GOVORNI, DIALOŠKI IN MULTIMODALNI JEZIKOVNI VIRI: PREGLED STANJA

Darinka Verdonik, Andrej Žgank, Simona Majhenič, Izidor Mlakar

Vsebinsko poročilo. HUMANIPA (J2-1737) in Napredne metode interakcij v telekomunikacijah (P2-0069)

Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru

Maribor, maj 2020

Vsebina

1 Govorni korpusi 2
2 Baze za avtomatsko razpoznavanje govora (ASR) 5
3 Strojno prevajanje govora v govor in tolmaški korpusi 8
4 Korpusi dialoških dejanj 10
5 Večmodalno označeni korpusi 12
6 Literatura 14

Ta dokument podaja pregled stanja govornih, dialoških in multimodalnih jezikovnih virov, potrebnih za razvoj govornih tehnologij in raziskave govorjene komunikacije.

1 Govorni korpusi

Govorni korpusi so obsežne zbirke zapisov govorjene rabe. Idealno je, da vključujejo tudi posnetek, čeprav večkrat ni tako.

Iz govornih korpusov lahko izluščimo informacije o jeziku, ki jih iz pisnih korpusov ne moremo dobiti. Pisni vir ne more v celoti ustrezno zastopati govorjene rabe. Iz samo 1-milijonskega govornega korpusa (GOS) kot dodatnega vira informacij ob 1-milijardnem pisnem korpusu (Gigafida) dobimo samo na področju besedja:

- do 100 dodatnih slovarskih iztočnic, ki jih sicer v slovar ne bi zajeli, so pa pogost del vsakdanjega govorjenega jezika
- do 1000 slovarskih iztočnic lahko dopolnimo z dodatnimi pomeni, funkcijami, frazami ali kolokacijami, značilnimi za govorjeno rabo
- do 2000 slovarskih iztočnic lahko opremimo z dodatnimi informacijami o izgovornih različicah, številne slovarske iztočnice pa dobijo povsem nove informacije o pragmatičnih in diskurznih razsežnostih rabe besedja, zlasti tam, kjer gre za metadiskurzno, semantičnega pomena izpraznjeno rabo

(Verdonik, Sepesy Maučec, 2017)

Posebnosti so tudi na skladenjski ravni, odpirajo se vsaj naslednja temeljna vprašanja:

- Kaj je osnovna enota? V govoru nimamo povedi, ki bi bila označena s piko
- Samopopravljanja – govorci se vračajo po sintagmatski osi vračajo in spreminjajo strukturi izjave za nazaj
- Nedokončane enote
- Ponavljanja
- Številne nestavčne enote

(Akinnaso, 1982)

Pomembno je tudi vprašanje označevalnikov za govorni korpus, oblikoslovnega in skladenjskega. Že oblikoslovni označevalnik je treba prilagajati govorjenemu jeziku (številni neverbalni in polverbalni glasovi, besedni fragmenti, za govor značilno besedje), vprašanje skladenjskega označevanja pa je še veliko bolj zahtevno, saj je skladnja govorjenega besedila zelo specifična v primerjavi s skladnjo pisnega besedila. Za slovenščino je bila na tem področju narejena samo manjša pilotna študija (Dobrovoljc, Martinc, 2018).

Kot pri vseh ostalih virih tudi pri govornih korpusih obstajajo največje baze za angleščino. Največje, kot sta recimo COCA in Bank of English, obsegajo tudi do 100 mio. besed ali več. Večina korpusov za angleščino in ostale evropske jezike se giblje med dvema do desetimi milijoni besed, na spodnji meji pa so korpusi z 1 mio. besed, kot je slovenski govorni korpus GOS (Verdonik et al., 2013). Ob tem je GOS zgleden zaradi skrbne uravnoteženosti, primerljiv z BNC (The British National Corpus) in drugimi

uravnoteženimi korpusi. Mnogi korpusi različnih evropskih jezikov namreč s ciljem po čim večjem obsegu:

- združujejo vse razpoložljive vire v en korpus (npr. češki),
- posegajo tudi po branih govorjenih besedilih in literarnih besedilih, npr. ruski in poljski, da hitreje dosežejo velik obseg korpusa,
- prevladuje tip govorjene rabe, ki ga je lažje dobiti, tipično govor medijev (npr. slovaški).

Prav tako referenčni govorni korpusi pogosto obsegajo samo pisni zapis govora, brez povezave na izvorni zvočni posnetek. To je tudi z vidika jezikoslovne analize velika slabost, hkrati pa je tak korpus neuporaben za akustično procesiranje.

