

Stöðuskýrsla M8

10.06.2022

Eftirfarandi skýrsla lýsir stöðu allra verkþætta máltækniáætlunar hvað varðar skil við M8-vörðu og inniheldur hlekkir á afurðir þar sem það á við. Skýrslan er ætluð sem fylgiskjal við skil M8-vörðu, en skil vörðunnar felast að mestu leyti í kynningarfundum fyrir menningarmálaráðuneyti og stjórn Almennaróms, sem fer fram 14. júní.

<i>Málföng</i>	2
<i>Stoðtöl</i>	3
<i>Vélpýðingar</i>	5
<i>Málrýni</i>	11
<i>Talgreining</i>	14
<i>Talgerving</i>	17

Málföng

G1 – Risamálheildin

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

Öllum markmiðum vörðunnar hefur verið náð. Viðbætur við samfélagsmiðlamálheildina hafa verið kláraðar og vinna við texta bóka og fræðigreina heldur áfram.

Samskipti við útgefendur héldu áfram á M8-tímabilinu, en þau gáfu af sér takmarkaðar afurðir. Þó að margir útgefendur hafi verið jákvæðir gagnvart því að senda inn texta hefur ferlið við að raunverulega senda þá inn gengið hægt. Nokkrir stórir útgefendur eru þó enn líklegir til þess að senda inn texta á næstu vikum, en það tekur þó sinn tíma.

Samningaumræður við Menntamálastofnun fóru fram varðandi að nota texta þeirra sem tilheyra höfundum sem tilheyra ekki Rithöfundasambandi Íslands eða Hagþenki, félagi höfunda fræðirita og kennslugagna. Stofnunin hefur verið mjög jákvæð varðandi að aðstoða okkur, og lét okkur nýlega hafa þær upplýsingar sem við þurfum til þess að fá leyfi frá fyrnefndum höfundum.

Vinna við texta bóka og fræðigreina heldur áfram. Á M8-tímabilinu var áhersla lögð á fræðigreinar, en með nýtilkomnu samkomulagi við Menntamálastofnun verður mögulegt að vinna þó nokkuð magn bóka á M9-tímabilinu.

Stoðtöl

I5 – Þáttarar

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

Öllum M8-markmiðum hefur verið náð. UD-varpari hefur verið aðlagður að trjágerð Greynis og gefinn út á CLARIN (<http://hdl.handle.net/20.500.12537/222>). Undirsetti Greynismálheildar hefur verið varpað yfir á UD-gerð og gefið út á CLARIN (<http://hdl.handle.net/20.500.12537/224>). Málheildin verður hluti af næstu útgáfu UD, útgáfu 2,11, sem kemur út næstkomandi nóvember.

Það undirsett Greynismálheildar sem var varpað og gefið út er gullhluti þess, sem hefur verið handyfirfarinn. Málheildin samanstendur af 5000 setningum og er skipt upp í þróunar- og prófunarhluta. Vörpunartólið má nota til að varpa fleiri hlutum GreynirCorpus, og það verður gert í framtíðinni.

I8 – BERT-lík mállíkon

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

Viðmiðið fyrir fjögur íslensk NLP-verk hefur verið undirbúið, en markmiðum hefur ekki verið náð að fullu. Verkin fjögur eru eftirfarandi:

Verk	Málheild	Forritasafn
Málfræðileg mörkun	MIM-GOLD	Transformers
Nafnakennsl	MIM-GOLD-NER	Transformers
Vélrænn útdráttur	IceSum	TransformerSum
Venslaþáttun	IcePaHC	DiaParser

Viðmiðið samanstendur af skriftu sem finstillir og metur forþjálfað mállíkan á hverju verki með því að nota viðeigandi forritasafn. Málheildum er halað niður annaðhvort handvirkt eða með því að nota skriftuna. Önnur skrifta býr til sérstakar skiptingar fyrir hverja málheild og varpar þeim yfir á snið sem samrýmist forritasafninu.

Ný útgáfa af MIM-GOLD-NER-málheildinni verður gefin út á næstu dögum (snemma í júní). Þessi útgáfa leysir ýmis misræmi við MIM-GOLD, sem málheildin er byggð á, og þá aðallega mun í tókun (t.d. skiptingu setninga og hvernig unnið er með ákveðin greinarmerki). Ætlunin okkar er að búa til skiptingar í þjálfunar- og prófunargögn sem eru eins og þær sem eru í MIM-GOLD. Þar sem skriftan sem við notum til þess að búa til skiptingar í þjálfunar- og prófunargögn er byggð á setningarvísunum, sem verða lítils háttar breyttir í tilvonandi útgáfu af MIM-GOLD-NER, finnst okkur viðeigandi að seinka útgáfu á viðmiðinu um nokkra daga, þar til nýja útgáfan hefur verið gefin út. Viðmiðið verður þar af leiðandi gefið út snemma í júní, stuttu eftir að nýja útgáfan af MIM-GOLD-NER kemur út.

Bráðabirgðaniðurstöður úr tilraunum með að nota flokkara til þess að sía burt lággæða vélþýddan texta hafa lofað góðu. Beðið er með frekari tilraunir þar sem líkön eru forþjálfuð á stórum söfnum af vélþýddum texta (bæði síuðum og ósíuðum) þar til ný vélþýdd málheild á

skjalastigi (En-Is) frá Miðeind er tilbúin, sem gerist vonandi á næstu vikum. Þar sem engin sérstök markmið eru varðandi tilraunir með gagnastækkun (þ.e. að bæta vélþýddum texta við forþjálfunarmálheildina okkar) ætlum við okkur að lýsa niðurstöðum okkar í ráðstefnugrein sem við gerum ráð fyrir að senda inn seinna á þessu ári. Við enda ársins munum við gefa út ELECTRA-Base-líkan sem hefur verið forþjálfað á málheild sem hefur verið stækkuð með vélþýddum texta.

Vélpýðingar

V2b – Bakpýðingar

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

Öllum M8-markmiðum hefur verið náð. Bakpýðingargögnum hefur verið safnað og þau pýdd. Búið er að bera saman aðferðir við ályktanabestun og sýna fram á að hægt er að minnka líkanið um $\frac{3}{4}$ hluta með litlu tapi í frammistöðugetu.

