Samstarf um íslenska máltækni
SÍM

Máltækniáætlun fyrir íslensku

Varða 9 – Lokaskýrsla

1. október 2022

Efnisyfirlit

Formáli

Þessi skýrsla dregur saman við vörðu 9 (M9) afurðir fyrir hvern verkþátt sem skilgreindur var í samningnum milli Almannaróms og SÍM (Samstarf um íslenska máltækni) frá september 2021, í verkefnaáætlun frá september 2021 og í endurskoðuðum vörðum frá janúar 2022. Hún þjónar því bæði hlutverki vörðuskýrslu og lokaskýrslu, líkt og lokaskýrslur um afurðir síðustu tveggja ára gerðu.

Þriðja ár er seinasta verkefnaár núverandi máltækniáætlunar. Mörgum vörðum hefur verið náð og fjölmargar afurðir eru nú aðgengilegar með opnu leyfi. Á þriðja ári hefur aukin áhersla verið lögð á að færa afurðir nær notendum, t.d. með hinum ýmsu viðbótum og tengingu þeirra við snjalltæki. Í samræmi við þetta hafa notendaprófanir haldið áfram, sem hafa skilað sér í góðri endurgjöf sem notuð er til að bæta virkni. Íslenska CLARIN-hirslan samanstendur nú af um 300 færslum og yfirlit yfir færslur var búið til í þeim tilgangi að auka aðgengi hirslunnar, en yfirlitið er aðgengilegt á https://clarin.is/gogn/. Færslurnar eru birtar eftir flokkum og útgáfunúmerum. Í framhaldi af máltækniáætluninni er mikilvægt að fyrrnefndum afurðum sé haldið við, svo þær haldist aðgengilegar og verði ekki úreldar. Einnig er mikilvægt að kynna almenningi allar þessar afurðir svo þær nýtist.

Markmiðum var náð í öllum verkþáttum. Eftirfarandi skýrsla inniheldur lýsingar, umræður og niðurstöður vinnu í öllum 26 verkþáttum, bæði á tímabilum M8 og M9.

G1 – Risamálheildin

Skýrsla: Stofnun Árna Magnússonar í íslenskum fræðum
Aðilar: Stofnun Árna Magnússonar í íslenskum fræðum

Markmið

Áframhaldandi stækkun RMH. Við lok annars árs greinist málheildin í sjö undirmálheildir. Allar þessar málheildir ættu að vera uppfærðar fyrir lok þriðja árs og meðhöndlaðar með nýjustu tólum. Ný útgáfa RMH ætti að innihalda meira en 1,8 milljarða tóka.

Varða 8 – lýsing

M8: Safna viðbótum við samfélagsmiðlamálheildina. Vinna texta bóka og fræðigreina. Vinna gögn sem bárust of seint fyrir fyrri útgáfu RMH.

Varða 9 – lýsing

M9: Öll gögn unnin og eftirstandandi undirmálheildir stækkaðar með nýjum gögnum frá fyrri gagnagjöfum. Ný útgáfa RMH gefin út.

Afurðir

Markmiðum vörðunnar var náð. Stækkun samfélagsmiðlaundirmálheildarinnar er lokið og söfnuð gögn hafa verið unnin. Ný útgáfa RMH, með 2.420 milljón orðum (2.699 milljón tókum), gefin út.

	Ómerkt útgáfa	**Merkt útgáfa**
Adjud	http://hdl.handle.net/20.500.12537/240	http://hdl.handle.net/20.500.12537/241
Books	http://hdl.handle.net/20.500.12537/249	http://hdl.handle.net/20.500.12537/250
Journals	http://hdl.handle.net/20.500.12537/245	http://hdl.handle.net/20.500.12537/246
Law	http://hdl.handle.net/20.500.12537/247	http://hdl.handle.net/20.500.12537/248
News1	http://hdl.handle.net/20.500.12537/236	http://hdl.handle.net/20.500.12537/237
News2	http://hdl.handle.net/20.500.12537/238	http://hdl.handle.net/20.500.12537/239
Parla	http://hdl.handle.net/20.500.12537/208	http://hdl.handle.net/20.500.12537/216
Social	http://hdl.handle.net/20.500.12537/242	http://hdl.handle.net/20.500.12537/243

Wiki	http://hdl.handle.net/20.500.12537/251	http://hdl.handle.net/20.500.12537/252

Skýrsla

Söfnun texta úr bókum og fræðiritum
Það reyndist erfitt í M8 að fá meiri gögn frá útgefendum, og sömuleiðis í M9. Eftir langar samningaviðræður gaf menntamálaráðuneytið okkur tengiliðaupplýsingar margra höfunda sinna, svo vonandi fáum við leyfi þeirra höfunda á komandi mánuðum. Í millitíðinni hafa þær bækur sem við fengum frá ráðuneytinu verið unnar og bíða útgáfu í náinni framtíð. Þessi pakki sem við fengum frá ráðuneytinu var sá alstærsti frá nokkrum útgefanda og þar sem hann hefur verið unninn eru allar safnaðar bækur tilbúnar til notkunar í RMH.

Öll söfnuð og nothæf fræðirit voru unnin fyrir RMH í M9.

Nýjar útgefnar málheildir
Á síðasta ári var RMH gefin út sem átta mismunandi málheildir, ein fyrir hvert undirsvið (tvær fyrir fréttir (News) þar sem þar eru tvö ólík leyfi). Þetta ár höfum við bætt við IGC-Wiki, textum frá íslenskri útgáfu Wikipedia. Ólíkt síðasta ári, þegar hver útgefinn pakki innihélt tvær málheildir, eina með ómerktum textum og aðra með merktum textum, ákváðum við nú að gefa þær út stakar. Þetta er gert svo fólk sem þarf aðeins ómerktu útgáfuna þurfi ekki að sækja allan pakkann, þar sem merktu útgáfurnar eru oft mjög stórar. Af þessari ástæðu gáfum við út 18 málheildir þetta árið, eins og sést í töflunni að ofan (Afurðir).

Samtals innihalda RMH-málheildirnar u.þ.b. 2.429 milljónir orða (tafla 1), sem er aukning um næstum 561 milljón orða frá síðustu útgáfu. Mesta aukningin kemur vegna viðbótar texta frá bland.is, spjallborði á netinu, við IGC-Social, sem inniheldur u.þ.b. 380 milljónir orða. Nýir textar frá fræðiritum eru samtals 3,6 milljónir orða og 740.000 orða frá bókum. Aukning vegna nýrra texta frá árinu 2021 (fréttir, dómsúrskurðir, þingskjöl og ræður) eru um 168 milljónir lesmálsorða.

IGC-News1 and IGC-News2
Samtals eru fréttamálheildirnar tvær meira en 53% af RMH. IGC-News2 er enn stærsta málheildin (37,1%), með texta með takmarkandi leyfi. Við höfum þegar haft samband við alla stærstu miðla á Íslandi, auk flestra smærri miðla, og bætt gögnum þeirra við fréttamálheildirnar tvær. Engum nýjum miðlum var bætt við í ár og aukningin í fréttamálheildunum tveimur (86,5 milljónir orða) er aðeins vegna nýrra texta frá síðasta ári.

IGC-Social
IGC-Social stækkaði umtalsvert og er næst stærsta málheildin, með næstum 30% allra texta í RMH. Engum nýjum bloggum var bætt við, og þau þrjú blogg sem málheildin inniheldur hafa verið óvirk í einhvern tíma svo engum nýjum textum var bætt við. Nýju færslurnar á spjallborðunum þremur innihalda samtals um 1,6 milljónir orða. Það eru einungis 0,3% allra

orða á spjallborðunum þremur, sem sýnir skýrt hvað vinsældir bloggspjallborða eru að dvína. Á sama tíma er aukning orða frá tístum um 23 milljónir.

Málheild	Setningar	Orð	Aukning frá 2021 (orð)	Hlutfall (%)
IGC-Adjud	3.187.434	69.310.007	12.714.210	2,58%
IGC-Books	1.005.330	1.005.330	740.494	0,04%
IGC-Journals	1.088.657	20.894.101	3.682.210	0,86%
IGC-Law	3.174.466	53.259.286	12.646.289	2,19%
IGC-News1	23.435.693	396.651.451	42.191.763	16,33%
IGC-News2	51.915.950	899.836.406	44.356.072	37,05%
IGC-Parla	12.438.041	254.120.759	31.589.700	10,46%
IGC-Social	59.226.104	724.999.194	405.039.944	29,85%
IGC-Wiki	580.756	8.497.031	8.497.031	0,35%
Samtals	**156.052.431**	**2.428.573.565**	**561.457.713**	**100%**

Tafla 1. Yfirlit yfir fjölda setninga og orða í öllum 9 undirmálheildum RMH, ásamt aukningu frá síðustu útgáfu og hlutfall hverrar undirmálheildar.

IGC-Parla, IGC-Adjud, IGC-Law
Opinberir textar, sóttir frá vefsíðu Alþingis (www.althingi.is) og vefsíðum dómstiganna þriggja á Íslandi, eru næstum 16% af heildarmagni texta í RMH. Nýir textar frá fyrri árum eru u.þ.b. 56 milljón orð.

IGC-Journals
Uppbyggingu IGC-Journal var breytt frá fyrri útgáfu, þannig að nú eru textar hvers fræðirits saman í undirmálheildum. ICG-Journal hefur því 23 undirmálheildir. Textum frá þremur nýjum fræðiritum var bætt við, samtals um 1,3 milljónir orða, og nýtt efni bættist við úr fræðiritum sem þegar hafði verið safnað, næstum 2,4 milljónir orða.

IGC-Wiki
Textar frá íslensku Wikipedia-síðunni voru áður hluti af RMH, en á síðasta ári þegar málheildinni var skipt í 8 aðskildar málheildir voru gögnin frá Wikipedia ekki tekin með. Nýjasta skammtinum frá Wikipedia hefur nú verið bætt við, samtals um 8,5 milljónir orða.

IGC-Books
Eins og fram kom að ofan voru samskipti við útgáfufyrirtæki erfið og reyndist erfiðara en búist var við að fá fleiri bækur. IGC-Books inniheldur nú 398 titla. 47 titlar eru í nýju málheildinni sem voru ekki í þeirri síðustu (u.þ.b. 740 þúsund orð), frá fjórum útgefendum, þar af 31 frá

menntamálaráðuneytinu. Við höfum 533 nothæfar bækur frá menntamálaráðuneytinu, þar af eru 162 í málheildinni. Við höfum enn ekki fengið leyfi frá höfundum eftirstandandi bóka.

I5 – Þáttarar

Skýrsla: Stofnun Árna Magnússonar í íslenskum fræðum
Aðilar: Stofnun Árna Magnússonar í íslenskum fræðum, Háskóli Íslands, Miðeind

Markmið

Þessi (litli) verkþáttur felur í sér áframhaldandi aukningu og uppfærslu reglubyggðra þáttara og tauganetsfullþáttara fyrir íslensku.

Auk þess verða útbúin nauðsynleg gögn til að þjálfa UD-þáttara (e. UD parsers), með UD-varpara (e. UD converter tool) sem getur varpað af liðgerðarformi (e. constituency structure) yfir á venslamálfræðiform (e. UD structure). Undirsetti Greynismálheildarinnar frá vörðu M5 í þessu undirverkefni verður varpað úr liðgerðartrjám (e. full constituency trees) yfir í venslamálfræðitré (e. UD trees), þar sem það mun gegna hlutverki frumþjálfunargagna.

Fullþáttun (e. full constituency parsing) er lykilforsenda málrýnieiningar. Þar er reglubyggður þáttari nauðsynlegur, vegna þess að mögulegt er að bæta sérstaklega merktum „villureglum“ við samhengisfrjálsa málfræði hans, þar sem algengum málfræðivillum er lýst. Reglubyggði þáttarinn er líka notaður til að framleiða þjálfunargögn fyrir tauganetsbyggða fullþáttarann. Þess má geta að utan máltækniverkefnisins er þegar farið að nota reglubyggða þáttarann í hugbúnaði, sem ekki eru á vegum máltækniáætlunarinnar, svo sem í talþjóni.

Reglubyggði þáttarinn hefur þegar verið gefinn út í heild sinni, en hægt er að setja hann upp sem Python-pakka ásamt meðfylgjandi skjalabúnaði, og sem opna GitHub-hirslu með MIT-leyfi. Þessi verkþáttur mun skila uppfærslum í Python-pakkann og hirsluna.

Tauganetsþáttarinn, ólíkt reglubyggða þáttaranum, umber málfræðivillur og útbýr áætlað þáttunartré jafnvel fyrir rangar setningar. Hann gæti því nýst betur en reglubyggði þáttarinn þegar kemur að því að þátta raddfyrirspurnir og annað inntak sem kemur beint frá notanda, eins og það kemur fyrir. Hins vegar er ekki hægt að aðlaga málfræði hans að sérstökum notkunardæmum; slík aðlögun krefst markverðs (e. nontrivial) þjálfunarsetts fyrir hvert notkunardæmi.

Varða 8 – lýsing

M8: UD-varpari hefur verið aðlagaður að trjágerð Greynis. Undirsetti Greynismálheildar hefur verið varpað yfir á UD-gerð og gefið út á CLARIN og UD-vefsíðunni.

Varða 9 – lýsing

M9: Tauganets- og fullþáttarar gefnir út á CLARIN. UD-þáttarar hafa verið gefnir út á CLARIN. Besta UD-þáttaranum hefur verið bætt við Korp.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.

UD-vörpunartól hefur verið aðlagað að trjágerð Greynis og gefið út á CLARIN (http://hdl.handle.net/20.500.12537/222). Undirsetti GreynirCorpus hefur verið varpað yfir í form venslamálfræði (e. UD structure) og gefið út á CLARIN (http://hdl.handle.net/20.500.12537/224). Málheildin verður hluti af næstu útgáfu Universal Dependencies, útgáfu 2.11, sem verður gefin út í nóvember.

Reglubyggði fullþáttarinn var uppfærður á CLARIN (http://hdl.handle.net/20.500.12537/269). Flestar breytingar eru tengdar villumálfræðireglum, sem eru notaðar í verkþætti L7 og útskýrðar þar. Undirliggjandi tilreiðari (úr verkþætti I3) var einnig uppfærður á CLARIN (http://hdl.handle.net/20.500.12537/262). Hann styður nú inntak með löngum texta, auk þess að þekkja fleiri gerðir greinarmerkjavillna úr verkþætti L7.

Tveir UD-þáttarar hafa verið gefnir út á CLARIN (http://hdl.handle.net/20.500.12537/273, http://hdl.handle.net/20.500.12537/272). Annar er nákvæmari venslaþáttari en hinn en skortir ákveðna eiginleika, eins og UPOS, XPOS og UFeats, og hinn nær aðeins verri árangri sem UD-þáttari en inniheldur fleiri eiginleika. Sá fyrri fær F-gildi 86,32 hvað varðar LAS (og 89,52 hvað varðar UAS), á meðan sá seinni fær F-gildi 85,97 hvað varðar LAS (og 89,13 hvað varðar UAS).

Leita má að textum sem hafa verið þáttaðir samkvæmt UD eftir UD-merkingu þeirra í Korp.

Skýrsla

UD vörpun

UD-vörpunartólið (http://hdl.handle.net/20.500.12537/169) var aðlagað að trjágerð Greynis. Upprunalega vörpunartólið var hannað til að varpa liðgerðatrjábönkum á formi PPCHE (Penn Parsed Corpora of Historical English), sérstaklega IcePaHC (http://hdl.handle.net/20.500.12537/62). Aðlögunin fól í sér að lesa liði í trjágerð Greynis á réttan hátt, sækja upplýsingar um tóka og uppflettimyndir, uppfæra markaskrá inntaksins og tengja mörk við rétt UD-mark, bæði fyrir tóka og liði.

Trjágerð Greynis inniheldur ýmsa margorða tóka, sem eru ekki leyfðir á UD-forminu. Það eru t.d. nöfn og forsetningar sem innihalda tvö eða fleiri orð, en eru ekki fastir margorða tókar. Bæta

þurfti meðhöndlun þessara tóka við vörpunartólið þar sem þeir koma ekki fyrir í IcePaHC. Lausnin var að sýna hvern tóka stakan, þar sem fyrsti tókinn er haus eftirfarandi tóka, sem hafa venslin *flat*, sem táknar að engar frekari upplýsingar um vensl þeirra séu í boði.

Eftir þessa aðlögun var undirsetti GreynirCorpus varpað yfir á venslamálfræðiform (e. UD structure) með tólinu. Gullstaðalsundirsett GreynirCorpus var notað í þessum tilgangi. Þetta undirsett hefur verið leiðrétt handvirkt og samanstendur af 5.000 setningum, skipt í þróunar- og prófunarsett. Þessari skiptingu var haldið í UD-útgáfu málheildarinnar, sem var gefin út á CLARIN og verður í næstu útgáfu UD.

Þróun íslenskra UD-þáttara

Tveir UD-þáttarar voru þróaðir sem hluti af vörðunum. Áður en frekari vinna gat farið fram þurfti að laga *UD Icelandic Modern* trjábankann með því að fjarlægja tvítekningar. Þær komu fyrir innan einstakra skjala auk milli þeirra, og fjöldi þeirra var umtalsverður. Núverandi útgáfu *UD Icelandic Modern* má finna hér:
https://github.com/UniversalDependencies/UD_Icelandic-Modern/tree/dev

Heildarstærð þjálfunar- og prófunargagnanna saman, *UD Icelandic Modern* og *IcePaHC*, er eftirfarandi:

dev.conllu: 148.259 orð
test.conllu: 148.387 orð
train.conllu: 768.872 orð

Fyrrnefndu þáttararnir tveir eru byggðir á transformer-líkaninu *electra-base-igc-is*, fínstillt fyrir UD-þáttun. Sá sem þjálfaður er með COMBO er á Korp, þar sem ákveðið var að nákvæmara úttak væri mikilvægara en örlítið betra LAS og UAS þess sem þjálfaður var með Diaparser. Niðurstöður þeirra eru eftirfarandi:

COMBO + electra-base-igc

Mælikvarði	Nákvæmni	Heimt	F1-gildi	AligndAcc
Tókar	99,50	99,66	99,58	
Setningar	100	100	100	
Orð	99,42	99,50	99,46	
UPOS	96,83	96,90	96,87	97,39
XPOS	93,49	93,57	93,53	94,04
UFeats	90,82	90,89	90,86	91,35
AllTags	86,60	86,66	86,63	87,10
Uppflettimyndir	95,60	95,67	95,63	96,16

UAS	89,07	89,19	89,13	89,67
LAS	85,91	86,03	85,97	86,49
CLAS	81,00	80,51	80,75	81,08
MLAS	68,60	68,18	68,39	68,67
BLEX	77,01	76,55	76,78	77,10

Diaparser + electra-base-igc

Mælikvarði	Nákvæmni	Heimt	F1-gildi	AligndAcc
Tókar	99,70	99,77	99,73	
Setningar	100	100	100	
Orð	99,62	99,61	99,61	
UPOS				
XPOS				
UFeats				
AllTags				
Uppflettimyndir				
UAS	89,53	89,51	89,52	89,87
LAS	86,32	86,31	86,32	86,65
CLAS	82,13	81,64	81,89	82,07
MLAS				
BLEX				

I8 – Merkingargreining – BERT-lík mállíkön

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að gera frekari tilraunir með margmála Transformer-líkön, ásamt gagnaukningu fyrir margmála málheildir. Stækkun einmála málheilda fyrir tungumál með litlu eða meðalmiklu magni gagna (e. low and medium-resource languages) er ekki alltaf fýsileg nálgun. Annar möguleiki er að nota margmála líkön í staðinn, en nóg er til af forþjálfunargögnum fyrir þau. Tilraunir á öðru ári hafa sýnt að stór, margmála líkön, svo sem mBert (sem er þjálfað á 104 tungumálum), standa sig ekki betur en álíka stór einmála líkön fyrir íslensku. Nýlegar rannsóknir benda til þess að fyrir tungumál með litlu eða meðalmiklu magni gagna, þá kunni að vera ákjósanlegra að framkvæma forþjálfun á jafnari margmála málheildum sem samanstanda af skyldum tungumálum.

Annar valkostur er að auka einmála málheildina með gagnaaukningaraðferðum. Slíkum aðferðum hefur verið beitt við verk á borð við vélþýðingar, með góðum árangri.

Gengið verður frá viðmiðum NLP-verka og þau gefin út opinberlega.

Varða 8 – lýsing

M8: Viðmið fyrir íslensk NLP-verk verður gefið út opinberlega.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.

- Viðmið fyrir íslensk NLP-verk (Icelandic NLP Benchmark) hefur verið gefið út á CLARIN á http://hdl.handle.net/20.500.12537/297.

Skýrsla

Viðmið fyrir íslensk NLP-verk
Fyrsta útgáfa viðmiðsins samanstendur af fjórum verkum/gagnasettum:

- Málfræðileg (POS) mörkun (MIM-GOLD 21.05)
- Nafnakennsl (NER) (MIM-GOLD-NER 2.0)
- Venslaþáttun (DP) (IcePaHC-UD, frá Universal Dependencies 2.10)
- Vélrænn útdráttur (ATS) (IceSum 22.09)

Viðmiðin innihalda skriftur til að sækja gagnasettin, skipta þeim í þjálfunar-, yfirferðar- og prófunarsett (þar sem þarf), og þjálfa og meta líkönin fyrir hvert verk. Við gefum einnig út nýja útgáfu IceSum-gagnasettsins, sem inniheldur minniháttar leiðréttingar ásamt þjálfunar- og prófunarskiptingum sem notaðar voru í greininni „Pre-training and Evaluating Transformer-based Language Models for Icelandic“[1]. Taflan að neðan sýnir árangur ýmissa Transformer-byggðra mállíkana sem voru forþjálfuð á íslensku Risamálheildinni (RMH/IGC), mældan með viðmiðunum. Við tökum með niðurstöður fyrir margmála mBERT-líkanið, sem var forþjálfað á Wikipedia-greinum á 104 tungumálum, þar á meðal íslensku. Viðmiðin mæla nákvæmni fyrir POS-verk, F_1 gildi fyrir NER, merktar tengingareinkunnir (e. labeled attachment scores, LAS) fyrir DP og ROUGE-2 heimt fyrir ATS.

Líkan	POS	NER	DP	ATS	Meðaltal
ConvBERT-Base	97,75%	94,14%	86,50%	71,09%	87,37%
ELECTRA-Base	97,72%	93,75%	86,20%	71,04%	87,18%
IceBERT-IGC	97,37%	93,04%	85,04%	69,14%	86,15%
ConvBERT-Small	96,88%	92,03%	84,75%	69,36%	85,76%
ELECTRA-Small	96,84%	91,23%	84,90%	69,29%	85,57%
mBERT	96,38%	91,31%	82,94%	69,09%	84,93%

Textasíun

Við mátum nokkra fyrirhugaða textasínuarflokkara á vefskrapaðri málheild sem samanstendur af u.þ.b. 2,3 milljörðum tóka. Málheildin fékkst með því að sameina íslenska hluta margmála málheildarinnar mC4 (https://huggingface.co/datasets/mc4) við einmála málheildirnar IC3 (https://huggingface.co/datasets/mideind/icelandic-common-crawl-corpus-IC3) og ICC (https://huggingface.co/datasets/jonfd/ICC). Til að þjálfa og meta flokkarana bjuggum við til málheild (TQ-IS) sem samanstendur af 2.000 skjölum sem voru handvirkt merkt sem annaðhvort há- eða lággæða, með 1.000 dæmi af hvorri gerð. Skjöl sem innihéldu mikið magn texta á erlendum tungumálum, ómálfræðilegt efni (t.d. Javascript- eða HTML-kóða), brenglaðan texta (t.d. vegna villna við stafakóðun) eða þjáðust af öðrum álíka vandamálum voru merkt sem lággæða.

Fyrri flokkarinn hendir skjölum með flækjugildi (e. perplexity score, PPL) undir ákveðnu marki, sem er ákvarðað út frá TQ-IS-málheildinni. Seinni flokkarinn var þjálfaður undir eftirliti á TQ-IS-málheildinni og reynir að flokka skjöl sem há- eða lággæða byggt á skjalagreypingum. Flokkararnir voru notaðir til að sía út lággæða skjöl í vefskröpuðu málheildinni. Við forþjálfuðum síðan ELECTRA-Small-líkön á síuðu málheildunum og fínstilltum fyrir tvö verk: málfræðilega mörkun (POS) og nafnakennsl (NER). Við bárum niðurstöðurnar saman við ELECTRA-Small-líkön sem voru forþjálfuð á RMH og ósíuðu vefmálheildinni.

[1] http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.804.pdf

Forþjálfunarmálheild	Fjöldi tóka	POS	NER
Íslenska Risamálheildin	1.690.847,741	96,84%	91,23%
Ósíuð vefmálheild	2.282.521.141	95,38%	90,18%
- Síuð (PPL)	1.292.744.629	96,71%	90,56%
- Síuð (undir eftirliti)	1.147.919.435	96,50%	90,51%

Mat okkar sýnir að flokkararnir bæta gæði vefskröpuðu málheildarinnar umtalsvert, sem orsakar allt að 29% lækkun villutíðni. Við höfum einnig metið flokkarana á safni 10.000 setninga sem voru vélþýddar frá ensku yfir á íslensku og merktar annaðhvort sem lág- eða hágæða þýðingar. Tilraunir okkar sýna að PPL-flokkarinn nær F_1-gildi 93,16% þegar hann er metinn á þessu gagnasetti. Niðurstöður okkar verða útskýrðar nánar í grein sem kemur brátt út, eftir að TQ-IS-gagnasettin og líkönin sem voru forþjálfuð fyrir tilraunir okkar verða gefin út.

V2b – Bakþýðingar

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Markmið

Lykilmarkmið í vélþýðingum á þriðja ári (sjá undirverkefni V4) er að styðja þýðingarlíkön fyrir vítt samhengi (á stigi margra setninga og/eða á skjalastigi). Búa þarf til nýjar bakþýðingagervimálheildir sem eru síðan notaðar sem hluti af þjálfun lokaútgáfu þýðingarlíkans á skjalastigi (e. document-level translation model).

Þetta undirverkefni felur einnig í sér hugbúnaðarvinnu til að þjappa/einfalda líkanið til að búa til þýðingar við ályktun (e. at inference time) á hraða sem endanlegir notendur geta sætt sig við, sérstaklega fyrir atvinnuþýðendur (sjá kafla um þýðingarvélar fyrir sérsvið) og fyrir notendur sem þurfa þýðingar í rauntíma á t.d. fréttum með viðbót í vafra eða sambærilegum aðferðum. Líkön á skjalastigi eru í eðli sínu stærri og hægari en líkön sem þýða aðeins setningu-fyrir-setningu, og þess vegna kalla þau á meiri hugbúnaðarvinnu fyrir hraða og samþjöppun.

Varða 8 – lýsing

M8: Bakþýðingarmálheild fyrir þýðingar á skjalastigi tilbúin til notkunar við þjálfun víðsamhengislíkans (e. long-context model) (sjá undirverkefni V4). Aðferðir við ályktanabestun valdar, byggt á tilraunum.

Varða 9 – lýsing

M9: Skjalastigslíkan (e. document-level model) þjálfað á bakþýðingum. Afkastageta líkansins bestuð og það meðtalið í afurðum undirverkefnis V4. Bakþýðingarmálheild fyrir þýðingar á skjalastigi gefin út á CLARIN.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.

- Bestunaraðferðir voru valdar og metnar og fyrstu málheildir fyrir bakþýðingar víðsamhengis valdar (sjá skýrslu að neðan).
- Upprunalega skjalastigsþýðingarlíkanið, þjálfað á bakþýðingum, er aðgengilegt á http://hdl.handle.net/20.500.12537/278.
- Bestaða líkanið er aðgengilegt á http://hdl.handle.net/20.500.12537/283.
- Bakþýðingarmálheildir hafa verið gefnar út á http://hdl.handle.net/20.500.12537/260.

Skýrsla

Sjá kafla um undirverkefni V4a fyrir ítarlegri upplýsingar um þjálfun upprunalega skjalastigslíkansins.

Bakþýðingar
Eftirfarandi einmála gögn voru þýdd til notkunar sem bakþýðingarmálheild með fyrstu ítrun skjalastigsþýðingarkerfisins okkar. Þetta kerfi notar samskeyttar þýðingar á setningastigi (e. concatenated sentence-level translations). Athugið að gögnin eru í einhverjum tilvikum höfundarréttarvarin.

Einmála gagnasett	Upprunatungumál	Tókar (milljónir)
Risamálheildin (án íþrótta) (IGC)	íslenska	1.461
The Icelandic Common Crawl Corpus (IC3)	íslenska	712
Ritgerðir nemenda	íslenska	322
Greynir News	íslenska	81
Wikipedia	íslenska	9
Íslendingasögur	íslenska	1,5
Íslenskar rafbækur	íslenska	1,9
Books3	enska	792
NewsCrawl	enska	787
Wikipedia	enska	751
EuroPARL	enska	59,7
Reykjavik Grapevine	enska	11
Iceland Review	enska	5,4

Eftir þjálfun fyrsta skjalastigsþýðingarkerfisins kom í ljós að minna magn gagna dugði til að þjálfa líkanið til meðhöndlunar víðara samhengis, svo við síuðum og prófuðum gögnin grimmt til að bæta meðalgæði. Við endurþýddum svo minni málheildina með nálægum setningum sem samhengi. Þar sem skjalaþýðingarlíkanið skapar stundum úttak með annan fjölda setninga en inntakið þurftum við að samraða þýðingunum aftur með BLEUAlign.[2] Við endurþýddum gögnin

[2] Við stefnum á að gera þessa virkni skilvirkari í næstu ítrun skjalaþýðingalíkana. Þegar sérstök merki milli inntakssetninga eru til staðar ættu þau merki að haldast í þýðingunni. Möguleiki á 1:1 samsvörun setninga er nytsamleg við gerð bakþýðinga.

einu sinni enn, og úr varð útgáfan er gefin út á CLARIN. Sjá töflu að neðan fyrir gögnin sem sú útgáfa inniheldur.

Einmála gagnasett	**Upprunatungumál**	**Tókar (milljónir)**
Risamálheildin (án íþrótta) (IGC)	íslenska	118,4
Wikipedia	íslenska	8,9
Íslendingasögur	íslenska	1,4
Íslenskar rafbækur	íslenska	1,6
NewsCrawl	enska	44,5
Wikipedia	enska	53,3
EuroPARL	enska	58,4

Við tökum fram að sum gagnasett sem notuð voru fyrir þjálfun þýðingarlíkansins má ekki dreifa vegna höfundarréttar.

Val á bestunaraðferðum
Samhliða þróun upprunalega skjalaþýðingalíkansins gerðum við tilraunir með bestunaraðferðir. Við sýnum fram á að mögulegt er að minnka líkanið um ¾ með litlum fórnum gæða.

Í tilraunum okkar við einföldun líkana einblíndum við á þýðingaráttina ensku yfir á íslensku og notuðum svo sömu aðferðir fyrir hina áttina. Við minnkuðum líkönin farsællega með litlum fórnum gæða á nokkuð almennu sviði, þ.e. fréttum. Niðurstöður tilrauna eru sýndar í töflunni að neðan. Upprunalega líkanið er 24 lög, en við gátum minnkað það í 7 lög við einföldun. Það er takmarkað tap á BLEU fyrir fréttatexta í WMT-2021-prófunargagnasettinu, frá 29,4 í 28,7. Fyrir hið sértæka svið EES-reglugerða (EEA) er tapið meira, frá 59,5 í 41,4, sem búast má við. Einnig gerðum við tilraunir með að skammta vægi líkansins í INT8 frá FP32, sem minnkar stærð minnis sem þarf til að keyra líkanið og getur flýtt ályktun á ákveðnum vélbúnaði (flestum örgjörvum og nýjustu skjákortum). Þessi skömmtun hafði engin slæm áhrif á árangur.

BLEU-gildi prófunar	Kennari	6-6 Nemandi	12-1 Nemandi	6-1 Nemandi
EEA	59,5	45,1	43,8	41,4
WMT-2021-prófunargögn	29,4	29,6	29,3	28,7
Hraði setningar/sek.				
CPU[3] B=32	1,67	3,74	4,70	4,98

[3] PyTorch 1.10, i7-7700 CPU sem notar 4 þræði. „B“ er skammtastærðin. Stór skammtastærð líkir eftir stóru ónettengdu þýðingarverki, á meðan skammtastærðin 1 líkir eftir litlu nettengdu þýðingarverki.

CPU B=1	0,33	0,67	0,87	0,88
CPU ONNX B=32	0,31	1,34	4,07	4,35
CPU Quantized B=32	3,46	6,59	7,21	9,83
GPU[4]	24,79	46,67	67,34	73,13

Bestun endanlegra líkana

Til að besta endanlegt líkan frá undirverkefni V4 notum við 6-1 nemendaaðferðina (e. Student method). Að auki takmörkum við inntak líkansins við hámark 256 inntakstóka og gróflega 3 setningar í senn vegna lakari afkasta fyrir lengri inntök. Líkanið er einfaldað frá 24-laga Transformer-líkani í 7 lög og er um þrisvar sinnum hraðara og skilar samsvarandi árangri og í tilraunum okkar. Það er að segja það er hærra fall á sértæku EEA-sviði, en fyrir svið frétta (WMT) og Wikipedia (Flores) er árangur líkansins sambærilegur kennara (e. teacher). Sjá eftirfarandi töflu fyrir beinan samanburð.

BLEU-gildi prófunar EN-IS	Kennari	6-1 Nemandi
EEA	60,3	39,7
WMT-2021 test	31,2	30,4
Flores devtest	27,1	28,2

BLEU-gildi prófunar IS-EN	Kennari	6-1 Nemandi
EEA	65,2	48,0
WMT-2021 test	35,0	33,2
Flores devtest	35,4	33,3

[4] PyTorch 1.10, Nvidia GeForce 1080 skjákort. Skammtastærð 64 nýtir skjákortið að fullu.