Tabela 1: Pregled govornih korpusov za angleščino, druge evropske in slovanske jezike

Korpus	**Jezik**	**Obseg v mio. besed (št. ur)**	**Vsebuje**
Angleški			
COCA – corpus of contemporary american english*	US	118	Neformalni pogovori ter radijski in TV intervjuji
Bank of English*	UK in US	41 – UK 20 – US	Velik del brana zapisana besedila
American National Corpus*	US	4	Tel. klici, neformalni pogovori, akademski govor
BNC – British National Corpus*	UK	10	Uravnotežen korpus
CANCODE – Cambridge and Nottingham Corpus of Discourse in English	UK	5	
Longman Spoken American Corpus	US	5	Vsakdanji pogovori, demografsko uravnotežen
Switchboard corpus	US	~2,4 (260 ur)	Telefonski pogovori
Drugi evropski			
Spoken Dutch Corpus / Corpus Gesproken Nederlands	nizozemski	9	Uravnoteženi različni govorjeni žanri
LANCHART – Corpus of Spoken Danish	danski	7	
CLIPS	italijanski	~1 (100 ur)	Vsakdanji govor iz 15 italijanskih mest
C-ORAL-ROM	italijanski francoski portugalski španski	1,5 2 1,5 1	
DGD – Datenbank fur Gesprocheness	nemški	12 (4000 ur)	Različni korpusi: akademski govor, dialektalni korpusi, standardna nemščina…

Deutsch			
CEFC – Corpus for the Study of Contemporary French	francoski	10 (350 ur)	Soočenja, šolski diskurz, literarni teksti…
Göteborg Spoken Language Corpus	švedski	1,4	Uravnoteženi različni žanri
LIA in NoTa-Oslo	norveški	2,5	Intervjuji in pogovori v dialektih in mestni govor
Corpus of spoken Estonian	estonski	1	Uravnotežen korpus
Slovanski			
Corpus of Spoken Russian	ruski	10	Literarni teksti, dialektalni govor
Czech National Corpus – SYN	češki	4	Združeni različni korpusi: neformalna govorjena češčina, demografsko uravnoteženo gradivo, šolski diskurz itd.
Polish National Corpus	poljski	30	Javni govori, TV diskusije, pogovorne oddaje, intervjuji, novice, neformalni spontani pogovori, leposlovje
Corpus of Spoken Slovak	slovaški	6,6 (714 ur)	Formalna, standardna slovaščina, mediji itd.
Korpus govorjene slovenščine – GOS	slovenski	1	Uravnotežen korpus
GOS Videolectures	slovenski	0,179 (22 ur)	Javna predavanja

* Govorni podkorpus je komponenta pisnega korpusa.

2 Baze za avtomatsko razpoznavanje govora (ASR)

Področje jezikovnih virov, povezanih z govorjenim jezikom, je avtomatsko razpoznavanje govora, pri čemer se nanašamo samo na razpoznavanje tekočega govora (Lee, 2003). Razpoznavanje izoliranih besed, kakršno npr. vključujejo nekatere sodobne naprave, so posebna, ločena tematika. ASR je tipična govornotehnološka aplikacija, ki je uporabna:

- kot samostojna funkcionalnost (npr. za narekovanje besedil)
- kot ena od komponent prevajanja govora v govor
- kot ena od komponent sistema dialoga v uporabniških vmesnikih naprav

Področje postaja eno od bolj perečih za slovenski jezik zadnji čas, veliko prav zaradi razširjenosti pogovornih aplikacij interneta stvari (IoT), kot je Alexa.

Čeprav že govorni korpusi predstavljajo potencial za govorno bazo, potrebno za razvoj avtomatskega razpoznavanja govora, razvoj velikokrat poteka ločeno (Žgank et al., 2014). Razlogi so lahko:

- Razpoznavanje govora ima svoje področno omejene potrebe po specifičnem gradivu, nekoliko drugačne, kot jih ima jezikoslovje. Če so potencialna področja uporabe npr. podnaslavljanje predavanj, video posnetkov, TV-vsebin, narekovalniki v zdravstvu ali pravosodju, knjige za slepe, posredovanje informacij, opravljanje nalog, kot so recimo rezervacije terminov ipd., potem je idealno, da je večina gradiva v govorni bazi za ASR tudi iz teh področij uporabe.
- Razvoj razpoznavanja govora v veliki meri poteka znotraj gospodarstva in potrebuje licence za komercialno rabo, medtem ko so obstoječi govorni korpusi pogosto celo nedostopni, če so že dostopni (Besacier et al., 2014), pa so večinoma samo za nekomercialno rabo – tako je tudi v primeru slovenskega govornega korpusa GOS.
- Za govorno bazo je nujen zvočni posnetek, zaželena je tudi ročna segmentacija na osnovne enote transkripcije (tipično izjave), to pa je pri govornih korpusih redko, čeprav zaželeno.
- Za govorno bazo je značilno, da morajo posnetki tudi iz vidika akustičnega okolja (šum ozadja, drugi izvori zvoka, karakteristike terminalska opreme in omrežja …) čim bolj ustrezati dejanskemu namenu uporabe, medtem ko je lahko izrazito šumno ozadje v primeru govornega korpusa prej ovira.
- Pogosto ni ustreznega pretoka informacij in sodelovanja.

Govorne baze, ki jih uporabljajo nekateri znani sistemi za ASR, obsegajo praviloma okrog 1000 ur govora ali več (Xiong et al., 2018; Zhang et al., 2014). To velja ne samo za jezike, kot je angleški, ampak tudi za nekatere manjše, po številu govorcev bolj primerljive s slovenščino, kot je recimo za slovaški (Staš in Juhar, 2015). Obseg govorne baze za ASR, manjši kot 1000 ur, ne omogoča razvoja ASR na kakovostni ravni, ki bi dosegala obstoječe standarde na tem področju.

Poleg govornih baz sta ključna jezikovna vira za ASR še:

- Velik pisni korpus, kakršen je za slovenščino korpus Gigafida (Arhar in Gorjanc, 2007; Erjavec in Berginc, 2012). Odprta dostopnost takega korpusa omogoča razvoj jezikovnega modela.
- Fonetični in oblikoslovni slovar, kakršen je za slovenščino (delno) Sloleks.

Tabela 2: Pregled govornih baz ASR za slovenščino, večjih od treh ur

baza	kanal	tip ASR	inštitucija	obseg (ur)	Transkripcije	Dostopnost
SpeechDat(II)	telefon	Izolirano/vezano	Siemens/ UM FERI	16	Ročno tvorjene	ELRA
Polidat	telefon	Izolirano/vezano	UM FERI	17	Ročno tvorjene	UM FERI
Gopolis	telefon	Vezano/tekoče	UL FE	15	Ročno tvorjene	UL FE
VNTV/VNRAD	studio	Vezano/tekoče	UL FE	6	Ročno tvorjene	UL FE
SNABI-tel	telefon	Vezano/tekoče	UM FERI	10	Ročno tvorjene	Clarin
SNABI-studio	studio	Vezano/tekoče	UM FERI	10	Ročno tvorjene	Clarin
UMB BNSI Broadcast News	TV	tekoče	UM FERI	36	Ročno tvorjene	ELRA
IETK-TV	TV	tekoče	UM FERI	30	Ročno tvorjene	UM FERI
SiBN	TV	tekoče	UL FE	36	Ročno tvorjene	UL FE
SloParl	TV	tekoče	UM FERI	100	Magnetogrami	UM FERI
SI TEDx-UM	dvorana	tekoče	UM FERI	54	Avtomatsko tvorjene (WER~ 50%)	CC-BY
SOFES	Letalske informacije	tekoče	UL FE	10		Clarin
GOS (javni)	TV	tekoče		42	Ročno tvorjene	Clarin
GOS-Videolectures	dvorana	tekoče	UM FERI	22	Ročno tvorjene	Clarin

Tabela 3: Pregled najpomembnejših govornih baz ASR za angleščino

baza	kanal	tip ASR	inštitucija	obseg (ur)	Transkripcije	Dostopnost
Wall Street Journal	studio	tekoče	LDC	80	Ročno tvorjene	LDC
ENG Broadcast News (HUB baze)	TV	tekoče	NIST/LDC	250	Ročno tvorjene	LDC
Switchboard	telefon	tekoče	LDC	300	Ročno tvorjene	LDC
Fisher	telefon	tekoče	DARPA/LDC	2000	Ročno tvorjene	LDC
LibriSpeech	studio	tekoče	John HopkinsUniversity	1000	Avtomatska poravnava	CC-BY
Mozilla Common	studio	tekoče	Mozilla/skupnost	780	Ročno tvorjene	CC-BY