Eftirfarandi gögn hafa verið pýdd til þess að nota sem bakpýðingargögn í pýðingarvél fyrir skjalastig:

Einmála gagnasett	Upprunatungumál	Tókar (milljónir)
Risamálheildin (án ípróttatexta) (IGC)	Íslenska	1.461
Íslenska Common Crawl-málheildin (IC3)	Íslenska	712
Lokaritgerðir nemenda	Íslenska	322
Greynir News	Íslenska	81
Wikipedia	Íslenska	9
Íslendingasögur	Íslenska	1,5
Íslenskar raðbækur	Íslenska	1,9
Books3	Enska	792
NewsCrawl	Enska	787
Wikipedia	Enska	751
EuroPARL	Enska	59,7
Reykjavík Grapevine	Enska	11
Iceland Review	Enska	5,4

Tekist hefur að minnka pýðingarlíkön án mikillar lækkunar í frammistöðugetu á almennu sviði. Niðurstöður þeirra tilrauna eru sýndar í töflunni að neðan. Upprunalega líkanið er 24 laga, en það var minnkað í 7 lög með eimingu (e. distillation). BLEU-skor, sem mælir nákvæmni vélpýðinga, minnkar lítillega fyrir fréttatexta úr WMT-2021-gagnasettinu, frá 29,4 í 28,7. Á sérsviði EES-reglugerðanna minnkar skorið meira, frá 59,9 í 41,4, eins og búast má við.

BLEU-prófunarskor	Kennari	6-6 nemandi	12-1 nemandi	6-1 nemandi
EES	59,5	45,1	43,8	41,4
WMT-2021 prófun	29,4	29,6	29,3	28,7
Flores devtest		28,2	28,1	27,7
Hraði setn./sek.				
CPU B=32	1,67	3,74	4,70	4,98
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	24,79	46,67	67,34	73,13

V4a – Opin þýðingarvél fyrir íslensku

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

Öllum M8-markmiðum hefur verið náð. Búið er að útfæra virkt þol bakþýddra upprunatexta og hægt er að þjálf þýðingarkerfið á skjalastigi. Fyrsta ítrun líkans hefur verið þjálfuð, en óstöðugleiki í þjálfun veldur eins og er frammistöðu sem er lægri en æskilegt er. Greint verður frá yfirgripsmiklum samanburði við fyrri aðferðir þegar leyst hefur verið úr þessu.

V4b – Opin þýðingarvél fyrir íslensku – þýðingarvél fyrir sérsvið

M8: Fínstillingarpípu pakkað og hún gefin út.

M8-markmiðum hefur verið náð. Pípa fyrir þýðingar fyrir sérsvið hefur verið sett upp og er bæði aðgengileg á GitHub (<https://github.com/mideind/domain-translation-pipeline>) og á CLARIN (<http://hdl.handle.net/20.500.12537/212>). Í M7-skýrslu kom fram samanburður á þýðingargæðum milli sérsviða. Líkanið hefur einnig verið sett upp fyrir samstarfsaðila í þýðingarmiðstöð utanríkisráðuneytisins gegnum Trados Studio-viðbót.

V4g – Opin þýðingarvél fyrir íslensku – Orðasambönd og þýðingar

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

M8-markmiðum hefur verið náð og um 2400 margorða segðum hefur verið safnað. Lista yfir meira en 4000 margorða segðir var safnað frá ISLEX (<https://islex.arnastofnun.is/>). Farið var yfir listann og hann unninn, og um 2400 margorða segðir voru metnar sem mikilvægastar og hæfar fyrir þýðingu. Þýðing á þeim segðum er hafin og mun standa yfir á M9-tímabilinu.

Margorða segðirnar voru unnar með eftirfarandi sniði:

ISLEX auðkenni segðar	00000
ISLEX auðkenni	00000
Upprunaleg segð	<NP1-nom> taka <NP2-acc> í bakariíð
Markmiðssegð	<NP1> take <NP2> to the cleaners
Upprunaleg merking	<NP1-nom> sigra <NP-ACC>
Dæmi um upprunalega segð	Jón tók Guðmund í bakariíð.
Dæmi um markmiðssegð	Jón took Guðmundur to the cleaners
Dæmi um upprunalega merkingu	Jón sigraði Guðmund
Dæmi um bókstaflega merkingu	Jón beat Guðmundur

Eftirfarandi myndir sýna hvernig margorða segðirnar voru unnar:

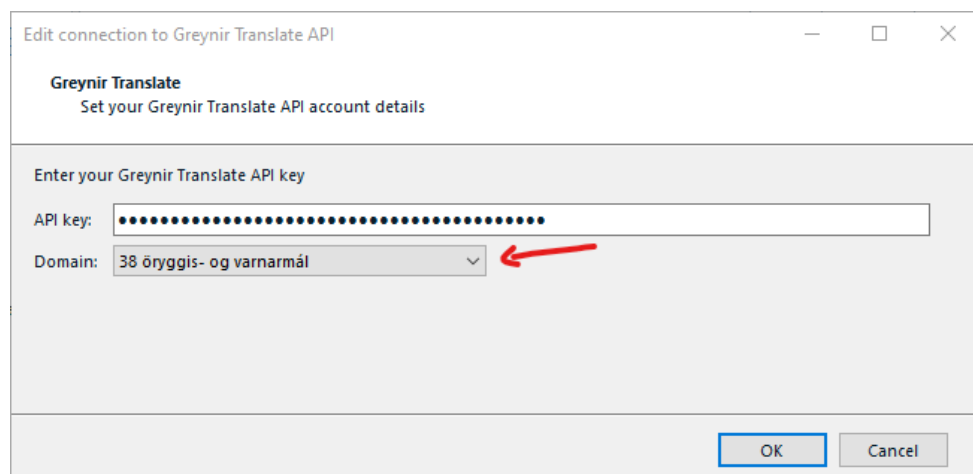
A	B	C	D	E	F	G
ISLEX idiom ID	ISLEX ID		Source idiom	Target idiom	Source sense	Source idiomatic example
278854	26687	leggja línurnar/lína	<NP1-nom> leggja línuna (fyrir <NP2-acc>)	<NP1> call the shots (for <NP2>)	<NP1-nom> ákveða stefnuna (fyrir <NP2-acc>)	Sigurður leggur línuna fyrir Guðmund.
267603	11857	það ferist fyrir í leikun	<NP1-exp> ferist fyrir í leikun	<NP1> the game is afoot	<NP1-exp> lína yfir hlutunum	það ferist fyrir í leikun
278853	26687	leita->þrsta á milli línanna	<NP1-nom> leita->þrsta á milli línanna	<NP1> read <NP2> between the lines	<NP1-nom> skilja það sem ekki er sagt beinum orðum	Sigurður les gagnrýni á milli línanna.
278852	26687	þetta eru hreinar línur	<NP1-nom> vera hreinar línur	<NP1> be clear cut	<NP1-nom> vera alveg klárt	málið er hreinar línur
278844	26764	það kom í ljós að <ég hafði misklíð þetta>	<NP1-nom> koma í ljós	<NP1> become apparent	<NP1-nom> sýna sig	sannleikurinn kemur í ljós
278843	26764	það rennur upp fyrir <þomum> ljós	ljós renna upp fyrir <NP1-dat>	dawn on <NP>	<NP1-nom> öðlast skilning	ljós rennur upp fyrir Sigurði
270024	9947	hella öllu á eldinn	<NP1-nom> hella öllu á eldinn	<NP1> add fuel to the fire	<NP1-nom> valda meiri óróa	Sigurður helta öllu á eldinn
278853	26872	gripa andann á lofti	<NP1-nom> gripa andann á lofti	<NP1> catch one's breath/ <NP1> gasp	<NP1-nom> taka andköf-<NP-nom> anda snögglega inn	Sigurður gripir andann á lofti
278828	26872	vera ekki hárl í loftinu	<NP1-nom> vera ekki hárl í loftinu	<NP1> be short in stature	<NP1-nom> vera lágvaxinn	Sigurður er ekki hárl í loftinu
278826	26872	vera mikill á lofti	<NP1-nom> vera mikill á lofti	<NP1> be putting on airs	<NP1-nom> vera góður með sig og láta á sér bera	Sigurður er mikill á lofti
535523	53540	<ta embættið>-á silfurfat	<NP1-nom> ta-<NP2-acc>-á silfurfat	<NP1> receive <NP2> on a silver platter	<NP1-nom> lá-<NP2-acc>-án þess að hafa fyrir þá	Sigurður fær embættið á silfurfat
785953	27787	troða marvaðinn	<NP1-nom> troða marvaðinn	<NP1> tread underfoot	<NP1-nom> halda sér á floti í vatni	Sigurður treður marvaðinn