V4a – Opin þýðingarvél fyrir íslensku

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Lýsing

Markmið þriðja árs fela meðal annars í sér að pakka og gefa út á GitHub og CLARIN þýðingarlíkan sem nota má áreiðanlega og án sérþekkingar umfram almenna forritunarfærni. Það er, þekking á tauganetum og gagnaúrvinnslupípum ætti ekki að vera skilyrði til að taka í notkun framleiðslukerfi með notkun þýðingarlíkans.

Til að kerfið verði gagnlegt, afkastamikið og samkeppnishæft að minnsta kosti næstu árin, höfum við stigið skref í átt að þýðingum sem spanna fleiri en eina setningu í einu og viðhalda þar með samhengi og bæta samheldni.

Að lokum verður tekið á þoli með því að skilgreina mælikvarða, grunnlínur og markmið fyrir afköst líkans á ófullkomnum heimildartexta (þ.e. textum með stafsetningar- eða málfræðivillum; textum skrifuðum af börnum, lesblindu fólki og íslenskumælandi fólki sem hefur annað móðurmál) og stilla greypingaraðferðir (e. embedding strategies) til að bæta mælanlegan árangur í slíkum dæmum.

Varða 8 – lýsing

M8: Þolaðferðir innleiddar. Fyrsta ítrun þýðingarkerfis á skjalastigi, með þolþáttum (e. robustness feature), er fær um að þýða texta.

Varða 9 – lýsing

M9: Þýðingarkerfi á skjalastigi, með auknu þoli, er fullgert, pakkað og skjalað fyrir almenna forritara, og útgefið á GitHub og CLARIN.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Þýðingarkerfi á skjalastigi (víðsamhengi) hefur verið þróað.

- Kerfið er aðgengilegt á CLARIN á http://hdl.handle.net/20.500.12537/278 (einnig með einfölduðu líkani, sjá V2b).
- Kerfið er aðgengilegt í gegnum pakkann *greynirseq* á PyPi (Python Package Index, pypi.org).

- Líkönin eru í gangi á vefsíðu sem opin er almenningi, https://velthyding.is.
- Það er aðgengilegt á GitHub á https://github.com/mideind/greynirseq.
- Líkönin eru þolin gagnvart stafsetningarvillum og suði í inntaki.
- Líkönin þýða hverja setningu með tengingu við margar fyrri setningar, sem gefur samhengi milli setninga.

Skýrsla

Líkanið var þjálfað með bakþýðingum. Sjá undirverkefni V2b fyrir nánari lýsingu á gögnum sem notuð voru. Við bættum þol líkansins gagnvart inntaksvillum með því að bæta suði við þjálfun.

Líkanið getur tekið við inntaki af lengd ~800 tóka, og úttaki af lengd ~800 tóka, og meðhöndlar margar setningar vel.

Val víðsamhengisgagna
Auk þeirra gagna sem nefnd voru í verkþætti V2b söfnuðum við samhliða skjölum með hágæða samröðun setninga. Fyrir það notuðum við eftirfarandi samhliða málheildir:

- Reglugerðir Evrópska efnahagssvæðisins (EES/EEA), skjalastig
- IPAC
- Stúdentablaðið
- Rannsóknarnefnd Alþingis
- Biblíuna
- Vottar Jehóva (JW300)

Þýðingar víðsamhengis
Við þjálfuðum tiltækt setningastigsþýðingarkerfi (þróað fyrir sama verkþátt í vörðu 6) á gögnum sem samanstóðu af samskeytingu margra[5] setninga í röð af heildarlengd upp í ~800 tóka.

Ýmis málfræðileg fyrirbæri birtast aðeins þegar víðara samhengi en setningastigið er skoðað. Meðal þessara fyrirbæra eru orðasamhengi með endurteknum orðum, einræðing með samhengi innan setningar og anafórum innan setningar. Að neðan eru nokkur dæmi um þessi fyrirbæri fyrir íslensku, með raunverulegum þýðingum úr kerfum okkar.

Fyrirbæri	Upprunatexti	Setningastig	Skjalastig	Skýring

[5] Þjálfunardæmin samanstóðu af jafnri blöndu 1, 2, 3, o.s.rfv., upp í 10 setningar í röð.

Einræðing	I'm fed up of bad candidates in these elections. Well, why don't you ***run***?	Ég hef fengið mig fullsadda af slæmum frambjóðendum í þessum kosningum. Af hverju ***hleypur*** þú ekki?	Ég er orðinn hundleiður á slæmum frambjóðendum í þessum kosningum. Af hverju ***gefurðu*** ekki *kost á þér*?	***Hleypur*** er einrætt á rangan hátt. Setningalíkanið þýðir það sem skylt hugtökunum {*hlaupa, spretta, þjóta*}
Einræðing	The flour's down, the dough's rolled out. Pass me the ***cutters***.	Mjölið er komið niður, deiginu er rúllað út. Réttu mér ***klippurnar***.	Mjölið er komið niður, deiginu rúllað út. Réttu mér ***skerin***.	***Cutter*** er einrætt á rangan hátt. Setningalíkanið þýðir það sem skylt hugtökunum {*klippur, skæri*}. Skjalalíkanið einræðir rétt en notar rangt kyn (***skerin*** vs ***skerann***)
Orðasamhengi	What's ***crazy*** about me? Is this ***crazy***?	Hvað er ***vitlaust*** við mig? Er þetta ***brjálæði***?	Hvað er ***brjálað*** við mig? Er þetta ***brjálað***?	Orðasafnsþýðing ***crazy*** er ekki samræmd í setningalíkaninu, en rétt í skjalalíkaninu.
Anafóra	Should I wash the plate? No, just bring ***it*** here, please.	Á ég að þvo diskinn? Nei, komdu bara með ***það*** hingað.	Á ég að þvo diskinn? Nei, komdu bara með ***hann*** hingað, gerðu það.	Vísað er ranglega í ***diskinn*** í hvorugkyni í stað karlkyns. Skjalalíkanið er rétt.
Anafóra	I told you to avoid any problems. Well, sometimes ***they*** find me.	Ég sagði þér að forðast vandamál. Jæja, ***þeir*** finna mig stundum.	Ég sagði ykkur að forðast öll vandamál. Stundum finna ***þau*** mig.	***Vandamál*** (e. problem) er ranglega vísað til í karlkyni í stað hvorugkyns. Skjalalíkanið er rétt.

Þol

Við innleiddum nokkrar þolaðferðir við þjálfun og mældum áhrif þeirra á BLEU-gildi með inntaksvillum. Til að auka þol bætum við suði við dæmin við þjálfun. Suðaðferðirnar sem við notuðum eru óháðar tungumáli, til að geta notað sömu útfærslu fyrir allar þýðingaráttir. Það sést að neðan að það að bæta við suði ákveðins tungumáls hefur ekki marktæk áhrif á þol.

Suðaðferðir okkar virka á textann eftir að hann hefur verið kóðaður á viðeigandi form til inntaks í líkanið. Þetta gefur möguleika á nokkuð einfaldri og skilvirkri útfærslu. Í eftirfarandi texta merkir „orðflís“ stafaklasi sem verður til við skiptingu orða, sérstaklega langra orða, í nokkra hluta til að gera þau tilbúin fyrir líkanakóðun. Aðferðirnar eru:

- Endurröðun á orða- og orðflísastigi: Skiptir stöku sinnum á samhliða orðum eða orðflísum. Þetta tekur á mörgum algengum innsláttarvillum þar sem orð og stafir skiptast á stöðum.
- Fjarlæging á orða- og orðflísastigi: Fjarlægir stöku sinnum orð eða orðflísar. Þetta tekur á algengum villum þar sem stafir og orð gleymast.
- Suðun orðflísakóðunar: Margar leiðir eru til þess að skipta orði í orðflísar. Viðbót suðs í kóðun orðflísa þýðir að nota ekki alltaf þá augljósustu. Þetta eykur þol við stafsetningarvillum með því að tryggja að líkanið reiði sig ekki um of á eina staka leið til að skipta orðum í orðflísar.
- Handahófskennd innskot á orðflísastigi: Handahófskenndum orðflísum, þ.e. stafaklösum, er bætt inn í textann. Þetta tekur á algengum innsláttarvillum þar sem aukastöfum er bætt við, og tryggir jafnframt að líkanið sjái alla mögulega stafi við þjálfun og læri að meðhöndla þá.

Áhrif þess að sjá alla mögulega stafi er vert að útskýra nánar. Líkön sem þjálfuð voru fyrir M6 tóku ekki á óþekktum stöfum á neinn sérstakan hátt. Þetta gat valdið því að þau voru alls óviss þegar þau þurftu að vinna með t.d. tjákn og gríska stafi.

Gamla líkanið brotnar stundum alveg þegar óvenjulegt inntak er gefið. Þó að það nýja geti ekki (skiljanlega) þýtt óvenjulegt inntak rétt nær það sér í ásættanlega stöðu á eftir.

Upprunatexti	**Gamla líkanið**	**Þolið líkan**	**Skýring**
engineer belives tha behind The model's immpresive skils might als0 lie a sentient mind	Verkfræðingur telur að þ á bak við rýrar skilgreiningar líkansins gæti als0 legið skynsamleg hugsun	Verkfræðingurinn telur að bak við líkanið geti órökstudd skilaboð líkansins legið í tilfinningaríkri hugsun.	Textinn er málfræðilega rangur án þols
I'm going to be a 🕵️ for Halloween.	Ég verð ö ö ö ö ö ö ö ö [repeats to max length]	Ég ætla að verða hljómsveitarstjóri á hrekkjavökunni.	Óséð tjákn
A small town in Greece called Καλαμπάκα is a popular tourist destination.	Lítill bær í Grikklandi sem kallast Godt 9 , 6 , 6 , 1 , 1 , 1 , 1 , 1 , [repeats to max length]	Lítill bær í Grikklandi sem kallast 9α–6 er vinsæll ferðamannastaður.	Gríska stafir, koma oft fyrir í Europarl og ensku Wikipedia

Áhrif suðs á BLEU-gildi

Til að meta áhrif inntaksvillna á gæði þýðingarlíkansins okkar endurnýttum við suðmyndunaraðferðir úr verkþætti L14 sem voru notaðar til myndunar gervivillugagna.[6] Nokkur dæmi myndaðra villna eru:

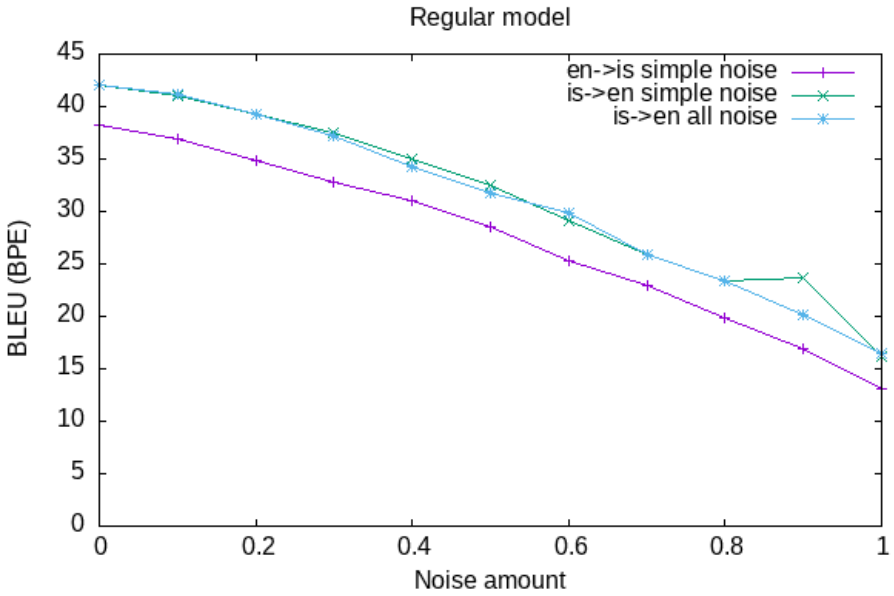
- Bæta við eða fjarlægja kommu, t.d. café → cafe

[6] Nákvæmar aðferðir suðs sem notaðar voru má sjá í kóðanum hérna https://github.com/mideind/GreynirSeq/blob/main/src/greynirseq/noising/ieg/spelling/rules.py og hér https://github.com/mideind/GreynirSeq/tree/main/src/greynirseq/noising/ieg/spelling/rules

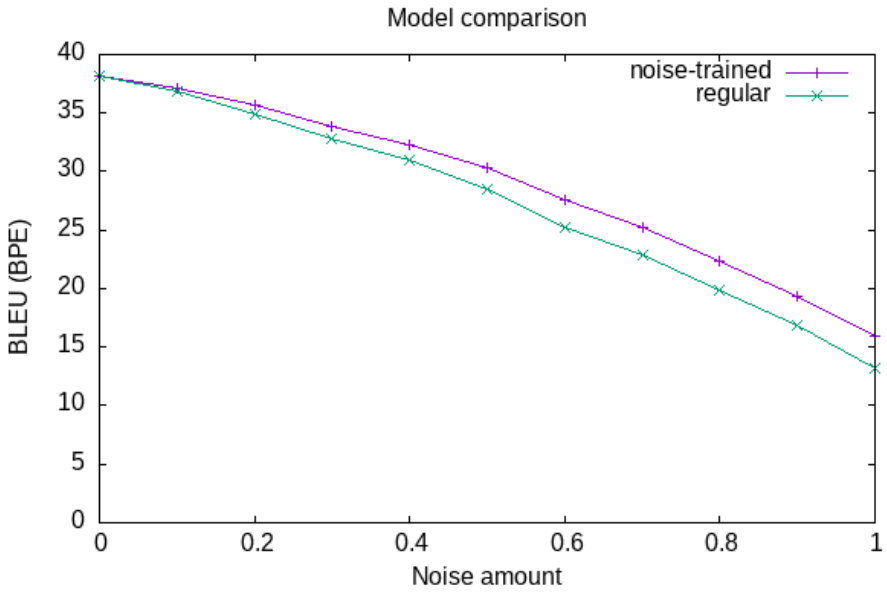
- Handahófskennd útskipting stafa, t.d. mouth → mouyh
- Handahófskennd viðbót sama stafs, t.d. food → foood
- Handahófskennt brottfall stafs, t.d. mouse → mose
- Handahófskennd skipting stafa, t.d. mouse → muose
- Algengar hljóðvillur, t.d. ll → gl, ung → úng
- Algengar villur þar sem íslenskir stafir eru ekki notaðir, þ.e. ð → d, æ → ae

Að neðan er línurit sem sýnir hvernig árangur upprunalega líkansins okkar fer lækkandi eftir því sem auknu suði er bætt við upprunahlið prófunarsettsins. Við sýnum báðar þýðingaráttir og tvær mismunandi suðaðferðir fyrir íslensku hliðina. Aðferðin *simple-noise* notast eingöngu við stafa- og orðasuðun óháð tungumáli, á meðan aðferðin *all-noise* bætir inn séríslensku suði.

Við sjáum að það er ekki greinilegur munur á árangri milli inntaka *simple-* og *all-noise*. Af þessu ályktum við að við þjálfun nægi notkun á suði óháð tungumáli.



Á næsta línuriti sjáum við að líkan þjálfað á gögnum með suði ná nokkrum BLEU-stigum betri árangri í návist suðs, jafnvel þó að árangurinn minnki.



Hér er dæmi um setningu með mismiklu suði:

is	Vísindamenn segja að sprengingin af völdum árekstursins hafi verið gríðarmikil.
en	Scientists say the explosion caused by the collision was massive.
noise-0.5	Scientists say tha explotion caused ba coolision theewas massove.
noise-1.0	SclentIsts syythhhe egsplysiyn caesed the bi sollision vas m8assive.

Árangur

Að neðan er ýmis samanburður á árangri líkansins á valin gagnasett með mismunandi stillingum. Fyrst berum við saman líkanið sem er þolið gagnvart suði (e. *noise-robust context model*) (Context) og fyrra líkanið sem var þróað fyrir vörðu M6 (No-context) á þýðingum einstakra setninga, þ.e. án nokkurs samhengis. Þetta gerir okkur kleift að meta hvort árangur líkansins hafi minnkað fyrir ákveðin gagnasett. Fyrir þennan samanburð gefum við upp BLEU-mælingar. Gildi er reiknað fyrir hverja setningu og svo er meðaltal reiknað yfir matsmálheildina.

Líkan	BLEU fyrir setningar				
	EEA[7]	os2018[8]	WMT 2021 dev	WMT 2021 test	Flores devtest
No-context EN-IS	59,5	30,5	29,1	29,4	27,6
Context EN-IS	60,3	25,7	30,3	31,2	27,1
No-context IS-EN	65,6	30,9	31,8	33,9	34,7
Context IS-EN	65,2	31,3	32,8	35,0	35,4

Við sjáum að árangur samhengislíkansins minnkar almennt ekki fyrir þýðingar á setningum, heldur eykst hann. Eina augljósa minnkunin er fyrir *OpenSubtitles* í áttinni EN-IS. Þetta kemur líklega til vegna þess að þetta svið samanstendur að mestu af tali í fyrstu persónu, sem birtist ekki í þjálfun fyrir þessa átt.

Nú sýnum við að samhengislíkanið getur þýtt margar setningar í einu og notum líkanið til að þýða mismunandi fjölda setninga í einu. Matsgagnasettin að ofan innihalda ekkert samhengi, að undanskildu gagnasettinu *WMT-2021*. Gagnasettið *WMT-2021* samanstendur af mörgum heilum fréttagreinum. Til að taka á því að þegar við þýðum margar setningar í einu er möguleiki að fá ekki sama fjölda setninga út þurfum við að reikna BLEU-gildin á annan máta. Í stað þess að reikna BLEU-gildi á hverja setningu reiknum við það yfir heila grein í senn, og tökum svo meðaltal yfir allar greinarnar. Við tökum einnig fram *BERTScore* sem reiknar líkindisgildi fyrir hvern tóka í mögulegri setningu miðað við hvern tóka í viðmiðssetningu. Það beitir forþjálfuðum samhengisháðum orðagreypingum úr textakóðara.[9] Við tökum þessar mælingar einnig fram, þar sem sýnt hefur verið fram á að þær samræmast mannlegu mati vel, og reiða sig ekki á orðaröð og meðhöndla samheiti. Að neðan skýrum við frá BLEU- og BERTScore-gildum með mismunandi fjölda inntakssetninga á WMT-2021-þróunarsettið.

Átt EN-IS	Fjöldi setninga inn	BLEU	BERTScore
No-context / Context	1	31,2 / 32,3	0,945 / 0,950
No-context / Context	2	29,7 / 32,7	0,946 / 0,950
No-context / Context	3	28,4 / 32,8	0,943 / 0,951
No-context / Context	7	16,5 / 32,5	0,920 / 0,950
Átt IS-EN	**Fjöldi setninga inn**	**BLEU**	**BERTScore**

[7] Innanhúss prófunarsett EES-reglugerða (EEA regulations).
[8] Innanhúss prófunarsett OpenSubtitles.
[9] IceBERT, 11. lag, fyrstu 512 tókar, án idf

No-context / Context	1	33,5 / 35,0	0,942 / 0,943
No-context / Context	2	33,2 / 34,8	0,943 / 0,942
No-context / Context	3	32,7 / 34,6	0,939 / 0,942
No-context / Context	7	16,9 / 32,1	0,914 / 0,942

Þessar niðurstöður gefa til kynna að líkanið geti stöðuglega þýtt margar setningar í einu án verulegra fórna, á meðan árangur líkansins No-context minnkar hratt.

V4b – Opin þýðingarvél fyrir íslensku – Þýðingarvél fyrir sérsvið

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Lýsing

Á þriðja ári einbeitum við okkur að umbótum á kerfinu með þýðingar á reglugerðum Evrópska Efnahagssvæðisins (EES) fyrir utanríkisráðuneyti Íslands í huga. Þetta samstarfsverkefni er eins konar leiðarljós til að læra að aðlaga og innleiða vélþýðingarkerfið fyrir raunverulega, faglega notkun. Það verður sniðmát á hvernig skal pakka fínstilltu pípunni til að nota á öðrum sviðum, og fínstilla „neðanstreymis“ (e. downstream) fyrir ákveðin aðgengileg notkunarsvið með lágmarks fyrirhöfn.

Umbætur á vélþýðingarkerfinu fela í sér að safna yfirförnum (e. post-edited) þýðingum frá þýðendum utanríkisráðuneytisins til frekari fínstillingar líkansins, og svo það styðji sérhæfðar þýðingar fyrir undirsvið innan EES-reglugerðasviða (fjármála-, umhverfis- o.s.frv.), þar sem hugtök kunna að hafa staðlaða, undir-sviðsbundna (e. sub-domain specific) þýðingu sem líkanið verður að læra.

Varða 8 – lýsing

M8: Fínstillingarpípu pakkað og hún gefin út.

Varða 9 – lýsing

M9: Bætt þýðingarkerfi tekið í notkun í utanríkisráðuneytinu með yfirförnum þýðingum.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Helstu niðurstöður eru:

- Fínstillingarpípa fyrir sviðsmerkingar (e. domain tags) í þýðingum hefur verið gefin út og gerð aðgengileg á CLARIN á http://hdl.handle.net/20.500.12537/212.
- Nýja kerfið er fínstillt með nýjum þýðingum úr utanríkisráðuneytinu, með yfirförnum þýðingum frá þeim notendum sem tóku þátt í vinnu fyrsta árs.
- Nýja kerfið er þjálfað til að taka við sviðsstillingu. Þýðandi getur valið t.d. *landbúnað* eða *fjármál* sem svið reglugerðarinnar sem þýdd er.

- Öllum þýðendum hjá utanríkisráðuneytinu var boðinn aðgangur að kerfinu og við sáum mikið hærri notkun en á fyrra ári.
- Þýðendur eru sammála um að kerfið hraði vinnu þeirra og bæti vinnuferli þeirra.

Skýrsla

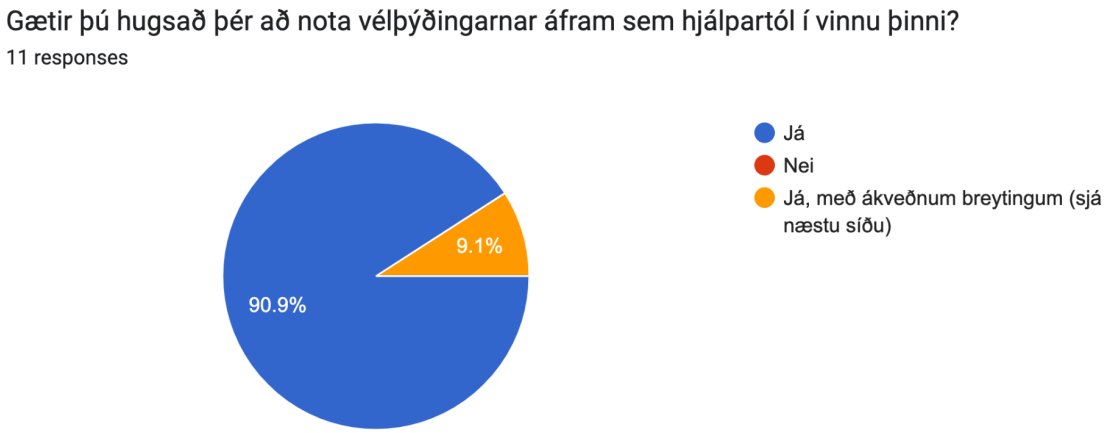
Þýðendur utanríkisráðuneytisins nota *Trados Studio* til að þýða evrópskar reglugerðir af ensku yfir á íslensku. Til að samþætta þýðingarkerfi okkar við vinnuferli þeirra var Trados-viðbót þróuð á öðru ári og bætt frekar á þriðja ári til að styðja val þýðingarsviðs. Við vísum í skýrslu M7 fyrir lista þeirra sviða og frekari lýsingu á uppsetningu.

Notendaprófanir á öðru ári voru takmarkaðar við nokkrar vikur og valda þýðendur hjá utanríkisráðuneytinu, en prófunartíminn var í þetta sinn mikið lengri, frá maí til september 2022, og prófanir voru opnar öllum þýðendum hjá utanríkisráðuneytinu sem höfðu áhuga. Á þessu ári höfum við séð um 13.000 umbeðnar vélþýðingar beint frá starfsfólki utanríkisráðuneytisins. Til samanburðar voru 6.500 slíkar þýðingar gerðar á síðasta ári í magnþýðingu sem notuð var til að búa til nýtt þýðingarminni með lægri forgang en það sem byggt er á fyrri þýðingum fólks.

Til að meta árangur kerfisins sendum við út spurningalista til þeirra þýðenda sem tóku þátt í prófunum fyrir viðbót Trados Studio. 11 svör bárust. Greint er frá niðurstöðum þessa spurningalista að neðan. Fimm þýðendanna sögðust hafa notað vélþýðingarviðbótina í meira en helming vinnu sinnar meðan á notendaprófunum stóð, á meðan hinir sex sögðust hafa notað hana af og til.

Ánægja í heildina
Heilt yfir sögðu 10 af 11 þýðendum að gæði vélþýðinganna væru betri en þeir hefðu búist við, og einn sagði gæðin hafa verið eftir væntingum.

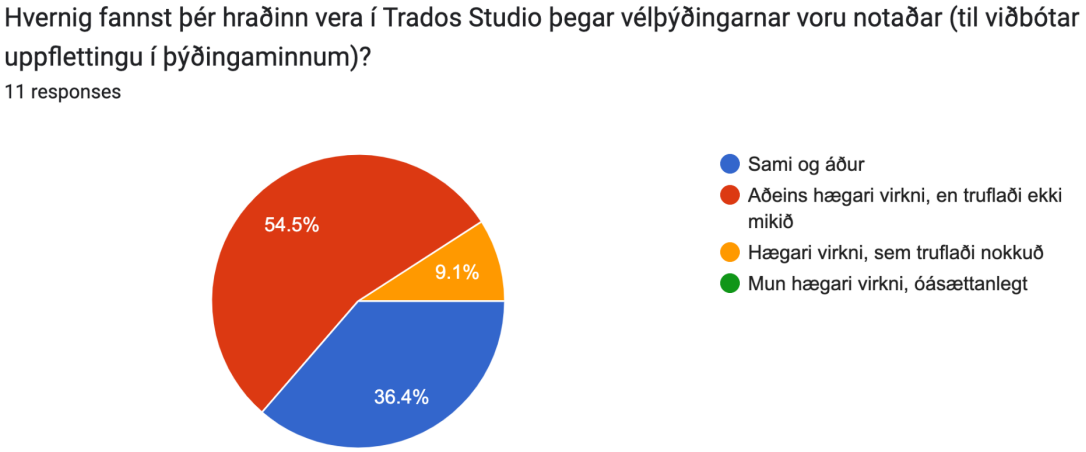
Spurðir hversu sáttir þýðendurnir væru með þá virkni að fá vélþýðingu þegar þýðingarminnin skila ekki niðurstöðu gáfu þeir meðaleinkunnina 4 af 5. Spurðir hversu jákvæðir eða neikvæðir þeir væru gagnvart vélþýðingum eftir tímabil notendaprófana gáfu þeir einnig einkunnina 4 af 5. Þýðendurnir voru á einu máli um að þeir hefðu áhuga á því að halda áfram notkun vélþýðinga sem auka hjálpartól við vinnu sína (Mynd 1).



Mynd 1. Gætir þú hugsað þér að nota vélþýðingarnar áfram sem hjálpartól í vinnu þinni? (1. Já, 2. Nei, 3. Já, með ákveðnum breytingum (sjá næstu síðu))

Hraði

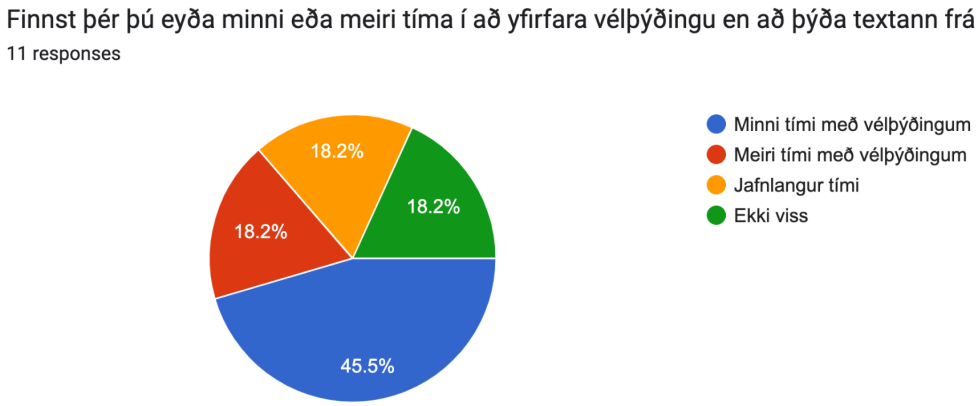
Leitað er í nokkrum stórum þýðingarminnum við þýðingu í Trados Studio-umhverfinu, og nokkurn tíma tekur að búa til tauganetsþýðingar svo ráðstafanir voru gerðar (notkun skjákorta og flýtiminnis) til að tryggja hraða svo að hver þýðingabútur yrði ekki vandamál. Allir þátttakendur nema einn voru sammála um að hraði Trados Studio með vélþýðingarviðbótinni hægði ekki á vinnu þeirra í áberandi mæli, á meðan einn sagði ferlið truflandi að einhverju leyti (Mynd. 2).



Mynd 2. Hvernig fannst þér hraðinn vera í Trados Studio þegar vélþýðingarnar voru notaðar (til viðbótar við uppflettingu í þýðingaminnum)? (1. Sami og áður, 2. Aðeins hægari virkni, en truflaði ekki mikið, 3. Hægari virkni, sem truflaði nokkuð, 4. Mun hægari virkni, óásættanlegt)

Þýðendurnir voru spurðir hvort þeir teldu ferlið við að laga þýðingu eftir vélþýðingu vera fljótlegra eða seinlegra en þýðing frá grunni. Fimm sögðu það spara tíma, tveir að það tæki sama tíma,

tveir að það tæki lengri tíma, og tveir voru óvissir (Mynd. 3). Síðar í spurningalistanum voru þeir spurðir hvort notkun vélþýðinga í vinnu þeirra hefði kosti í heildina, og flestir nefndu aukinn hraða.



Mynd 3. Finnst þér þú eyða minni eða meiri tíma í að yfirfara vélþýðingu en að þýða textann frá grunni? (1. Minni tími með vélþýðingum, 2. Meiri tími með vélþýðingum, 3. Jafnlangur tími, 4. Ekki viss)

Íðorð

Þýðendur utanríkisráðuneytisins þurfa að fara eftir ákveðnum lista samþykktra íðorða við þýðingar sínar. Eiginleika fyrir svið var bætt við í núverandi þýðingarlíkan til að koma til móts við athugasemdir úr notendaprófunum síðasta árs, þar sem þýðendur tóku fram að ólík svið þörfnuðust mismunandi íðorða (til dæmis getur hugtakið *performance* haft minnst sjö mismunandi þýðingar í texta reglugerða, eftir sviði).

Muninn á þýðingum íðorða eftir sviði getur verið erfitt að bera kennsl á, svo þýðendurnir gátu ekki staðfest ótvírætt hvort þessi eiginleiki virki. Þó voru svör þeirra varðandi hversu vel kerfið meðhöndlaði íðorð mun jákvæðari en á öðru ári, og vandamál íðorða sem þeir lentu í virtust að mestu hafa orsakast af öðru. Við tökum þessu sem merki um að virkni þýðingasviðs gefi rétt íðorð samkvæmt völdu sviði.

Notkun orðalista er ennþá algengasta kvörtunin þegar litið er á heildina, sem er skiljanlegt þar sem strangar kröfur eru um notkun íðorða í textum reglugerða, og stöðugt er bætt við nýjum íðorðum. Þýðingarkerfið skilar ekki endilega sömu þýðingu fyrir ákveðið hugtak í hvert skipti, svo þýðendur báðu um eiginleika til að geta beðið þýðingarvélina um að nota sama hugtakið yfir allt skjalið, sem við viljum gjarnan geta innleitt.

Virkni viðbótar

Þýðendurnir stungu upp á nokkrum endurbótum fyrir viðbótina sjálfa, eins og að geta virkjað eða slökkt á vélþýðingunum fyrir ákveðna hluta þýðinga, eða að geta betur skipt milli sviðsmerkinga. Þó að Trados Studio takmarki að einhverju leyti stillingarmöguleika viðbóta þriðju aðila munum við skoða hvernig við getum stillt viðbótina frekar með þetta að leiðarljósi.

Kostir og gallar
Að lokum voru þýðendurnir spurðir um helstu kosti og galla þess að nota vélþýðingar í vinnu sinni. Heilt yfir nefndu þeir þörfina á ítarlegri yfirferð og hættuna á því að yfirsjást villu í þýðingunni, þar sem augljóst samræmi textans getur verið blekkjandi. Sumar setningar hentuðu vélþýðingu síður, sem lækkar gæðin og getur krafist þýðingar frá grunni. Þeir nefndu að þetta ferli yfirferðar eftir á krefðist aðeins öðruvísi nálgunar en hefðbundið ferli þýðingar+yfirferðar, sem tæki tíma til að venjast.

Eins og nefnt var að ofan var kosturinn sem oftast var nefndur aukinn þýðingarhraði, sérstaklega við þýðingu texta með engum samsvörunum í þýðingarminnum. Vegna eðlis þýðingarvinnu er erfitt að mæla þýðingartíma nákvæmlega og bera saman tíma þess að þýða með og án vélþýðinga. Þetta er ástæða þess að langtímamat líkt og þetta er nauðsynlegt, til að þýðendurnir og utanríkisráðuneytið fái tilfinningu fyrir því hvernig vélþýðingar hjálpa þeim til langs tíma. Frekara mat myndi hjálpa þeim að öðlast betri skilning á hversu mikinn tíma það sparar þeim.

Þýðendurnir tóku fram að þýðingarnar væru hnökralausar og sýndu oft sniðugar lausnir og nýjar hugmyndir. Einn ófyrirséður kostur sem þeir tóku fram var það hvernig vélþýðing þýðir alla hluta bútar, sem minnkar líkur á því að eitthvað gleymist úr langri og flókinni setningu, sem er algeng villa í mennskri þýðingu. Svo jafnvel þó að vélþýðingar þurfi að fara yfir eftir þýðingu veit þýðandinn að allir hlutar upprunatexta eru til staðar.