Voice						
TED-LIUM(1,2,3)	dvorana	tekoče	LIUM	777	Avtomatsko tvorjene (WER~ 5%)	CC-BY
Baidu ASR	Studio / TV	tekoče	Baidu	18000	Ročno / avtomatsko	Baidu interno
Google ASR	Dom / Studio / TV	tekoče	Google	40000	Ročno / avtomatsko	Google interno

Tabela 4: Nekaj izbranih govornih baz ASR za ostale jezike

baza	**Jezik**	**kanal**	**tip ASR**	**inštitucija**	**obseg (ur)**	**Transkripcije**	**Dostopnost**
Baidu Mandarin	kitajski	Studio/TV	tekoče	Baidu	22000	Ročno/avtomatsko	Baidu interno
MAGICDATA-CC	kitajski	studio	tekoče	MAGICDATA	755	Ročno/avtomatsko	CC-BY
MAGICDATA	kitajski	studio	tekoče	MAGICDATA	10567	Ročno/avtomatsko	MAGICDATA interno
AISHELL-1	kitajski	studio	tekoče	Shell-Shell	170	Ročno/avtomatsko	CC-BY
Google Italian	italijanski	Dom / Studio / TV	tekoče	Google	15000	Ročno/avtomatsko	Google interno
Google French	francoski	Dom / Studio / TV	tekoče	Google	15000	Ročno/avtomatsko	Google interno
Google Brazilian	portugalski	Dom / Studio / TV	tekoče	Google	15000	Ročno/avtomatsko	Google interno
Google Russian	ruski	Dom / Studio / TV	tekoče	Google	15000	Ročno/avtomatsko	Google interno
Russian Open Speech	ruski	mešano	tekoče	skupnost	7000	Ročno/avtomatsko	CC-BY, PD
Mozilla Common Voice	nemški	studio	tekoče	Mozilla / skupnost	325	Ročne	CC-BY
Parlament Parla	katalonski	dvorana	tekoče	Col-lectivaT	320	Ročno	CC-BY
Mozilla Common Voice	katalonski	studio	tekoče	Mozilla/skupnost	107	Ročne	CC-BY

3 Strojno prevajanje govora v govor in tolmaški korpusi

Aktualno področje govornih tehnologij je tudi strojno prevajanja govora v govor. Razen razpoznavalnika ta vključuje še dve ključni komponenti:

- prevajanje govorjenega besedila
- sintezo govora

Ker deluje prevajalnik v dve smeri, potrebujemo skupno dva krat tri module, torej šest.

Tukaj se osredotočimo samo na komponento prevajanja. Za učenje te so ključni viri veliki vzporedni ali tudi primerljivi korpusi prevodov. Zaradi potrebe po veliki količini podatkov so ti viri praviloma samo pisni, zato jih kot takih v tem pregledu ne obravnavamo posebej. Za govorjeno obliko prevajanja, torej tolmačenje, pa so viri izredno skromni. Toda posebnosti govorjenega jezika so pomembne tudi za razvoj tehnologij, ki skušajo prevajati iz govora v govor. Sistem za prevajanje pisnih besedil še ne pomeni učinkovite rešitve za strojno prevajanje govora v govor brez dodatnih prilagoditev. Te se morajo nanašati na:

- posebnosti govorjene rabe (obilica metadiskurza, skladenjske posebnosti, večmodalnost – del komunikacije poteka tudi nejezikovno…)
- napake, ki nastajajo pri strojnem razpoznavanju govora

Tolmaški korpusi pri tem predstavljajo jezikovnih vir z izrednim potencialom informacij tako za razvoj strojnega prevajanja govora v govor kot za razvoj tolmačeslovja kot vede. Vendar je v nasprotju z govornimi korpusi in govornimi bazami za ASR pri tolmaških korpusih stanje tudi v mednarodnem, ne samo slovenskem prostoru, izredno skromno.

Najbolj znana tolmaška korpusa v evropskem prostoru, EPIC in njegova nadgradnja EPTIC, ki vključujeta govore v Evropskem parlamentu, obsegata okrog 175.000 besed oz. 18 ur (Bernardini idr., 2016) (Bendazzolli in Sandrelli, 2005). EPTIC sicer obsega tudi komponento za slovenski jezik EPTIC-SI, ki pa nastaja kot individualna akademska pobuda v okviru zaključnih del študentov in njen obseg je temu primeren – 77.000 besed (Lampe, 2019).

Večina ostalih tolmaških korpusov za evropske jezike se giblje v obsegu okrog 20 ur. Večje vire je zaslediti za neevropske jezike, npr. arabskega (korpus WAW), kjer zasledimo obsege korpusov tudi čez 100 ur (Abdelali et al., 2018).