H	I	J	K
Target idiomatic example	Source sense example	Target literal example	Keywords
Sigurður calls the shots for Guðmundur.	Sigurður ákveður stefnuna fyrir Guðmund	Sigurður decides how Guðmundur will proceed	lína
the game is afoot	það lífnar yfir hlutunum	things are getting exciting	fjör,leikur
Sigurður reads between the lines of the review	Sigurður skilur það sem ekki er sagt beinum orðum	Sigurður comprehends the implied meaning	lína
the case is clear cut	málið er alveg klárt	the case is clearly understood	lína
the truth becomes apparent	sannleikurinn sýnir sig	the truth becomes known	ljós
it dawns on Sigurður	Sigurður öðlast skilning	Sigurður gains understanding	ljós
Sigurður added fuel to the fire	Sigurður veldur meiri óróa	Sigurður causes greater unrest	eldur
Sigurður caught his breath Sigurður gasped	Sigurður tekur andköf,Sigurður andar snögglega inn	Sigurður gasps for breath,Sigurður takes a sudden breath of air	andi,loft
Sigurður is short in stature	Sigurður er lágvaxinn	Sigurður is short	loft
Sigurður is putting on airs	Sigurður er góður með sig og lætur á sér bera	Sigurður acts pompous and wants to be noticed	loft
Sigurður received the office on a silver platter	Sigurður fær embættið án þess að hafa fyrir því	Sigurður receives the office without having to exert himself	silfurfat
Sigurður treads water	Sigurður heldur sér á floti í vatni	Sigurður keeps himself afloat in the water	marvaði

V5 – Vefviðmót fyrir vélþýðingar

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

M8-markmiðum hefur verið náð, og þá sérstaklega hvað varðar að þjóna þýðingarvél fyrir sérsvið (V4b). Viðmót Trados 2021-viðbótarinnar hefur verið aðlagð að þýðingu fyrir sérsvið (V4b). Fellivalmynd er núna í boði í viðbótinni, eins og sýnt er á myndinni að neðan.



Almennar endurbætur á stöðugleika hafa einnig verið gerðar.

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

M8: Forþjálfað margmála mállíkan valið og finnstílt.

M8-markmiðum hefur verið náð. Tilraunir með fjölmála líkan með mjög mörg tungumál skiluðu sér ekki í hágæða þýðingum. Upprunalega var DeltaLM-líkanið notað og tvö-til-margra (2-to-many) og margra-til-2 (many-to-2) aðferð, þar sem þýtt er úr og í ensku og íslensku á sama tíma. Talsverðum fjölda gagna var safnað, en fyrir sum tungumál, eins og íslensku-pólsku, vantaði fleiri samhliða gögn.

Eftir að hafa farið yfir gögnin varð ljóst að hægt var að samraða gögnum með því að bera saman enskar setningar í t.d. íslensk-ensku og ensk-pólsku samhliða gögnunum. Þetta skilaði sér í ágætlega góðu íslensk-pólsku gagnasetti með meira en 500.000 setningapörum. Þessar setningar virðast að mestu leyti vera EES-reglugerðir, læknisfræðilegar lýsingar frá Lyfjastofnun Evrópu og trúartextar.

Með öll þessi gögn er ljóst að svipaðar aðferðir og notaðar voru í íslensk-enska verkefninu eru mögulegar. Við færðum okkur frá DeltaLM-líkaninu fyrir þetta, og fórum aftur í mBART, eins og gert var fyrir íslensk-enska parið. Í stað þess að nota mBART-25-líkanið, sem notað var fyrir íslensk-enskar þýðingar, notum við mBART-50, sem er að hluta til þjálfað á pólskum gögnum. Fyrstu niðurstöður lofa góðu og eru sýndar í töflunni að neðan.

Prófunargögn	Ís-pól BLEU	Pól-ís BLEU
Sérsvið	26,49	26,70
Flores	9,73	11,10

Dæmi um þýðingar frá pólsku yfir á íslensku, með gögnum frá þjálfunarsérsviðinu, eru hér að neðan.