Að lokum var endurgjöfin frá þýðendum utanríkisráðuneytisins mjög nytsamleg og gríðarlega jákvæð gagnvart notkun vélþýðinga Miðeindar, og báðir aðilar hafa lýst yfir von um að halda þessu samstarfi áfram í einhverju formi.

V4g – Opin þýðingarvél fyrir íslensku – Orðasambönd og þýðingar

Skýrsla: Stofnun Árna Magnússonar í íslenskum fræðum
Aðilar: Stofnun Árna Magnússonar í íslenskum fræðum

Markmið

Markmiðið er að safna margra orða segðum, eða orðasamböndum, svo að hægt sé að merkja þau í upprunalegum textum, hugsanlega til að beita umorðun fyrir þýðingu. Orðasambönd geta verið föst eða með breytilegum liðum, svo upplýsingar um gerð þarf til að gefa bestu upplýsingar. Áherslan verður á margra orða segðir með breytilegum liðum, þar sem bókstafleg merking er ekki ætluð merking og er því ekki hægt að þýða orð fyrir orð.

Varða 8 – lýsing

M8: Margra orða segðum með ólíkar bókstaflegar merkingar hefur verið safnað og forgangsraðað.

Varða 9 – lýsing

M9: Færslur fyrir safnaðar margra orða segðir tilbúnar. Gögn gefin út á CLARIN.

Afurðir

Markmiðum M8 og M9 var náð, og færslur fyrir 1.000 orðasambönd hafa verið kláraðar og gefnar út á CLARIN (http://hdl.handle.net/20.500.12537/275).

Skýrsla

Listi yfir 4.000 orðasambönd var fenginn frá ISLEX (https://islex.arnastofnun.is/). Þessi listi var yfirfarinn og unninn, og að lokum voru 1.000 orðasambönd, með samsvarandi þýðingum, gefin út á CLARIN.

Orðasamböndin voru unnin með eftirfarandi sniðmáti:

Upprunalegt orðasamband	<NP1-nom> horfa aðdáunaraugum á <NP2-acc>
Orðasambandsmarkmið	<NP> gaze at <NP2> in admiration

Upprunaleg merking	<NP1-nom> horfa á <NP2-acc> með aðdáun
Merkingarmarkmið	<NP1> look at <NP2> admiringly
Dæmi upprunalegs orðasambands	Sigurður horfði aðdáunaraugum á Guðmund
Dæmi markmiðs orðasambands	Sigurður gazed at Guðmundur in admiration
Dæmi upprunalegrar merkingar	Sigurður horfði á Guðmund með aðdáun
Dæmi merkingarmarkmiðs	Sigurður looked admiringly at Guðmundur
Lykilorð	aðdáunarauga

Skjáskotin að neðan sýna hvernig orðasamböndin voru unnin:

Source idiom	Target idiom	Source sense	Target sense
<NP1-nom> horfa aðdáunaraugum á <NP2-acc>	<NP> gaze at <NP2> in admiration	<NP1-nom> horfa á <NP2-acc> með aðdáun	<NP1> look at <NP2> admiringly
<NP-nom> vera afrendur að afli	<NP> be strong as an ox	<NP-nom> vera geysilega sterkur	<NP> be extremely strong
<NP1-nom> þekkja <NP2-acc> af afspurn	<NP1> know of <NP2>	<NP1-nom> þekkja <NP2-acc> af umsögn anna	<NP1> know <NP2> through reports from other
<NP-nom> rísa upp á afturfæturna	<NP> put up a fight|<NP> make a stink	<NP-nom> láta í ljós andstöðu	<NP> express one's opposition
<NP-nom> vera úti að aka	<NP> be asleep at the wheel|<NP> be out to lun	<NP-nom> vera utan við sig	<NP> be absent-minded
<NP-nom> stytta sér aldur	<NP> take one's own life	<NP-nom> fremja sjálfsmorð|<NP-nom> svipta	<NP> commit suicide
<NP-nom> vera í algleymingi	<NP> be in full swing	<NP-nom> standa sem hæst	<NP> be at its most exciting point
<NP-nom> vera á almannavitorði	<NP> be public knowledge	<NP-nom> vera alkunnugur	<NP> be common knowledge
<NP-nom> koma fyrir almenningssjónir	<NP> be publicly displayed|<NP> be open to th	almenningur getur séð <NP-acc>	the public will be able to view <NP>

Source idiomatic example	Target idiomatic example	Source sense example	Target sense example	Keywords
Sigurður horfði aðdáunaraugum á Guðmund	Sigurður gazed at Guðmundur in admiration	Sigurður horfði á Guðmund með aðdáun	Sigurður looked admiringly at Guðmundur	aðdáunarauga
Sigurður var afrendur að afli	Sigurður was strong as an ox	Sigurður er geysilega sterkur	Sigurður was extremely strong	afl
Sigurður þekkti Guðmund af afspurn	Sigurður knew of Guðmundur	Sigurður þekkti Guðmund af umsögn annarra	Sigurður knew Guðmundur through reports from	afspurn
Sigurður reis upp á afturfæturna	Sigurður put up a fight|Sigurður made a stink	Sigurður lét í ljós andstöðu	Sigurður expressed his opposition	afturfótur
Sigurður var úti að aka	Sigurður was asleep at the wheel|Sigurður was o	Sigurður var utan við sig	Sigurður was absent-minded	aka
Sigurður stytti sér aldur	Sigurður took his own life	Sigurður framdi sjálfsmorð|Sigurður svipti sig lí	Sigurður committed suicide	aldur
gleðskapurinn var í algleymingi	the party was in full swing	gleðskapurinn stóð sem hæst	the party was at its most exciting point	algleymingur
fjárskorturinn er á almannavitorði	the deficit is public knowledge	fjárskorturinn er alkunnugur	the deficit is common knowledge	almannavitorð
listaverkið kemur fyrir almenningssjónir	the artwork will be publicly displayed|the artwo	almenningur getur séð listaverkið	the public will be able to view the artwork	almenningssjónir

V5 – Vefviðmót fyrir vélþýðingar

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Lýsing

Markmiðið á þriðja ári er að tryggja langlífi vélþýðingarlausnanna, umfram máltækniáætlunina, með því að skjala og pakka biðlara fyrir þýðingarlíkanið fyrir beina notkun hönnuða og rannsakenda í gegnum skipanalínuviðmót og sem Python-einingu. Á sama tíma halda áfram að styðja veflægt þýðingarkerfi (viðmiðunarútfærsla sem er opinber á velthyding.is) með áframhaldandi umbótum, sem tól fyrir almenning til að nota við styttri þýðingarverkefni og sem sýnidæmi um getu kerfisins.

Varða 8 – lýsing

M8: Viðvarandi uppfærslur á veflæga viðmótinu, með viðbótum úr öðrum undirverkefnum.

Varða 9 – lýsing

M9: Þýðingarpakki (skipanalínutól og Python-þýðingarbiðlaraeining) fyrir íslensku – ensku (báðar áttir) aðgengilegur á Python Package Index (PyPI) til einfaldrar uppsetningar með pip, sem og á CLARIN.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.

- Viðmótinu á https://velthyding.is (https://github.com/mideind/Velthyding) hefur verið haldið við og það uppfært til að styðja nýjar þýðingarátir.
- Þýðingarpakkinn er samþættur *greynirseq*-safninu sem er aðgengilegt á Python Package Index (PyPi, pypi.org) og má setja upp með `pip install greynirseq` í hvaða umhverfi sem styður Python 3.
- GreynirSeq (https://github.com/mideind/GreynirSeq) pakkinn sem inniheldur skipanalínuviðmót (CLI) og kóða til stuðnings hefur verið gefinn út á CLARIN á http://hdl.handle.net/20.500.12537/257.

Skýrsla

Skipanalínuviðmót (CLI)
Til að setja upp Python-pakkann mælum við með því að setja upp nýtt sýndarumhverfi og keyra

```
$ pip install greynirseq
```

Hér sýnum við hvernig skal nota skipanalínuviðmótið (CLI) til að þýða beint úr ensku í íslensku.

```
$ echo "This is an awesome test that shows how to use a pretrained translation
model." | greynirseq translate --source-lang en --target-lang is

Þetta er æðislegt próf sem sýnir hvernig nota má forprófað þýðingarlíkan.
```

Að neðan sýnum við hvernig skal nota skipanalínuna til að þýða beint frá íslensku yfir á ensku.

```
$ echo "Þetta er æðislegt próf sem sýnir hvernig nota má forprófað þýðingarlíkan."
| greynirseq translate --source-lang is --target-lang en

This is an awesome test that shows how a pre-tested translation model can be used.
```

Ef líkanið/líkönin hafa ekki verið sótt áður eru þau sótt sjálfkrafa af vefnum. Ef þegar hefur heppnast að keyra `pip install greynirseq` þarf ekki að gera neitt annað. Athugið að skrárnar sem sóttar eru eru stórar (þó nokkur gígabæti).

Python-viðmót
Þýðingarlíkönin má einnig nota beint með Python-viðmóti með því að flytja inn viðeigandi einingar úr *greynirseq*-pakkanum. Fyrir dæmi um kóða sem notar viðmótið vísum við í útfærslu skipanalínuviðmótsins í
https://github.com/mideind/GreynirSeq/blob/main/src/greynirseq/cli/greynirseq_main.py

Dæmi um hvernig pakkann má eingilda og nota fyrir ályktun er sýnt að neðan.

```
from argparse import NameSpace
from greynirseq.cli.greynirseq_main import TranslateBart

model_args = NameSpace(command="translate", target_lang="is", source_lang="en",
model_name="mbart25-cont-ees-enis")
handler = TranslateTransformer(device="cpu", batch_size=16, show_progress=True,
max_input_words_split=100, model_args)
source_text = "Some text for translating"
translation = handler.run([source_text])
```

V6 – Grunnlína fyrir opnar margmála þýðingar yfir á og frá íslensku

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Lýsing

Markmið þessa undirverkefnis er að koma á fót og sýna almenna skilvirkni þeirra aðferða sem þróaðar hafa verið fyrir þýðingar milli íslensku og ensku, með því að þýða margmála líkön til að þýða milli íslensku og annarra tungumála en ensku. Mörg gagnasöfn eru til í þessum tilgangi, frá sömu gagngjöfum og voru nýttir við þýðingar milli ensku og íslensku. Þó við spáum því að ekki náist sömu gæði og fyrir ensku, þá mun þetta verkefni tryggja að þekkingin og aðferðirnar sem þróaðar eru innan Máltækniáætlunar hafi eins mikil áhrif og mögulegt er á framtíð vélþýðingarverkefna á og frá íslensku. Afurðirnar verða stökkpallur fyrir slík framtíðarverkefni og þróun fleiri tungumálapara til viðbótar. Grunnvinnan sem innt er af hendi í þessu undirverkefni gæti einnig liðkað fyrir framtíðarfjármögnun máltækniverkefna í gegnum ESB/EES-styrki.

Sem dæmi, og líklega forgangsmál, myndi opið vélþýðingarkerfi milli íslensku og pólsku hafa mikil áhrif og mikinn samfélagslegan ávinning, þar sem nærri 10% vinnuafls á Íslandi í dag hefur pólsku að móðurmáli.

Varða 8 – lýsing

M8: Forþjálfað margmála mállíkan valið og fínstillt.

Varða 9 – lýsing

M9: Þýðingarlíkan/-líkön fyrir íslensk-pólskt tungumálapar gefið út á CLARIN.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Við höldum áfram að þjálfa mBART50[10] líkanið sem inniheldur pólsku í forþjálfunarverki sínu.

- Líkan sem getur þýtt milli íslensku og pólsku (í báðar áttir) hefur verið þróað og gefið út á CLARIN á http://hdl.handle.net/20.500.12537/259.

[10] https://github.com/facebookresearch/fairseq/tree/main/examples/multilingual#mbart50-models

Skýrsla

Líkanið var þjálfað með fairseq með þýðingu undir hálfu eftirliti, fyrst með mBART50-líkaninu. Líkanið var svo þjálfað með margra verka skrá til að læra fyrst að afsuða setningar, líkt og markmið forþjálfunar fyrir t.d. BART-líkönin sem við höfum notað í öðrum verkefnum V4. Síðan var líkanið þjálfað til að þýða með notkun samraðaðs samhliða texta. Að lokum fær líkanið einmála texta bæði á íslensku og pólsku, sem það notar til að búa ítrað til bakþýðingar. Þessi uppsetning þjálfar því í báðar áttir á sama tíma, yfir þessi þrjú verk (afsuðun, samhliða þýðing, bakþýðing).

Þjálfunargögn
Þjálfunargögnin sem voru notuð eru fengin með því að sækja allar tiltækar ensk-pólskar þýðingar í almennum gagnasettum. Þetta skapaði um 600.000 íslensk-pólsk þýðingarpör. Fyrir setningar sem áttu sér ekki hliðstæðu í ensk-íslensku gagnasetti bakþýddum við ensku hliðina með EN-IS-þýðingarlíkani til að búa til hágæða gervibakþýðingardæmi (snúnar bakþýðingar (e. pivoted backtranslation)).

Fyrir einmála gögn notum við Risamálheildina og pólska hluta mC4-gagnasettsins sem er aðgengilegt á Hugging Face.

Fyrir gervisamhliða gögn (EN-PL-setningapör send í gegnum EN-IS-þýðingarlíkan):

- Digital Corpus of European Parliament (DCEP)
- Lyfjastofnun Evrópu (EMEA)
- Europarl (málflutningar þingmála)
- JRC-Acquis (texti reglugerða)
- OpenSubtitles (OS2018)

Niðurstöður
Þar sem niðurstöður úr þjálfun líkana undir hálfu eftirliti fara að miklu leyti eftir stillingu verkskrár þjálfuðum við líkanið þrisvar sinnum til að sjá hvaða bætingar við getum fengið.

Fyrir fyrstu tvær ítranirnar leituðum við aðallega að viðeigandi ytri stikum (e. hyperparameters) (eins og verkþyngd, skammtastærð og námshraða), auk viðeigandi verkskrár (hve langur er fasi aðeins-afsuðunar, o.s.frv.). Við ítrun þrjú bættust við tvær endurbætur: við bættum við gervisamhliða gögnunum sem nefnd eru að ofan (eftir síun), auk notkunar bættrar bakþýðingaraðferðar (sem byggist á færanlegu meðaltali líkansins). Við ítrun fjögur þjálfum við líkanið áfram með nokkrum breytingum ytri stika til að ná sem bestum árangri.

Við viljum benda sérstaklega á að *Wikipedia*, sem er grunnurinn í Flores-prófunarsettinu, er hvorki hluti af samhliða gögnunum né einmála gögnunum sem notuð voru til að þjálfa líkanið. Við notum *Flores*-gagnasettið sem mælistiku á *alhæfingarhæfni utan sviðs* og við búumst því við töluvert verri niðurstöðum en á gögnum innan sviðs. Við hefðum getað bætt *Wikipedia* við þjálfunargögnin okkar, en þá hefði *Flores* hætt að mæla algjöra *alhæfingarhæfni utan sviðs*.

Þannig að þrátt fyrir að við sjáum lækkun fyrir prófunargögn innan sviðs milli ítrana tvö og fjögur teljum við það ásættanlegt þar sem við fáum bættan árangur á Flores og ályktum að þetta bendi til betri *alhæfingarhæfni* líkansins.

Líkan	**BLEU fyrir PL-IS**	
	Heldni frá raunverulegum samhliða þjálfunargögnum[11]	**Þróunarsett Flores**
Undir eftirliti (bara samhliða gögn)	26,49	9,73
Ítrun 1 (verkskrá, leit ytri stika)	25,70	13,40
Ítrun 2 (verkskrá, leit ytri stika)	28,30	14,40
Ítrun 3 (viðbætt gervisamhliða gögn)	26,70	14,80
Ítrun 4 (frekari þjálfun með smávæg. breytingum ytri stika)[12]	27,60	15,30
Líkan	**BLEU fyrir IS-PL**	
	Heldni frá raunverulegum samhliða þjálfunargögnum	**Þróunarsett Flores**
Undir eftirliti (bara samhliða gögn)	26,70	11,10
Ítrun 1	25,90	11,10
Ítrun 2	28,30	11,44
Ítrun 3	27,10	13,10
Ítrun 4	27,70	13,30

[11] Þetta samanstendur af handahófsvali *EMEA,* nokkurra *EES-reglugerða*, *Biblíunnar* og *Vottum Jehóva*.
[12] Við jukum m.a. verkþyngd bakþýddra verka

Að neðan má sjá *BLEU*-gildi fyrir ýmsar undirmálheildir. Fyrir ítrun þrjú voru sumar þeirra þegar hluti af samhliða þjálfunargögnum á meðan aðrar voru nýjar viðbætur, sem orsakar stökkið frá ítrun tvö til ítrunar þrjú. Þó að líkanið virðist ekki alhæfast vel á óséð svið (*Wikipedia* textar - *Flores,* sbr. að ofan), lærir það á sviðin sem eru í samhliða og gervisamhliða þjálfunargögnunum á fullnægjandi hátt. Á þessari stundu erum við ekki viss hvort niðurstöður Flores megi skýra með stílfræðilegum mun eða umræðuefni (orðaforða) þýðingarparanna.

	BLEU fyrir PL-(Gervi-IS)				
Líkan	**DCEP**	**EMEA**	**Europarl**	**JRC-Acquis**	**OS2018**
Ítrun 2	25,9	38,9	20,4	29,9	14,0
Ítrun 3	37,2	40,3	29,2	41,5	18,0
Ítrun 4	41,2	42,6	32,1	45,6	19,6
	BLEU fyrir (Gervi-IS)-PL				
Líkan	**DCEP**	**EMEA**	**Europarl**	**JRC-Acquis**	**OS2018**
Ítrun 2	22,1	36,4	16,1	26,1	11,5
Ítrun 3	33,8	37,5	25,2	36,7	14,2
Ítrun 4	37,2	39,4	27,5	40,0	15,6

L2 – Sérhæfðar villumálheildir

Skýrsla: Háskóli Íslands
Aðilar: Háskóli Íslands

Markmið

Safna textum frá tilteknum undirhópum innan þess hóps fólks sem notar íslensku sem annað mál, og stækka lesblindumálheildina. Stærðarmarkmiðum annars árs hefur verið náð en vinna við þau hefur sýnt að málheildirnar þarf að stækka enn frekar.

Vinna á öðru ári hefur sýnt að til að afla upplýsinga um villur sem tilteknir hópar sem nota íslensku sem annað mál gera, er þörf á frekari gögnum frá þessum hópum. Af þessum sökum verður villumálheild annars máls málhafa stækkuð með textum skrifuðum af höfundum innan stærstu innflytjendahópa á Íslandi, einkum pólskum. Með þessari nálgun náum við fullri yfirsýn yfir algengustu villurnar sem þessir hópar gera, sem gagnast við kerfisbundna þróun málrýnis. Af þessu leiðir að málrýnirinn mun á endanum nýtast meirihluta innflytjenda á Íslandi.

Villutíðnir í sérhæfðu villumálheildunum sýna að flestar villur eru í textum sem skrifaðir eru af lesblindum, sem þýðir að þessi notendahópur þarf mest á málrýni að halda. Villudreifing innan málheildarinnar er öðruvísi en í almennu villumálheildinni, til dæmis inniheldur hún hærra hlutfall réttritunarvillna (e. orthographic errors). Lítil málheild er líkleg til að vera hlutdræg, villur sem einskorðast við tiltekinn höfund geta verið of áberandi og af því leiðir að málheildin endurspeglar ekki raunverulega villudreifingu í textum sem skrifaðir eru af lesblindum. Af þessum ástæðum verður lesblindumálheildin stækkuð.

Varða 8 – lýsing

M8: Söfnun texta frá stóru innflytjendahópunum lokið og merkingar hafnar. Söfnun texta frá fullorðnum með lesblindu hafin.

Varða 9 – lýsing

M9: Allir textar hafa verið merktir og lokaútgáfan gefin út á CLARIN. Minnst 10.000 villur greindar og merktar frá textum skrifuðum af fullorðnum með lesblindu.

Afurðir

Flestum markmiðum vörðunnar var náð. Við greindum og merktum 8.436 villur úr textum skrifuðum af fullorðnum einstaklingum með lesblindu, sem þýðir að við náðum ekki 10.000 villna markmiðinu. Við erum þó með nægilegt magn villna til eigindlegrar greiningar með marktækan mun á hlutfallslegri tíðni villutegunda, og getum við því ályktað hvers konar villur fullorðnir einstaklingar með lesblindu gera, og hvernig þær eru frábrugðnar villum frá þeim sem eru ekki með lesblindu. Athugið að Íslenska lesblinduvillumálheildin er stærsta tiltæka lesblinduvillumálheildin í dag, bæði hvað varðar fjölda orða og villna[13].

- Við söfnuðum 101 texta frá annarsmálshöfum íslensku með samtals 29.948 villum. Málheildin hefur verið gefin út á CLARIN á http://hdl.handle.net/20.500.12537/280.
- Við söfnuðum 35 textum frá fullorðnum með lesblindu með samtals 8.436 villum, og gáfum út á CLARIN á http://hdl.handle.net/20.500.12537/281.

Skýrsla

Íslenska L2 villumálheildin
Íslenska L2 villumálheildin er safn 101 texta, aðallega ritgerðum nemenda, skrifuðum af 44 annarsmálshöfum íslensku með 17 ólík móðurmál, og inniheldur alls 24.948 villutilvik. Við höfum 17 mismunandi móðurmál þar sem við skilgreindum ekki í byrjun ákveðin tungumál til að leggja áherslu á. Markmið um söfnun texta frá helstu hópum innflytjenda var skilgreint seinna í ferlinu.

Í töflu 1 eru 10 algengustu undirflokkar Íslensku L2 villumálheildarinnar sýndir. Algengustu villurnar eru *wording (orðalag)*, 10,96% villna. *Punctuation (greinarmerki)* og *inflection (beygingar)* fylgja fast á eftir með 9,60% og 9,04%. Athyglisvert er að þó að *inflection* sé þriðja algengasta villan í L2-málheildinni er hún ekki meðal 10 algengustu villna í Íslensku villumálheildinni. Villutíðnin í Íslensku L2 villumálheildinni er 153,93 á hver 1.000 orð.

Undirflokkar	**Yfirflokkar**	**Tíðni**	**Hlutfall (%)**
wording	style	2735	10,96
punctuation	orthography	2396	9,60
inflection	grammar	2256	9,04
miscellaneous	other	1895	7,59
agreement	grammar	1524	6,11
prep	grammar	1452	5,82
definitiveness	grammar	1186	4,75

[13] Næststærst er BDAC-málheildin. Aðrar tiltækar málheildir eru allar töluvert minni. Sjá https://aclanthology.org/W17-1309.pdf

typo	orthography	1153	4,62
syntax	grammar	1146	4,59
insertion	vocabulary	1133	4,54

Tafla 1. Tíðni 10 algengustu villna í Íslensku L2 villumálheildinni, eftir undirflokkum.

Íslenska lesblinduvillumálheildin

Íslenska lesblinduvillumálheildin samanstendur af 35 skrám með 38.891 orð, þar sem 5.075 endurskoðanir eru merktar og 8.436 villur merktar (sjá töflu 2). Villutíðni Íslensku lesblinduvillumálheildarinnar er 216,91 villa á hver 1.000 orð.

Undirflokkar	**Yfirflokkar**	**Tíðni**	**Hlutfall (%)**
punctuation	orthography	722	10,57
typo	orthography	714	10,46
wording	style	634	9,29
nonword	orthography	617	9,04
miscellaneous	other	449	6,58
spacing	orthography	411	6,02
syntax	grammar	401	5,87
insertion	vocabulary	337	4,94
inflection	grammar	311	4,56
prep	grammar	296	4,34

Tafla 2. Tíðni 10 algengustu villna í Íslensku lesblinduvillumálheildinni, eftir undirflokkum.

Íslenska lesblinduvillumálheildin svipar til Íslensku villumálheildarinnar[14] að því leyti að meðal algengustu villna eru *punctuation* og *wording*, en ólíkt hinum málheildunum hefur lesblindumálheildin hærra hlutfall innsláttarvillna (typos); sérkenni sem búist er við frá lesblindum. Flokkurinn *nonword* er einnig nokkuð stór miðað við aðrar málheildir.

Eftir söfnun og útgáfu málheilda verður ítarleg skýrsla gefin út um allar sérhæfðar villumálheildir, þ.e. Íslensku L2 villumálheildina, Íslensku lesblinduvillumálheildina og Íslensku barnamálsvillumálheildina. Hún mun innihalda tölfræði um villutíðni og villutegundir hvers

[14] Sjá http://hdl.handle.net/20.500.12537/105

undirhóps, auk almennrar lýsingar á textasöfnun, og hvernig var tekið á vandamálum sem komu upp.

L7 – Kerfisbundin þróun málrýnis

Skýrsla: Miðeind
Aðilar: Miðeind, Háskóli Íslands

Markmið

Verkefnið má skilgreina gróflega út frá þremur þáttum: A) villuflokkar, B) notendahópar, og C) villumeðhöndlun/prófunaraðferðir.

Í vinnu fyrsta og annars árs var mest áhersla lögð á A), til að tryggja sem besta dekkun. Við byrjuðum á samhengisfrjálsum villum, því næst samhengisháðum villum og að lokum málfræðivillum. Til að auðvelda þetta var B) skorðað við notendahópa sem koma fyrir í almennu villumálheildinni. C) var ekki skorðað við villugreiningu en það varð aðal mælistikan fyrir vinnu okkar.

Í vinnu þriðja árs verður C) í brenndiepli. Gerðar verða tilraunir fyrir mismunandi notendahópa í undirverkefni L14. Til að fá nytsamlegustu og heildstæðustu afurðina fyrir stærsta notendahópinn fyrir lok áætlunarinnar er markmiðið á þriðja ári að bæta leiðréttingar, tillögur og lýsingar fyrir villuflokka sem áður hafa verið meðhöndlaðir í almennu villumálheildinni. Slíkar upplýsingar eru notendum mjög dýrmætar.

Tveir almennir notendahópar eru skilgreindir; fjölmiðlafyrirtæki, og höfundar fræðilegra ritgerða, rannsóknargreina o.s.frv. Á öðru ári ljúkum við notendaprófunum fyrir fjölmiðlafyrirtæki. Til að mæta kröfum annarra hópa verða gerðar notkunarprófanir meðal háskólanema sem eru að skrifa lokaritgerðir.

Varða 8 – lýsing

M8: Prófunum byggðum á lokaritgerðum lokið. Almennum lýsingum og viðmiðum hefur verið bætt við fyrir alla villuflokka, og ítarlegri lýsingum hefur verið bætt við valið sett flokka/villna.

Varða 9 – lýsing

M9: Tauganetsflokkarinn frá L14 (sem ákvarðar hvort setning er líklega rétt og þarfnast ekki frekari vinnslu) bætt við sem valkosti við málrýninn. Forgangsröðuð vandamála úr prófunum byggðum á lokaritgerðum hafa verið lagfærð. Endanleg útgáfa málrýnis gefin út á CLARIN.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.

- Prófunum byggðum á textum úr ritgerðum nemenda er lokið, og viðeigandi gögn hafa verið gefin út á CLARIN á http://hdl.handle.net/20.500.12537/258.
- Allir villuflokkar innan ramma hafa nú almennar lýsingar og viðmið.
- Tauganetssetningaflokkarinn er nú aðgengilegur sem valkostur í kerfinu.
- Forgangsröðuð vandamál úr prófunum byggðum á ritgerðum nemenda hafa verið leyst.
- Endanleg útgáfa málrýnis er aðgengileg á CLARIN á http://hdl.handle.net/20.500.12537/270 og á GitHub á https://github.com/mideind/GreynirCorrect.
- BinPackage var einnig uppfærður og er aðgengilegur á CLARIN á http://hdl.handle.net/20.500.12537/267.
- Vefviðmótið er aðgengilegt á CLARIN á http://hdl.handle.net/20.500.12537/266 og á vefnum á https://yfirlestur.is/.

Skýrsla

1. Yfirlit

Áherslan fyrir M9 var stuðningur við notendur, svo sem betri viðmið, og skil á endanlegri afurð. Þess utan voru endurbætur gerðar á kerfinu í heild sinni.

2. Prófanir ritgerða

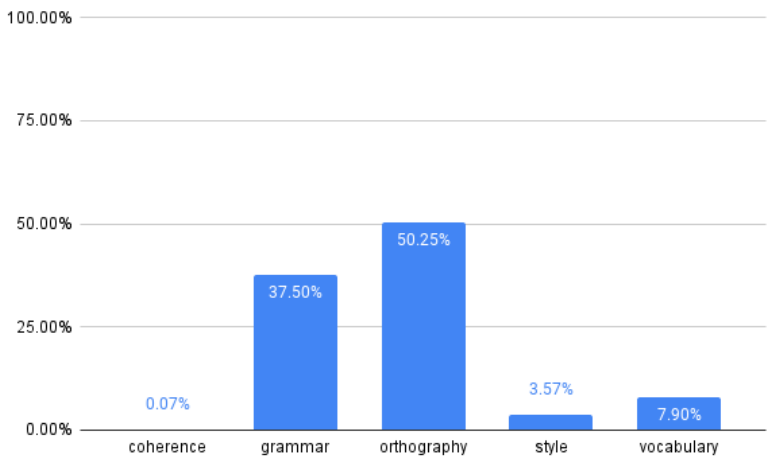
Notendaprófanir byggðar á ritgerðum nemenda voru gerðar. Flestar ritgerðir frá Háskóla Íslands eru gefnar út á https://skemman.is á PDF formi með lokuðum leyfum.

Ein þessara ritgerða, sem vitað var að innihéldi margar villur, var valin sem prófunardæmi. PDF-skránni var varpað yfir á .txt-skrá með PDFMiner (https://pypi.org/project/pdfminer/). Til að bera málrýninn saman við gullstaðal var ritgerðin leiðrétt handvirkt. Prófarkalesari, sem vann áður við villumálheildir í verkþáttum L1 og L2, leiðrétti ritgerðina, en aðeins ótvíræðar villur og engar stílvillur. Óleiðrétta útgáfa ritgerðarinnar var svo leiðrétt með málrýninum og úttakið borið saman við útgáfu prófarkalesarans.

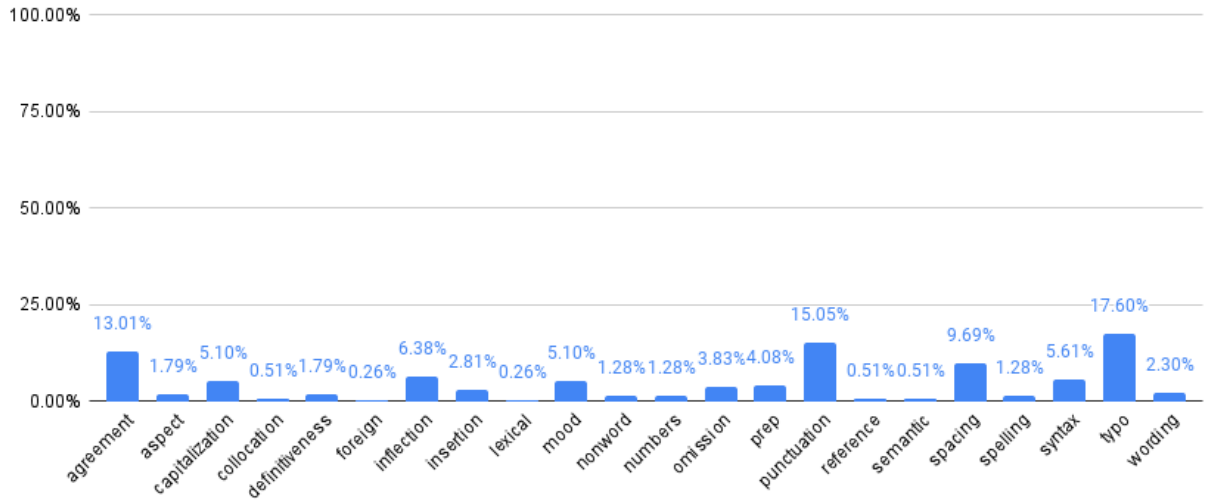
Vegna leyfismála er ekki hægt að gefa út flestar ritgerðir nemenda undir opnu leyfi. Það sem er gefið út í þessum hluta er listi handvirkra leiðréttinga sem gerðar voru á upprunalegri skrá. Þær leiðréttingar innihalda ekki samhengi vegna ofangreindra ástæðna, en listinn inniheldur línunúmer upprunalegrar villu, upprunalega mynd villunnar og handvirka leiðréttingu hennar. Listinn er útgefinn á CLARIN á http://hdl.handle.net/20.500.12537/258.

575 villur voru annaðhvort ekki greindar eða rangt leiðréttar af málrýninum, miðað við handvirkt leiðrétta útgáfu. Ekki var hægt að flokka allar þessar villur handvirkt, en um 70% þeirra voru flokkaðar með notkun yfirflokka og undirflokka úr mörkunarskema sem notað er í Íslensku villumálheildinni (https://github.com/antonkarl/iceErrorCorpus/blob/master/errorCodes.tsv). Þessi flokkun gefur yfirsýn yfir það hvers konar villur málrýnirinn greinir ekki, og var notuð til að stýra frekari þróun málrýnisins.

Mynd 1 að neðan sýnir tíðni yfirflokka í flokkuðu villunum. *Orthography* og *grammar* eru algengustu flokkarnir, um 87% allra villna. Mynd 2 sýnir hvernig tíðni og dreifing undirflokka er nokkuð jöfn. *Typo*, *punctuation* og *agreement* eru algengustu undirflokkarnir, og þeir einu sem eru með hærri tíðni en 10%.



Mynd 1. Tíðni yfirflokka.



Mynd 2. Tíðni undirflokka.

3. Tillögur og lýsingar

Allir villuflokkar innan ramma hafa almennar lýsingar, með ítarlegri lýsingar fyrir ákveðna flokka.