Za razvoj strojnega prevajanja so zlasti zanimivi intermodalni korpusi, kot recimo korpus WAW ali EPTIC, saj besedilo v izvirnem jeziku primerjajo s simultanim tolmačenjem in kasneje opravljenim prevodom. Prednost intermodalnih korpusov je primerjava različnih prevodnih strategij, denimo krajšanje, povzemanje, izpusti ali dodajanje, s čimer bi lahko izboljšali strojno prevajanje.

Tabela 5: Pregled tolmaških korpusov

Korpus	**Jeziki**	**Obseg v mio. besed (št. ur)**	**Vsebuje**	**Tema**
EPIC	EN–IT–ES	>18 ur >177.000 besed	transkripcijo, zvočni in video posnetek	govori v Evropskem parlamentu
EPTIC	EN–IT–FR–PL–SL	>> 175.000 [1]besed	transkripcijo, zvočni in video posnetek	govori v Evropskem parlamentu
EPTIC-SI	EN→SL	77.000 besed	transkripcijo, zvočni in video posnetek	govori v Evropskem parlamentu
WAW	EN→AR	119 oz. 63 ur[2]	transkripcijo, zvočni posnetek	inovacije v izobraževanju in zdravstvu, raziskovanje in razvoj
DIRSI[3]	EN↔IT	20 ur oz. 130.000 besed	transkripcijo, zvočni posnetek	medicina
FOOTIE[4]	IT, FR, ES, EN	/	/	nogomet
K2[5]	DE, TR, PT, ES	25 ur oz. 160.000 besed	/	pogovori med zdravnikom in pacientom
K6[6]	PT→DE	5 ur oz. 35.000 besed	/	okoljevarstvo

[1] Brez tolmačenih besedil.
[2] Posnetih je sicer 119 ur govora, vendar jih je analiziranih le 63.
[3] (Bendazzoli in Sandrelli, 2009, Bendazzoli, 2016).
[4] (Sandrelli, 2012).
[5] (Bendazzoli in Sandrelli, 2009).
[6] (Meyer, 2008).

4 Korpusi dialoških dejanj

Dialoška dejanja pomenijo posebno raven označevanja, ki se uvršča v široko polje semantično-pragmatičnih razsežnosti jezika. Pomen, izražen skozi jezik, ima številne plasti in nemogoče ga je opisati z eno samo označevalno shemo. Dialoška dejanja se navezujejo na akcijsko oz. vplivanjsko razsežnost jezika, tj. kaj skozi jezik naredimo, kako z njim vplivamo na druge, spreminjamo svet: konkretno, skozi jezik svetujemo, prosimo, predlagamo, zahtevamo, informiramo, obljubljamo, poizvedujemo po informacijah, pozdravljamo, obtožujemo, hvalimo…

Označevanje dialoških dejanj se na področju govornih tehnologij uporablja zlasti pri razvoju pogovornih agentov (Siri, Alexa, Cortana, Google Assistant) oz. sistemov dialoga, v jezikoslovju pa se uvršča na področje korpusne pragmatike. Oboje je tudi v mednarodnem prostoru še slabo razvito.

Označevalne sheme dialoških dejanj so različne. Zaenkrat še ni široko sprejetega standarda, čeprav se v zadnjih letih avtorji sheme ISO 24617-2 (2012) trudijo, da bi to postala ta shema. Splošno uporabne, in ne domensko specifične, so sicer zlasti štiri sheme: DAMSL (Allen, Core, 1997) in njene izpeljanke SWBD-DAMSL (Jurafsky et al., 1997), MRDA (Dhillon et al., 2004); shema projekta AMI (2005), shema ISO 24617-2 (2012) ter DART (Weisser, 2019).

Večina dialoško označenih korpusov obstaja za angleški jezik, bodisi za otoško bodisi ameriško različico. Posamezni od teh obsegajo tudi že več kot 100 ur govora, mnogi pa tudi samo po nekaj deset ur. Gradivo, ki ga označujejo, so večinoma studijski oziroma poustvarjeni dialogi, pogosto na temo posredovanja informacij, izvajanja rezervacij, opravljanja nalog ali delovnih sestankov. Avtentično gradivo je uporabljeno izjemoma. Za neangleške jezike redkeje naletimo na korpuse dialoških dejanj – najdemo jih za nizozemski, nemški in italijanski jezik. Za slovenski jezik je bil konec 2019 razvit prvi enourni korpus dialoških dejanj GORDAN (Verdonik, 2020).