Pólska	Þýdd íslenska	Upprunalegur texti
Albert Muchanga, komisarz ds. handlu i przemysłu w Unii Afrykańskiej, ogłosił przyłączenie się Beninu do bloku.	Albert Muchanga, framkvæmdastjóri viðskipta og iðnaðar í Afrikusambandinu, tilkynnti að Benín skyldi falla í flokkinn.	Albert Muchanga, stjórnarmaður verslunar- og iðnaðarráðs Afrikusambandsins, tilkynnti að Benín ætlaði að taka þátt.
Właściwe odżywianie nie gwarantuje doskonałych wyników, ale może znacząco wpłynąć na ogólną formę młodych sportowców.	Rétt mataræði getur ekki tryggt fullkominn árangur en það getur haft veruleg áhrif á almenna framvindu íþróttamanna.	Réttar næringaraðferðir einar og sér veita ekki úrvalsframmistöðu. Þær geta samt haft veruleg áhrif á heildarvellíðan ungra íþróttamanna.
Zarówno Łotwa, jak i Słowacja odkładały procedurę przystąpienia do ACTA.	Bæði Lettland og Slóvakía lögðu niður málsmeðferðina við þátttöku í ACTA-samþykktinni.	Lettland og Slóvakía hafa bæði tafið ferlið við inngöngu í ACTA.
Bazą narracyjną gry jest druga bitwa o Faludzę, zaciekle starcie między oddziałami amerykańskimi a siłami Iraku.	Siðari stríðið gegn Faludsiu, sem átti sér stað á milli bandarískra hersveita og hershöfðingja Íraks, er aðalpersóna leiksins.	Leikurinn er byggist á siðari orrustunni um Fallujah, sem var hræðilegur bardagi á milli bandarískra og íraskra hersveita.
Silna paszcza mogła być przydatna dla triceratopsa, gdy chciał oberwać liście, zanim zabrał się do zjadania pnia.	Það getur verið gott fyrir triceratopsa að hafa sterka hreyfingu áður en hann sækir fuglinn.	Nashyrningseðla gæti hafa notað sterkeyggðan gogginn til að fjarlægja laufblöðin áður en hún át trjábolinn.
Większość przedmiotów, które pochowano razem z Tutenchamonem, było dobrze zachowanych, wliczając tysiące artefaktów zrobionych z metali szlachetnych i rzadkich kamieni.	Meirihluti þeirra hluta, sem grafnir voru ásamt Tutenchamon, var vel varðveittur, þar á meðal þúsundir steina úr góðum og sjaldgæfum málmi.	Flestir munirnir sem grafnir voru með Tutankhamun hafa verið vel varðveittir, þar á meðal þúsundir muna sem unnir voru úr góðmálmum og sjaldgæfum steinum.
Sikhizm jest religią, która narodziła się na subkontynencie indyjskim w regionie Pendżabu w XV wieku wskutek rozłamu wewnątrz hinduskiej tradycji.	Sikhismi er trú sem fæddist á Indlandi á Pendžabusvæðinu á 15. öld eftir að hindúar höfðu brotið niður erfðavenjur sínar.	Sikhismi eru trúarbrögð frá Indlandi. Þau eiga uppruna sinn í Punjab-héraðinu á 15. öld eftir trúarklofningu innan hindúatrú.
Obcasy muszą być niskie i szerokie. Lepszą przyczepność uzyskuje się poprzez rozsypywanie piasku, żwiru lub soli (chlorku wapnia) na drogach i ścieżkach.	Lítið og breitt svæði verður að vera til viðhalds með því að dreifa sandi, kvoðu eða salti (kalsíumklóríð) á vegum og göngum.	Hælar eiga að vera lágir og breiðir. Sandi, mól eða salti (kalsíumklóríð) er oft dreift á vegi og slóða til að auka grip.
Zaobserwowano niecałe tysiąc przypadków wśród ludzi, lecz niektóre okazały się śmiertelne.	Aðeins þúsund tilfelli hafa komið fram hjá mönnum en sumar hafa verið banvænar.	Færri en þúsund tilfelli í mönnum hafa verið skráð, en sum þeirra hafa leitt til dauða.
Victoria Falls jest miastem położonym w zachodnim regionie Zimbabwe, nieopodal miasta Livingstone znajdującego się tuż po drugiej stronie granicy, w	Victoria Falls er borg í vesturhluta Simbabve, nálægt Livingstone sem er rétt við landamærin í Zambí og í náinni	Victoria Falls er bær í vestuhluta Zimbabve, á móti Livingstone í Zambíu hinum megin landamæra og nálægt Botswana.

Zambii, a także w bliskiej odległości od granicy z Botswaną.	fjarlægð frá landamærunum við Botsvan.	
--	--	--

Málrýni

L2 – Sérhæfðar villumálheildir

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

Öllum M8-markmiðum hefur verið náð. Söfnun á textum frá fjórum innflytjendahópum, Pólverjum, Filippseyingum, Þjóðverjum og Litháum, er lokið og flokkun er hafin. Söfnun á textum frá einstaklingum með lesblindu er í gangi og gengur vel. Flokkun á þeim textum er hafin.

Við enduðum með ásættanlegan fjölda texta frá innflytjendahópunum, 11 texta eftir Pólverja, 12 eftir Filippseyinga, 13 eftir Þjóðverja og 6 eftir Litháa. Textarnir eru mjög misjafnir að lengd og eins er mjög misjafnt af hvaða getustigi höfundarnir eru, svo það er góð breidd í textunum. Í annarsmálmálheildinni í heild, þar sem textar frá höfundum af fleiri þjóðernum eru, eru 102 textar sem innihalda 162.824 orð og 17.223 leiðréttingarspannir.

Í lesblindumálheildinni eru núna 35 textar sem innihalda samtals 38.919 orð og 5.085 leiðréttingarspannir.

L7 – Kerfisbundin þróun málrýnis

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

M8-markmiðum hefur ekki verið náð að öllu leyti, þ.e. lokaritgerðarprófunum er ekki lokið. Að öðru leyti hefur markmiðum verið náð. Búið er að skipuleggja prófanir, sem verða gerðar á lokaritgerðum sem hýstar eru á Skemmunni. Ritgerðirnar verða prófarkalesnar og þær útgáfur bornar saman við ritgerðir sem hafa verið leiðréttaðar af málrýninum. Þessi vinna verður unnin á næstu vikum og mun því ekki hafa mikil áhrif á M9-markmið þessa verkþáttar.

Ný útgáfa af GreynirCorrect er komin á CLARIN (<http://hdl.handle.net/20.500.12537/218>). Almennar leiðbeiningar eru nú í boði fyrir alla villuflokka sem eru höndlaðir í kerfinu. Auk þess fylgja stórum hluta villnanna ítarlegri leiðbeiningar, svo sem ákveðnum málfræðivillum sem auðvelt er að einangra.