Fyrir villugögn úr *Ritmyndum* (óstaðlaðar orðmyndir sem rætt er um í M6-skýrslunni fyrir verkþátt G4), bættum við sérstökum villuskilaboðum við sem byggð eru á gefnum villuflokki í *Ritmyndum* og skjölun í upprunalegri gagnaheimild fyrir hvern flokk.
Skilaboðin eru geymd sem f-strengir, sem leyfir innskot strengja sem standa fyrir upprunalegt gildi, leiðrétt gildi, og uppflettimyndina á ákveðna staði. Þetta gerir okkur kleift að birta notandanum villuskilaboð sem innihalda bæði upplýsingar um villuflokk og tiltekna orðmynd.

Að auki eru sumir villuflokkar í *Ritmyndum* tengdir sértækum reglum í íslenskum málstaðli í upprunalegri skjölun. Við bættum þeim strengjum við skilaboðsgögnin fyrir viðeigandi villuflokk og útfærðum aðgerð sem sækir rétta slóð fyrir hvern streng, í flestum tilfellum https://ritreglur.arnastofnun.is. Þetta gerir okkur kleift að sýna notandanum hlekki á tilvísanir fyrir hverja villu innan villuskilaboðanna í vefviðmótinu.

Þar sem þessi gögn eru sótt frá þriðja aðila (innan SÍM) og reglulega uppfærð útfærðum við lítið hjálparforrit sem athugar hvort nýjum flokkum hefur verið bætt við *Ritmyndir*, og hvort einhverjir flokkar hafi verið gerðir úreldir. Þetta auðveldaði viðhald villuskilaboðanna fyrir gögn *Ritmynda* til muna og keyrist við hverja uppfærslu.

Sumar aðrar villur á tókastigi eru meðhöndlaðar á sértækan hátt eftir eðli þeirra, til dæmis hástafavillur. Hér er hvert tilfelli meðhöndlað í tilteknum hluta kóðans, sem gerir okkur kleift að skilgreina villuskilaboð með öllum viðeigandi upplýsingum.

Fyrir önnur villugögn á tókastigi sem ekki hafa verið flokkuð, og eru meðhöndluð á sjálfvirkari máta, höfum við ekki til taks allar viðeigandi upplýsingar til að sýna notanda. Sjálfvirk flokkun með tauganetum er ein möguleg leið, en það er ekki innan ramma þessa verkefnis. Eins og er birtast villuskilaboð fyrir hverja villu, sem útskýra ranga mynd og rétta mynd.

Málfræðivillurnar eru að mestu meðhöndlaðar með nokkuð sértækum leitarmynstrum. Þetta þýðir að við höfum í flestum tilfellum allar nauðsynlegar upplýsingar til að veita notanda haldbærar upplýsingar um hvers vegna villan var gerð og hvernig skal komast hjá henni. Vegna þessa innihalda allar málfræðivillur sem kerfið meðhöndlar sértæk villuskilaboð.

Sum villuskilaboð voru gerð ítarlegri. Sem dæmi koma villuskilaboð óþáttaðra setninga frá þáttaranum og gefa aðeins takmarkaðar upplýsingar frá sjónarhorni þáttunar, til dæmis hvers vegna þáttun mistókst. Við bættum upplýsingum við varðandi málfræðiyfirferð, til að leiðbeina notanda betur.

Tegund	**Setning**	**Lýsing**
Ritmyndir	Jóni veiðimanni lýst ekki á þetta mál.	Stafsetningarvilla; **lýst** -> **líst** *Rita ætti **í** í stað **ý** í **lýst**, svo: **líst**.*

		[Hlekkir í tilvísanir]
Hástöfun	Hann var Félags- og barnamálaráðherra.	Rita á **félags- og barnamálaráðherra** með lágstaf.
Hástöfun	Við borðuðum Júní í júní.	Í dagsetningunni **Júní** á mánaðarnafnið að byrja á lágstaf
Sjálfvirkar aðferðir	Hún borðaði fyskinn.	Rita á **fyskinn** sem **fiskinn**
Málfræðivilla	Börnin voru út á túni allan daginn.	Hér á líklega að vera **úti** í stað **út** *Í samhenginu **út á túni** er rétt að nota atviksorðið **úti** í stað **út**.*
Málfræðivilla	Ég hef búið á Hafnarfirði alla mína tíð.	Rétt er að rita **í Hafnarfirði** *Ýmist eru notaðar forsetningarnar **í** eða **á** með nöfnum staða, bæja og borga. Í tilviki Hafnarfjarðar er notuð forsetningin **í**.*
Óþáttanleg setning	Víst að hafði gaman hvernig hvar sem gengur.	Ekki tókst að þátta setningu; mögulega felst villa í henni

Tafla 1. Dæmi um lýsingar fyrir mismunandi villuflokka.

Til að gefa réttar upplýsingar var staðfest að öll villumeðhöndlun innihéldi viðeigandi upplýsingar; rétt gildi þar sem við á, rétta spönn (þetta er sérstaklega mikilvægt fyrir málfræðivillur, þar sem leitarmynstrin leita oft víða), o.s.frv.

4. Tauganetssetningaflokkari

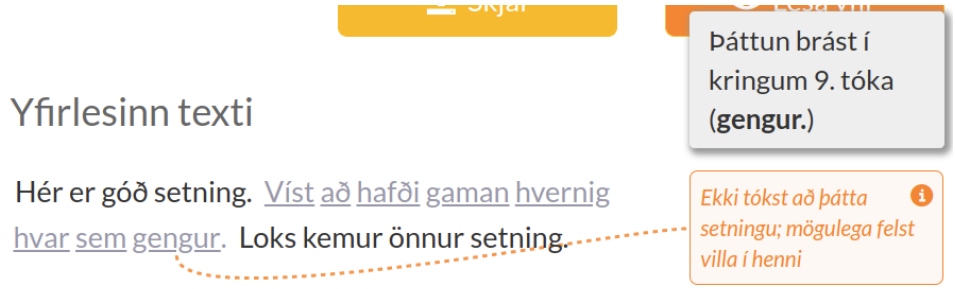
Tvígilda setningastigsflokkaranum sem rætt var um í skýrslunni fyrir verkþátt L14 var bætt við sem valkostur. Markmiðið er að skoða hvort setning sé líkleg til að innihalda villu. Ef svo er skoðar málrýnir hana nánar. Ef svo er ekki leitar málrýnir ekki að villum í setningunni með þáttun. Hugmyndin er að þar sem flestar setningar innihalda ekki villur hraði þetta yfirferðinni þar sem sleppa má dýra þáttunarfasanum.

Við bættum við einföldum hjúp (e. wrapper), *classifier.py*, fyrir flokkarann. Við bættum valkostinum [*sentence_prefilter*] við *correct* skipunina á skipanalínunni. Þetta gerir okkur kleift að sía setningar í inntakinu og senda aðeins þær sem hugsanlega eru rangar í málrýninn. Flokkaranum var bætt við sem valkvæðum innflutningi í GreynirCorrect. Þetta var talin besta hönnunin, þar sem grunnpakkinn getur áfram verið nokkuð léttur í notkun með takmörkuðum tilföngum, en einnig stutt við þyngri notkun.

5. Endurbætur

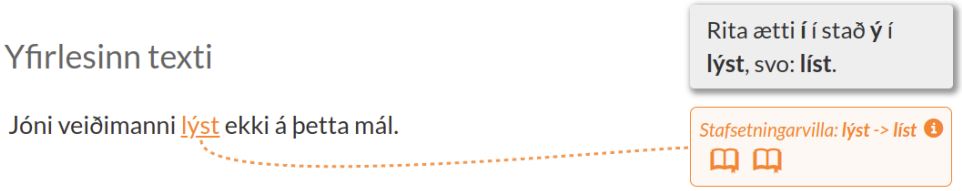
5.1 Forritaskil (API)
Stuðningur við valkostina sem rætt var um í skýrslunni fyrir vörðu M7 var útfærður í forritaskilunum, ásamt valkostunum fyrir setningaflokkarann sem rætt var um í kafla 4 og valkost þess að útiloka ákveðna villuflokka eins og nefnt var í kafla 5.3. Víðtækum prófunum var bætt við fyrir nýju eiginleikana, auk skjölunar fyrir villuflokka, virkni forritaskila og studdra stöðuvísa/valkosta.

5.2 Vefviðmót
Ásamt því að uppfæra undirliggjandi forritaskil var vefviðmótið slípað til. Setningar sem ekki er hægt að þátta birtast nú gráar en ekki rauðar, þar sem þær innihalda ekki endilega villur. Þetta skilur þær betur frá villum, sem bætir notendaupplifun.



Mynd 3. Setningar sem ekki er hægt að þátta birtast á annan hátt en setningar með villum.

Hönnun hlekkja í Íslenskan málstaðal var einnig slípuð, eins og nefnt var í hluta 3.



Mynd 4. Tilvísunarhlekkir birtast nú sem bókartákn.

5.3 Fölsk jákvæðni
Helsta áherslan fram að síðustu vörðu var á dekkun, bæði í villugreiningu og leiðréttingu. Þetta hefur gengið vel en hefur leitt til ýmissa falskra jákvæðra, þar sem sumar reglur meðhöndlunar voru of víðtækar. Sú staðreynd að GreynirCorrect og IceErrorCorpus hafa ekki sama merkingarskema (vegna eðlismunar) þýddi að niðurstöður sjálfvirkra prófana sýndu okkur ekki hvaða flokkar í GreynirCorrect ollu þessu. Til að fá betri sýn á hvaða villuflokkar innan GreynirCorrect voru að valda þessum vandamálum og þurfti að laga skoðuðum við fréttagreinar úr gagnasafni Greynis með GreynirCorrect. Við skoðuðum 20 villur úr hverjum villuflokki, ákvörðuðum hlutfall sannra/falskra jákvæðra og hvaða ráðstafanir við gætum tekið til að bæta niðurstöður fyrir flokkana með verstu niðurstöðurnar. Þessi skoðun reyndist mjög mikilvæg þar sem hún gaf okkur annað sjónarhorn heldur en dekkunarbyggðu árangursmælingarnar.

Við bættum valkosti við kerfið til að sleppa merkingum fyrir valda villuflokka. Þetta hjálpaði bæði við að sleppa flokkum sem þurfti enn að laga og bætti upplifun fyrir ýmsa notendur.

Við bættum meðhöndlun fyrir marga flokka byggt á þessum niðurstöðum og má þar nefna:

- Lýsingarorð með hástöfum eru í mörgum tilfellum hluti nafneininga (Danska ríkisútvarpið, Íslensk erfðagreining), sem kerfið þekkir ekki endilega, en eru líka algengar villur. Þær eru nú aðeins merktar sem villur ef valkosturinn [*suppress_suggestions*] er stilltur á „falskt“. Á svipaðan hátt eru orð með hástaf í byrjun ekki lengur merkt sem villur nema tillögð orðmynd sé einnig þekkt sem hástafað orð.
- Greining og leiðrétting bannorða er mjög viðkvæmt mál. Listinn sem við notum er upprunalega frá þriðja aðila, en er nú haldið við innan GreynirCorrect. Þetta gerir okkur kleift að eyða færslum sem ættu í heild sinni ekki að vera merktar sem bannorðavillur. Dæmi um þetta er „hommi“, sem hægt er að nota á niðrandi hátt, en hefur verið endurheimt og er mjög algengt í notkun og talið hlutlaust orð. Annað dæmi eru orðmyndir sem má túlka sem bannorð, en eiga sér einnig mjög almenna hlutlausa merkingu. Dæmi um þetta er 'nýbúinn [að gera eitthvað' – 'just finished [doing something]', sem getur einnig táknað örlítið niðrandi nafnorðið 'nýbúi' – 'immigrant'. Þessum tilteknu orðmyndum hefur verið eytt úr gögnunum.
- Villumeðhöndlun fyrir margar málfræðivillur var uppfærð til að taka á óalgengum tilvikum. Prófunum var bætt við fyrir bæði sönn jákvæð og mögulega falsk jákvæð til að tryggja rétta meðhöndlun.

5.4 Stoðtól

Undirliggjandi tilreiðari var uppfærður til að styðja betur langan inntakstexta. Þetta hefur hverfandi áhrif á málrýninn sjálfan, þar sem þáttunarferlið er margfalt lengra en tilreiðing, en aðstoðar setningaflokkarann sem rætt er um í kafla 4, þar sem hann krefst þess að textanum sé skipt í setningar með tilreiðaranum, og hjálpar þetta því málrýnikerfinu í heild sinni. Meðhöndlun fleiri gerða af greinarmerkjavillum var einnig bætt við.

Einhverjar breytingar voru gerðar á BinPackage, BÍN-byggða leitartólinu úr verkþætti G4, aðallega uppfærsla orðaforða frá BÍN og lagfæringar á böggavillum.

Breytingar voru gerðar á GreynirPackage, undirliggjandi þáttaranum, sérstaklega hvað varðar villumálfræðireglurnar.

5.5 Aðrar endurbætur

Fyrstu málfræðimynstrin fyrir málfræðivillur voru mjög sértæk svo við bættum við fullyrðingarsetningu ef við fundum samsvörun en enga villu. Fyrir næstu mynstur reyndist árangursríkast að leita vítt til að finna samsvaranir, og skera svo niður lista samsvarana (þetta hjálpaði líka með falsk jákvæð). Í samræmi við þessa aðferð breyttum við fullyrðingarsetningu í if-lykkjuprófanir þar sem við átti.

Að auki voru nýjar útgáfur villugagna sóttar, t.d. *Ritmyndir*.

6. Mat

Niðurstöðurnar sem birtar voru fyrir vörðu M7 voru verri en búist var við, þó að þær stæðust markmið, sérstaklega fyrir málfræðivillur. Á meðan á vinnu við vörður M8 og M9 stóð tókum við eftir því að aðvaranir (villur með kóðaendingu '\w', sem merkja aðeins tillögur og ekki sjálfvirkar leiðréttingar) voru ekki með við prófanir. Þessi stilling varð til við vinnu á villum á tókastigi, þar sem '\w' mark er geymt fyrir ólíklegri merkingar sem ætti bara að benda á, ekki leiðrétta sjálfkrafa. Þegar verkefnið færði sig yfir í málfræðivillur og stílfræðilegar villur, þar sem tillaga til notanda hentar oft betur, fengu mjög margar villur merkinguna '\w'. Þegar þessar villur voru ekki lengur undanskildar reyndust niðurstöður mikið betri.

Yfirflokkar	**Tíðni**	**Greining**	**Leiðrétting**
coherence	4	7,81	7,81
grammar	182	30,29	15,69
orthography	1165	63,47	45,18
other	256	22,44	9,06
style	8	6,94	0
vocabulary	47	37,07	18,55
Samtals	1662	52,36	35,33

Tafla 2. Niðurstöður fyrir yfirflokkana.

Orthography-yfirflokkur	**Tíðni**	**Greining**	**Leiðrétting**
capitalization	114	52,92	39,33
nonword	74	43,67	40,12
punctuation	498	48,07	21,29
spacing	209	75,47	56,28
spelling	60	89,65	79,19
typo	210	93,27	86,04

Tafla 3. Niðurstöður fyrir undirflokka innan yfirflokksins *orthography (stafsetning)*.

Grammar-yfirflokkur	**Tíðni**	**Greining**	**Leiðrétting**
agreement	76	41,12	24,42

aspect	4	7,35	0
case	4	62,5	62,5
inflection	13	47,62	47,62
mood	27	4,03	2,79
prep	45	22,14	1,23
syntax	13	25,59	0

Tafla 4. Niðurstöður fyrir undirflokka innan yfirflokksins *grammar (málfræði)*.

Annað sem skal athuga er að áherslan hefur verið á að bæta viðmið, sem er erfitt að mæla með sjálfvirku mati. Þessar niðurstöður eru settar fram til að geta gert samanburð við aðrar skýrslur en hefur lítið með vinnu fyrir þessar vörður að gera.

L8 – Málrýni í snjalltækjum

Skýrsla: Grammatek
Aðilar: Grammatek, Miðeind

Markmið

Að gera málrýni mögulega á íslenskum lyklaborðum, sem og sjálfvirka útfyllingu (e. autocompletion), á Android- og iOS-tækjum. Á öðru ári munum við byrja að undirbúa þetta undirverkefni með því að skoða hugbúnaðarhögun og umhverfi og hvernig þarf hugsanlega að aðlaga kjarna málrýnis til að hægt verði að nota hann í snjalltækjum. Þessi vinna mun einnig nýtast öllum öðrum verkefnum sem þurfa að keyra á snjalltækjum og verður hún unnin í nánu samstarfi við H9 (Talgreining í snjallsímum). Vegna minnkandi tilfanga munum við einbeita okkur að því að afhenda lausnir fyrir Android-tæki, hins vegar verða kröfur og vegvísir fyrir iOS-innleiðingu afhent sömuleiðis fyrir innleiðingu í framtíðinni.

Varða 8 – lýsing

M8: Frumgerð málrýnieiningar byggð á vefþjónustu GreynirCorrect hönnuð og beta-útgáfa gefin út á Google Play. Anysoft Keyboard útfært með bættri íslenskri orðabók og togbeiðni í upprunalega hirslu opnuð.

Varða 9 – lýsing

M9: Eining sjálfvirkrar útfyllingar (e. auto completion) fyrir Android byggð á Lucene hönnuð, innbyggð í Anysoft Keyboard-kvíslina og gefin út á Google Play. Málrýnieining 1.0 gefin út á Google Play.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Forritið Réttritun 1.1.0 hefur verið gefið út á Google Play-versluninni. Við höfum innleitt íslenskan tungumálapakka í aðallínu AnySoftKeyboard, og við höfum samþætt Lucene inn í AnySoftKeyboard sem sjálfvirkan útfylli fyrir íslensku með notkun orðabókarinnar sem við höfum þegar innleitt í okkar eigin kvísl af hirslunni.

- Réttritun, Android-app tengt við forritaskil Yfirlesturs: https://play.google.com/store/apps/details?id=org.grammatek.simacorrect. Gefið út á CLARIN á http://hdl.handle.net/20.500.12537/284.
- AnySoftKeyboard með íslenskum tungumálapakka (mainlined): https://github.com/AnySoftKeyboard/AnySoftKeyboard.

- AnySoftKeyboard, kvísl með sjálfvirkri útfyllingu næsta orðs fyrir íslensku: https://github.com/grammatek/AnySoftKeyboard/releases/tag/v1.0.0-gt. Gefið út á CLARIN á http://hdl.handle.net/20.500.12537/290.

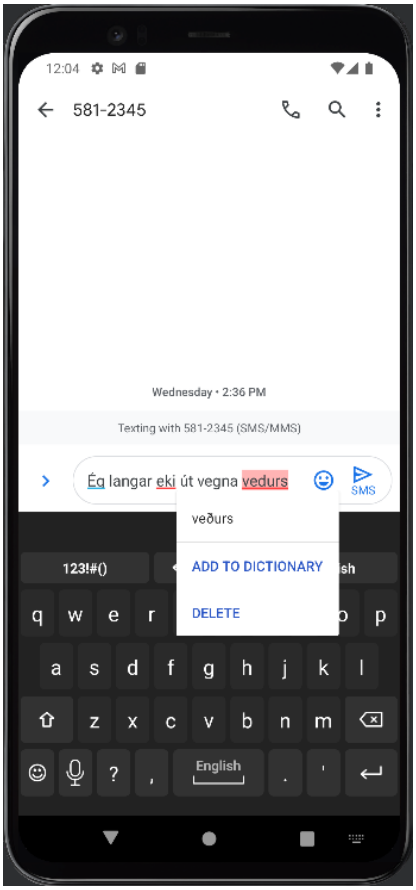
Skýrsla

Réttritun
Í fyrsta hluta verkefnisins útfærðum við íslenskan tungumálapakka fyrir opna lyklaborðið „AnySoftKeyboard“. Þetta lyklaborð er mjög vinsælt á Android og styður nú þegar mörg mismunandi tungumál. Fyrir íslensku settum við saman lista af 100.000 færslum úr Risamálheildinni og bjuggum til hugbúnaðarpakka til að bæta íslensku við sem stillingarmöguleika. Hugbúnaðurinn var gerður aðgengilegur verkefninu í formi togbeiðni á GitHub og var farsællega innleitt inn í aðalkvísl verkefnisins. Þar sem AnySoftKeyboard er forritað á Java notuðum við einnig Java-forritunarmálið fyrir íslenska tungumálapakkann.

Á sama tíma hófum við vinnu við Android-forritið Réttritun, sem veitir leiðréttingu íslenskrar stafsetningar á Android. Þetta forrit var alfarið skrifað á Kotlin-forritunarmálinu, sem er það forritunarmál sem mælt er með að nota fyrir Android í dag, en Kotlin má einnig nota í öðrum tækjum. Fyrir utan að bjóða upp á leiðréttingarþjónustu íslenskrar stafsetningar býður Réttritun upp á lítið smáforrit til að prófa leiðréttingu á stafsetningu/málfræði í textareit og upplýsingar um forritið sjálft.

Réttritun sendir öll innslegin orð til vefþjónustu fyrir leiðréttingu, sem skilar merktum listum leiðréttra orða sem birtast notendum þegar þeir ýta á rauð undirstrikuð orð í textareit. Frá og með Android 12 gefa blá undirstrikuð orð til kynna málfræðivillur. Þessar villur eru einnig studdar af Réttritun og vefþjónustunni. Vefþjónustan sjálf er byggð á opnu vefþjónustunni Yfirlestur, sem er líka afurð máltækniáætlunarinnar. Við bættum við OpenAPI 3.0 skilgreiningu fyrir þessa vefþjónustu og gátum þá auðveldlega átt samskipti við hana með því að skapa sjálfvirkt útgáfuhæfan Kotlin-kóða með OpenAPI-smiðnum: https://github.com/OpenAPITools/openapi-generator. Við veitum aðgang að vefþjónustunni á vefsíðunni https://yfirlestur.grammatek.com. Forritið notar þessa þjónustu, en getur notað hvaða samhæfðu þjónustu sem er ef það er endurþýtt.

Gögnin sem vefþjónustan fær frá notandanum geta verið viðkvæm. Við tryggjum því að tengingin við þjóna okkar sé dulkóðuð samkvæmt ströngum kröfum og að engar upplýsingar notanda séu geymdar á þeim. Að auki gerum við IP-tölur nafnlausar. Hægt er að lesa nánar um persónuverndarstefnu okkar hér: https://www.grammatek.com/legal#r%C3%A9ttritun.



Mynd 1. Dæmi um hefðbundna notkun með blöndu af stafsetningar- og málfræðivillum

Samhliða mánuðum af þróun höfum við haldið notendaprófanir fyrir forritið Réttritun og svarað endurgjöf notenda með endurbótum. Núverandi útgáfa á Google Play er útgáfa 1.1.0, en hún virkar einnig vel með AnySoftKeyboard.

Samþætting Lucene fyrir sjálfvirka útfyllingu inn í AnySoftKeyboard

AnySoftKeyboard hefur sinn eigin sjálfvirka útfylli, sem byggir á orðabók og orðtíðnilistum sem hún geymir. Sjálfvirka útfyllingin er gerð með tveimur aðskildum ferlum: þegar notandinn byrjar að skrifa er stóra (einstæðu- (e. unigram)) orðabókin notuð til að gefa tillögur. Þegar notandinn hefur lokið einu orði, annaðhvort með því að slá inn bil eða velja tillögu frá lyklaborðinu, hefst aftur ferli tillaga sem spáir fyrir um næsta orð byggt á því sem var ritað. Tillögur næsta orðs reiða sig aðeins á orðasamsetningar sem notandinn hefur þegar ritað og eru geymdar sér á tækinu. Þessi nálgun hefur ákveðna galla: Í fyrsta lagi er stærð þessarar orðabókar takmörkuð við 900 orð (miðað við t.d. 100–200 þúsund í kjarnaorðabókum tungumálapakkanna), og í öðru lagi tekur smá tíma fyrir þessar orðabækur að fyllast, þannig að í byrjun eru engar eða fáar tillögur gerðar.

Við útfærðum aðferð til að leggja til næsta orð í AnySoftKeyboard byggt á stöðuferjaldsútfærslu (e. Finite-State-Transducer, FST) Lucene. Tvístæðulisti var dreginn úr Risamálheildinni[15] og sannreyndur með því að keyra hvern tvístæðutóka á móti íslensku orðabókinni sem við höfðum þegar bætt við í íslenska tungumálapakkann fyrir lyklaborðið. Eins og er notum við 50.000 tvístæður, en listann mætti stækka ef frekari prófanir sýna fram á bættan árangur með meiri gögnum. Lucene er mjög skilvirkt, þannig að ferli þess að hlaða þessum lista inn og búa til stöðuferjald skapar enga augljósa töf þegar skipt er yfir í íslenska lyklaborðið. Leit að tillögum er einnig virkilega hraðvirk. Taka skal fram að algengustu tvístæður eru að sjálfsögðu mjög stutt orð (*til að, af því, sem er, o.s.frv.*). Stöðuferjaldið er ekki samsett með algildum þyngdum eða tíðnum, en atriðum er skipt í *fötur.* Forvinnsla tvístæðulistanna samanstendur því af greiningu á tíðnidreifingu tvístæðna og viðeigandi skiptingu þeirra í fötur, með lögmál Zipf til hliðsjónar (þ.e. nokkrar tvístæður eru mjög tíðar, og flestar tvístæður eru sjaldséðari). Því munu föturnar með hæstu stuðlunum hafa færri atriði en lægstu. Við leit er atriðum raðað fyrst eftir nákvæmri samsvörun, svo eftir stuðli fötu og að lokum í stafrófsröð.

Við viljum ekki missa sértæka spá notenda fyrir næstu orð, svo við höldum hefðbundna ferlinu við geymslu notaðra tvístæðna í 900 orða tvístæðuorðabók notanda. Við hverja hleðslu er þessi skrá sameinuð stóra tvístæðulistanum, og notkunarteljarar eru notaðir til að hækka stuðul fötunnar í stöðuferjaldinu. Tvístæður frá notandanum sem ekki eru í stóra listanum er bætt við með notkunartalningu þeirra sem stuðull fötu. Því má keyra uppfærslu á stóra tvístæðulistanum hvenær sem er (sem fylgir sama sniði fötu) án þess að missa gögn um notkun.

Þar sem þetta skref er grundvallarstefnubreyting frá fyrri nálgun og er ekki hægt að stækka auðveldlega fyrir fyrir öll önnur tungumál sem AnySoftKeyboard styður eins og er bjuggum við til okkar eigin kvísl af verkefninu og innleiddum þessa lausn eingöngu þar.

[15] Við notuðum 2021 útgáfu RMH, drógum út gögn sem höfðu CC BY 4.0-leyfið og slepptum Alþingi og lagatextum, sem skildi þá eftir að mestu frétta- og miðlatexta auk Wikipedia.

L9 – Málrýni í ritvinnslukerfum

Skýrsla: Háskóli Íslands
Aðilar: Háskóli Íslands

Markmið

Að eiga heildarsamþættingu af íslenska málrýninum í Google Docs, MS Office og MacOS. Verkþátturinn felur einnig í sér notendaprófanir háskólanema á Google Docs-samþættingunni.

Varða 8 – lýsing

M8: Notendaprófunum Google-Docs viðbótar lokið og fyrsta útgáfa MS/MacOS Office-samþættingar tilbúin.

Varða 9 – lýsing

M9: Google Docs-viðbótin bætt í samræmi við niðurstöður notendaprófana og lokaútgáfa Google Docs og MS/MacOS Office-viðbóta hafa verið gefnar út á CLARIN.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.

- Lokaútgáfur Google Docs- og MS/MacOS Office-viðbótanna eru tilbúnar.
- Frumkóði Google Docs-viðbótarinnar er aðgengilegur á CLARIN (http://hdl.handle.net/20.500.12537/288) og GitHub (https://github.com/hinrikur/Yfirlestur-Docs).
- Frumkóði MS/MacOS Office-viðbótarinnar er aðgengilegur á CLARIN (http://hdl.handle.net/20.500.12537/289) og GitHub (https://github.com/hinrikur/Yfirlestur-Word).
- Viðbótin samanstendur af viðmóti sem sýnir merktar villur í íslenskum textaskjölum sem notandinn hefur gagnvirkt aðgengi að.
- Viðbótin er þróuð og prófuð með opnu forritaskilunum Yfirlestur.is, sem þróuð eru af Miðeind, en er ekki gefin út með þeim sem sjálfvalinn endapunktur.

Skýrsla

1. Yfirlit
Stærsta verkið fyrir vörður M8–M9 var framkvæmd notendaprófana á Google Docs. Auðveldara var að bæta við eiginleikum eftir endurgjöf notenda.

2. Notendaprófanir fyrir Google Docs-viðbót

2.1 Skipulag
Upphaflega var gert ráð fyrir því að notendaprófanir á Google Docs-viðbótinni yrðu opinn, megindlegur prófunarfasi, helst með nokkur hundruð þátttakendum. Hins vegar var ákveðið að gera þetta ekki, bæði vegna takmarkaðs notkunarsviðs viðbótarinnar og vegna tafa við skipulag og framkvæmd notendaprófana. Þess í stað var smærri, eigindleg nálgun tekin.

Söfnun endurgjafa fór fram í formi spurningalista, sem prófararnir fylltu út eftir að hafa notað viðbótina í a.m.k. nokkrar mínútur.

2.2 Spurningalisti
Spurningalistinn byrjar á almennum spurningum um notandann. Þetta eru t.d. spurningar um aldur, núverandi starf eða námsleið, hvaða tæki og umhverfi þeir nota venjulega til að meðhöndla texta á íslensku og álit þeirra á innbyggðri villuleiðréttingu og meðhöndlun Google Docs. Notendurnir voru beðnir um að svara þessum spurningum áður en þeir prófuðu viðbótina.

Notendur voru spurðir um almennar hliðar upplifunar þeirra á viðbótinni, m.a. hvort þeir gátu notað hana án þess að lenda í forritsvillum, hvort notkun viðbótarinnar væri augljós og auðveld og hvort þeir myndu mæla með henni til samstarfsaðila í núverandi mynd. Eftir þetta voru notendurnir beðnir um að gefa einkunn fyrir tiltekin atriði viðbótarinnar á skalanum 1 til 5. Þessi atriði voru:

- Stærð texta
- Letur notað
- Litaskema
- Myndir og tákn notuð
- Stærð viðmóts
- Leiðbeiningar til notanda
- Villuskilaboð

Fyrir utan einkunnagjöf var hverjum þátttakanda gefinn kostur á að gefa aukaendurgjöf fyrir hvert atriði á textaformi. Að lokum voru notendur beðnir að gefa heildarlýsingu á upplifun sinni á viðbótinni, og nefna einn eiginleika sem þeir myndu vilja sjá í endanlegri afurð.

2.3 Framkvæmd
Til að veita aðgengi að Google Docs-viðbótinni var upprunalega áætlunin að hýsa viðbótina til bráðabirgða á Google Marketplace, þar sem prófendur gætu nálgast og sett upp viðbótina með gefnum hlekk. Þetta virkaði ekki á endanum vegna strangra reglugerða Google varðandi viðbætur fyrir Google Docs, sem hefði getað tekið vikur að standast samkvæmt svörum frá starfsfólki Google. Til að lágmarka frekari tafir var síðri nálgun valin, sem virkaði þó, þar sem hverjum prófara var bætt handvirkt við þróunarskjal viðbótarinnar, og þannig var komist hjá vandamálinu með hýsingu á Google Marketplace.

Að lokum voru 23 þátttakendur skráðir til formlegra notendaprófana, og af þeim tóku 15 þátt á þann hátt sem lýst var að ofan og skiluðu viðeigandi niðurstöðum.

2.4. Endurgjöf

Meðalaldur þátttakenda var 30, og flestir þeirra voru nemendur við Íslensku- og menningardeild Háskóla Íslands. Þar sem eigindleg nálgun var notuð fyrir notendaprófanir var besta endurgjöfin í formi textasvara frá notendunum. Samantekin gefa þessi textasvör nokkuð góða mynd af notkun hvers notanda, með samhengi sem er mikilvægt fyrir endursköpun villna og villuleit. Þetta ógildir þó ekki einkunnir fyrir tiltekin atriði viðmóts viðbótarinnar, sem eru birtar að neðan.

Atriði viðmóts	**Meðaleink. (1-5)**
Stærð texta	3,00
Letur notað	4,27
Litaskema	4,07
Myndir og tákn notuð	3,67
Stærð viðmóts	3,20
Leiðbein. til notanda	3,67
Villuskilaboð	3,40

Tafla 1: Meðaleinkunnir fyrir tiltekin atriði viðmóts frá endurgjöfum notenda.

Þessar einkunnir gáfu góða yfirsýn yfir hvaða atriði viðmóts viðbótarinnar þörfnuðust mestrar vinnu á þeim tíma, en þær gáfu engar upplýsingar um ástæður einkunna. Eins og nefnt var að ofan gátu notendurnir gefið aukaendurgjöf fyrir ákveðin atriði viðbótarinnar, fyrir utan einkunn milli 1 og 5. Í ljós kom að þetta var sá hluti endurgjafarinnar sem gat best stýrt áframhaldandi þróun Google Docs-viðbótarinnar. Þó að stundum hafi lýsingar stangast á („Of ítarlegar“, „Of óljósar“ og „Passlegar“ voru notuð til að lýsa leiðbeiningum fyrir notendur viðbótarinnar), voru ólíkar skýringar notenda fyrir einkunnum sínum ómissandi í vinnunni sem kom á eftir.

Heilt yfir voru notendurnir ánægðir með almenna virkni viðbótarinnar. Tæknilegri samþættingu forritaskila Yfirlestur.is var almennt vel tekið, þó að hún væri ókláruð. Helsta áhersla almennrar endurgjafar notendanna sneri að augljósum villum í notendaviðmótinu, og beinum samskiptum notenda við viðbótina. Í einu stöku tilfelli gat notandi ekki keyrt viðbótina yfir höfuð, sem síðar kom í ljós að var vegna rangra stillinga notandans. Þrátt fyrir þetta sögðust flestir þátttakendur myndu mæla með Google Docs-viðbótinni, jafnvel með viðmóti í óklæraðri mynd.