Tabela 6: Pregled dialoško označenih korpusov

Korpus	**Jezik**	**Obseg**	**Vsebuje**	**Označevalna shema**
Večje				
Switchboard dialogue act corpus	US	1,4 mio. besed	Telefonski pogovori; setup dialogues	SWBD-DAMSL
ICSI-MRDA	Angleški – US	75 ur	Sestanki (avtentično gradivo, sorodno z vsakdanjim pogovorom)	MRDA – prilagojena SWBD-DAMSL
AMI	angleški	100 ur	Sestanki – set-up (http://www.amiproject.org/ami-scientific-portal/meeting-corpus.html)	AMI
DialogBank	Nizozemski, angleški	Ni podatka, ocenjeno na 150 ur	HCRC, Switchboard, TRAINS, DBOX, DIAMOND, Schipol, OVIS, Dutch Map Task	ISO 24617-2
SPAAC	UK	182.300 besed	Klicni centri britanskega telekoma in železnic	DART
Manjše				

SPAADIA (Speech Act Annotated Dialogues) Corpus	UK	35 dialogov	Poizvedovanje po voznih redih, rezervacije	Različica DART
TRAINS	US	55.000 besed	Opravljanje nalog – set-up	DAMSL
COCONUT	US	35 dialogov	Pogajalni govor, kupovanje pohištva – set-up	prilagojeni DAMSL
Monroe	US	20 dialogov	Obvladovanje urgentnih situacij – set-up	Samostojna shema
HCRC Map Task Corpus	UK	15 ur (128 dialogov)	Reševanje problema z zemljevidi – set-up	Samostojna shema
DIAMOND corpus	nizozemski	(majhen)	Reševanje problema s faks napravo – set-up	DIT++
DBOX	angleški			ISO 24617-2
Verbmobil corpus	nemški, japonski, angleški	1172 dialogov	Dogovarjanje za sestanek, podajanje informacij (turizem)	Verbmobil
MATE – verjetno MapTask corpus		100 dialogov		DAMSL
MALTUS – korpus ICSI-MRDA				MALTUS – prilagojena MRDA
ADAM	italijanski	58.000 besed (7 ur)	Dialogi v turizmu	Prilagojena DAMSL in SWBD-DAMSL
GORDAN/GODI – govorni korpus dialoških dejanj	slovenski	1 ura	Uravnotežen izbor GOS	

5 Večmodalno označeni korpusi

Govorna komunikacija je vedno izražena večmodalno: ne samo prek jezika, ampak tudi preko drugih verbalnih vzvodov. kot so intonacija, poudarki, premori, za katere uporabljamo skupni termin prozodija; in neverbalnih vzvodov, kot so premikanja glave, gibanja telesa, mimike obraza, kar lahko imenujemo s skupnim izrazom govorica telesa oziroma gestikulacija. Toda raziskovanje rabe nejezikovnih in neverbalnih elementov in vloge njihovega sopojavljanja z jezikovnimi elementi sporočila je izredno zahtevno in časovno nehvaležno. Zato se mu raziskovalci radi izogibajo. V jezikoslovju se tovrstne študije, če že, pojavljajo zlasti v sociološko orientiranih vejah analize diskurza, in tudi na tehnološkem področju še nismo bili soočeni z aplikacijami, ki bi znale prepričljivo razpoznavati spremljevalno nejezikovno komunikacijo ali jo sintetizirati v naravni jezik. Vendar pa področje zadnji čas dobiva velik zagon in atraktivnost. Tudi od mednarodnih korporacij lahko verjetno prej ali slej pričakujemo, da bodo Siri, Cortana, Alexa in sorodne aplikacije prej ali slej dobile tudi podobo, ne samo glasu.

Osnova za raziskovanje in razvoj na tem področju so večmodalno označeni korpusi oz. baze, ki temeljijo na video posnetkih različnih govornih situacij, ki poleg transkripcije govora vsebuje tudi oznake drugih modalnosti, tako na ravni prozodije kot na ravni govorice telesa. Tovrstni korpusi obstajajo za večino večjih evropskih jezikov, njihova velikost je prevladujoče do 100 ur, redkeje več. V veliko tovrstnih bazah so posnetki poustvarjeni oz. studijsko posneti, redkeje so avtentični. Vsebine so zelo raznovrstne, od vsakdanjih razgovorov do formalnih razgovorov in sestankov do Youtube posnetkov. Tudi načini in vrste modalnosti, za katere so dodane oznake oz. interpretacije, so zelo raznovrstni.