Villuflokkar sem koma úr Ritmyndum (gögnum úr Stórasniði BÍN) hafa nú verið tengdir við Ritreglur, þannig að í viðmóti getur notandi smellt á tákn og hann færist þannig á viðeigandi færslu innan vefsíðunnar <https://ritreglur.arnastofnun.is>. Þetta hefur verið gert fyrir helstu villuflokka upp úr gögnum, en er erfiðara fyrir villur sem eru sjálfvirkir leiðréttaðar, þar sem þær eru aðeins leiðréttaðar en engar frekari upplýsingar koma fram. Skoða mætti flokkun á þeim með tauganetsaðferðum síðar meir. Fyrir utan þetta hafa farið fram endurbætur til að lagfæra villur og annað sem hefur fundist við prófanir.

L8 – Málrýni í snjalltækjum

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.

M8-markmiðum var náð að mestu leyti. Búið er að þróa frumgerð málrýnieiningar fyrir Android sem byggð er á vefþjónustu GreynirCorrect, en hún hefur ekki verið gefin út á Google Play. Hún er aðgengileg á GitHub (<https://github.com/grammatek/simacorrect/tree/M8>) og

kallast Símacorrect. Íslensk orðabók fyrir AnySoft Keyboard var bætt og togbeiðni í upprunalega hirslu hefur verið opnuð.

Málrýnieiningin Símacorrect var þróuð og málfræðiskýringum fyrir setningar hefur verið bætt við. Það eru þó enn vandamál við að varpa innbyggðu Android-tókuninni yfir á tókana sem Yfirlestursforritaskilin skila. Verið er að vinna úr þessu og var því ákveðið að gefa ekki betaútgáfu málrýnieiningarinnar út strax.

Nýr íslenskur tungumálapakki fyrir AnySoft Keyboard var búinn til fyrir sjálfvirka útfyllingu (e. auto completion) og pakinn hefur verið sameinaður aðalgeymslu hugbúnaðarins. Samsvarandi togbeiðni á GitHub má finna hér: <https://github.com/AnySoftKeyboard/AnySoftKeyboard/pull/3389>. Tungumálapakinn inniheldur íslenska orðabók sem byggð er á Risamálheildinni, með um 219.000 orð, og normaðar tíðnir sem eru samsvarandi sniðinu sem AnySoft Keyboard þarfnast.

L9 – Málrýni í ritvinnslukerfum

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

M8-markmiðum hefur ekki verið náð að öllu leyti, og á þetta bæði við um notendaprófanir Google Docs-viðbótarinnar og þróun MS/MacOS Office-innleiðingar.

Notendaprófanir á Google Docs-viðbótinni standa yfir og ganga vel. Notendurnir sækja viðbótina á Google Workspace Marketplace til prófunar. Stuttu fyrir M8-skil lokaði Google fyrir aðgang að viðbótinni á vettvanginum, þar sem villur komu upp í skráningu viðbótarinnar á vettvanginum. Þetta hefur verið lagfært en mun taka nokkra daga að færast inn aftur. Upplýsingasíða um viðbótina sjálfa er því ekki aðgengileg á þessari stundu en bakendakóðinn er aðgengilegur á GitHub hér: <https://github.com/hinrikur/Yfirlestur-docs>.

Til þess að tryggja nægilegt magn af svörum frá notendum munu notendaprófanir standa yfir til 12. júní, en þessi seinkun mun ekki hafa mikil áhrif á vinnu við M9-vörðu.

Þróun MS/MacOS Office-innleiðingar gekk vel í fyrstu en vegna seinkunar sem varð á framgangi notendaprófana á Google Docs-viðbótinni var ákveðið að leggja áframhaldandi þróun til hliðar þangað til notendaprófanir væru komnar á strik. Við leggjum því til að fresturinn til að skila M8-markmiðum fyrir MS/MacOS Office verði færður til 1. júlí, svo hægt sé að ganga frá fyrstu útgáfu.

L11 – Villumódel fyrir ljóslestur

M8: Aðferðafræði komið á.

M8-markmiðum hefur verið náð. Þrjár tilraunir hafa verið gerðar á samhliða gögnunum sem í boði eru, en þau innihalda 50.000 línur.

Fyrsta tilraunin var með reglubýggð stafavíxl ásamt leit í orðabók. Þessi aðferð er hentug fyrir litlar, fyrirsjáanlegar villur sem finnast í ljóslesnum textum af ágætum gæðum. Þessa aðferð mætti nota sem stuðning, eða eftir aðra villuleiðréttingu, en býður að öðru leyti upp á lítið.

Önnur tilraunin var með Seq2Seq LSTM-líkan. Það var þjálfað með 10.000, 20.000, 30.000, 40.000 og 50.000 línur af texta. Þó að þessi aðferð hafi í fyrstu lofað góðu bættust niðurstöðurnar ekki bersýnilega.

Þriðja tilraunin var með Seq2Seq Transformer-líkan ásamt gervivillum. Aðferðin skilaði fyrst betri niðurstöðum en LSTM-líkanið, en batnaði ekki bersýnilega á milli 10.000 og 50.000 línur af texta. Um 2.500 stafavíxl (t.d. $p \leftarrow \rightarrow p$ og $m \leftarrow \rightarrow rn$) voru dregin út úr 80.000 línur af samhliða ljóslesnum texta.

Engar viðmiðunarprófanir hafa verið gerðar, en síðastnefnda aðferðin skilar efnilegum niðurstöðum og verður því notuð í framhaldinu. Auk gervistafavíxlanna ætti öðrum algengum ljóslestursvillum, t.d. óviljandi orðaskiptingum og -samrunum, að vera bætt við þjálfunargögnin.

L14 – Málrýni með djúpum tauganetum

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 villumálheildir hafnar.

Öllum M8-markmiðum hefur verið náð. Gervivillumálheild var búin til úr almennu villumálheildinni, og hún inniheldur 35 milljónir lína. Villuhlutfall málheildarinnar er mjög hátt, þar sem það kom best út í þjálfunum.

Tilraunir með þjálfun ýmissa þýðingarneta voru framkvæmdar. Ýmsar tilraunir með þjálfun mBART-líkans, sem kallast *mbart2-dae-enis-gec-icelandic*, fyrir þýðingu milli brenglaðrar og réttar íslensku, voru framkvæmdar. Einnig var prófað að þjálfu með þróunarhluta almennu villumálheildarinnar og auka vægi þeirra gagna í þjálfuninni. Auk þess var önnur útfærsla prófuð til þess að sjá samanburð. Það er líkanið ByT5, sem er ekki þjálfuð með orðflísaskiptingu, líkt og gert er í hinum líkönunum, heldur er textanum skipt upp í bæti. Þetta líkan lofar líka góðu og verður þjálfuð frekar í framhaldinu.