Mikið áhyggjuefni varðandi notendaprófanafasa fyrir Google Docs-viðbótina í heild sinni var að notendurnir myndu einblína á villuleiðréttingarvirkni frá forritaskilum Yfirlestur.is og ekki viðbótina sjálfa. Leiðbeiningar okkar til prófendanna tóku að mestu á þessu, og aðeins örfá svör einstaklinga minntust á efni villuleiðréttinganna sjálfra.

3. Þróun viðbótar

3.1 Google Docs
Eftir nokkur bakslög og tafir fyrir vörðu M7 var staðan á Google Docs-málrýniviðbót fyrir íslensku nokkuð góð við M8. Einhverjar tafir fylgdu skipulagi og framkvæmd notendaprófunarfasa, en það takmarkaði ekki endanlega söfnun gagna.

Eftir vörðu M8 og þegar nær dró M9 var megináherslan lögð á að bæta tillögum notenda við og laga böggavillur sem bent var á. Þetta endaði á að taka lengri tíma en búist var við, sem hafði neikvæð áhrif á endanlega útgáfu MS Office-útfærslunnar vegna tímaþröngar.

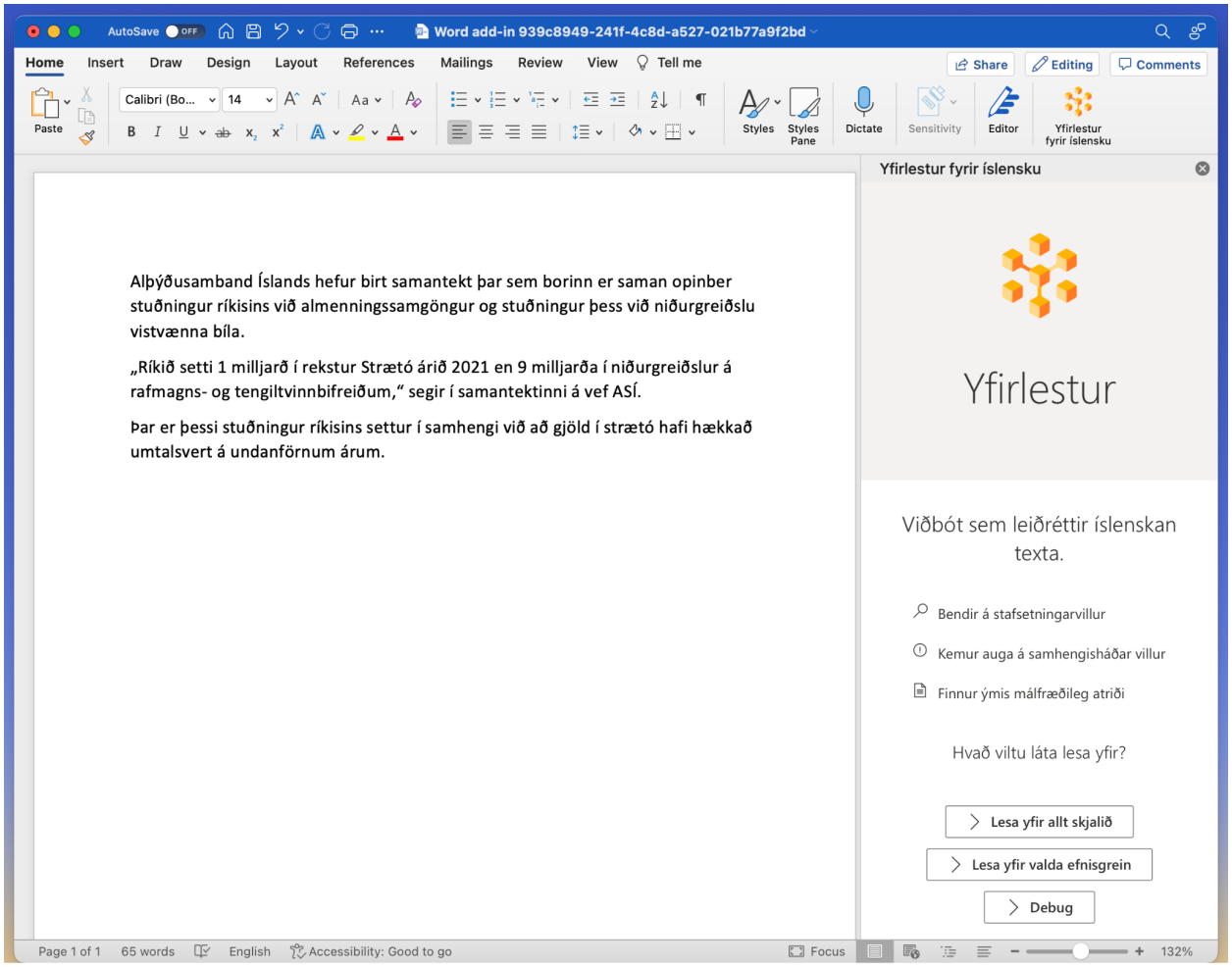
3.2 MS/MacOS Word
Eins og áætlað var í upprunalegum vegvísi fyrir verkefnið í M7 var unnið með Node.js, sem bjó til viðbót í formi vefforrits, sem er keyrt af MS Office-forritinu. Uppsettning einfaldrar MS Office-viðbótar á þennan hátt er mjög aðgengileg og studd af formlegri skjölun Microsoft.

Stuðningur þvert á kerfi var tryggður með því að taka nálgunina sem lýst er að ofan, þar sem vefbyggðar viðbætur virka bæði í MS Office-uppsetningum á MS Windows og MacOS.

Upprunalega var möguleiki þess að útfæra íslenska málfræðileiðréttingu í flestan MS Office-hugbúnað hafður opinn, með MS Word sem byrjunarreit þróunar. Þetta breyttist í það að MS Word varð eina forritið sem viðbótin var útfærð fyrir (hér eftir vísað til sem MS Word-viðbót) vegna tímaþröngar meðal annarra ástæðna.

Þar sem MS Word-viðbótin hafði engan formlegan notendaprófunarfasa var endurgjöfin úr notendaprófunum Google Docs notuð, þar sem við átti, fyrir frekari þróun MS Word-viðbótarinnar. Þetta var aðeins framkvæmt að hluta; Þróun Google Docs-viðbótarinnar eftir notendaprófanir tók lengri tíma en búist var við. Því fór minni tími og athygli í útfærslur endurgjafar og almenna villuleit í endanlega útgáfu MS Word-viðbótarinnar.

3.3 Núverandi staða
Báðar viðbótirnar virka og hafa verið undirbúnar fyrir notkun í raunheimi. Dæmi úr MS Word-viðbótinni má sjá á mynd 1. Viðbæturnar er þó ekki hægt að setja upp í viðeigandi ritvinnsluforritum fyrir almenna notkun, heldur aðgengilegar í formi kóða á viðeigandi hirslum á CLARIN og GitHub. Þetta gefur þriðju aðilum kost á að nota afurðir þessa verkefnis til að búa til og hýsa viðbæturnar eftir vild.



Mynd 1. MS/MacOS Word-viðbótin keyrð á MacOS-tölvu.

Taka skal fram að þótt viðbæturnar hafi verið þróaðar með forritaskilum Yfirlesturs frá Miðeind sem bakenda málfræðileiðréttinga eru þær ekki gefnar út með þetta í huga.

L11 – Villumódel fyrir ljóslestur

Skýrsla: Stofnun Árna Magnússonar í íslenskum fræðum
Aðilar: Stofnun Árna Magnússonar í íslenskum fræðum

Markmið

Að eiga sérhæft villulíkan/sérhæfð villulíkön fyrir texta sem hefur verið skannaður með ljóslestri (e. Optical Character Recognition software, OCR). Villurnar sem slíkar einingar taka á eru villur við stafgerð (e. digitizing) bókstafa frekar en stafsetningar- og málfræðivillur.

Slíkur leiðréttingarhugbúnaður nýtist kannski ekki í daglegri vinnu fyrirtækja og stofnana, en á sérstaklega við fyrir gríðarstór verkefni eins og stafræna safnið timarit.is (https://timarit.is/about?lang=en), og Landsbókasafn Íslands og öll verkefni sem snúa að færa texta á pappír yfir í stafrænt form, t.d. mögulegt „bókahilluverkefni“ fyrir íslensku (sjá norska verkefnið „bokhylla“ við Landsbókasafn Noregs).

Helsti galli slíkra stafrænna textasafna er að ekki er auðvelt að leita í þeim, sem minnkar notagildi þeirra. Þessi verkþáttur miðar að því að þróa villulíkön sem bæta leit í stafrænum textum með ljóslestri.

Á fyrri helmingi þriðja árs munum við fá innsýn í málrýni með tauganetum (sjá undirverkefni L14). Að auki mun yfirstandandi verkefni innan Stofnunar Árna Magnússonar (utan máltækniáætlunarinnar), sem lýkur við enda árs 2021, skila gögnum um gerðir villna sem núverandi hugbúnaður til ljóslesturs gerir. Á seinni helmingi þriðja árs munum við byggja á þessari reynslu og vinna með tengdum verkefnum, eins og ljóslestursleiðréttingu fyrir finnsku.

Varða 8 – lýsing

M8: Aðferðafræði komið á.

Varða 9 – lýsing

M9: Hugbúnaðarpakki sem les OCR-úttak, leiðréttir OCR-villur og skrifar leiðréttan texta á sama sniði. Einföld forritaskil fyrir vefþjónustu verða innleidd sömuleiðis, ef slíkt nýtist endanlegum notendum.

Afurðir

Öllum markmiðum hefur verið náð. Talið var að notendur með þörf fyrir þetta tól myndu keyra það á miklu magni gagna. Því var þörfin á vefþjónustu með forritaskilum talin mjög lítil, og við lögðum áherslu á góða skjölun til að setja tólið upp á eigin vél. Kóðann, líkön og tiltæk gögn má finna á CLARIN á http://hdl.handle.net/20.500.12537/271.

Skýrsla

Niðurstöðurnar eru hófsamar og má bæta frekar. Líkönin voru þjálfuð á 50.000 línum af ljóslesnum (e. OCR) texta og handvirkt leiðréttum hliðstæðum þeirra, ásamt ca. 900.000 línum úr RMH, með viðbættum gervivillum úr ljóslesnum/handvirkt leiðréttum hliðstæðum, vegna takmarkaðra tiltækra gagna og hve tímafrekt er að afla þeirra. Áhrif þess að bæta við um 5.000 línum af ljóslesnum texta eru sýnd í lok þessarar skýrslu.

Ljóslesnu textarnir eru að mestu frá 19. öld, og textarnir úr RMH eru frá ýmsum áratugum 20. aldarinnar. Þessir textar eru notaðir fyrir þjálfun, yfirferð og prófanir. Tvö líkön eru afhent, annað þjálfað með PyTorch og tilreitt með WordPiece, og hitt þjálfað með fairseq og tilreitt með SentencePiece.

Ályktun er hægt að keyra með öðru líkaninu, eða báðum, ásamt valkvæðri leit í orðasafni (sem bætir niðurstöðurnar á nútímagögnum en lækkar þær á 19. aldar gögnum) í BÍN og Ritmálssafni. Ef ályktun er keyrð með báðum líkönum reyna þau hvert um sig að leiðrétta ljóslesnu línurnar. Betri línan er valin eftir því hvort línan inniheldur fleiri orð sem eru til staðar í fyrrnefndu orðasöfnunum tveimur.

Niðurstöður eru sýndar að neðan. ERR er lækkun villutíðni (e. error rate reduction).

	PyTorch	fairseq	samanlagt
chrF	96,84	96,35	96,85
chrF ERR	41,25	32,22	41,48
BLEU	98,44	98,52	98,45
BLEU ERR	44,45	47,24	44,74

Tafla 1. 19. aldar prófunargögn, leit í orðasafni óvirk.

	PyTorch	fairseq	samanlagt
chrF	96,80 (-0,04)	96,35 (+0,0)	96,84 (-0,01)
chrF ERR	40,69 (-0,56)	32,23 (+0,01)	41,28 (-0,2)
BLEU	98,43 (-0,01)	98,49 (-0,03)	98,45 (+0,0)
BLEU ERR	43,77 (-0,68)	45,92 (-1,32)	44,74 (+0,0)

Tafla 2. 19. aldar prófunargögn, leit í orðasafni virk (breytingar vegna leitar í sviga).

	PyTorch	fairseq	samanlagt
chrF	96,83	96,58	96,85
chrF ERR	33,79	28,63	34,19
BLEU	98,45	98,54	98,46
BLEU ERR	31,92	36,14	32,27

Tafla 3. 20. aldar prófunargögn, leit í orðasafni óvirk.

	PyTorch	fairseq	samanlagt
chrF	96,85 (+0,02)	96,62 (+0,04)	96,87 (+0,02)
chrF ERR	34,12 (+0,33)	29,37 (+0,73)	34,52 (+0,33)
BLEU	98,46 (+0,01)	98,57 (+0,03)	98,47 (+0,02)
BLEU ERR	32,48 (+0,56)	37,17 (+1,03)	32,83 (+0,56)

Tafla 4. 20. aldar prófunargögn, leit í orðasafni virk (breytingar vegna leitar í sviga).

PyTorch-líkanið nær örlítið betri árangri með 5.348 línum bætt við af rétt ljóslesnum texta/handvirkt leiðréttri hliðstæðu hans. Þessi prófun var aðeins keyrð á nútímatextunum og PyTorch+WordPiece-líkaninu, og niðurstöður eru birtar í töflunni að neðan. Tölurnar í svigum tákna bætingar sem orsakast af viðbættum línum.

	PyTorch	PyTorch+LL
chrF	96,88 (+0,05)	96,90 (+0,05)
chrF ERR	34,89 (+1,10)	35,31 (+1,19)
BLEU	98,54 (+0,09)	98,56 (+0,10)
BLEU ERR	35,89 (+3,97)	36,73 (+4,25)

L14 – Málrýni með djúpum tauganetum

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Markmið

Meginmarkmið þessa verkþáttar er að þjálfa tauganet til að þýða „ófullnægjandi“ íslenska texta yfir á réttari texta. Annmarkarnir geta verið margra flokka, þar á meðal textar með „dæmigerðum“ stafsetningar- og málfræðivillum sem fullorðnir móðurmálhafar gera, textar skrifaðir af lesblindum, textar skrifaðir af börnum, og textar eftir fólk sem hefur annað móðurmál en íslensku.

Málrýnar byggðir á tauganetum munu líklega skila betri afköstum en reglubyggðir rýnar (sbr. L7), og viðbúið er að þeir ráði betur við hærri villutíðni, t.d. í mjög ófullnægjandi textum. Aftur á móti verður geta þeirra til að útskýra leiðréttingar takmörkuð. Þeir henta því vel til að leiðrétta mikið magn af texta og þar sem ekki er þörf á útskýringum.

Þetta undirverkefni byggir á reynslu sem safnast hefur saman í undirverkefnum um vélþýðingar. Við höfum komist að því að vélþýðingaafkóðarar (e. MT decoders) sem búa til íslenskar setningar úr enskum gjafatexta (e. source text), með hjálp vel þjálfaðs mállíkans, eru nokkuð góðir að búa til rétta íslenska stafsetningu og málfræðilega rétta texta. Við erum því bjartsýn á að net sem þýðir frá „slæmri“ íslensku yfir á „góða“ íslensku gæti náð góðum árangri, að því gefnu að kóðarahlutinn af netinu sé þjálfaður með viðeigandi hætti, þar á meðal með vandlega stilltu brottfalli (e. dropout) og með greypingum sem eru að hluta á bókstafsstigi (e. character level) (öfugt við orðflísar/BPE eingöngu).

Netin verða þjálfuð fyrir hvert villuóðal (almennt, lesblindra, barna, útlendinga) með blöndu af raunverulegum gögnum sem hefur verið safnað í öðrum undirverkefnum og gervigögnum sem verða búin til með því að líkja eftir villumynstrum sem koma fyrir í raunverulegu gögnunum.

Aðferðir og reynsla úr þessu undirverkefni gæti orðið til þess að bæta niðurstöður við umritun texta með talgreini (sbr. undirverkefni H7), sem má flokka sem eina gerð af þýðingu frá „ófullnægjandi“ íslensku til „góðrar“ íslensku.

Aukamarkmið þessa undirverkefnis er að þjálfa flokkara ofan á leiðréttingarkóðarann. Þessi flokkari úthlutar líkum á því að inntakssetning sé málfræðilega rétt. Ef setningin er mjög líklega rétt, þá þarf ekki að þátta hana og athuga með (hægari) reglubyggða rýninum, og þannig er ferlinu flýtt – sem er sérstaklega mikilvæg fyrir lengri skjöl, svo sem skýrslur, lokaritgerðir og bækur.

Varða 8 – lýsing

M8: Aðferðir til að búa til gervivillumálheildir tilbúnar; þjálfun á „þýðingar“-neti hafin. Tauganetsflokkari fyrir valda villuflokka tilbúinn og gefinn út á CLARIN. Tilraunir með sérhæfðar villlumálheildir hafnar.

Varða 9 – lýsing

M9: Tauganet fyrir málrýni mismunandi flokka „ófullnægjandi“ texta tilbúið og gefið út á CLARIN.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.

- Aðferðir til að búa til gervivillumálheildir eru tilbúnar og þeim lýst að neðan. Kóði er aðgengilegur á https://github.com/mideind/GreynirSeq/tree/main/src/greynirseq/noising.
- Flokkari fyrir almenna villuflokka hefur verið þróaður og gefinn út á CLARIN á http://hdl.handle.net/20.500.12537/183.
- Flokkari hefur verið innleiddur í GreynirCorrect[16] sem valkvætt ákvæði fyrir forsíun inntakssetninga. Flokkarinn sem notaður er hefur verið gefinn út á CLARIN á http://hdl.handle.net/20.500.12537/256.
- Tauganet sem breytir „slæmum“ texta í „góðan“ texa hefur verið gefið út á CLARIN á http://hdl.handle.net/20.500.12537/255 og Hugging Face á https://huggingface.co/mideind/yfirlestur-icelandic-correction-byt5.

Skýrsla

Málrýnilíkan var þjálfað og metið fyrir íslensku. Það er bitastigsþýðingarlíkan (e. byte-level) þjálfað á samhliða gervigögnum og raunverulegum villumálheildum, og þýðir lakan íslenskan texta yfir á rétta íslensku með frábærum árangri.

Gervivillugögn fyrir leiðréttingu málræðivillna
Þó að iceErrorCorpus[17] og sérhæfðu málheildirnar sem fylgja nýtist vel við fínstillingu og mat tauganeta sem greina og leiðrétta íslensku þarf mikið meira magn gagna til að þjálfa transformer-net að fullu fyrir slíkt verk.

Stór gervivillumálheild var búin til með því að hanna villugjafa (e. error generator) sem líkir eftir algengum málfræði- og stafsetningarvillum í íslenskum texta. Villurnar sem verða til eru m.a. villur sem algengt er að almenningur, lesblindir, annarsmálshafar og börn geri. Þessar villur eru m.a.:

[16] Sjá https://github.com/mideind/GreynirCorrect/
[17] Sjá https://github.com/antonkarl/iceErrorCorpus

- Rangt málfræðilegt fall í nafnorðum
- Framsöguhætti breytt í viðtengingarhátt
- Þágufallshneigð (e. dativitis)
- Bilum bætt inn í orð
- Bilum eytt milli orða
- Kommum eytt úr setningum
- Orðaröð breytt
- Orð tvítekin
- Rétt stafsettum orðum skipt út fyrir þekktar stafsetningarvillur (af ýmsum listum algengra stafsetningar- og málfræðivillna)
- Stöfum skipt út samkvæmt einföldum reglum (eins og *y/i*, *ing/íng*, *n/nn*, *ó/'o*, *ýi/ýji*)
- Bókstafur með broddum eða broddar fjarlægðir (*a/á, I/Í*)
- Stafir endurteknir
- Handahófskennt suð stafa

Við sköpun villna er hægt að stilla hlutfall villna í hverri setningu, auk hlutfalls milli villutegunda.

Við bjuggum til 35 milljón línur samraðaðra gagna, með hæsta villuhlutfalli þar sem suð er til staðar í flestum setningum, með því að nota texta síaða úr Risamálheildinni (RMH) sem uppsprettu af fjölbreyttri og nokkuð réttri íslensku. Úr þessu urðu til verulega suðuð þjálfunardæmi, eins og:

Uppruni (gervivillum bætt við)	**Markmið (upprunalegur texti frá RMH)**
Þú skalt hafa þetsa frélssí fræam yfir alla aðra aðyla í þjóðfélagið, eð verðaað ptttaka ými tll, ti tíllíti eða verða að berjast á ýmsan mátum.	Þú skalt hafa þetta frelsi fram yfir alla aðra aðila í þjóðfélaginu, sem verða að taka ýmis tillit eða verða að berjast á ýmsan máta.
En En meginhluuti auðs Ólafs Ólafssonar liggur í Kaupþíngi bánka enda Ólafur einn a8 leiðani fjárfestumm í einkavæðíngju Búnaðarbankans sme síðarsámeináðíst Kaupþingi.	En meginhluti auðs Ólafs Ólafssonar liggur í Kaupþingi banka, enda Ólafur einn af leiðandi fjárfestum í einkavæðingu Búnaðarbankans sem síðar sameinaðist Kaupþingi.

Þessi gögn voru notuð sem grunnur til þjálfunar transformer-líkans fyrir villuleiðréttingu. Tilraunir með mismunandi hlutfall villna sýndu að líkönin læra mest frá verulega röngum dæmum eins og þessum að ofan.

mBART

Fyrstu tilraunir voru gerðar með forþjálfuða margmála líkaninu mBART, sem hafði verið þjálfað frekar á enskum og íslenskum textum. Þetta er orðflísatilreiðingarlíkan, og mat á ólíkum prófunarsettum sýndi að það er nokkuð viðkvæmt fyrir suði þar sem það getur ekki alltaf fundið út rétta greypingu fyrir suðaðan orðflísatóka. Það getur því gefið falskar villur í sjaldgæfum en rétt rituðum orðum þar sem annar orðflísatóki er líklegri, eins og þegar staðarheiti er skipt út fyrir annað sem hljómar svipað (*lögreglan í Bodmin -> lögreglan í Boston*).

byT5
Við beindum athygli okkar því að *byT5*[18], bæti-til-bæti útgáfu T5-högunarinnar. Þetta líkan, ólíkt flestum forþjálfuðum máll íkönum, notar ekki orðflísatilreiðara, heldur vinnur beint á bætum, og sleppur því við tilreiðingarvandamálið. Forþjálfaða líkanið (byt5-base) var fínstillt beint á 35 milljón línum gervivillugagnanna, sem notaði um mánuð af A100-skjákorti.[19] Líkanið var fínstillt frekar á ~60þ línum í gögnum málheildar með raunverulegar villur, úr íslenskum villumálheildum sem aflað var í verkþætti L2.

Mat
Líkönin voru metin með mismunandi prófunarsettum bæði gervigagna og raunverulegra gagna. Mæliaðferðir sem notaðar voru fyrir mat voru GLEU[20], sem er lítillega breytt útgáfa BLEU-mæliaðferðarinnar, en hún er notuð fyrir leiðréttingu málfræðivillna, og tíðni þýðingarvillna (e. translation error rate, TER), reiknuð út frá stafskiptafjölda (e. edit distance) milli grunnsannleiksins (e. ground truth) og tilbúins texta.

Niðurstöður

Líkan	iec-test	iec-dyslexic	iec-L2	iec-child	RÚV	synthetic-dativitis	synthetic-noise
mBART2 - gervi- og raun.gögn	0,833131	0,444523	0,462842	0,417238	0,833036	0,899545	0,965873
byT5-base - gervigögn	0,839163	0,452672	0,463352	0,425454	0,842439	**0,967322**	**0,972725**
byT5-base - gervi- og raun.gögn	**0,862917**	**0,535711**	**0,537112**	**0,523965**	**0,888931**	0,936812	0,943656

Taflan að ofan sýnir GLEU-gildi fyrir þrjú af þjálfuðu líkönunum. Í fyrsta lagi sýnir hún mBART-orðflísalíkan þjálfað á gervigögnum og villumálheildunum, í öðru lagi byT5-líkanið þjálfað eingöngu á gervigögnum, og að lokum byT5-líkanið þjálfað á gervigögnum og fínstillt frekar á raunverulegum villugögnum. Prófunarsettin sem greint er frá hér eru iceErrorCorpus prófunarsettið, setningar sem teknar voru til hliðar úr sérhæfðu málheildunum, lítið brot fréttatexta (RÚV) sem innihalda eðlilega myndaðar villur, og tvö sett setninga með gervivillum.

Niðurstöðurnar sýna stórt stökk byT5-líkansins sem fínstillt er á raunverulegum gögnum, á öllum prófunarsettum nema gervigögnum (við þessu er búist þar sem líkanið leiðréttir líka önnur hugsanleg vandamál í gagnasettunum, þar sem „grunnsannleikurinn“ er tekinn af handahófi úr RMH og hefur ekki verið yfirfarinn handvirkt og gæti því innihaldið ýmis vandamál sem líkanið hefur lært að leiðrétta).
Í lýsingu verkefnisins segir að net verði þjálfuð fyrir hvert svið leiðréttinga (almennt/lesblindu/barna/L2). Allar tilraunir okkar hafa þó sýnt að besta lausnin er eitt líkan sem getur meðhöndlað öll svið, í stað fjögurra líkana sem eru fínstillt á mismunandi gagnasettum. Afurðin er því fjölhæft líkan sem nýtur góðs af þjálfun á öllum tiltækum villugögnum.

[18] Sjá https://arxiv.org/abs/2105.13626
[19] Þetta var keyrt á HPE 8xA100 GPU-kerfi Miðeindar.
[20] Sjá https://aclanthology.org/P15-2097

Dæmi

Eins og við var búist nær líkanið frábærum árangri við að leiðrétta texta sem inniheldur margar villur, t.d. þann sem notandi með lesblindu skrifar, og leiðréttir réttilega mismunandi gerðir villna, án þess að leiðrétta um of. Hér eru nokkur dæmi um setningar úr óséðum prófnargögnum og úttak byT5-líkansins. Takið t.d. eftir fyrstu tveimur dæmunum, þar sem villa er til staðar í nær hverju orði:

Upprunaleg setning	**Leiðrétt af byT5**	**Uppruni**
Fretastofa Bilgjunar og Stöðfar tvöö fylgjist með þróun mála í allann dag.	Fréttastofa Bylgjunnar og Stöðvar tvö *fylgjast* með þróun mála í allan dag.	Gervigögn
Kvitu fiðrildinn fljua firir utan gluggan.	Hvítu fiðrildin fljúga fyrir utan gluggann.	Lesblindum álheild
Batman borgaði sig úr fangelsi svo að úlfar fékk sér eina góða máltíð aður enn hann fór í fangelsi, það var góð leðurblöku steik frá nýjum veitingastað í kambódíu.	Batman borgaði sig úr fangelsi svo Úlfar fékk sér eina góða máltíð áður en hann fór í fangelsi, það var góð leðurblökusteik frá nýjum veitingastað í Kambódíu.	Barna-málheild
Út af þessum framförum hefur verið fleiri og fleiri tekið þá ákvörðun að læra kínversku tunguna.	Út af þessum framförum hafa fleiri og fleiri tekið þá ákvörðun að læra kínversku tunguna.	L2-málheild
Ég held þetta er ekki góður tími fara í heimsókn.	Ég held þetta sé ekki góður tími til að fara í heimsókn.	L2-málheild
Sæl Þórunn, hér er viðhengi og þar er reikningu sem þú getur haft í bókhaldið, og þar finnur þú upplýsingar um innborgunar reikning, ég vona að krafa sem var stofnð á þig sé falli úr heima bankanum þínum, þú mátt láta mig vita ef svo er ekki.	Sæl Þórunn, hér er viðhengi og þar er reikningur sem þú getur haft í *bókhaldinu*, og þar finnur þú upplýsingar um innborgunarreikning, ég vona að krafa sem var stofnuð á þig sé fallin úr heimabankanum þínum, þú mátt láta mig vita ef svo er ekki.	Notandi með lesblindu

Greining málfræðivillna með tauganeti

Annað markmið þessa verkþáttar var að þjálfa setningaflokkara sem hægt er að nota sem hluta af GreynirCorrect/Yfirlestur.is til að flýta vinnslu. Við bjuggum til flokkara sem gefur upp hvort setningar eru líklega málfræðilegar eða ekki, og tengdum hann við skipanalínutól GreynirCorrect sem valkost, sérstaklega fyrir magnþýðingu, þar sem margar setningar eru líklega lausar við villur (þetta er hluti af undirverkefni L7).

Tveggja flokka flokkari var þjálfaður ofan á IceBERT[21]-líkanið á raunverulegum villumálheildum. Niðurstöðurnar á yfirferðinni og prófunargögnunum ollu vonbrigðum, með F1 0,715 mælt á prófunargögnum iceErrorCorpus. Lækkun á þröskuldi til að fækka falsk neikvæðum bætti ekki niðurstöður, og ekki heldur tilraunir okkar með að þjálfa flokkarann á gervigögnum eða blöndu af raunverulegum gögnum og gervigögnum.

Önnur nálgun var prófuð: Þjálfun margverka byT5-líkans til að gera bæði leiðréttingu og flokkun með notkun forskeyta. Þetta líkan náði sama árangri og einverka byT5-líkönin við leiðréttingu texta, og árangurinn við flokkun var aðeins betri, 0,728 F1 á prófunarsettinu.

Það virðist vera nokkuð erfitt verk að læra hvort setning inniheldur villu eða ekki, a.m.k. þegar metið er á tiltækum prófunargögnum. Textavillur geta verið mjög fjölbreyttar, allt frá því að punkt eða kommu vanti yfir í flókin málfræði- eða samhengisvandamál, eða vandamál með stíl sem eru ekki strangt til tekið villur. Hugsanlegt er líka að tauganetið greini ekki einhver flóknari vandamál sem það var ekki þjálfað til að greina, en GreynirCorrect gæti leiðrétt. Til samanburðar gerðum við tilraunir með að keyra villuleiðréttingarlíkanið og gá hvort einhverjar breytingar voru gerðar, en niðurstöðurnar voru ekki mikið betri, og þessi nálgun er augljóslega slæm fyrir afköst.

Samantekt
Villuleiðréttingarlíkanið fyrir málfræðivillur er virkilega árangursríkt í því að leiðrétta fjöldann allan af stafsetningar- og málfræðivillum, auk fjölda annarra villna sem fólk gerir. Notkun bæti-til-bæti líkans reyndist besta nálgunin, með síðari þjálfun með mjög suðmiklum gervidæmum, og svo fínstillingu á raunverulegum villumálheildum.

Líkanið hefur verið þjálfað til að leiðrétta villur gerðar af lesblindum og L2-notendum auk barna, og hefur lært að móta málfræði- og setningafræðilega rétta íslensku mjög vel. Sem möguleg verk í framtíðinni gæti líkanið haft gagn af fleiri þjálfunardæmum og stillingum fyrir forþjálfunarverkin, til að læra á og nota merkingu setningar í víðara samhengi þegar tillögur leiðréttinga eru metnar.

Auk þess að leiðrétta texta skrifaðan af fólki má einnig nota þetta líkan til að eftirvinna úttak frá talgreiningarkerfum, þar sem það hefur lært mikið um greinarmerki og setningagerð. Það má einnig nota fyrir aðra magnhreinsun texta, t.d. suðað inntak notanda, áður en önnur málvinnsluverk neðanstreymis eru leyst.

[21] A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models, Snæbjarnarson et al. 2022, http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.464.pdf

H1 – Upptökur með Samrómi

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að ljúka skilgreindri söfnun upptaka stuttra segða fyrir þjálfun talgreinis. Verkvangurinn verður notaður til að safna gögnum frá fullorðnum, unglingum, málhöfum með íslensku sem annað mál, og gagnasöfnuninni verður haldið áfram til að afla gagna frá þessum hópum. Enn fremur hefur söfnunarappið (e. crowd-sourcing app) verið aðlagað til að safna radddæmum frá fólki sem les ekki skrifaðar setningar, en þannig ætti að vera hægt að safna radddæmum frá ungum börnum.

Varða 9 – lýsing

M9: Gefa út öll tiltæk gögn úr Samrómi.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Tiltækum gögnum sem eftir voru var skipt í þrjú mismunandi gagnasett og þau gefin út.

1. Samrómur Unverified 22.07 → http://hdl.handle.net/20.500.12537/265
2. Samrómur L2 22.09 → http://hdl.handle.net/20.500.12537/263
3. Samrómur Mimic 22.09 → http://hdl.handle.net/20.500.12537/264

Skýrsla

Undirbúningur gagna
Samanlagður fjöldi segða í safni Samróms er 2.845.000. Við höfum þegar gefið út þrjú gagnasett í fyrri vörðum.

Gagnasett	Fjöldi segða
Samromur 21.05	100.000
Samromur Children 21.09	137.000
Samromur Queries 21.12	17.000

Tafla 1. Áður útgefin gagnasett.

Að auki hafa samtals 161.000 segðir verið greindar sem ógildar í sjálfvirkri og handvirkri vinnu. Þá eru eftir um 2.430.000 óútgefnar segðir. Upprunalegu gögnin koma frá fjölmörgum ólíkum tækjum, hvert með eigin hljóðnemauppsetningu. Því voru allar segðir staðlaðar á 16 kHz hljóðtökutíðni, 16 bita línulega PCM, 1 rás, *.flac-skráarform.

Segðir með engu eða of lágu hljóði fyrir talgreiningu voru síaðar burt (37.000 segðir).

Eftirstandandi segðum var skipt í þrjá flokka: Segðum safnað með endurteknu hljóði (einnig kallað *Herma* áður og *Mimic* í útgefna gagnasettinu), segðum safnað frá mælendum með annað móðurmál en íslensku (L2) og allt annað (Unverified). Sjá dreifingu að neðan.

Gagnasett	**Segðir**
Samromur Unverified 22.07	2.159.000
Samromur L2 22.09	143.000
Samromur Mimic 22.09	91.000

Tafla 2. Gagnasett gefin út við þessa vörðu.

Gagnasettin eru gefin út með lýsigögnum sem innihalda eftirfarandi.