Za slovenski jezik obstaja večmodalni korpus EVA (Mlakar et al., 2019) v obsegu 1 ure govornega posnetka, označenega na številnih nivojih, od gest in mimike na fizični ravni do njihove semiotične interpretacije, oznak prozodije, sentimenta itd.

Tabela 7 predstavlja podatke o relevantnejših večmodalnih korpusih, ki povezujejo uporabo vseh treh kanalov, govornega, jezikovnega in vizualnega v bolj ali manj spontanih situacijah (Strauss, Minker, 2010; Knight, 2011).

Tabela 7: Pregled dialoško označenih korpusov

Korpus	**Jezik**	**Obseg**	**Vsebuje**
AMI	angleški	100 ur	Sestanki – set-up (AMI, 2005)
SmartKom	nemški	13 ur	Vizard-of-Oz, izvajanje nalog
HuComTech	madžarski	50 ur	Formalni (zaposlitveni razgovor) in neformalni, vsakdanji dialogi – studio
VACE Meeting Corpus	angleški	5 scenarijev	Sestanki, domena vojska, 5 sodelujočih v petih scenarijih ciljno usmerjenega dialoga v kontroliranem okolju.
REPERE	francoski	60 ur	(ELDA)
IFADV	nizozemski	5 ur	prost diskurz med prijatelji/sodelavci
MOSI	Angleški, native in non-native	Ca. 5,5 ure	93 Youtube videov
Niki and Julie Corpus	US ang.	600 dialogov	Opravljanje nalog, dialog z robotom
InSight	Nizozemski,	5 ur	Prosti zasebni razgovor med prijatelji,

Interaction	flamski		studijsko snemanje
CID	francoski	8 ur	Zasebni pogovor
NOMCO	Danski	12 ur	Zasebni pogovori, poustvarjeni
Large-Scale Multimodal Movie Dialogue Corpus	japonski	1100 ur	Izseki dialogov iz filmov, avtomatsko izvlečeni, ni označeno! http://www.ice.tohtech.ac.jp/~inoue/paper/yasuhara16icmi_demo_edited.pdf
NIST corpus	Angleški	15 ur, 19 sestankov	Delno spontani in delno realni scenariji s po 3-9 sodelujočimi
D64	Angleški	8 ur	Dve seji, s po 4-5 sodelujočimi v spontanem pogovoru brez scenarija v domačem okolju
Fruits Cart	Angleški	104 posnetki (4-8 minut na posnetek)	Ciljno usmerjen diskurz v akademskem okolju
Goteborg Spoken Language	Švedski	1.2 miljiona izgovorjenih besed	Posnetki v različnih socialnih kontekstih brez ciljnih scenarijev
SaGA	Nemški	280 minut	Opisi opazovanega okolja v virtualni resničnosti
SK-P 2.0	Nemški	86 sodelujočih v 172 krajših sejah	Ciljno usmerjen diskurz skozi katerega se izvede natančno opredeljena naloga

6 Literatura

Abdelali, A., Temnikova, I., Hedaya, S. in Vogel, S. (2018). The WAW Corpus: The First Corpus of Interpreted Speeches and Their Translations for English and Arabic.

Akinnaso, F. Niyi. (1982). On the differences between spoken and written language. *Language and Speech* 25(2), 97-125.

Allen, J., Core, M. (1997). *Draft of DAMSL: Dialog Act Markup in Several Layers.* Pridobljeno 4. 9. 2019 s strani https://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/.

AMI (2005). *Guidelines for Dialogue Act and Addressee Annotation Version 1.0.* Pridobljeno 4. 9. 2019 s strani http://groups.inf.ed.ac.uk/ami/corpus/Guidelines/dialogue_acts_manual_1.0.pdf.

Arhar, Š. in Gorjanc, V., (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. Jezik in slovstvo 52/2, 95--110.

Bendazzoli, C. in Sandrelli, A. (2009). Corpus-based Interpreting Studies: Early Work and Future Prospects. Pridobljeno 4. 9. 2019 s strani https://www.raco.cat/index.php/Tradumatica/article/view/154835.

Bendazzoli, C. in Sandrelli, A. (2005). An Approach to Corpus-based Interpreting Studies: Developing EPIC (European Parliament Interpreting Corpus). V H. Gerzymisch-Arbogast in S. Nauert (ur.) MuTra 2005 – Challenges of Multidimensional Translation: Conference Proceedings, str. 149—160.

Bernardini, S., Ferraresi, A. in Miličević, M. (2016). From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective. Target, 28(1), str. 61—86.