Líkönin eru metin á prófunarsettum úr gervivillugögnunum, á prófunarhluta almennu villumálheildarinnar og á sérhæfðu villumálheildunum. Niðurstöðurnar sýna að netin ráða vel við að afbrensla texta með mjög mikið af villum, sérstaklega einföldum villum, en einnig að þeim takist að laga sumar flóknari villur eins og þágufallshneigð eða rangan hátt sagna.

Tauganetsflokkari fyrir valda villuflokka er kominn út á CLARIN hér: <http://hdl.handle.net/20.500.12537/217>.

Talgreining

H5 – Samröðun á stórum málheildum

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

Öllum M8-markmiðum hefur verið náð. Til þess að prófa samröðunarferlið voru gögn úr tveimur hlaðvörpum stöðluð, og hrá gögn eru nú samtals 135 klst. Við höfum einnig fleiri gögn úr nokkrum öðrum hlaðvörpum, sem hafa ekki verið stöðluð, en þau eru samtals rétt rúmlega 100 klst. Heildarfjöldi hrárra gagna sem hefur verið safnað er því 235 klst., sem er gott magn af göngum til þess að búa til gagnasett. Þar sem einhver hluti gagnanna gæti tapast við samröðun mun lokaútgáfa gagnanna líklega innihalda um 115 klst. Við viljum hins vegar enn safna meiri gögnum þannig að fjöldi hrárra gagna ætti að aukast fyrir M9-vörðu. Skriftur til að staðla gögn eru á GitHub hér: <https://github.com/cadia-lvl/speech-corpora-toolkit>.

Fyrsta útgáfa sneiðara og samraðara, MAFIA (Match-Finder Aligner), er tilbúin og hefur verið gefin út á CLARIN hér: <http://hdl.handle.net/20.500.12537/215>. Þessa útgáfu má nota til þess að vinna hrá gögn sem safnast á M9-tímabilinu, en enn má bæta þessa útgáfu af MAFIA og hægt verður að vinna að bestun hugbúnaðarins á næsta tímabili. Ef þörf er á verður einnig unnið að því að laga smávægilegar villur.

Við höfum einnig búið til vefforrit (<https://podromur.samromur.is>) þar sem gjafar geta hlaðið upp skriftum sínum og bætt við hlekk á hlaðvörp þeirra. Hlekkur á þetta vefforrit hefur verið sendur á þrjá gjafa sem sagst hafa ætla að láta gögn af hendi. Undirliggjandi hugbúnaður þessa vefforrts er á GitHub hér: <https://github.com/cadia-lvl/upload-to-s3>.

H9 – Talgreining í snjallsímum

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.

M8-markmiðum hefur verið náð. Alfa-útgáfa af Heyra hefur verið gefin út á Google Play Store (<https://play.google.com/store/apps/details?id=is.tiro.heyra>), og prófun hennar er opin öllum. Útgáfan inniheldur útfærslu á talgervilsþjónustu fyrir Android-kerfi og forrit sem getur meðhöndlað talgreiningarbeiðnir um inntak frá forritum sem senda slíkar beiðnir, t.d. AnySoft Keyboard.

H10 – Raddstýring og fyrirspurnir

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

Öllum M8-markmiðum hefur verið náð. Gögnum hefur verið safnað og líkön verið búin til fyrir tvö verkefni, stuttar spurningar (e. trivia questions) og einingavörpun (e. unit conversion). Auk þess hefur gögnum verið safnað fyrir eftirfarandi verkefni: heimilisföng, símanúmer, kennitölur og full nöfn.

Stuttar spurningar. 33.000 nothæfum spurningum var safnað frá þremur stöðum: <https://spurningar.is>, Is-Trivia (<https://github.com/sveinn-steinarsson/is-trivia-questions>) og Gettu Betur. Spurningarnar úr Gettu Betur eru því miður ekki opnar almenningi og þær verða ekki gefnar út, en þær voru notaðar í þjálfun líkansins. Við notuðum Regina (https://github.com/cadia-lvl/regina_normalizer) til þess að skrifa út tölur og skammstafanir, auk regex-reglna fyrir frekari normun. Hluti spurninganna sem opinn er almenningi er

aðgengilegur á GitHub (<https://github.com/cadia-lvl/is-trivia-questions>) og líkan fyrir þetta verkefni er aðgengilegt á GitLab (https://gitlab.com/tiro-is/h10/create_models/-/blob/master/models/trivia-2022-05-31_3g.tar.gz).

Einingavörpun. Skrifta sem býr til setningar með algengum einingavörpunum var búin til. Nokkrar grunnsetningar voru búnar til, sem innihalda mælieiningar eða rúmeiningar og tölur sem skrifaðar eru út. Dæmi um vörpun fjarlægðar eru: „Hversu margar tommur eru í núll komma þrem kílómetrum?“ og „Hversu margir sentimetrar eru í fjögur hundruð fjórutíu og fjórum kílómetrum?“. Dæmi um vörpun rúmeininga eru: „Hvað eru fjögur hundruð og fimmtíu grömm mörg kílógrömm?“ og „Hvað eru sjö hundruð fimmtíu og níu desílítrar margir peli?“. Þó nokkur vinna fór í að búa til skriftu sem skrifar út tölur í réttu falli, kyni og beygingu. Þar sem þetta er nokkuð algengt verkefni var ákveðið að búa til úr þessari skriftu Python PIP-pakka sem kallast Númi (<https://pypi.org/project/numi/>). 170.000 mismunandi strengir voru búnir til en ætlunin er að bæta við þá með meiri einingavörpun. Gögnin ásamt skriftunni sem notuð var til þess að búa til gögnin eru aðgengileg á GitLab (<https://gitlab.com/tiro-is/h10/unit-conversion>) og einnig líkan fyrir þetta verkefni (https://gitlab.com/tiro-is/h10/create_models/-/blob/master/models/unit-conversion-170522_3g.tar.gz).

Mállíkönin eru þjálfuð með staðlaðri Kaldi-skriftu (https://gitlab.com/tiro-is/h10/create_models). Tiltæk líkön og gögn eru áföst með Git LFS. Öll gögn og líkön sem opin eru almenningi verða gefin út á CLARIN við M9-vörðu.

Heimilisföng. Stórt mengi mögulegra heimilisfanga var búið til með því að nota öll kunn götuheiti og hverfi. Heimilisföngin eru normuð í tveimur föllum, nefnifalli og þágufalli, í samræmi við hvernig þau eru venjulega notuð. Gögnin hafa ekki verið gefin út en verða gerð aðgengileg við næstu vörðu, ásamt Kaldi-forskrift.