Metadata	**Lýsing**
id	7 stafa auðkenni fyrir segðina
speaker_id	6 stafa auðkenni fyrir mælanda segðarinnar
filename	Samsett auðkenni mælanda og segðar sem myndar nafn hljóðskrár fyrir segðina
sentence	Setningin eins og hún birtist notandanum
sentence_norm	Setningin normuð
gender	Kyn sem notandinn gaf upp
age	Aldur sem notandinn gaf upp
native_language	Móðurmál sem notandinn gaf upp
dialect	Nokkrar mismunandi framburðarmállýskur voru hluti af sérstöku söfnunarátaki, ef við á er það tekið fram hér
created_at	Dagsetning upptöku
marosijo_score	Sumum segðum var gefin einkunn með þvingunarsamraðaranum (e. forced aligner) Marosijo sem birtist hér

is_valid	1 ef yfirfarin, NAN ef óljóst
duration	Tímalengd segðar
sample_rate	Hljóðtökutíðni segðar
size	Stærð á skrá segðar
user_agent	Inniheldur upplýsingar um hvers konar tæki notandinn notaði við upptöku á segðinni

Tafla 3. Lýsigögn með lýsingum.

Tölfræði gagna

	Samromur Unverified	Samromur L2	Samromur Mimic
Segðir	2.159.000	143.000	91.000
Klukkustundir	2.233	1.521	676
Mælendur	17.000	2.000	1.364
Gildar segðir	84.000	4.957	0
Segðir með Marosijo-einkunn	668.000	42.000	0

Tafla 4. Almenn tölfræði gagnasetta.

Aldurs- og kynjadreifing

	Samromur Unverified	Samromur L2	Samromur Mimic
0–19:	48,7%	58,4%	45,7%
20–39:	17,6%	14,5%	32,1%
40–59:	28,9%	24,4%	18,0%
60–79:	4,6%	1,1%	3,7%
80+	0,1%	0,6%	0,1%
Kvenkyns:	68,6%	67,9%	52,6%
Karlkyns:	30,5%	29,9%	46,9%
Önnur:	0,8%	2,2%	0,5%

H5 – Samröðun á stórum málheildum

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að safna miklu magni af hljóðupptökum frá opinberum og einkareknum stofnunum. Opinbert svið (e. public domain) gæti innihaldið upptökur af Alþingisræðum, opnum nefndarfundum og réttarhöldum. Einkarekið svið (e. private domain) inniheldur einnig gögn sem gætu verið aðgengileg, svo sem hljóðbókasöfn og símafyrirspurnir. Þessa gagngjafa þarf að bera kennsl á og afla viðeigandi leyfa.

Varða 8 – lýsing

M8: Gagna aflað og þau skilgreind með tilbúinni byrjunarútfærslu samraðara.

Varða 9 – lýsing

M9: Gögn samröðuð og þau gerð hentug fyrir talgreinisþjálfun.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.

- MAFIA Aligner: Þetta er samraðarinn sem var þróaður til að standast afurðir H5. Ef umritanir sem MAFIA fær eru ekki nákvæmar ályktar MAFIA þær með sjálfvirkri talgreiningu. Útgefinn á CLARIN á http://hdl.handle.net/20.500.12537/215.
- RADDRÓMUR Corpus: Þetta er mikilvægasta afurð H5. Þetta er málheild sem samanstendur af útvarpshlaðvörpum, að mestu frá RÚV. Það voru tiltæk textahandrit fyrir hvert hlaðvarp svo mögulegt var að samraða handritum sjálfvirkt með talskránum með notkun MAFIA Aligner. Útgefin á CLARIN á http://hdl.handle.net/20.500.12537/286.

Skýrsla

Gagnaöflun

Gögnin sem notuð voru til að búa til málheildina Raddróm samanstanda af nokkrum útvarpshlaðvörpum, samtals 159 klukkustundir og 46 mínútur, dreift yfir 185 hljóðskrár. Eftir vinnslu allra inntaksgagnanna er heildarstærð Raddrómsmálheildarinnar 49k09m, dreift yfir 13.030 hljóðskrár. Ítarlegri listi yfir uppruna hvers hlaðvarps er eftirfarandi:

Rokkland (20 þættir / 36k23m)
Höfundur: Ólafur Páll Gunnarsson
Hlaðvarp/Útvarpsþáttur á vegum RÚV.

Á tónsviðinu (29 þættir / 26k47m)
Höfundur: Una Margrét Jónsdóttir
Hlaðvarp/Útvarpsþáttur á vegum RÚV.

Í ljósi sögunnar (121 þættir / 80k36m)
Höfundur: Vera Illugadóttir
Hlaðvarp/Útvarpsþáttur á vegum RÚV.

Neðanmáls (4 þættir / 1k35m)
Höfundar: Elísabet Rún Þorsteinsdóttir og Marta Eir Sigurðardóttir.
Elísabet Rún Þorsteinsdóttir og Marta Eir Sigurðardóttir.

Leikfangavélin (14 þættir / 14k48m)
Höfundur: Atli Hergeirsson
Sjálfstætt hlaðvarp/Útvarpsþáttur

Samröðunartól

MAFIA er skammstöfun á Match-Finder Aligner. Samraðarinn MAFIA er hugbúnaðartól hannað til að búa sjálfvirkt til talgreiningarmálheild úr talskrám, ásamt textahandritum sem spegla það sem sagt er í talskránum. Ef textinn í handritunum er ekki algjörlega nákvæmur ályktar MAFIA umritun með sjálfvirkri talgreiningu.

Inntaksgögn

MAFIA tekur við tal- og textaskrám. Talskrárnar verða að vera á sniðinu WAV á 16khz@16bit/mono. Búist er við að handritsskrár og talskrár deili sameiginlegri rótarmöppu. Báðar verða að hafa saman nafn, en ólíkar endingar: „txt“ fyrir handrit og „wav“ fyrir talskrár. Fjöldi skráarpara, þar sem handrit og tal fara saman, má vera hver sem er.

Hljóðlíkan

Í fyrstu vann MAFIA-samraðarinn með hljóðlíkani sem búið var til fyrir verkþátt H16. Þetta líkan var búið til í NeMo[22] og gefið út á CLARIN í júní 2022[23]. Tafla 1 sýnir árangur þessa líkans, prófað með ólíkum íslenskum málheildum og með mismunandi mállíkönum.

Corpus		Word Error Rate (WER) with H16 Model		
	Portion	No LM	LM 5GB	LM 10GB
Althingi	Dev	62.291%	42.338%	40.947%
Malrómur	Dev	48.0%	29.207%	28.824%
Samrómur Children	Dev	44.567%	26.267%	24.181%
Samrómur 21.05	Dev	35.6%	19.9%	18.45%
Althingi	**Test**	61.66%	42.146%	40.707%
Malrómur	**Test**	48.0%	29.207%	28.824%
Samrómur Children	**Test**	56.554%	39.635%	38.431%
Samrómur 21.05	**Test**	38.968%	22.414%	20.841%

Tafla 1. Árangur NeMo-líkansins úr H16.

Eins og sjá má í Töflu 1 nær líkanið betri árangri með lesnum gögnum og verri með málheild Alþingis (Althingi), sem er ekki lesin. Þessi árangur gæti valdið vandræðum í H5 þar sem gögn sem þarf að samraða eru sjálfsprottið tal.

Endurbætur hljóðlíkansins

Til að bæta hljóðlíkönin fyrir MAFIA-samraðarann voru 875 klst. íslenskra gagna notaðar til að búa til fínstillt nýtt líkan úr ensku forþjálfuðu líkani frá NVIDIA. Eins og stendur er þetta besta hljóðlíkanið sem við höfum sem byggt er á NeMo. Þetta nýja líkan nær betri árangri en þau sem við bjuggum til fyrir H16, jafnvel á á prófunarsettunum Málrómur, Althingi og Samrómur Children. Nýja líkanið er byggt á Jasper-höguninni[24], og byggist upp af 10 bálkum og 4 endurtekningum hvert; þess vegna nefndum við það JasperFT10x4 („FT“ fyrir „fine-tuned“, fínstillt).

Tafla 2 sýnir árangur JasperFT10x4-líkansins með sömu málheildum og í Töflu 1. Eins og sjá má eru niðurstöður í Töflu 2 betri en í Töflu 1, sérstaklega fyrir Althingi og Samrómur Children.

[22] NVIDIA NeMo: https://developer.nvidia.com/nvidia-nemo
[23] Samrómur NeMo forskrift 22.06: http://hdl.handle.net/20.500.12537/228
[24] Fyrir frekari upplýsingar um Jasper sjá: https://arxiv.org/pdf/1904.03288.pdf

Word Error Rate (WER) with JasperFT10x4 (NEW MODEL)				
Corpus	**Portion**	**No LM**	**LM 5GB**	**LM 10GB**
Althingi	Dev	24.121%	21.104%	21.376%
Malrómur	Dev	18.5%	19.699%	18.13%
Samrómur Children	Dev	17.853%	21.639%	14.88%
Samrómur 21.05	Dev	19.244%	22.946%	17.456%
Althingi	**Test**	23.913%	21.248%	21.237%
Malrómur	**Test**	18.5%	19.699%	18.13%
Samrómur Children	**Test**	28.257%	29.41%	26.146%
Samrómur 21.05	**Test**	21.647%	25.47%	19.692%

Tafla 2. Árangur NeMo-líkansins JasperFT10x4.

Raddrómsmálheildin

Opinbert nafn þessarar málheildar er „Raddrómur Icelandic Speech 22.09“ og hún hefur verið gefin út á CLARIN.

Málheildin Raddrómur er ætluð fyrir talgreiningu og byggist upp af útvarpshlaðvörpum sem fengin voru að mestu frá RÚV[25]. Hlaðvörpin voru valin vegna þess að þeim fylgja textahandrit sem samsvara að mestu því sem sagt er í þættinum. Eftir sjálfvirka bútun þáttanna voru umritanirnar ályktaðar með skriftunum ásamt þvingaðri samröðunaraðferð. Bæði bútun og ályktun er gerð með MAFIA-samraðaranum.

Til að skilja frá umritanir með færri viðbúnum mistökum var gæðamælingu „MAFIA Score“ bætt við lýsigögnin sem fylgja málheildinni. MAFIA Score nálægt núll bendir til betri gæða umritunarinnar.

[25] https://www.ruv.is/

H9 – Talgreining í snjallsímum

Skýrsla: Tiro ehf.
Aðilar: Tiro ehf.

Markmið

Að þróa talgreinisviðmót fyrir Android-lyklaborð. Á fyrri hluta þriðja árs verður miðpunktur vinnunnar þróun á frumgerð Android-lyklaborðs með einföldu talviðmóti. Unnið verður í samstarfi við L8 - Málrýni í snjalltækjum (3 mánuðir úr H9 og 3 mánuðir úr L8). Á þriðja ári munum við kanna aðferðir til að bæta talgreini og málrýni við snjallsímalyklaborð, bæði sem vefþjónustu og sem einingu í tæki. Við munum búa til frumgerð af lyklaborði fyrir Android sem getur breytt rödd í texta (e. voice-to-text facility) með því að nota vefþjónustu. Vegna minnkaðra tilfanga munum við einbeita okkur að því að gefa lausnina út fyrir Android, en einnig skilgreina kröfur fyrir iOS fyrir framtíðarþróun (sjá einnig L8).

Varða 8 – lýsing

M8: Gefa út alfa-útgáfu Heyra, sem er útfærsla á talgreiningarþjónustu Android (e. The Android Speech Recognition Service), með því að nota vefþjónustu á Google Play.

Varða 9 – lýsing

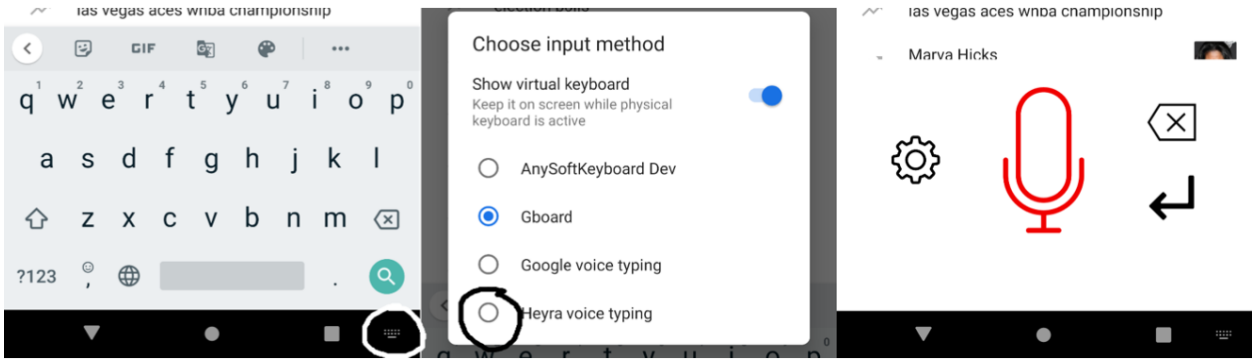
M9: Útgáfa 1.0 af Heyra gefin út á Google Play. Taltextun (e. speech-to-text) útfærð í AnySoftKeyboard með útfærslu innsláttaraðferðar (e. input method implementation).

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Alfa-útgáfa Heyra, sem inniheldur bæði talgreiningarþjónustu fyrir Android og ætlunarmeðhöndlara (e. intent handler) fyrir talgreiningu, var gefin út á Google Play Store til opinna prófana við enda M8. Heyra 1.0, sem inniheldur líka útfærslu innsláttaraðferðar fyrir talgreiningu, hefur nú verið gefið út á Google Play Store[26]. Kóðinn er aðgengilegur á GitHub á https://github.com/tiro-is/heyra/releases/tag/M9, á GitLab á https://gitlab.com/tiro-is/heyra/-/tags/1.0 og á CLARIN á http://hdl.handle.net/20.500.12537/274.

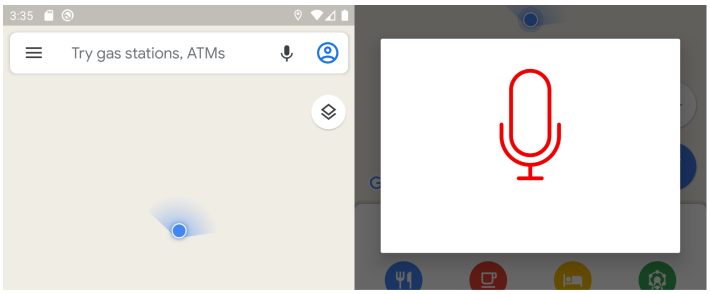
[26] Heyra: https://play.google.com/store/apps/details?id=is.tiro.heyra

Við höfum útfært innsláttaraðferð (e. input method editor, IME) fyrir Android með þjónustunni Heyra, sem hægt er að nota til raddritunar, annaðhvort með því að ræsa hana úr studdu forriti eða sem sjálfstæða innsláttaraðferð. Venjulega myndi notandi vilja innsláttaraðferð Heyra sem viðbót við hefðbundnara lyklaborð. Þetta er notanda mögulegt með því að skipta yfir í lyklaborð Heyra, með notkun lyklaborðsskiptis eða á auðveldari máta með hljóðnematákninu í forriti sem skynjar innsláttaraðferð Heyra. Við gerðum breytingar á AnySoftKeyboard[27], sem studdi aðeins Google Voice Input, til að gera notanda kleift að velja Heyra sem raddinntak þegar notandi ýtir á hljóðnematáknið. Við lokun innsláttaraðferðar Heyra skiptist alltaf aftur yfir í virka innsláttaraðferð, svo að hún verði ekki sjálfgefin.



Mynd 1. Skipt yfir í Heyra með innsláttarskipti kerfisins.

Við útfærðum einnig aðgerðarætlunarmeðhöndlara (e. action intent handlers) fyrir raddleit á vefnum og talgreiningu. Þetta gerir forritum kleift að biðja um raddinntak, oftast stuttar fyrirspurnir eins og heimilisfang í Google Maps eða nöfn forrita eða tengiliða. Þegar Heyra er uppsett á tæki notandans og beðið er um raddinntak birtist gluggi sem biður um raddinntak. Notandinn talar þá, forritið greinir lok raddvirkninnar og sendir niðurstöðurnar aftur til forritsins sem bað um raddinntak.



Mynd 2. Meðhöndlun talgreiningarbeiðni með Heyra.

Allar breytingar kóðans sem er ýtt til aðalkvíslar GitLab-hirslunnar sem hýsir Heyra koma af stað útgáfu af innri prófunarútgáfu forritsins. Þegar nýir eiginleikar og/eða lagfæringar hafa verið

[27] AnySoftKeyboard: https://github.com/anysoftkeyboard/anysoftkeyboard

prófaðar innbyrðis er breyting merkt með útgáfunúmeri sem orsakar uppfærslu á útgáfunni sem aðgengileg er almenningi á Google Play Store.

H10 – Raddstýring og fyrirspurnir

Skýrsla: Háskólinn í Reykjavík, Tiro ehf.
Aðilar: Háskólinn í Reykjavík, Tiro ehf.

Markmið

Að þróa Kaldi-forskriftir fyrir raddstýringar- og spurningasvörunarkerfi á íslensku. Raddstýring vísar til einhliða samskipta milli manns og kerfis, með það að markmiði að biðja um einhverja aðgerð. Frá sjónarhóli talgreiningar krefst þetta þess að mállíkanagerðinni sé veitt sérstök athygli, þar sem ekki er víst að segðin lúti hefðbundnum reglum samfellds máls. Talgreining fyrir spurningasvörun er háð sambærilegum skilyrðum með sérhæfða uppbyggingu segða, en í spurningaformi. Þetta þýðir að markmiðuð tungumálalíkön mætti þróa fyrir verkefnið. Nálgunin hér snýst um að skilgreina sértæk verkefni innan almennra flokka talgreinikerfa og búa til Kaldi-forskriftir fyrir þau verkefni. Væntur fjöldi þróaðra verkefna er 2-3 í hverjum flokki.

Varða 8 – lýsing

M8: Málfanga aflað fyrir öll verkefni og Kaldi-forskriftir tilbúnar fyrir að minnsta kosti tvö verk.

Varða 9 – lýsing

M9: Kaldi-forskriftir tilbúnar fyrir öll skilgreind verkefni.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.
Tiltæk gögn og gagnasköpunarskriftur hafa verið gefnar út hér:

- Umreikningur milli mælieininga: https://github.com/tiro-is/unit-conversion
- Númi (normari fyrir tölur): https://pypi.org/project/numi/
- fstring2fst (tól til að búa til stöðuferjöld (e. Finite State Transducer) frá Python-sniði strengja til að búa til gögn): https://github.com/thdg/fstring2fst
- Skapari upplýsingagagna: https://github.com/cadia-lvl/lm-is-forms
- Spurningar úr spurningaleikjum: https://github.com/cadia-lvl/is-trivia-questions

Forskriftir eru aðgengilegar hér: https://github.com/tiro-is/create_models
Líkön eru aðgengileg hér: https://github.com/tiro-is/create_models/tree/master/models

Afurðin er einnig aðgengileg á CLARIN á http://hdl.handle.net/20.500.12537/295.

Skýrsla

Markmið þessa verkþáttar var að þróa Kaldi-forskriftir fyrir raddstjórnun og svörun spurninga fyrir íslensku. Með raddstjórnun er átt við einhliða samskipti milli persónu og kerfis til að biðja um aðgerð. Við skilgreindum sex verk og annaðhvort bjuggum til eða söfnuðum gögnum fyrir hvert þeirra, normuðum gögnin og þjálfuðum svo Kaldi-mállíkön.

Afurðirnar hafa verið gefnar úr á CLARIN. Þær samanstanda af líkönum fyrir hvert verk ásamt fónemasetti til að nota megi viðeigandi Kaldi-hljóðlíkan með líkönunum. Þau gögn sem voru búin til eða safnað hafa einnig verið gefin út, fyrir utan einkaleyfisvarinn hluta spurninga úr spurningaleikjum. Skrifturnar sem notaðar voru til að búa til líkönin og gögnin voru einnig gefin út og vísað er til þeirra í afurðum.

Þar sem þetta verkefni fól í sér að búa til líkön fyrir hugsanleg notagildi þar sem endanlegur notandi, eða jafnvel forritari, er ókunnur lögðum við áherslu á að búa til og hreinsa gögnin svo auðvelt sé að nota þau fyrir sérsniðin líkön í framtíðinni, hvort sem er viðbót við tiltækt líkan eða með fínstillingu líkans.

Verkin sem við bjuggum til eða söfnuðum gögnum fyrir eru umreikningur milli mælieininga (kílógrömm í pund, o.s.frv.), spurningar úr spurningaleikjum, heimilisföng, íslenskar kennitölur, nöfn og íslensk símanúmer.

Umreikningur mælieininga
Markmiðið hér var að búa til líkan fyrir almennar fyrirspurnir með mörgum umreikningum milli mælieininga. Við bjuggum til þjálfunargögn byggð á dæmasetningum þar sem mælieiningu (fjarlægð, gjaldmiðill, rúmmál o.s.frv.) og tölu var bætt við með réttri beygingu. Nokkur dæmi:

> Ísl: Hversu margar íslenskar krónur fæ ég fyrir einn dollara?
> Ens: How many Icelandic kronur do I get for one dollar?
>
> Ísl: Hvað er einn desilítri margir lítrar?
> Ens: How many liters is one deciliter?[28]
>
> Ísl: Breyttu fimm mílum í kílómetra
> Ens: Convert five miles into kilometers

Setningagjafaskriftan er á GitHub á https://github.com/tiro-is/unit-conversion. Fyrir þessa skriftu þurftum við lista af tölum á stækkuðu sniði ásamt beygingu. Vegna mikils fjölda beyginga fyrir hverja tölu er þetta ekki einfalt verk, og því bjuggum við til tól sem kallast Númi (https://pypi.org/project/numi/), sem skrifar tölurnar út ef það fær heiltölu og ákveðna beygingu, t.d.:

[28] Röð röklíða er ekki sú sama á ensku og íslensku.

Inntak: 92, ft_hk_ef
Úttak: níutíu og tveggja

Inntak: 121, ft_kk_þgf
Úttak: „eitt hundrað tuttugu og einum“, „hundrað tuttugu og einum“

Kóðinn fyrir þetta tól er á GitHub á https://github.com/tiro-is/numi og er aðgengilegt sem pip pakki, „pip install numi“.

Spurningar úr spurningaleikjum
Við söfnuðum gögnum frá þremur stöðum; frá spurningahöfundum Gettu betur, frá opnu safni spurninga úr spurningaleikjum (https://github.com/sveinn-steinarsson/is-trivia-questions) og frá spurningasvörunarumhverfi sem kallast Spurningar.is (https://spurningar.is). Við gátum ekki fengið leyfi til að birta spurningar frá höfundum Gettu betur opinberlega, svo þau gögn voru aðeins notuð við þjálfun líkansins.

Dreifing gagnanna er eftirfarandi:

Uppruni	*Fjöldi setninga*
is-trivia-questions	11.309
Gettu Betur	4.184
Spurningar.is	18.304

Gögnin voru normuð hvað varðar tölur og styttingar með Regina (https://github.com/cadia-lvl/regina_normalizer), ásamt handvirkri þáttun að einhverju leyti. Skrifturnar og gögnin fyrir opnu hlutana eru aðgengileg á GitHub á https://github.com/cadia-lvl/is-trivia-questions.

Upplýsingagögn
Næstu fjögur verk taka á algengum atriðum sem krafist er til ýmissa nota, sem þarfnast sérstakrar athygli varðandi orðaforða og uppbyggingu mállíkana. Við völdum atriði sem eru algeng fyrir mörg verk, með fjölbreytni atriða í huga. Með þessu er mögulegt að endurstilla eða aðlaga líkönin til ýmissa nýrra verka. Atriðin fjögur sem voru valin eru heimilisföng, full nöfn, kennitölur og símanúmer.

Heimilisföng
Þjóðskrá er með lista yfir öll (lögleg) heimilisföng á Íslandi. Þessi gögn eru aðgengileg á vefsíðunni https://www.skra.is. Heimilisföng voru normuð með handskrifuðum reglum fyrir hverja gerð heimilisfanga. Kóðinn fyrir normun er aðgengilegur á GitHub á https://github.com/cadia-lvl/lm-is-forms.

Nöfn
Til að búa til nöfn notuðum við BÍN-gagnasafnið til að safna eiginnöfnum, millinöfnum og eftirnöfnum (bæði föður- og móðurnöfnum). Kóðinn til að búa til nöfnin er aðgengilegur á GitHub á https://github.com/cadia-lvl/lm-is-forms. Ekkert normunarskref þarf fyrir þetta atriði.

Kennitölur
Allt fólk og öll fyrirtæki á Íslandi eiga sér kennitölur sem eru notaðar til ýmissa auðkenningar. Tölurnar fylgja ákveðnum sannreynanlegum mynstrum og má lesa upp á ýmsan máta. Til dæmis má lesa kennitöluna „241270 2329“ upp á alla eftirfarandi máta:

- tveir fjórir einn tveir sjö núll tveir þrír tveir níu
- tuttugu og fjórir tólf sjötíu tuttugu og þrír tuttugu og níu
- tuttugu og fjórir tólf sjötíu tuttugu og þrír tveir níu
- tuttugu og fjórir tólf sjötíu tveir þrír tuttugu og níu
- tuttugu og fjórir tólf sjötíu tveir þrír tveir níu
- tuttugu og fjórir tólf sjö núll tveir þrír tveir níu
- tuttugu og fjórir tólf sjö núll tveir þrír tuttugu og níu
- tuttugu og fjórir tólf sjö núll tuttugu og þrír tveir níu
- tuttugu og fjórir tólf sjö núll tuttugu og þrír tuttugu og níu
- tveir fjórir tólf sjö núll tuttugu og þrír tuttugu og níu
- tuttugu og fjórir einn tveir sjötíu tuttugu og þrír tuttugu og níu

Við bjuggum til langan lista af mögulegum kennitölum og normuðum hverja kennitölu á allt að 11 mismunandi máta til að leyfa ólíkan lestur hverrar tölu. Kóðinn til að búa til og norma kennitölurnar er aðgengilegur á GitHub á https://github.com/cadia-lvl/lm-is-forms.

Símanúmer
Símanúmerin voru búin til og normuð á svipaðan máta og kennitölurnar. Við bjuggum til lista af mögulegum símanúmerum og stöfuðum hvert og eitt símanúmer á níu mismunandi máta til að gefa kost á mismunandi upplestri talnanna. Kóðinn til að búa til og norma símanúmerin er aðgengilegur á GitHub á https://github.com/cadia-lvl/lm-is-forms.

Líkönin
Við þjálfuðum líkan fyrir hvert verk með tólum úr Kaldi. Mállíkanið er fjórstæðulíkan þjálfað með KenLMl. Þar sem þetta eru líkön sérhæfð fyrir ákveðin verk og ættu aðeins að dekka takmarkaðan orðaforða var prófunarsettið búið til með því að taka setningar af handahófi úr þjálfunargögnum. Við bjuggum til dæmi með talgervilsröddunum á https://tts.tiro.is. Gögn sem búin eru til af talgervlum geta innihaldið galla sem geta haft áhrif á talgreininn, svo við berum niðurstöðurnar saman við stórt almennt orðaforðalíkan.

Niðurstöðurnar eru bornar saman með orðvillutíðni (e. Word Error Rate, WER). Orðvillutíðni er mæliaðferð þar sem heildarfjölda skiptinga-, innsetninga- og fjarlægingavillna er deilt með fjölda orða í viðmiðssetningu. Fjöldi villna getur því verið hærri en lengd setningar, eins og sést á orðvillutíðni almenna líkansins fyrir nöfn.

Líkan	Orðvillutíðni (%) sérhæfðs líkans	Orðvillutíðni (%) almenns líkans
Heimilisföng	37,38	80,75
Kennitölur	15,81	44,89
Nöfn	22,46	104,17
Spurningaleikir	19,60	48,15
Símanúmer	16,66	40,01
Umreikningur mælieininga	5,37	27,48

Eins og við var búist ná líkönin sem eru sérhæfð fyrir ákveðið verk mikið betri árangri en almenna líkanið. Nöfnin, sem eru til staðar í verkum heimilisfanga og nafna, eru sérstaklega erfið fyrir almenna líkanið. Það eru líklega einhver orð utan orðaforða til staðar í almenna líkaninu.

H11 – Sérhæfð talgreining

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að þróa Kaldi-forskriftir fyrir börn, unglinga og þau sem hafa íslensku sem annað mál. Þessi vinna er framhald á þróun talgreinaforskrifta fyrir íslensku sem beinist að notendahópum sem þarf að aðlaga bæði hljóð- og mállíkön fyrir með einhverjum sérstökum hætti.

Varða 8 – lýsing

M8: Kaldi-forskrift fyrir unglinga.

Varða 9 – lýsing

M9: Kaldi-forskrift fyrir annarsmálshafa.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Kaldi forskriftir fyrir öll verk eru aðgengilegar á CLARIN og GitHub. Að auki bjuggum við til forskrift fyrir mælendur undir lögaldri (underage), sem inniheldur bæði börn og unglinga:

Unglignar (13 til 17 ára):
CLARIN: http://hdl.handle.net/20.500.12537/221
GitHub: https://github.com/cadia-lvl/samromur-asr/tree/s5_adolescents_lstm

Annarsmálshafar:
CLARIN:http://hdl.handle.net/20.500.12537/291
GitHub: https://github.com/cadia-lvl/samromur-asr/tree/s5_l2_speakers

Skýrsla

Kaldi forskrift fyrir unglinga (og undir lögaldri)

Undir lögaldri (4 til 17 ára) (auka):
https://github.com/cadia-lvl/samromur-asr/tree/s5_underage_lstm

Forskriftirnar fyrir unglinga og mælendur undir lögaldri voru afhentar við M7. Greint var frá niðurstöðum mats og prófana í lokaskýrslu M7 en við höfum þær hér til samanburðar.

Forskrift	Aldursbil	Orðvillutíðni HMMs	Orðvillutíðni LSTMs
S5_children_lstm (dev)	4 – 12 ára	22,08%	11,24%
S5_children_lstm (test)	4 – 12 ára	34,95%	18,06%
s5_adolescents_lstm (dev)	13 – 17 ára	17,13%	10,03%
s5_adolescents_lstm (test)	13 – 17 ára	29,53%	15,26%
S5_underage_lstm (dev)	4 – 17 ára	18,12%	9,46%
S5_underage_lstm (test)	4 – 17 ára	33,81%	15,27%

Kaldi-forskrift fyrir annarsmálshafa
Sama forskriftin og notuð var fyrir mælendur undir lögaldri var aðlöguð fyrir annarsmálshafa. Við keyrðum forskriftina á öllum tiltækum gögnum annarsmálshafa frá Samrómi – 125k55m af þjálfunargögnum auk 4k16m þróunargagna og 10k55m prófunargagna.
Niðurstöðurnar eru birtar í töflunni að neðan. Niðurstöðurnar fyrir LSTM bárust ekki í tæka tíð fyrir skýrsluna. Fyrri verkefni (taflan að ofan) gefa til kynna að við getum búist við betri árangri frá þeim. Núverandi niðurstöður eru aðgengilegar á GitHub (https://github.com/cadia-lvl/samromur-asr/blob/s5_l2_speakers/s5_l2_speakers_lstm/RESULTS) og verða uppfærðar þegar þær berast. Að því sögðu þarf meiri gögn fyrir gott líkan eða aðrar aðferðir, eins og að aðlaga almennt talgreiningarlíkan að gögnum annarsmálshafa.

Forskrift	Orðvillutíðni HMMs
S5_l2_speakers (dev)	47,46%
S5_l2_speakers (test)	46,73%

H16 – Talgreiningarforskriftir í öðrum kerfum

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að þróa forskriftir talgreinikerfa í öðrum kerfum en Kaldi. Þessi vinna gefur verkefninu kost á því að aðlagast annarri tækni eftir því sem hún þróast. Afurðirnar sem eru aðgengilegar hér verða notaðar til að aðlaga Kaldi forskriftir sem þróaðar hafa verið í verkefninu að nýrri útgáfu Kaldi og/eða til að þróa forskriftir í nýjum kerfum eins og Nvidia NeMo eða HTK.

Varða 8 – lýsing

M8: Forskrift almenns talgreinis hönnuð í umhverfi utan Kaldi.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð; tvær forskriftir talgreiningarkerfa annarra en Kaldi voru farsællega þróaðar og gefnar út á CLARIN. Önnur forskriftin er byggð á NVIDIA-NeMo, og hin er byggð á DeepSpeech.

- **Samrómur NeMo Recipe 22.06:** Þetta er kóðaforskrift ætluð til að sýna hvernig skal nota málheildina „Samromur 21.05“[29] og „Íslenskt 6-stæðu mállíkan fyrir NeMo (Binary útgáfa) 22.06“[30] til að búa til talgreini með NVIDIA-NeMo-umhverfinu[31]. Gefin út á CLARIN á http://hdl.handle.net/20.500.12537/228.
- **Samrómur DeepSpeech forskrift 22.06:** Þetta er kóðaforskrift sem sýnir hvernig má nota „Samromur 21.05“ og „DeepSpeech matsgjafi fyrir íslensku 22.06“[32] til að búa til talgreini með Mozilla DeepSpeech recognizer[33]. Gefin út á CLARIN á http://hdl.handle.net/20.500.12537/229.
- **Íslenskt 6-stæðu mállíkan fyrir NeMo (Binary útgáfa) 22.06**: Þetta er n-stæðu mállíkan byggt á orðum á bitasniði til að nota með talgreinum sem eru búnir til í NVIDIA-NeMo-umhverfinu. Gefið út á CLARIN á http://hdl.handle.net/20.500.12537/226.

[29] https://www.openslr.org/112/
[30] http://hdl.handle.net/20.500.12537/226
[31] https://developer.nvidia.com/nvidia-nemo
[32] http://hdl.handle.net/20.500.12537/227
[33] https://deepspeech.readthedocs.io/en/r0.9/

- **DeepSpeech matsgjafi fyrir íslensku 22.06**: Matsgjafi sem hentar talgreinum byggðum á DeepSpeech-talgreini Mozilla. Matsgjafi er stök skrá notuð til að mynda mállíkön. Hún samanstendur af tveimur hlutum, KenLM-mállíkani og trjálíkani (e. trie) sem inniheldur allan orðaforða í tungumálinu. Gefin út á CLARIN á http://hdl.handle.net/20.500.12537/227.