Besacier, L., Barnard, E., Karpov, A. in Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56. str. 85-100.

Dhillon, R., Bhagat, S., Carvey, H., Shriberg, E. (2004). *Meeting Recorder Project: Dialog Act Labeling Guide.* ICSI Technical Report TR-04-002. Pridobljeno 4. 9. 2019 s strani http://www1.icsi.berkeley.edu/ftp/pub/speech/papers/MRDA-manual.pdf.

Dobrovoljc, K., Martinc, M. (2018). Er ... well, it matters, right? On the role of data representations in spoken language dependency parsing. V: *Proceedings of the workshop*. Second Workshop on Universal Dependencies (UDW 2018), November 1, 2018, Brussels. Strasbourg: Association for Computational Linguistics. Str. 37-46.

Erjavec, T., in Berginc, N.L. (2012). Referenčni korpusi slovenskega jezika (cc) Gigafida in (cc) KRES. In Zbornik Osme konference Jezikovne tehnologije. Ljubljana: Jožef Stefan Institute, str. 57-62.

ISO 24617-2 (2012). *Language resource management – Semantic annotation framework (SemAF): Part 2: Dialogue acts.* Reference number ISO 24617-2:2012(E). Geneva.

Jurafsky, D., Shriberg, E., Biasca, D. (1997). *Switchboard SWBD-DAMSL shallow-discourse-function annotation. Coders manual, draft 13.* University of Colorado at Boulder & +SRI International. Pridobljeno 4. 9. 2019 s strani https://web.stanford.edu/~jurafsky/ws97/manual.august1.html.

Knight, D. (2011). Multimodality and active listenership: A corpus approach. A&C Black.

Lampe, L. (2019). Primerjava medpovednega in v korpusu tolmačenih in prevedenih besedil Evropskega parlamenta EPTIC-SI. Magistrsko delo. Univerza v Ljubljani.

Lee, C.H. (2003). On Automatic Speech Recognition at the Dawn of the 21st Century, IEICE Trans. Inf. Syst., vol. E86-D, No. 3, str. 377-396.

Meyer, B. (2008). Interpreting Proper Names: Different Interventions in Simultaneous and Consecutive Interpreting? Trans-kom, 1(1), str. 105-122.

Mlakar, I., Verdonik, D., Majhenič, S., & Rojc, M. (2019). Towards Pragmatic Understanding of Conversational Intent: A Multimodal Annotation Approach to Multiparty Informal Interaction–The EVA Corpus. In International Conference on Statistical Language and Speech Processing (pp. 19-30). Springer, Cham.

Sandrelli, A. (2012). Interpreting Football Press Conferences: The FOOTIE Corpus. V C. J. Kellett Bidoli (ur.) Interpreting across Genres: Multiple Research Perspectives, str. 78--101. Trst: EUT.

Staš, J., Juhár., J. (2015). Modeling of Slovak language for broadcast news transcription. Journal of Electrical and Electronics Engineering 8, no. 2 (2015): 43.Strauss, P. M., & Minker, W. (2010). Proactive spoken dialogue interaction in multi-party environments. Springer Science & Business Media.

Verdonik, D. (2020). *Dialogue act annotated spoken corpus GORDAN 1.0 : (transcription)*. Maribor: Faculty of Electrical Engineering and Computer Science, University, 2020. CLARIN.SI data & tools. Pridobljeno 4. 9. 2019 s strani https://www.clarin.si/repository/xmlui/handle/11356/1291.

Verdonik, D., Sepesy Maučec, M. (2017). A speech corpus as a source of lexical information. *International journal of lexicography*, 30(2), str. 143-166.

Verdonik, D., Kosem, I., Zwitter Vitez, A., Krek, S., Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus : the case of the Slovene corpus GOS. *Language resources and evaluation*, 47(4), str. 1031-1048.

Weisser, M. (2019). *The DART Taxonomy v. 3*. Pridobljeno 4. 9. 2019 s strani http://martinweisser.org/DART_scheme.html.

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., in Stolcke, A. (2018). The Microsoft 2017 conversational speech recognition system. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), str. 5934-5938. IEEE, 2018.

Zhang, X., Trmal, J., Povey, D., in Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, str. 215-219.

Žgank, A., Zwitter Vitez, A., Verdonik, D. (2014). The Slovene BNSI broadcast news database and reference speech corpus GOS: towards the uniform guidelines for future work. Proc. of the LREC'14, Reykjavik, Islandija, str. 2644-2647.