Símanúmer. Stór listi yfir 7 stafa símanúmer var búinn til, en númerin fylgja reglum yfir íslensk símanúmer frá Fjarskiptastofu. Númerin eru normuð og eru öll skrifuð út á nokkra mismunandi vegu til þess að fylgja óreglulegu mynstri sem símanúmer hafa í töluðu máli. Gögnin hafa ekki verið gefin út en verða gerð aðgengileg við næstu vörðu, ásamt Kaldi-forskrift.

Kennitölur. Stórt mengi kennitalna var búið til, en þær fylgja stöðlum um kennitölur fólks og fyrirtækja. Kennitölurnar eru normaðar á svipaðan hátt og símanúmerin, þar sem hluti kennitölnunnar sem vísar í ártal og síðustu fjórir tölustafir eru annaðhvort skrifuð út sem einstakar tölur eða sem samsettar tölur. Gögnin hafa ekki verið gefin út en verða gerð aðgengileg við næstu vörðu, ásamt Kaldi-forskrift.

Full nöfn. Stórt mengi mögulegra íslenskra nafna var búið til, en þau eru byggð á leyfilegum nöfnum samkvæmt mannanafranafnd og eftirnöfnum sem koma fram í BÍN. Nöfnin þurfa ekki að vera normuð þar sem þau innihalda ekki nein sértákn. Gögnin hafa ekki verið gefin út en verða gerð aðgengileg við næstu vörðu, ásamt Kaldi-forskrift.

H11 – Sérhæfð talgreining

M8: Kaldi-forskrift fyrir unglinga.

M8-markmiðum hefur verið náð. Kaldi-forskrift fyrir unglinga er aðgengileg á GitHub (https://github.com/cadia-lvl/samromur-asr/tree/s5_adolescents_lstm/s5_adolescents_lstm) og á CLARIN (<http://hdl.handle.net/20.500.12537/221>).

H16 – Talgreiningarforskriftir í öðrum kerfum

M8: Forskrift almenns talgreinis hönnuð í umhverfi utan Kaldi.

M8-markmiðum hefur verið náð. Nokkrar talgreiningarvélar voru prófaðar, en ekki voru allar tilraunirnar árangursríkar. Til dæmis er WAV2LETTER++, sem gerð er af rannsóknarhópi Facebook (<https://github.com/flashlight/wav2letter>), ekki við hæfi fyrir Terra, umhverfið sem notað er til að keyra þjálfanir, þar sem hún þarfnast mjög nákvæmra hugbúnaðarútgáfa sem ekki er hægt að setja upp í sýndarumhverfi, og þær eru því ekki samrýmanlegar við núverandi stillingar Terra.

K2 (<https://github.com/k2-fsa/k2>) er Python-útgáfa af Kaldi. Vandamálið við K2 er að hann er háður geymslu sem kallast Lhotse (<https://pythonlang.dev/repo/lhotse-speech-lhotse/>). Með Lhotse getur gagnaundirbúningsskrefið verið aðeins nokkrar línur af kóða, en ef málheildin er ekki nú þegar geymd þar verður erfitt að nota K2. Hvað varðar ESPNet gátum við keyrt grundvallarþjálfunarskref með Samrómsgagnasettinu, en nokkrar villur komu upp í kerfinu sem ekki var hægt að komast hjá. Þær villur ollu slæmum áhrifum á orðavillutíðni, sem var mjög há.

Tilraunir með NVIDIA-NeMo (<https://developer.nvidia.com/nvidia-nemo>) og DeepSpeech (<https://deepspeech.readthedocs.io/en/r0.9/>) voru árangursríkar. Tvær kóðaforskriftir voru búnar til með sama sniði og fyrri forskriftir og báðar eru aðgengilegar á GitHub. Gagnasettið sem notað var til að þjálfra forskriftirnar er Samrómur 21.05. NVIDIA-NeMo-forskriftin er aðgengileg hér: https://github.com/cadia-lvl/samromur-asr/tree/n5_samromur/n5_samromur og DeepSpeech-forskriftin hér: https://github.com/cadia-lvl/samromur-asr/tree/d5_samromur/d5_samromur.

Talgerving

T4 – Vefgáttir fyrir talgervla

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

Öllum M8-markmiðum hefur verið náð. Vefgáttin er virk og keyrð á <https://tts.tiro.is>. Frumkóðinn er aðgengilegur á GitHub (<https://github.com/tiro-is/tiro-tts/releases/tag/M8>) og á CLARIN (<http://hdl.handle.net/20.500.12537/225>).

Tveir aðalnotendur þjónustunnar eru verkþættir T6 (Veflesarinn WebRICE) og T5 (Talgervlar í snjallsímum, Símarómur). Við höfum fengið stöðuga endurgjöf og svarað henni með því að laga eða bæta það sem þarf. Þjónustan er nú virk og keyrð með stuðningi til að senda inn bæði hráan texta og SSML-skjöl með studdum tögum. Með SSML er hægt að tilgreina framburð ákveðinna orða með hljóðanstaginu, stjórna hljómfalli (talhraða, tónhæð og hljóðstyrk), skipta út orðum fyrir gervingu og tilgreina hvernig tölustafir og tákn eiga að vera túlkuð.

T5 – Talgervlar í snjallsímum

M8: Símarómur fyrir Android virkar til þess að hoppa yfir og stoppa talþætti sem búið er að búa til.

M8-markmiðum hefur ekki verið náð. Eins og útskýrt var í nýlegri tillögu um endurskilgreindar vörður þessa verkþáttar þarf að breyta eða víkka hugbúnaðarhögun Símaróms að einhverju leyti. Breytingarnar eru flóknari en fyrst var talið og þar af leiðandi þarf nokkrar vikur í viðbót til þess að útfæra þær.

Breyta þarf samtímavinnslunni sem nú er notast við í vinnslu sem gerist ekki samtíma. Gerst getur að segð sem hefur verið stoppuð hafi ekki þegar verið unnin, en að vinna eigi nýja segð á sama tíma. Hvað varðar raddirnar sem geymdar eru á netinu þarf þar af leiðandi að marka hverja segð um sig, þannig að hægt sé að varpa talgervingarbeiðni yfir á hljóðsvarið sem er skilað. Búið er að bæta við talgervingarþjónustuna (verkþáttur T4) til þess að hún virki í þessum tilvikum.

