.2.9 Nestandardna slovenščina

Čeprav slovar razpoznavalnika govora vsebuje tudi nekatere nestandardne besede (kot je npr. šiht v spodnjem primeru), pa te predstavljajo problem za strojni prevajalnik. Prevajanje nestandardnih jezikovnih elementov tako pogosto privede do vsebinskih napak.

FE, Sinigoj	*"Vat študij je **šiht**, to je tako kot osem ur na **šiht hodil**"*	*"Vat study is **shit**, it's like eight hours of **shit walking!**"*

2.2.10 Ostalo

Pri izpisu in prevajanju se pojavljajo tudi besede oziroma besedne zveze, ki za register predavanja niso primerne.

"Dude", "perverted snow"

V nekateri primerih razlogi za prevajalske napake niso tako očitni:

FE, Sinigoj	*"Ne gre, vse v življenju je **garanje**."*	*"It doesn't work, everything in life is a **guarantee**."*

Težavo predstavljajo tudi števila, zlasti z decimalnimi mesti, ki jih razpoznavalnik razpoznava ločeno, prevajalnik pa nato tudi prevaja ločeno:

FDV, Lozar-Manfreda	"Zdaj točkovna ocena bi bila neka ocena, ki je podana z eno samo vrednostjo, z eno samo točko, in če mi recimo izračun, medtem ko intervala, je seveda podana z intervalom vrednosti, torej v tem primeru napovedovanja izidov v županskih volitev, če bi mi na osnovi tistega vzorca, kjer smo dobili vzorčni **delež, nič, cela pet, osem, tri,** rekli, da je delež volivcev, ki bo dovolil, ki volijo tega kandidata na populaciji enak 58,3 odstotka. To bi bila točkovna	"Now a point estimate would be some estimate given with a single value, with a single point, and if we say a calculation, while an interval is of course given with a value interval, so in this case predicting the results in a mayoral election if we, based on that sample where we got the sample **share, zero, a whole five, eight, three,** said that the share of voters who will allow those who vote for this candidate in the population is equal to 58.3 percent. That would be a point score, but you don't think it's good to do so."

	ocena, ne pa se vam zdi dobro, da tako naredimo."	
FE, Sinigoj	"Tukaj je minus, **4 cilje, osem**, plus, **4, cilj osem**; minus, **4, cilje osem**; se pravi, minus **4, cele osem,** Milji! Kolona."	"Here is minus**, 4 goals, eight**, plus, **4, goal eight**; minus, **4, goals eight**; that is, minus **4, a whole eight,** Miles! Column."

3 Ostala opažanja

3.1 Predlogi za izboljšave na področju dostopnosti za študente z omejitvami na področju slušnega in vidnega razumevanja

V naslednjih fazah projekta bo na platformi dodana možnost spreminjanja barve ozadja, tako bo branje olajšano za posameznike z disleksijo in drugimi primanjkljaji na področju vizualnega procesiranja informacij.

4 Povzetek ključnih ugotovitev

Pri vseh evalviranih posnetkih pilotnih predavanj so se pojavljale podobne napake, v nadaljevanju projekta pa bomo posebno pozornost posvetili naslednjim težavam:

- **Segmentacija**. V govorjenem besedilu so meje med povedmi težko določljive. Z dodatnimi popravki in nadgradnjami modela lahko izboljšamo razpoznavo mej in tako poskrbimo, da so prelomi na pravih mestih, segmenti pa niso predolgi.
- **Postavljanje ločil**. V poročilu se postavljanju (nekončnih) ločil nismo posvečali. V nadaljevanju bo izboljšano tudi postavljanje vejic, kar bo pripomoglo k boljši segmentaciji, predvsem pa omogočilo boljše prevode.
- **Napačno razpoznane besede**. Do napak pride zaradi omejenega nabora besed v slovarju in težav z homonimijo oziroma bližnjo homonimijo. Dobro bi bilo, da se v nadaljevanju projekta nabor besedja še dodatno razširi in tako poskrbi, da bo na voljo vse besedje, ki ga predavatelji uporabljajo pri svojih predavanjih.
- **Diskurzni označevalci**. Ker so diskurzni označevalci pogost element v spontanem govoru, bo v nadaljevanju projekta pomembno, da se poišče načine za njihovo prepoznavo in izpust v transkripciji. Težavo predstavlja predvsem »(a/ali) ne«, ki vpliva na pomen.
- **Napake pri prevajanju**. V poročilu so omenjene tudi ostale napake, ki se pojavljajo pri prevajanju. Gre za elipse, deiktičnost, homonimijo, nepovezanost besede, ki jo sistem prevaja, s sobesedilom in za ostale napake, ki so posledica

nestandardne slovenščine, nekonsistentne rabe terminologije ipd. V nadaljevanju projekta bo sistem, ki slovenska besedila prevaja v angleščino (in ostale tuje jezike), testiran še na drugih prevajalnikih (trenutno je bil uporabljen le Googlov prevajalnik; poleg prevajalnika, ki je naučen specifično na slovenskih besedilih in njihovih prevodih, velja npr. preizkusiti še DeepL). Tisti, ki se bo izkazal za najboljšega, bo v nadaljevanju postal del sistema ON.

Pilotska raziskava kaže, da lahko sistem za strojno prevajanje predavanj z dopolnjevanjem izboljšamo do te mere, da je splošno uporaben pri učnem procesu. Pri tem je treba upoštevati določene omejitve: pripraviti je treba priporočila predavateljem, v katerih so pojasnjene omejitve pri uporabi sistema in navodila, kako lahko pripomorejo k boljšemu delovanju sistema (npr. z razločnejšim govorom, z izogibanjem pretirani rabi diskurznih označevalcev). Evalvacijo delovanja sistema bomo po potrebi opravili večkrat, da ocenimo vpliv morebitnih izboljšav in spremljamo napredek. V evalvacijo bomo vključili tudi končne uporabnike (tuje študente), da ocenimo, kako delovanje sistema doživljajo oni in v kolikšni meri lahko sledijo predavanju. Posebno pozornost bomo namenili tudi študentom s posebnimi potrebami in poskrbeli, da bo sistem kar se da uporabniško prijazen tudi zanje.

prof. dr. Marko Bajec

vodja projekta Online Notes