Skýrsla

Tilraunir með mismunandi kerfi

Til að ná markmiðum H16 gerðum við tilraunir með nokkur talgreiningarkerfi. Því miður stóðust þau ekki öll væntingar okkar. Eftirfarandi eru lýsingar á þeim vandamálum sem upp komu í hverju kerfi.

WAV2LETTER++
Þetta er enda-til-enda (e. End-to-End) talgreiningarkerfi þróað af Facebook Research Group og er byggt á vélnámssafni sem kallast flashlight[34]. Við komumst að því að gríðarlega erfitt er að samþýða WAV2LETTER++ þar sem það krefst ákveðinna útgáfa af miklum fjölda hugbúnaðartóla. Það krefst stjórnunarréttinda til að vera þýtt (e. compiled), það er ekki hægt að setja upp að fullu í conda-umhverfi og ekki er mælt með því að nota það í fjarumhverfi af þeim ástæðum.

ESPNet
Þetta er enda-til-enda talgreiningarkerfi þróað af Johns Hopkins-háskólanum. Það er auðvelt í uppsetningu og kemur í svipuðu formi og Kaldi. Við gátum ekki búið til forskriftir með ESPNet vegna vandræða með CUDA.

K2
K2 er rammi skrifaður í Python til að þróa talgreiningarkerfi. Hann var búinn til að sumum þeirra sem bjuggu til Kaldi. K2 stendur fyrir „Kaldi 2“, en er þó ekki bara Python-útgáfa af Kaldi. Hann er rammi af því að hann veitir tól til að þróa margar gerðir talgreiningarkerfa. Hann notfærir sér Lhotse[35], sem er Python-safn sem miðar að því að gera undirbúning tal- og hljóðgagna sveigjanlegri, og Icefall[36], sem er hugbúnaðartól sem býr til forskriftir fyrir K2 með Lhotse.

Forskriftirnar í Icefall eru hannaðar til að undirbúa tilraunir þínar með talgreini með aðeins nokkrum línum af kóða, sem þýðir að þær forskriftir búa á þjónum sem Icefall á. Þetta er stórt vandamál fyrir okkur því að ef við viljum búa til nýja forskrift fyrir K2 getum við ekki notað Lhotse eða Icefall þar sem enn eru engar forskriftir neinna Samrómsmálheilda á þjónum þeirra. Af þessari ástæðu ákváðum við að nota ekki K2 til að búa til forskriftir fyrir H16.

Valin kerfi

[34] https://github.com/flashlight/flashlight
[35] https://github.com/lhotse-speech/lhotse
[36] https://github.com/k2-fsa/icefall

Þau kerfi sem okkur fannst þægilegast að vinna með voru NVIDIA-NeMo (NeMo, til styttingar) og DeepSpeech. Bæði eru nútímaleg enda-til-enda kerfi, það fyrra er byggt á hlykkjóttum tauganetum og það seinna er byggt á ítrekuðum tauganetum.

NVIDIA-NeMo
NVIDIA-NeMo er opinn rammi til að búa til, þjálfa, og fínstilla skjákortshröðuð tal- og málskilningslíkön (e. natural language understanding, NLU) með einföldu Python-viðmóti. Með NeMo geta hönnuðir búið til nýjar gerðir af líkanahögun og þjálfað þær með blandaðri nákvæmni reiknaðar á þinkjörnum (e. Tensor Cores) í NVIDIA-skjákortum með forritaskilum (APIs) sem auðvelt er að nota.

Auðvelt er að setja upp NVIDIA-NeMo á þjóni í gegnum Anaconda-umhverfi.

DeepSpeech
DeepSpeech er opin talgervingarvél, sem notar líkan þjálfað með vélnámsaðferðum sem byggja á vísindagrein Baidu's Deep Speech[37]. DeepSpeech-verkefnið notar TensorFlow frá Google til að gera útfærsluna auðveldari.

Líkt og NeMo er auðvelt að setja upp DeepSpeech á þjóni í gegnum Anaconda-umhverfi.

Niðurstöður

Í þessum hluta lýsum við afurðum sem voru búnar til fyrir H16 ásamt orðvillutíðni sem fékkst með hverju kerfi, nánar tiltekið í þróunar- og prófunarhlutum málheildarinnar „Samromur 21.05".

Báðar forskriftirnar eru settar fram í Kaldi-stíl, sem þýðir að notandinn getur keyrt öll skref forskriftarinnar (forvinnslu gagna, sköpun mállíkans, þjálfun og afkóðun) í gegnum aðalskriftu í bash. Forskriftin inniheldur README-skrá með ítarlegum leiðbeiningum til að setja upp þann hugbúnað sem þarf til að keyra hverja forskrift.

Niðurstöður NeMo
NeMo-forskriftin er byggð á högun sem kallast QuartzNet15x1[38]. Hún var þjálfuð fyrir 50 mynstrasöfn með ~80 klukkustundum úr þjálfunarhluta málheildarinnar „Samromur 21.05". Fyrir afkóðunarhlutann var 6-stæðumállíkan þjálfað með um 5GB af íslenskum texta. Þetta mállíkan var gefið út á CLARIN sem aðskilinn hlutur (http://hdl.handle.net/20.500.12537/226).

Tafla 1 sýnir niðurstöður varðandi orðvillutíðni. Eins og sést gefur NeMo kost á mælingu orðvillutíðni (WER) með og án mállíkans. Þetta er ekki mögulegt í Kaldi, eða a.m.k. ekki með hefðbundnum leiðum.

[37] https://arxiv.org/abs/1412.5567
[38] Fyrir frekari upplýsingar um QuartzNet sjá greinina: „Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions", https://arxiv.org/abs/1910.10261.

Hluti	Orðvillutíðni án málíkans	Orðvillutíðni með málíkani
Þróun	69,19%	20,23%
Prófun	80,38%	22,83%

Tafla 1. Orðvillutíðni með notkun NeMo-forskriftarinnar á „Samromur 21.05“-málheildinni.

DeepSpeech-niðurstöður

DeepSpeech-forskriftin er mjög svipuð forskrift NeMo; eini munurinn er að DeepSpeech krefst matsgjafa í stað n-stæðumállíkans. Matsgjafi er stök skrá notuð til að búa til málíkön. Hún samanstendur af tveimur hlutum, KenLM-málíkani og trjálíkani (e. trie) sem inniheldur allan orðaforða í tungumálinu. Þessi matsgjafi var búinn til með sama íslenska textanum og í NeMo-forskriftinni og hann var einnig gefinn út sem aðskilinn hlutur (http://hdl.handle.net/20.500.12537/227).

Tafla 2 sýnir niðurstöður orðvillutíðni. Eins og sést er ekki mögulegt að greina frá niðurstöðum án málíkans, en DeepSpeech gefur upp stafvillutíðni (e. Character Error Rate, CER) ásamt orðvillutíðni (WER).

Hluti	Orðvillutíðni	Stafvillutíðni
Þróun	30,65%	14,56%
Prófun	34,74%	17,17%

Tafla 2. Orðvillutíðni með notkun DeepSpeech-forskriftarinnar á „Samromur 21.05“-málheildinni.

T4 – Vefgáttir fyrir talgervla

Skýrsla: Tiro ehf.
Aðilar: Tiro ehf.

Markmið

Að búa til vefþjónustu fyrir talgervla. Frumgerðin sem gefin var út á öðru ári tekur við stöðluðu inntaki fyrir stutta texta og fyrsta (opinbera/útgefna) útgáfan inniheldur forvinnsluskref, m.a. tilreiðslu og textanormun. (Síðari útgáfur, gefnar út á þriðja ári munu taka við lengri textum og stjórna hljómfalli.)

Varða 8 – lýsing

M8: Vefþjónustan virk, ásamt skilgreiningum, og öllum notendaprófunum lokið.

Afurðir

Markmiðum afurða fyrir M8 hefur verið náð. Vefþjónustan er komin í notkun á https://tts.tiro.is, og er notuð af undirverkefnum T6 og T5. Kóðinn er aðgengilegur á https://github.com/tiro-is/tiro-tts/releases/tag/M9 og á CLARIN á http://hdl.handle.net/20.500.12537/268.

Skýrsla

Tveir helstu notendur þjónustunnar eru veflesarinn WebRICE (T6) og Símarómur (T5). Við höfum fengið stöðuga endurgjöf og svarað henni með því að laga eða bæta það sem þarf. Þjónustan er í gangi og styður nú sendingar hrárra texta og SSML-skjala með viðeigandi studdum mörkum. Með SSML er mögulegt að tilgreina framburð ákveðinna orða með fónemsmarki, með því að stjórna hljómfalli (hraði, tónhæð og hljóðstyrkur tals), og staðgengilsorðum fyrir talgervingu og skilgreina hvernig skal túlka tölur og tákn. Tiro hýsir þjónustuna með 8 röddum á https://tts.tiro.is, ásamt OpenAPI v2-skilgreiningu viðmótsins. Vefútgáfa til að prófa raddirnar er aðgengileg á https://talgervill.is.

T5 – Talgervlar í snjallsímum

Skýrsla: Grammatek
Aðilar: Grammatek, Tiro ehf., Háskólinn í Reykjavík

Markmið

Að hanna talgervla sem styðja aðgengisviðmót snjallsíma (e. smartphone accessibility interface). Eldri Android-raddir, Karl og Dóra, er verið að úrelda og leggja af, þannig að framtíðarnotendur skjálesara munu ekki geta notað þær. Enn fremur vantar iOS stuðning fyrir íslensku svo það verður að leysa, annaðhvort með því að hanna hann sjálf, eða með samstarfi við erlend fyrirtæki.

Vefbyggðir talgervlar geta nýtt sér meira vinnsluafl en talgervlar í tæki og þar af leiðandi búið til betra tal. Að bæta hágæða vefröddum við báða talgervlana er því ákjósanlegt í notkunardæmum þar sem náttúrulegir eiginleikar eru mikilvægari (í samvinnu við T4).

Varða 8 – lýsing

M8: Símarómur fyrir Android virkar til að hoppa yfir og stoppa þætti tals sem þegar hafa verið búnir til.

Varða 9 – lýsing

M9: Allt kapp hefur verið lagt á að skila mjög hraðvirkri röddu á tæki fyrir Android, ítarleg skýrsla verður afhent ef það mistekst. Leiðbeiningar um notkun Símaróms hafa verið gefnar út.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Android-app Símaróms, útgáfa 1.2.0, hefur verið gefið út á Google Play í opninni prófunarbraut: https://play.google.com/apps/testing/com.grammatek.simaromur. Kóða hugbúnaðarins má nálgast í gegnum https://github.com/grammatek/simaromur/tree/v1.2.0.

- Símarómur á CLARIN: http://hdl.handle.net/20.500.12537/287.
- Leiðbeiningar hafa verið gefnar út á https://github.com/grammatek/simaromur/wiki/ ásamt hljóðleiðbeiningum fyrir sjónskerta notendur.
- Hraðvirka einingavalsröddin hefur verið gefin út á https://github.com/grammatek/simaromur_voices/releases/tag/0.1 og byggingaruppskriftirnar eru aðgengilegar á https://github.com/grammatek/lvl_is_flite/releases/tag/0.1.

Skýrsla

Eftir að við ráðfærðum okkur við Blindrafélagið var frekari áhersla lögð á hraða og áreiðanleika í þróun Símaróms. Annars vegar bættum við skyndiminnisstigi (e. caching level) í forritið fyrir M8, sem þýðir að raddhljóð sem hefur þegar verið búið til er ekki búið til að nýju ef kallað er eftir því aftur. Hins vegar er nú mögulegt að stöðva sköpun raddar sem þegar er hafin, svo að það valdi ekki óþörfum töfum þegar hratt er farið yfir.

Þrátt fyrir það eru tauganetsraddir á tæki enn hægar, og enn verri á hægari tækjum. Fyrir M9 notuðum við eldra form raddmyndunar sem er ekki af jafn háum gæðum, en hraðari. Fyrir þetta notuðum við hugbúnaðarpakkann Flite og bættum við leið til að búa til kviklega hlaðanlegar hugbúnaðarraddir fyrir „Unit Selection“-aðferðina.

Hugbúnaðarröddin sem myndast með þessari aðferð er nokkuð stór. Því var nauðsynlegt að gera hana aðgengilega til að sækja og útfæra bæði fyrirspurn og endurheimt þessara radda í forriti Símaróms. Útfæra þurfti leið til að greina nýjar útgáfur raddanna og gefa notendum kost á að uppfæra þær staðbundið eftir þörfum.

Afleiðingin af öllu þessu er að notendur hafa nú val um hæga en hágæða rödd á tæki, og meðalgæða en hraða rödd á tæki. Auk þess hafa notendur val um hágæða netraddir fyrir minni biðtíma, samanborið við hágæðarödd á tæki.

Viðbót háhraða Flite-raddar (einingaval (e. Unit selection))

Þar sem enginn innbyggður stuðningur var fyrir hleðslu einingavalsradda fyrir Flite-rammann þurftum við að finna raunhæfa leið til að hlaða Flite-einingavalsrödd eftir að hún hefur verið sótt á Android-tækið. Við breyttum því hvernig þessar raddir eru búnar til og bættum við skrefi þar sem sameiginlegt safn er búið til, auk raddrekilslags sem hjúpar allar innri einingar raddar. Þar sem fjórar mismunandi gerðir högunar Android eru studdar (armv7a, aarch64, x86, x86_64) þurftum við að búa til fjórar raddskrár fyrir eina útgáfu raddar og vegna þess þurftum við einnig að búa til öll hæði þessarar gerðar Android-högunar. Þó að við höfum einbeitt okkur að því að búa aðeins til eina aðgengilega rödd getum við nú sótt hvaða rödd sem er sem búin er til á þennan máta vegna þess að byggingarskref og rekilslagið eru almenn.

Fyrir vefþjóninn ákváðum við að nota eiginleika GitHub-útgáfa til að samþætta niðurhleðslu á Flite-rödd í Símaróm. GitHub-útgáfur geta geymt skrár upp að 1GB að stærð, sem nægir fyrir flest raddlíkön. Við höfum búið til raddhirslu á https://github.com/grammatek/simaromur_voice og bætt Flite-raddlíkönunum þar í gegnum GitHub-útgáfur https://github.com/grammatek/simaromur_voices/releases/tag/0.1. GitHub veitir umfangsmikil RESTful-forritaskil sem hægt er að nota t.d. til að finna auðveldlega nýjar útgáfur eða atriði innan þeirra. Frekari uppfærslur Flite-raddarinnar má ýta þangað án þess að þurfa að uppfæra Símarómsforritið sjálft. Við teljum einnig að GitHub sé framtíðarhelt umhverfi fyrir slíkar þarfir.

Við höfum bætt við möguleika til að sækja Flite-röddina í gegnum GitHub-útgáfur í Símaróm. Þó að þetta hafi verið nauðsynlegt vegna stærðar einingavalsraddarinnar annars vegar – er það enn nauðsynlegra núna þar sem við þurfum ný að styðja fjórar gerðir högunar, sem fjórfaldar stærðarkröfurnar – og stærðartakmarkanir Android-forrita hins vegar væri hægt að nota þennan eiginleika í framtíðinni fyrir þær tauganetsraddir sem nú eru innifaldar. Kostur þess að aðskilja forritið frá röddunum er að hægt er að uppfæra hvort tveggja sjálfstætt og aðeins raddirnar eru sóttar í tækið sem er í raun notað af notandanum.

Símarómur sækir aðeins þau raddgögn sem eru samhæfð einni af fjórum gerðum fyrrnefndrar högunar sem forritið er að keyra á. Hverri útgáfu raddar fylgir lýsigagnaskrá á JSON-formi sem er hönnuð á almennan hátt svo að hún geti stutt öll núverandi íslensk raddlíkön. Einnig höfum við undirbúið Símaróm fyrir uppfærslur á útgáfum radda. Þar sem notandi gæti aðeins beðið um 60 færslur á klukkustund til GitHub útfærðum við takmörkun beiðna til raddútgáfunnar á GitHub í forritið.

Leiðbeiningar

Við höfum búið til leiðbeiningar á vefnum fyrir notkun Símaróms á https://github.com/grammatek/simaromur/wiki/. Þessu var óskað eftir af notendahópum, þar sem virkjun talgervilskerfis þriðja aðila eins og Símaróms þarfnast nokkurra ítarlegra skrefa sem eru ekki öllum notendum auðveld og hafa því leitað til samtaka sinna (eins og lesblindir) til stuðnings. Leiðbeiningarnar sýna hvernig skal sækja, virkja og setja upp Símaróm í einföldum skrefum með skjáskotum og viðbættum sjónrænum þáttum. Við bjuggum einnig til hljóðskrá með þessum leiðbeiningum með talgervilsröddinni Diljá með vefforritinu *„Skjalalestur“* sem við bjuggum til fyrir verkþátt T9 á https://tts.grammatek.com, svo þessi skref séu einnig útskýrð með hljóði.

T6 – Veflesari

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að gera talgervil aðgengilegan fyrir lestur á vefsíðum. Fyrsta markmiðið er að þróa tól sem gerir talgervil á íslensku aðgengilegan vefhönnuðum til að nota á vefsíðum sínum. Tvö aukamarkmið eru að búa til viðbót fyrir vafra og að kynna opna íslenska talgervla sem verða í boði fyrir utanaðkomandi hönnuði veflesara.

Seinkuð varða 7 – lýsing

M7: SSML stuðningur í WebRICE.

Varða 9 – lýsing

M9: Stuðningur við fjölmarga talgervilsbakenda, bæta við talgervilsbakenda og stuðningi við margar raddir (notendaviðmót).

Afurðir

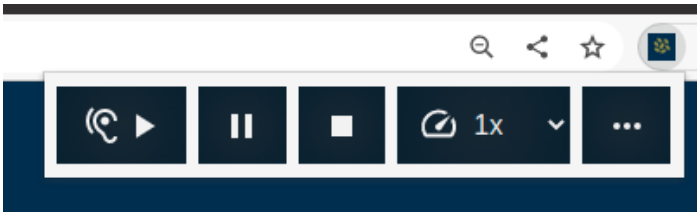
Öllum markmiðum vörðunnar hefur verið náð. Chromium-viðbót (virkar í öllum Chromium-vöfrum eins og Edge og Chrome) hefur verið gefin út á Chrome Extension-búðinni á https://chrome.google.com/webstore/detail/webrice/mmijkiiefioinbdgbadgghcchfilmlmp. Viðbótin styður marga TTS-bakenda, margar raddir og SSML.

Skýrsla

Chrome-viðbót

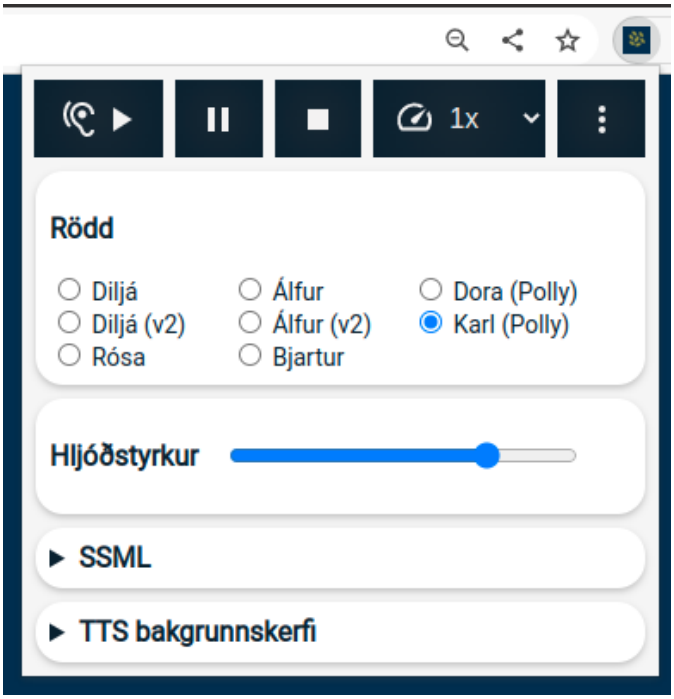
Þessi útgáfa WebRICE er í boði sem Chrome-viðbót. Viðbótin hefur fengið nýja eiginleika við þessa vörðu til að ná markmiðum.

Að neðað en aðalnotendaviðmót WebRICE:

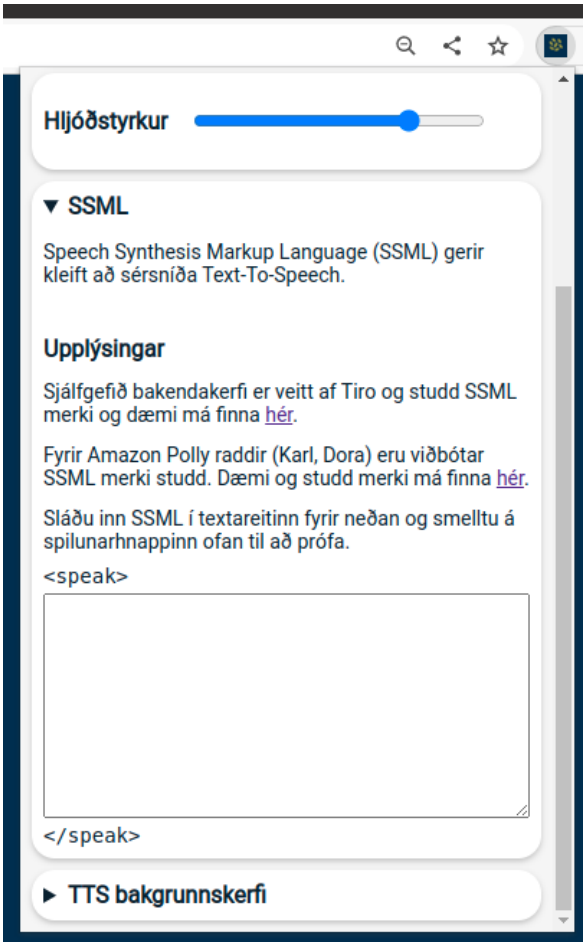


Viðmótið inniheldur grunnstýringar til að spila, gera hlé, stoppa og breyta hraða spilunar. Notandinn getur valið texta og opnað svo viðbótina til að spila talgervilslesinn texta. Ef enginn texti er valinn er allur texti á núverandi síðu spilaður.

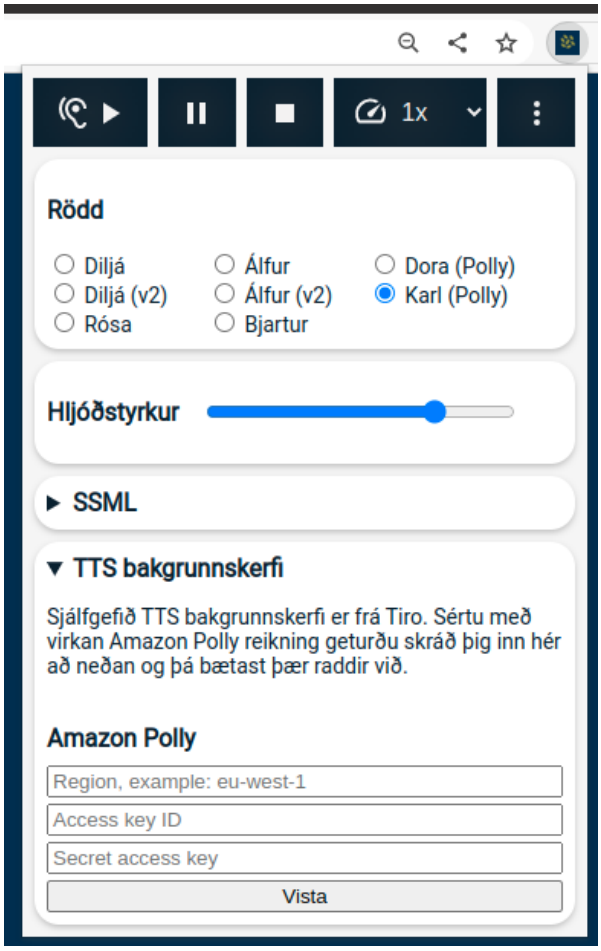
Ef ýtt er á „…“ merkið opnast frekari stillingarmöguleikar, sýnt á myndinni að neðan. Notandinn getur valið hvaða rödd skal nota og stillt hljóðstyrk.



Ef SSML (Speech Synthesis Markup Language) hlutinn er stækkaður getur notandinn gert tilraunir með SSML, líkt og sýnt er á myndinni að neðan.



Ef notandinn hefur aðgang að Amazon Web Services (AWS) og vill nota raddir Amazon Polly getur hann stækkað hlutann „TTS-bakgrunnskerfi“ (e. TTS backend) og slegið inn AWS-auðkenni sín. Þetta afhjúpar Polly-raddirnar Karl og Dora, eins og sést á myndinni að neðan.



Tiro útvegar sjálfvalinn bakenda. Viðbótin hefur einnig verið bætt með því að streyma hljóðinu frá bakendanum. Þetta gefur notendum betri upplifun þar sem minna þarf að bíða eftir að talgervilsrödd byrji að spilast.

T8 – Mat á gæðum talgervla

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík, Grammatek ehf.

Markmið

Að veita nauðsynlegan stuðning við þróun talgervla með því að meta frammistöðu fyrir ýmsar útfærslur talgervla. Eftir bakgrunnsrannsókn verða þrjár til fimm aðferðir valdar og innleiddar í hugbúnaði. Prófanir eru svo skipulagðar með því að undirbúa raddirnar, tímasetja prófanir og fá hlustendur. Niðurstöðurnar eru einnig metnar og greint frá þeim. Hugbúnaður „neðanstreymis“ (e. downstream) eins og t.d. afurðir T5, eru notendaprófaðar og skýrslur gerðar fyrir hverja og eina.

Varða 8 – lýsing

M8: Huglægt mat á (T13. M8) Notkunarmat fyrir T5.M7, ásamt ítarlegri skýrslu um nauðsynlegar umbætur. Matsverkvangurinn bættur, byggt á reynslu notenda.

Varða 9 – lýsing

M9: Huglægt mat á (T13. M9). Notendamat T5.M8, ásamt ítarlegri skýrslu um nauðsynlegar umbætur.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.

Huglægt mat var gert á afurðum T13 fyrir M8 og M9. Þetta var framkvæmt í formi hlustunarprófa með 47 þátttakendum, og úr urðu meðalmatseinkunnir fyrir full-, afritstalgervðar segðir og upptökur tals til samanburðar.

Notendaprófanir voru gerðar á afurðum T5 fyrir M8 og M9. Endurgjöf var fengin frá notendum Símaróms svo hægt væri að bæta hann fyrir endanlega afurð.

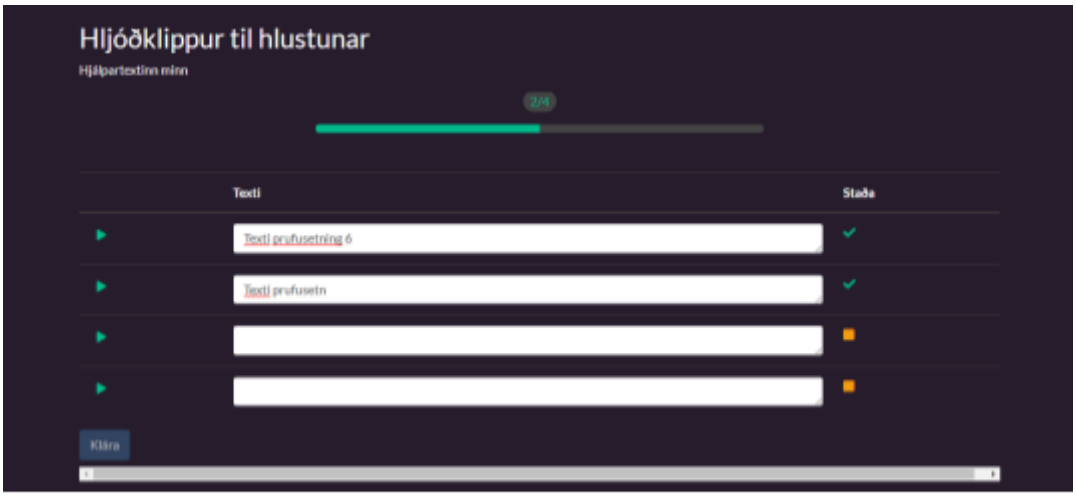
Skýrsla

Matsvettvangur (MOSI)
Við náðum öllum markmiðum okkar fyrir matsvettvang talgervla.

Við höfðum þegar útfært möguleika á framkvæmd MOS-prófana og AB-prófana í MOSI. Þessi virkni var bætt frekar. Niðurstöðusíður beggja voru stækkaðar umtalsvert svo nú er auðveldara

og skýrara en áður að skoða niðurstöður fyrir hverja prófun. Við létum línuritin birtast á skýrari og þýðingarmeiri hátt og þau má aðlaga frekar. Við bættum við möguleika á að bera saman raddir eða líkön eða bæði fyrir MOS-prófun. Umsjón notenda og öryggis var gerð auðveldari með því að gefa hverjum notanda hlutverk sem tengist síðum sem þeir hafa aðgang að.

Við bættum við möguleika á SUS-prófunum. Þessi gerð prófunar er aðallega notuð til að mæla skiljanleika talgervilslíkans. Í SUS-prófun sendir rannsakandi inn sett af hljóðbútum og samsvarandi textum. Þátttakandinn þarf svo að hlusta og skrifa niður þær setningar sem heyrast. Dæmi um þetta má sjá í myndinni að neðan.



Mynd: Skjáskot af glugga þar sem þátttakendur verða að skrifa hvað þeir heyra í upptökunni

Rannsakandinn getur svo sannreynt þetta inntak frá þátttakanda á annarri síðu til að meta hvort textinn er réttur eða ekki. Myndin að neðan sýnir þennan eiginleika.

Mynd: Skjáskot af glugga þar sem rannsakendur geta yfirfarið svör þátttakenda

Niðurstöðurnar birtast svo á síðu fyrir niðurstöður.

Vettvangurinn í heild sinni fékk góða notkun og við gátum lagað nokkrar villur. Þetta var aðallega að samræma þróun vettvangsins við það sem þurfti í raun og að gera upplifunina í heild sinni hnökralausa bæði fyrir rannsakendur og þátttakendur. Við skjöluðum mikinn hluta kóðans og gerðum aðgengilegan á GitHub fyrir hvern sem er á https://github.com/cadia-lvl/MOSI/.

Huglægt mat á aðlögunarhæfum talgervli
Gæði raddanna sem voru búnar til í T13 eru metin með meðalmatseinkunn (MOS) úr hlustunarprófunum. Uppsetningu talgervla er lýst undir verkþætti T13. Hér lýsum við matsaðferðafræði og niðurstöðum sem fengust. Við tökum einnig saman niðurstöður og drögum helstu ályktanir.

Matsaðferðafræði
Fjögur hönnunaratriði voru skoðuð, og úr varð prófunarsett 144 segða. Þau atriði voru:

- *Tvö* karl- og kvenkyns talgervilsraddlíkön, þjálfuð á 18 og 17 af raddgjöfum í jafnskipta 40 mælenda settinu Talrómur 2 (Sjá T2).
- *Tvær* gerðir radda voru búnar til eftir því hvort þær voru notaðar í þjálfunarsettinu („séð“) eða ekki („óséð“).
- *Þrjú* stig talgervingar: grunnsannleikur, afritstalgerð og fulltalgerð. Segðir grunnsannleiks eru einfaldlega upptökurnar úr gagnasafninu Talrómur 2, afritstalgerðar segðir voru búnar til með upptöku úr gagnasettinu (og ekki setningatexta) og viðeigandi x-vigur markmiðsmælanda, og fulltalgerðar segðir voru búnar til með öllu aðlögunarhæfa talgervilskerfinu.
- *Tólf* setningar með málfræðilega fjölbreytt hljóð. Engin þeirra var notuð við þjálfun.

Við fengum 47 hlustendur, og hver þeirra mat 24 segðir tekin með aðferð rómversks kvaðrats (e. latin-square) úr lista 144 segða. Setningarnar eru birtar þannig að hver þátttakandi ætti aðeins að sjá hvert textaboð einu sinni. Fjórum textaboðum var deilt meðal kven- og karlkyns radda, svo að 10 karlkyns raddir eru með mismunandi textaboð, 10 kvenkyns raddir með mismunandi textaboð og 4 sameiginleg karl- og kvenkyns textaboð. Samanlagt eru þetta 24 segðir sem hver hlustandi metur.

Hlustendur voru beðnir um að gefa hverri segð einkunn frá 1–5 eftir því hversu eðlileg hún hljómar, þar sem 1 er mjög óeðlileg og 5 mjög eðlileg. Nákvæmar leiðbeiningar voru á íslensku: *„Gefðu röddinni einkunn á kvarðanum 1 til 5 eftir því hversu eðlileg þér þykir hún þar sem 1 er mjög óeðlileg og 5 er mjög eðlileg.“*

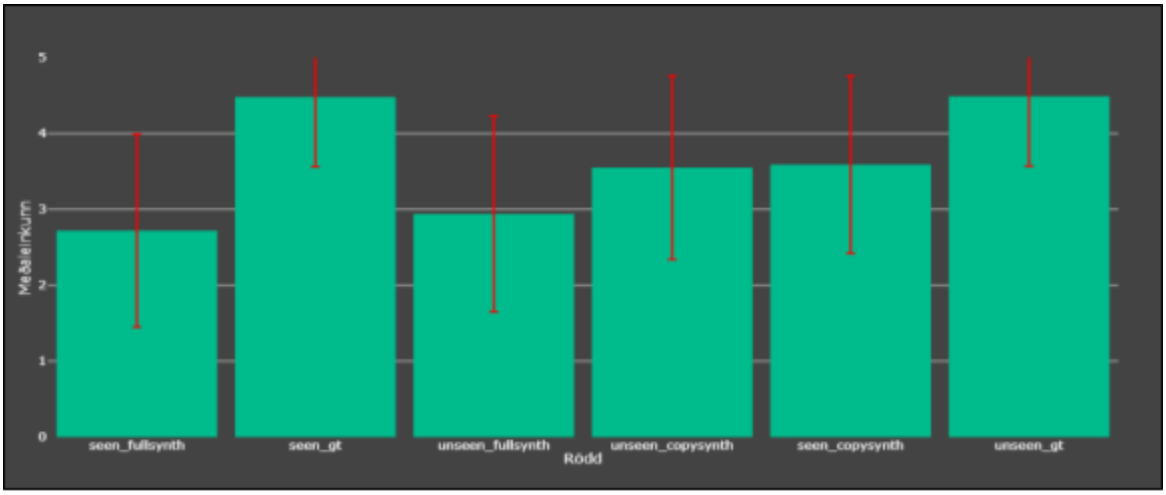
Þátttakendur gátu notað hvaða tæki sem er, en voru hvattir til að nota góð heyrnartól. Þeim var ætlað að hlusta á alla raddupptökuna og hlusta á hana eins og þeir þurftu til að meta hversu eðlileg hún væri.