Vandamálið liggur í röddum sem geymdar eru á tækinu. Aðeins er hægt að vinna eina segð í einu þar sem hömlur á tilföngum eru í tækinu. Þar að auki getur notandi ekki stoppað vinnslu raddar, þ.e. eins og er er ekki hægt að hoppa yfir textabút þegar raddir sem geymdar eru á síma eru notaðar. Hlutalausn fyrir raddir sem geymdar eru á síma er að byrja að vista tímabundið hljóð sem hefur þegar verið búið til, og þannig er a.m.k. hægt að hoppa fljótlega yfir texta sem er spilaður oft. Þetta er hægt þar sem flókin vinnsla á tækinu er ekki nauðsynleg fyrir hljóð sem hefur þegar verið vistað.

Raunveruleg lausn og lokalausnin til þess að hoppa yfir raddir sem geymdar eru á tæki er að samþætta léttu Flite-röddina við tæki, líkt og gert er ráð fyrir að gera fyrir M9-vörðu. Með henni er vinnslan hröð og hægt er að keyra mörg tilvik af röddinni í einu, þannig að mögulegt verður að koma á virkni sem vinnur með talbeiðnir sem koma á sama tíma, á sama hátt og gert er fyrir raddirnar sem geymdar eru á netinu.

T6 – Veflesari

Engin M8-markmið eru í þessum verkþætti, en veflesarinn var gefinn út sem Google Chrome-viðbót hér:

<https://chrome.google.com/webstore/detail/webprice/mmijkiiefioinbdgbadgghchfilmlmp>.

Unnið er að því að koma á SSML-stuðningi í veflesarann.

T8 – Mat á gæðum talgervla

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

M8-markmiðum hefur að hluta til verið náð. Endurbætur á matsverkvangnum hafa verið gerðar eftir niðurstöður notendaprófana. Huglægt mat á T13.M8 seinkaðist þar sem þjálfun líkana tók lengri tíma en gert var ráð fyrir. Verkvangurinn er þó tilbúinn um leið og líkönin eru tilbúin, og markmiðinu verður náð á næsta tímabili.

Nokkrar smávægilegar lagfæringar á MOSI, matsverkvangnum, voru gerðar og mismunandi útfærslur á því hvernig niðurstöður eru sýndar voru gerðar. Upptökur eru nú valdar úr fellivalmynd, sem gerir notanda auðveldara að velja setningar fyrir AB-prófanir. Kerfisstjórasíðan var hreinsuð, og núna er auðveldara að stýra hvaða notendur fá aðgang. Tími sem fer í að hlaða upp niðurstöðum var einnig minnkaður og töflur með niðurstöðum voru stækkaðar fyrir hverja rödd.

SUS-prófanir hófust á tímabilinu, en í þeim þarf notandi að geta skrifað inn texta og skipuleggjendur þurfa að geta hlaðið upp texta og hljóði. Til þess að bæta þessari virkni við var farið í gagnagrunnsvinnu og uppkast að uppsetningu síðunnar var gert.

Nokkur líkön voru þjálfuð til þess að geta metið talgervla, og þá sérstaklega Tacotron 2-líkan sem er skilyrt x-vigri. Það líkan getur búið til hljóðþætti fyrir marga mismunandi mælendur. Þegar það líkan er notað, ásamt raddkóðaralíkönnum sem túlka mælendur, er komin fullgerð uppsetning sem má nota til þess að búa til tal fyrir marga mismunandi mælendur með einu líkani. Þó þarf að endurþjálfa líkönin með gögnum sem hefur verið skipt almennilega niður í þjálfunar-, þróunar- og prófunargögn til þess að tryggja áreiðanleika matsúttaks áður en prófanirnar fara fram.

T9 + T10 – Framendapípa talgervils

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.

Öllum M8-markmiðum hefur verið náð. Frumgerð af framendapípu talgervils fyrir sérsvið hefur verið gefin út á GitHub (<https://github.com/grammatek/tts-frontend>), auk þjónustu fyrir framendapípuna (<https://github.com/grammatek/tts-frontend-service/tree/abn-frontend-pipeline>). Framendapípan tengir saman og samþættir sjálfstæðu einingarnar sem hafa verið þróaðar innan verkþátta T9 og T10.

Skriftu fyrir beinan aðgang að þjónustunni hefur verið skilað til þeirra aðila sem prófa framendapípuna, þ.e. aðila hjá Hljóðbókasafninu. Notendaviðmóti á netinu verður skilað fyrir M9-vörðu. Framburðarorðabókin hefur verið stækkuð eftir innri og ytri prófanir á talgervingarkerfinu.

Kvistur, einingin fyrir stofnhlutagreiningu, var notaður til þess að auka við gagnagrunn yfir samsett orð, en hann er notaður fyrir atkvæðaskiptingu í g2p-einingunni. Fjöldi aðalorða var aukinn, úr 5.694 orðum í 49.845 orð, og fjöldi ákvæðisorða fór úr 4.389 í 21.003. Lítil bæting í atkvæðaskiptingu var mæld við þetta, en nýja og eldri útgáfan eiga það til að gera mismunandi villur. Orðvillutiðni atkvæðaskiptingu og áherslumerkinga yfir heildina, mælt á

1.000 atriða prófunarsetti, er 6,9% fyrir grunnlínuna og 6,5% fyrir nýju, stækkuðu listana yfir orðhluta samsettra orða.

T13 – Stikaðir talgervlar

M8: Textanormari úr T9.M6 felldur inn í SPSS-framenda. Frumgerð af eftirlitslausu taugatalgervilskerfi fyrir hljómfall og áherslu, frumgerð af raddblöndunar- og/eða mælendaaðlögunar.

M8-markmiðum þessa verkþáttar hefur ekki verið náð. Frumgerð af talgervingarkerfi með mælendaaðlögun verður kláruð, en fyrst þarf að bæta það kerfi til þess að tryggja hágæða úttak fyrir mælendur utan prófunarsettsins. Þjálfun á íslensku talgervilskerfi fyrir hljómfall stendur yfir og náðist því ekki að skila því fyrir lok þessarar vörðu. Þjálfun seinkaðist því margar óþekktar villur komu upp við þjálfun talgervilskerfisins á VCTK-gagnasafninu. Búið er að komast hjá þeim villum og vonir standa til að með því að nota söma þjálfunarskriftu verði hægt að skila íslensku talvgervilskerfi fyrir hljómfall fyrir næstu vörðu.

*Þórunn Arnardóttir
verkefnisstjóri SÍM*