MOS-prófunin var send iðkendum í máltækni og starfsfólki Hljóðbókasafnsins, sem hefur fyrri reynslu af hlustunarprófunum á talgervlum. Alls voru 49 þátttakendur, 23 karlkyns hlustendur og 26 kvenkyns hlustendur, öll með íslensku að móðurmáli.

Niðurstöður
Við reiknum út samanlagða meðaleinkunn yfir líkönin, séð og óséð, og grunnsannleika. Niðurstöður líkana má sjá í töflunni að neðan.

Rödd	Meðaleinkunn	Staðalfrávik
Séð fulltalgerð	2,72	1,27
Séður grunnsannleikur	4,48	0,92
Séð afritstalgerð	3,59	1,17
Óséð fulltalgerð	2,94	1,29
Óséður grunnsannleikur	4,49	0,92
Óséð afritstalgerð	3,55	1,21

Með því að færa þessar niðurstöður í stöplarit fæst eftirfarandi rit með villustöplum sem staðalfrávik. Við sjáum að grunnsannleikurinn nær hárri einkunn, eins og vera ber. Afritstalgerða líkanið nær svipaðri einkunn, um 3,5 meðaleinkunn. Áhugavert þykir að séða fulltalgerða líkanið nær lægri einkunn en óséða fulltalgerða líkanið.



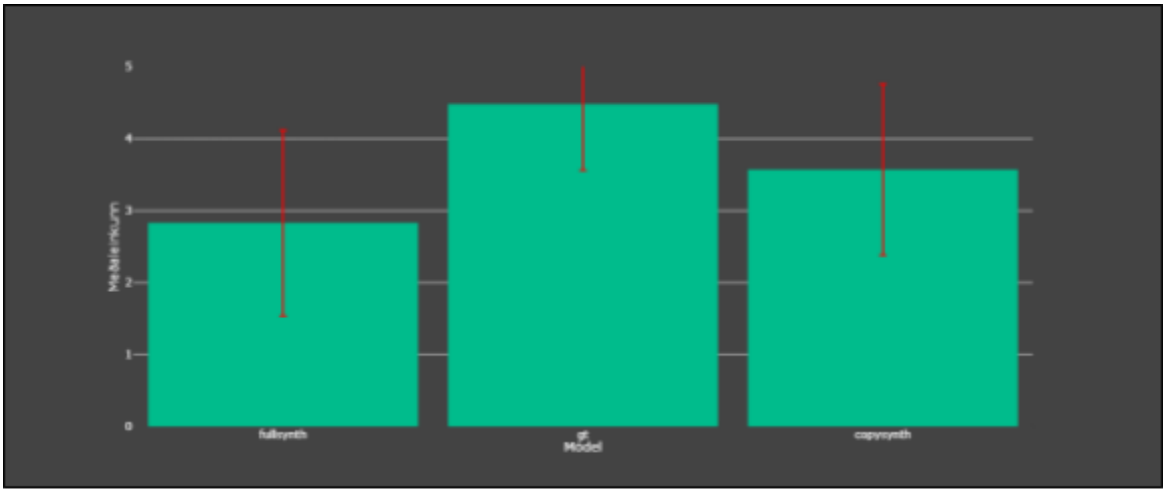
Mynd: Samanburður á niðurstöður MOS-prófana fyrir séð og óséð líkön. Villustöplar tákna staðalfrávik.

Ef við leggjum saman séðar og óséðar breytur fáum við niðurstöðurnar í töflunni að neðan.

Líkan	Meðaleinkunn	Staðalfrávik

Fulltalgerving	2,83	1,29
Afritstalgerving	3,57	1,19
Grunnsannleikur	4,48	0,92

Með því að færa þessar niðurstöður í stöplarit fæst eftirfarandi rit með villustöplum sem staðalfrávik. Við getum séð að grunnsannleikurinn nær aftur hárri einkunn, eins og má búast við. Afritstalgervingin nær einkunn í kringum 3,5 meðaleinkunn og fulltalgerving nær um 2,9.



Mynd: Niðurstöður MOS-prófana bera saman líkön tekin saman yfir séð og óséð gögn. Villustöplar tákna staðalfrávik

Niðurstöður og ályktanir
Eins og við var búist nær grunnsannleikurinn bestum árangri, með 4,48 meðalmatseinkunn. Talan samsvarar einkunn grunnsannleika sem mælst hefur í annarri vinnu, sem gefur til kynna að upptökur okkar séu af álíka gæðum og þær sem eru notaðar í öðrum talgervilskerfum. Töluvert fall er í matinu fyrir afritstalgerðar raddir, sem gefur til kynna að talgervilskerfið mætti bæta. Þetta kemur hins vegar einnig fram í öðrum talgervilskerfum.

Frekari niðurstöður eru birtar á: https://cadia-lvl.github.io/TTS-evaluation/index.html.

Mat notenda fyrir T5, með skýrslu sem skýrir nauðsynlegar endurbætur.

Endurgjöf frá notendum Símaróms-forrits
Talgervilskerfin sem þróuð voru í verkefninu hafa verið innleidd í Android-forrit Símaróms (sjá verkþátt T5), og við höfum safnað endurgjöf frá tveimur notendahópum: notendum frá Blindrafélaginu og frá Félagi lesblindra.

Almenna endurgjöfin er að karlkyns röddin *Álfur* sé þægileg við hlustun, en kvenkyns röddin *Diljá* síður svo. Notendahóparnir tveir, blindir eða verulega sjónskertir og lesblindir, hafa ólíkar þarfir fyrir daglega notkun, óskylt því hvort þægilegt er að hlusta á raddirnar. Blindir krefjast

hraða: þeir vilja geta hraðað spilun radda og þurfa mjög skjót viðbrögð frá talgervilskerfinu. Lesblindir eru umburðarlyndari gagnvart biðtíma, þar sem þeir velja oft hluta texta til að lesa og þurfa ekki allt á skjánum lesið. Núverandi biðtími ~3 sekúndna er því þolanlegur fyrir þennan hóp en talinn ónothæfur af sjónskertum. Fyrir leshraða í rauntíma er einnig æskilegt að hafa fleiri hlé. Þetta gerir hlustun minna þreytandi.

T9 + T10 – Framendapípa talgervils

Skýrsla: Grammatek
Aðilar: Grammatek, Háskólinn í Reykjavík, Hljóðbókasafnið

Markmið

Við lok annars árs munum við eiga mjög nákvæmt textanormunarkerfi fyrir íslensku. Verkefni þriðja árs eru tvíþætt: 1) þróun tóls fyrir áherslur (e. phrasing) og 2) notendaverk (e. use case task) sem felur í sér texta sérsviða. Til að merkja áherslur og hlé (e. pause) í texta sem inntak fyrir talgervilinn má nota þáttara til að sjá til þess að hlé birtist á skynsamlegum stöðum. Við munum nota þáttarana úr undirverkefni I5, en talið er að grunnþáttari sé nóg fyrir þetta verk. Ef aðrar lausnir virðast vænlegri verða þær greindar og gerð grein fyrir þeim. Enn fremur munum við skilgreina notkunardæmi í samstarfi við Hljóðbókasafnið. Kennslubækur eru mjög mikilvægt sérsvið hljóðbóka, t.d. í stærðfræði og efnafræði. Annað sérsvið eru opinberar skýrslur sem innihalda margar töflur, línurit o.s.frv. Við munum skilgreina sameiginlegt notkunardæmi sem er frábrugðið fréttatextum og koma á fót pípu úr lestrarefninu frá útgefanda til normaðs texta fyrir talgervilskerfið.

Varða 8 – lýsing

M8: Frumgerð framendapípu talgervils fyrir sérsvið, með notendaviðmóti fyrir aðila notendaprófana. Niðurstöður um bætta atkvæðaskiptingu og merkingu áherslna með notkun Kvists, einingar fyrir stofnhlutagreiningu (e. compound analysis module). Skýrsla um vankanta g2p í samræmi við prófanir og upplifun notenda.

Varða 9 – lýsing

M9: Greint frá notkunarprófunum framendapípu talgervils fyrir sérsvið. Umbætur innleiddar í samræmi við ábendingar notenda. Pakkar fyrir hljóðritun gefnir út, þar á meðal orðabókareining, tungumálagreining, samsetningagreining og tengingar við reglubyggt og LSTM g2p, bæði fyrir íslensku og íslenskan framburð enskra orða.

Afurðir

- Framendapípa með skipanalínuviðmóti fyrir félagsvísindasvið.
- Uppfærð g2p-eining.
- Vefforritið Skjalalestur á https://tts.grammatek.com/.
- CLARIN-hlekkir:
    - Framendi talgervils: http://hdl.handle.net/20.500.12537/279
    - G2P: http://hdl.handle.net/20.500.12537/296

    - Skjalalestur: http://hdl.handle.net/20.500.12537/282
- Git-hirslur fyrir vefforritið, framendapípu og mismunandi einingar eru eftirfarandi:
    - Vefforrit: https://github.com/grammatek/tts_webapp.
    - Framendaþjónusta: https://github.com/grammatek/tts-frontend-service.
    - Framendapípa: https://github.com/grammatek/tts-frontend.
    - G2P (einnig gefið út á PyPI): https://github.com/grammatek/ice-g2p.
    - Áherslur (Phrasing): https://github.com/grammatek/phrasing-tool, kvíslað frá upprunalega verkefninu hjá HR: https://github.com/cadia-lvl/phrasing-tool.
    - Normari (Normalizer): https://github.com/grammatek/regina_normalizer, kvíslað frá upprunalega verkefninu hjá HR, sem er einnig útgefið á PyPI: https://github.com/cadia-lvl/regina_normalizer.
    - Textahreinsari: https://github.com/grammatek/text-cleaner.

Öllum markmiðum vörðunnar hefur verið náð. Nú er g2p-einingin **ice-g2p** aðgengileg á PyPI. Framendapípan (forvinnsla texta fyrir talgervingu) er tiltæk sem eining og sem gRPC-vefþjónusta, og allir hlutar pípunnar eru tiltækir sem stakar, sjálfstæðar einingar. Vefforritið **Skjalalestur** hefur verið keyrt og prófað fyrir texta ákveðinna sviða af höfundum hljóðbóka hjá Hljóðbókasafninu.

Skýrsla

Á seinni hluta þriðja árs hefur áherslan verið á eftirfarandi þrjú atriði fyrir textavinnslu talgervils: 1) bæta g2p-einingunа 2) tengja allar einingar saman til að mynda samfellda pípu en halda þó sjálfstæðum eiginleikum allra eininga, og 3) þróa vefforrit sem les skjöl á hljóðbókaformi og skilar .mp3-skrá með samsvarandi talgerðu tali.

ice-g2p
Eitt verk fyrir endurbætur g2p-einingar var að nota utanaðkomandi greiningareiningu samsetninga til að bæta atkvæðagreiningu og áherslumerkingar. Kvistur, greiningareining samsetninga, er Python-eining sem var þróuð fyrir tveimur árum og gerð opin á þessu ári (https://github.com/jonfd/kvistur). Hún er byggð á BiLSTM, og þjálfuð á handyfirförnum samsetningagögnum frá Árnastofnun. Einingin ice-g2p inniheldur greiningareiningu samsetninga sem notar gagnasafn til að skipta inntaksorðum, og upprunalega hugmyndin var að skipta þessari einingu út fyrir Kvist. En þar sem Kvistur krefst ákveðinnar útgáfu Python og Tensorflow sem virka ekki með ice-g2p og öðrum hlutum framendapípunnar (sem nota nýrri útgáfur), tókum við aðra nálgun. Núverandi eining reiðir sig mjög á gagnagrunnsfærslur fyrir höfuð samsetninga og ákvæði, og því ákváðum við að nota Kvist til að búa til fleiri færslur fyrir þetta gagnasafn. Við létum Kvist greina orð úr tíðnilista sem gerður var úr Risamálheildinni (sjá G1), alls 500.000 orð af lengd > 6 stafa. Listar höfða og ákvæða sem urðu til voru sannreyndir með því að keyra þá í gegnum BÍN og að lokum var sannreyndum listum bætt við upprunalega lista frá g2p-einingunni. Með þessu stækkuðum við listana frá 5.700 færslum í næstum 50.000 færslur (höfuð), og frá 4.400 í 21.000 (ákvæði). Athygli vekur að þessi margföldun á færslum fyrir greiningu samsetninga bætti niðurstöður aðeins lítillega fyrir atkvæðagreiningu og áherslumerkingar: síðasta útgáfa g2p fyrir stækkun gagnasafnsins í gegnum Kvist hefur í

heildina 6,9% orðvillutíðni á 1.000 færslna prófunarsetti, á meðan nýja útgáfan nær 6,5% orðvillutíðni.

Í gegnum endurgjöf hefur framburðarorðabókin sem er í g2p-einingunni verið stækkuð til að innihalda orð sem g2p umritar ekki rétt, t.d. enskar merkingar í stillingum síma.

Einnig hefur valkostum verið bætt við eininguna. Til dæmis eru nú fjögur mismunandi hljóðritunarstafróf úttaks í boði, og notendur geta valið hvort þeir vilja atkvæðagreiningu og/eða áherslumerkingar, og hvaða mállýsku skal nota. Einingunni hefur verið pakkað og send til PyPI.

Í tengslum við þennan verkþátt viljum við nefna mjög jákvæð viðbrögð sem við höfum fengið frá talgervilsteyminu hjá Microsoft. Þau hafa skoðað afurðir SÍM fyrir talgervingu og fyrstu svör voru að framburðarorðabókin (verkþáttur G6) sé meðal þeirra bestu sem þau hafa séð hvað varðar nákvæmni umritana og dekkun. Við vonum að orðabókin og aðrar afurðir máltækniáætlunarinnar geti reynst þeim vel í frekari þróun nýju íslensku talgervilsraddar þeirra.

Framendapípa talgervils
Við útfærðum tileinkaðan framendastjórnanda (e. *FrontendManager*) til að tengja allar textavinnslueiningar í samræmda pípu. Helstu markmið voru að halda sjálfstæðum eiginleikum hverrar einingar, þ.e. að breyta ekki hverri undireiningu fyrir pípuna, og að viðhalda öllum upplýsingum frá upprunalegum tókum í gegnum pípuna, sem endar í hljóðritaðri birtingarmynd. Þetta þýðir að framendastjórnandi verður að sníða inntak og úttak við hvert stig, eftir því hvaða vinnslueining er næst, og jafnframt halda öllum upplýsingum úr fyrri einingu. Dæmi um birtingarmynd tóka úr setningu: *Suðaustan 10-18 **m/s***:

```
Token:
Original: m/s, Clean: m/s,
Tokenized: ['m/s'],
Normalized: [metrar, nkfn, á, af, sekúndu, nveþ]
Transcribed: ['m E: t r a r', 'au:', 's E: k u n t Y']
index: 2, 16, 19
```

Normað úttak inniheldur málfræðileg mörk sem þarf í áherslueininguna, og innvær birtingarmynd inniheldur einnig stöðuvísi sem segir hvort normaður tóki hefur breyst vegna stafsetningarleiðréttingar. Áherslueiningin getur svo sett inn *TagTokens* í lista tóka þar sem bið skal bæta inn. Heiltölurnar þrjár sem liggja undir „index“ tákna vísi tókans í setningunni (byrjar á núll) og hinar tvær tákna spönn stafsins í upprunalegu setningunni (frá 16 til 19). Þessar upplýsingar eru mikilvægar t.d. til að litamerkja upprunalegan texta á meðan talgervill talar.

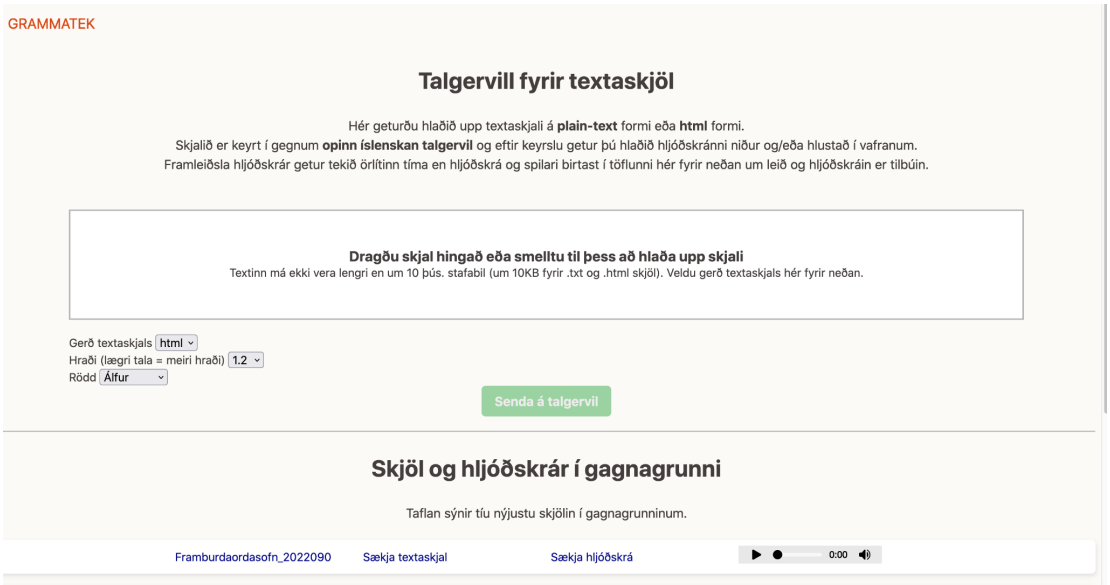
Forritaskilin eru hönnuð á þann máta að notandinn getur valið að nota aðeins hluta pípunnar, t.d. bara normun með eða án stafsetningarleiðréttingar og/eða áherslna. Pípuna má keyra í gang sem þjónustu með notkun framendaþjónustu talgervilsins, sem útfærir gRPC-þjónustu fyrir

framendapípuna. Ásamt talgervilsþjónustunni (T4) byggja þessar þjónustur upp heildstæða talgervingarkerfisþjónustu.

Vefforritið Skjalalesari (e. Document reader)

Vefforritið Skjalalesari (e. *Document reader)* les skjöl og sendir þau til talgervilsþjónustunnar sem nefnd er að ofan. Forritið útfærir biðraðarkerfi til að koma í veg fyrir að forritið þurfi að bíða eftir að talskrár séu búnar til, sem gæti tekið einhvern tíma fyrir lengri skjöl. Um leið og .mp3-skráin er tilbúin fær notandinn tilkynningu og getur spilað skrána beint í vafranum og/eða sótt hana. Á hverjum tíma eru tíu nýlegustu skrárnar sýndar í vafranum. Forritið getur lesið hreinar textaskrár, html-skrár (á EPUB-formi sem Hljóðbókasafnið notar) og pdf-skrár (engin þáttun er þó gerð á efni á pdf-skráa, aðeins einfaldur útdráttur texta).

Ein mikilvæg endurbót fyrir framleiðendur hjá Hljóðbókasafninu frá fyrstu útgáfu var að geta hægt á grunnhraða raddanna. Við bættum svo við tveimur möguleikum: fellilista til að velja rödd og fellilista til að stilla hraða, með gildum frá 0,8–1,5.



Vefforritið “Skjalalestur”.

Notkunardæmið sem valið var fyrir sérsvið fyrir prófanir er námsefni í félagsfræði. Slíkar bækur innihalda gjarnan töflur með ýmsum upplýsingum, og við notum ákveðnar leitaraðferðir til að ákveða hvernig skal lesa töflurnar. Kerfið ákveður hvenær skal endurtaka dálk- eða línuhaus og í hvaða röð töflurnar eru lesnar. Að lokum þarf alltaf að stjórna niðurstöðunni; þar sem það er stundum ómögulegt að greina sjálfkrafa í hvaða röð skal lesa töfluna er það jafnvel ekki einfalt fyrir manneskju að ákveða hvernig skal lesa stóra töflu á þann hátt að hlustandi geti fylgst með upplýsingunum.

Greining á endurgjöf frá Hljóðbókasafninu hefur sýnt styrki og veikleika núverandi kerfis, og hún sýnir einnig að einhver undirbúningsskref mun alltaf þurfa áður en hljóðbók er búin til með talgervilskerfi. Nokkur atriði:

- Núverandi kerfi er með góða dekkun á styttingum og stækkar þær oftast á rétt málfræðilegt form.
- Sjálfvirk hljóðritun óþekktra enskra orða er almennt góð, sem sýnir að bæði aðferð tungumálsgreiningar og enska hljóðritunarlíkansins virka vel.
- Leitarnám sem notað er fyrir vefhlekki og hástafatóka virkar ekki nógu vel. Við verðum að útfæra aðferðir til að greina á milli tilfella þar sem ætti að stafa hlekki og hástafatóka (eins og er nú sjálfgefið) og hvenær ætti að bera þá fram sem orð. Hástafatókar upp að ákveðnum stafafjölda eru taldir vera skammstafanir, og venjulega er ekki slæmt að bera þá fram staf fyrir staf. Þó höfum við rekist á tilfelli í hljóðbókum þar sem t.d. kaflaheiti eru rituð með hástöfum og gætu einnig hugsanlega innihaldið skammstafanir (svo ekki er nóg að stíla kaflaheiti alltaf sem venjulegan texta). Einnig er mjög tímafrekt að hlusta á hlekki sem innihalda venjuleg orð staf fyrir staf, til dæmis ‘heilsuvera.is’ sem ‘h e i l s u v e r a punktur is’.
- Jafnvel þó að við lögum almenna vinnslu mögulegra skammstafana áður en bók innan sérsviðs er unnin, eins og námsbók, þarf sérfræðingur að yfirfara framburð mikilvægra skammstafana samkvæmt almennri notkun innan tiltekins sérsviðs.
- Hið sama á við um mikilvæg erlend nöfn; þau skal umrita handvirkt til að tryggja að þau séu borin fram á skiljanlegan máta. Jafnvel þó að enska hljóðritunareiningin okkar virki nokkuð vel eru oft erfið nöfn meðal þeirra erlendu nafna sem nemendur þurfa að þekkja. Eitt dæmi: *Jean Piaget* sem ætti að bera fram nær frönskum framburði frekar en enskum.

T13 – Stikaðir talgervlar

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að eiga fyrsta flokks stikaðan talgervil fyrir íslensku. Niðurstöðurnar eftir annað ár innihalda ítarlegar rannsóknir, tilraunir og undirbúningsvinnu fyrir hönnun hágæðakerfis. Niðurstöðurnar sýna að taugatalgervlar (e. neural TTS) hafa þróast á þá leið að fýsilegt er að innleiða þá fyrir íslensku. Á þriðja ári verður áhersla lögð á að hraða þróun taugatalgervils, byggðum á FastSpeech2, sem og að kanna raddblöndun og mælendaaðlögun með því að nota gögn úr T2.

Varða 8 – lýsing

M8: SPSS-framendi samþættir textanormara úr T9.M6. Frumgerð af eftirlitslausu taugatalgervilskerfi fyrir hljómfall + áherslu, frumgerð af raddblöndun- og/eða mælendaaðlögun.

Varða 9 – lýsing

M9: Hugbúnaðarútgáfa gefin út af bættri fyrsta flokks SPSS-nálgun fyrir íslensku byggt á notendaprófunum í T8. SPSS-forskrift gefin út ,fær um raddblöndun og mælendaaðlögun.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Við höfum útfært forskriftir og gefið út þjálfuð líkön fyrir fjölbreytt söfn og högunargerðir fyrir talgervla.

- FastSpeech2 (m/ áherslu og hljómfalli): https://github.com/cadia-lvl/FastSpeech2
- FastSpeech2-líkön fyrir Talróm 1: http://hdl.handle.net/20.500.12537/201
- GlowTTS-forskrift: https://github.com/cadia-lvl/coqui-ai-TTS/releases/tag/M9
- GlowTTS-líkan fyrir Talrómur 1: http://hdl.handle.net/20.500.12537/293
- Margmælenda GlowTTS-líkan fyrir Talrómur 2: http://hdl.handle.net/20.500.12537/292
- Tacotron2 m/ GST (enska): https://github.com/cadia-lvl/espnet/tree/feature/prosody-vctk
- Tacotron2- og Fastspeech2-forskriftir sem við bjuggum til voru gefnar út í opinberri útgáfu ESPnet: https://github.com/espnet/espnet/releases/tag/v.202207
- Talrómur 1 ESPnet Fastspeech2- og Tacotron2-líkön: http://hdl.handle.net/20.500.12537/294
- X-vector Tacotron2 ESPnet-forskrift: https://github.com/cadia-lvl/espnet/tree/talromur2

- Talrómur 2 ESPnet Tacotron2-líkan: https://huggingface.co/espnet/talromur2_xvector_tacotron2

Skýrsla

Samþætting framenda
Við höfum aðlagað coqui-ai/TTS-safnið (upprunalega Mozilla/TTS) til að styðja að fullu íslensku. Við útfærðum íslenskan textahreinsara, normara (T9) og tilreiðara auk fulls stuðnings við rittákn-til-fónems (g2p) fyrir íslensku (T10). Allar forskriftir í coqui-ai/TTS-safninu hafa sjálfkrafa fullan aðgang að þessum tólum. Kóðinn er aðgengilegur á GitHub með Mozilla 2.0-leyfi og verður gefinn út á CLARIN þegar það leyfi verður mögulegt þar. Á meðan má nota þennan hlekk: https://github.com/cadia-lvl/coqui-ai-TTS/releases/tag/M9.

Bætt fyrsta flokks SPSS-nálgun fyrir íslensku (byggð á notendaprófunum í T8)
Við höfum útfært forskriftir fyrir ýmsar mismunandi högunargerðir með tiltækum talgervilssöfnum og högunum:

- GlowTTS:
  - Forskrift: https://github.com/cadia-lvl/coqui-ai-TTS/releases/tag/M9
  - Líkön: http://hdl.handle.net/20.500.12537/293
- Fastspeech2:
  - Forskrift: https://github.com/cadia-lvl/FastSpeech2
  - Líkön: http://hdl.handle.net/20.500.12537/201
- ESPnet Fastspeech2:
  - Forskrift: https://github.com/espnet/espnet/releases/tag/v.202207
  - Líkön: http://hdl.handle.net/20.500.12537/294
- ESPnet Tacotron2:
  - Forskrift: https://github.com/espnet/espnet/releases/tag/v.202207
  - Líkön: http://hdl.handle.net/20.500.12537/294

Notendaprófanir gefa til kynna að líkönin sem við höfum metið virki almennt vel, en að þau glími öll við nokkur sameiginleg vandamál, til dæmis erfiðleika við að búa til setningar sem innihalda aðeins eitt orð, og erfiðleika við að bera fram tölur. Að auki eiga sum líkönin erfitt með samsett orð, klúðra annaðhvort áherslu eða framburði. Ráðstafanir hafa verið gerðar til að taka á þessum vandamálum. Í fyrsta lagi er það gert með því að nota nýlegri g2p-umbreyta (T10) og reyna aðrar gerðir högunar sem sleppa g2p-skrefinu. Í öðru lagi með því að bæta stuttum setningum og tölum við þjálfunargögnin (T2). Við höfum verið að gera fleiri MOS-prófanir og aðrar svipaðar matsprófanir til betri fínstillingar fyrir ákveðnar gerðir högunar, og munum halda því áfram á næstu mánuðum.

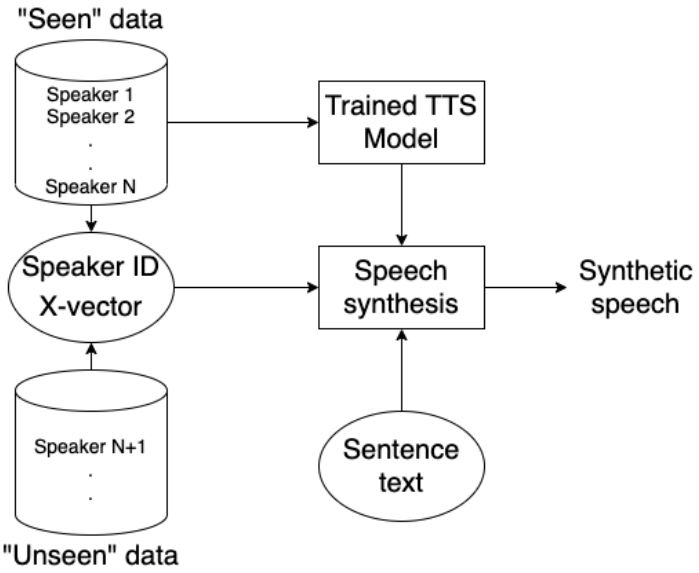
SPSS-forskrift sem getur blandað röddum og/eða aðlagað mælanda

Margmælenda GlowTTS

GlowTTS-forskriftin getur verið þjálfuð á mörgum mælendum. Þegar líkanið er keyrt má velja einn mælanda fyrir talgervingu. GlowTTS-forskriftin er aðgengileg á GitHub með Mozilla 2.0-leyfi og verður gefin út á CLARIN þegar það leyfi verður mögulegt þar. Á meðan má nota þennan hlekk: https://github.com/cadia-lvl/coqui-ai-TTS/releases/tag/M9
Líkan þjálfað að hluta til hefur verið gefið út á CLARIN: http://hdl.handle.net/20.500.12537/292

X-vigra-skilyrt margmælenda Tacotron TTS
Ólíkt GlowTTS-forskriftinni, sem notar auðkenni mælanda til að velja rödd fyrir talgervingu, notar X-vigraforskriftin sérþjálfað X-vigramælandaauðkenni sem inntak til að talgera. Þetta gerir mögulegt að talgera tal óséðra mælenda með því að búa til X-vigur með litlu magni gagna frá nýjum mælanda. Útkoman getur búið til talgert tal fyrir hvaða mælanda sem er sem þú hefur lítið magn gagna fyrir.
Forskriftin er aðgengileg hér, og verður á endanum ýtt til aðalhirslunnar: https://github.com/cadia-lvl/espnet/tree/talromur2.

Líkanið er gefið út á Hugging Face, í stað CLARIN, þar sem Hugging Face hefur möguleika þess að sækja og keyra líkanið innan forskriftarinnar. Hlekkur á líkanið: https://huggingface.co/espnet/talromur2_xvector_tacotron2



Frumgerð hljómfalls og áherslu
Grunnlíkön talgervingar í Tacotron2 og Fastspeech2 hafa innbyggða tilhneigingu til að læra einhverja mynd hljómfalls og áherslu. Sem dæmi um áherslu getur Fastspeech2 lært stutta og langa bið ef slíkt er til staðar í þjálfunargögnunum, og getur svo bætt þeim við hvar sem er í talgervingu, sem orsakar eðlilegra flæði þegar komma eða punktur eiga við. Fastspeech2 getur einnig framkvæmt einfalda stjórnun hljómfalls með því að breyta tónhæð og styrk, auk þess að

hægja eða hraða á tali. Þetta hefur verið útfært með ESPnet og er aðgengilegt á GitHub, sem hluti opinberu útgáfunnar: https://github.com/espnet/espnet/releases/tag/v.202207.

Við gerðum tilraunir með notkun GST (e. global style tokens) fyrir hljómfall og áherslu, ásamt Tacotron2-höguninni[39] frá ESPnet. Í þessari tilraun var líkan þjálfað með VCTK[40], opnum mælendagögnum á ensku, útfærð með 125 tókum. GST er eftirlitslaus flokkari sem er notaður ásamt Tacotron2 við þjálfun. Í tilrauninni lentum við í tveimur vandamálum: eitt vegna Tacotron2 og annað vegna takmarkaðs magns samhliða tjáningartaldæma í málheildinni.

Tacotron2-högunin sem við notuðum fyrir tilraunina innihélt ekki GST-veljara og við gátum ekki valið tjáningartóka (e. expressional tokens) þegar líkanið hafði verið þjálfað. Vegna þess gátum við ekki valið mismunandi GST-tjáningar fyrir gefið talgert taldæmi. Annað vandamál kom upp vegna þess að þetta er eftirlitslaust nám, og við verðum því að mata það með fjölbreyttum gögnum (í þessu tilviki tjáningartalmálsheild) svo við sem mannfólk getum greint mun. Þegar GST er beitt á talmálsheild án tjáninga sýna niðurstöður að það getur ekki búið til aðgreinanlega hljómfallstóka.

Við höfum greint þrjú atriði til að hafa í huga fyrir frekari rannsóknir byggðar á vinnu okkar með hljómfall:

- Fastspeech2 er betri valkostur fyrir GST en Tacotron2 þar sem hægt er að skeyta inn í það GST völdum af notanda.
- Frekar en að reiða okkur á stóru talmálsmálheildina þurfum við að hafa eða búa til vel skilgreinda talmálsmálheild áður en við getum bætt frekara hljómfalli við.
- Þar sem hljómfall er mjög stórt fræðisvið verðum við fyrst að finna út hvaða gerð hljómfalls við viljum búa til áður en hægt er að framkvæma það. Einnig ættum við að einbeita okkur að einu atriði frekar en að reyna að leysa mörg atriði í einu.

[39] https://github.com/cadia-lvl/espnet/tree/feature/prosody-vctk
[40] https://datashare.ed.ac.uk/handle/10283/3443