Samstarf um íslenska máltækni
SÍM

Máltækniáætlun fyrir íslensku

Varða 3 – lokaskýrsla

1. október 2020

Efnisyfirlit

Formáli

Fyrirvari: Þessi texti er þýðing úr enskri útgáfu M3-skýrslunnar og skal hafa ensku útgáfuna til viðmiðunar ef misræmi er á milli textanna.

Þessi skýrsla dregur saman fyrsta verkefnisár fyrir hvern verkþátt sem skilgreindur er í samningi milli Almannaróms og SÍM (Samstarf um íslenska máltækni) frá september 2019. Skýrslan byggir á skýrslunni sem afhent var við vörðu 1 í febrúar 2020 og inniheldur vísanir í vörðu 2 þar sem á við. Varða 2 var afhent í júní 2020 með kynningarfundi fyrir Almannaróm og mennta- og menningarmálaráðuneytið, þar sem gagnasöfn og frumgerðir hugbúnaðar voru kynntar. Slíkur fundur er gott tækifæri fyrir alla aðila (fulltrúa ráðuneytisins, stjórn Almannaróms og þátttakendur í SÍM) til að ræða eitt og annað tengt verkefninu: máltækni fyrir íslensku almennt, yfirstandandi verkefni í heild sinni auk undirverkefna. Við leggjum því til sama fyrirkomulag fyrir annað verkefnisár, þ.e. að staðin verði skil á vörðum 4 og 6 með ítarlegum skýrslum og útgáfum, en á vörðu 5 með kynningarfundi.

Á fyrsta ári hafa allir aðilar innan SÍM stækkað teymi sín í þróun máltækni. Um 50 manns hafa unnið við verkefnin á fyrsta ári og auk þess hafa nokkrir nemendur tekið þátt í sumarvinnu, við verk eins og yfirferð gagna og merkingar. Í upphafi var megináhersla á að byggja sterk teymi í hverju kjarnaverkefni, en nú erum við að vinna enn frekar í auknum samskiptum og samvinnu milli teyma. Þar sem möguleikar á staðfundum hafa minnkað mjög vegna COVID-19 faraldursins, afhendum við nú skýrslur allra verkþátta reglulega á sameiginlegum vettvangi, auk þess að halda reglulega fjarfundi. Þetta er gert til þess að stuðla að samræðum milli teyma um þróunina og til þess að deila framgangi annarra verkþátta. Einnig áformum við að setja upp eina miðlæga hirslu á GitHub fyrir SÍM, þar sem má finna öll okkar hugbúnaðarverkefni á einum stað.

Meirihluti verkþáttanna er á áætlun við M3 fyrir 1. október 2020, þrátt fyrir að öll teymi séu undir auknu álagi við að laga sig að nýjum aðstæðum af völdum COVID-19. Það eru þó nokkrir verkþættir sem ná ekki öllum markmiðum á réttum tíma, en áætlað er að þeim verði öllum lokið vel innan þess þriggja mánaða svigrúms sem veitt er til viðbótar fyrir endanleg skil skv. samningnum. Faraldurinn hafði veruleg áhrif á tvö gagnasöfnunarverkefni (og hefur enn áhrif á annað þeirra), en öðrum verkþáttum seinkar lítillega, t.d. vegna seinkaðrar byrjunar í verkefninu. Í öllum þessum tilvikum heldur vinna áfram á öðru ári, en er ekki áformuð fyrir allt árið, þannig að seinkun á afhendingu M3 mun ekki hafa áhrif á endanlega afhendingu við M6. Þróun og afurðum fyrir hvern verkþátt er lýst í eftirfarandi skýrslu. Við höfum almennt stefnt á hnitmiðaðar skýrslur vegna fjölda verkþátta, en sumar skýrslnanna vitna í ítarlegri skýrslur og skjölun í gagnahirslum verkþátta.

Verkefnið hefur fengið mikla jákvæða umfjöllun á Íslandi á fyrsta verkefnisári. Í samstarfi við Almannaróm stefnum við að því að byggja á þeim jákvæðu viðbrögðum og halda áfram að auka vitund um kosti máltækni fyrir íslensku. Fyrstu viðtöl fyrir stöðu atvinnulífsfulltrúa verða haldin á næstu dögum. Atvinnulífsfulltrúi mun vinna náið með verkefnastjóra, og verkefnastjórnun á öðru ári mun því leggja aukna áherslu á samvinnu milli teyma og hvernig má best nýta afurðir verkefnisins í þróun hugbúnaðar í fyrirtækjum og stofnunum.

G1 – Risamálheildin

Skýrsla: Stofnun Árna Magnússonar í íslenskum fræðum
Aðilar: Stofnun Árna Magnússonar í íslenskum fræðum, Háskólinn í Reykjavík

Vörður 2 & 3 – lýsing

M2: Öll málheildin var mörkuð og lemmuð með nýjustu tólum. Ný útgáfa gefin út, með textum til 2019.
M3: Ljúka handvirkri yfirferð matssetts. Meta áreiðanleika málheilda og valdra undirmálheilda. Niðurstöður viðræðna við leyfishafa og nýjum textum bætt í málheildina.

Afurðir

Öllum markmiðum vörðunnar var náð.

Ný útgáfa Risamálheildarinnar (IGC) var mörkuð og lemmuð með nýjustu tólum og skilað til CLARIN:

- IGC 1 (með MIM leyfi): https://repository.clarin.is/repository/xmlui/handle/20.500.12537/41
- IGC 2 (með CC-BY): https://repository.clarin.is/repository/xmlui/handle/20.500.12537/33

Matssettið inniheldur samtals 101.267 tóka og hefur verið yfirfarið handvirkt og skilað til CLARIN:

- https://repository.clarin.is/repository/xmlui/handle/20.500.12537/51

Viðræðum við leyfishafa er lokið. Viðræðum við útgefendur varðandi söfnun texta lýkur bráðlega. Textum úr útgefnum bókum verður bætt við í næstu útgáfu Risamálheildarinnar.
Verkþátturinn heldur áfram á öðru ári verkefnisins.

Skýrsla

Risamálheildin

Nýjasta útgáfa málheildarinnar, IGC-20.05, inniheldur texta til ársloka 2019. IGC-2019 inniheldur samtals 1.555 milljón runuorð, á meðan IGC-2018 inniheldur 1.400 milljón orð. 155 milljón orðum hefur því verið bætt við frá síðustu útgáfu. Textum sem innihalda u.þ.b. 57 milljón runuorð var bætt við frá 15 nýjum vefmiðlum, öllum með CC BY leyfi (sjá Töflu 1) og 98 milljón runuorðum var bætt við tiltæk textasöfn.

Eins og tekið var fram í skýrslu fyrir vörðu 1 höfðum við vonast til að bæta 215 milljón tókum við málheildina (sjá Töflu 2 í M1 skýrslu, bls. 8). Við bættum við mun meiri texta frá nýjum vefsíðum en áformað var, 57 milljón runuorðum í stað 10 milljóna, en við tókum engar bloggsíður með í þetta skiptið. Útgefnum bókum hefur ekki verið bætt við þar sem viðræðum við útgefendur lauk ekki í tæka tíð (sjá neðar). Við höfum sótt gögn frá stærsta spjallborði Íslands (bland.is) en vegna lagalegrar óvissu höfum við ekki bætt þeim við málheildina.

Allir textar voru tilreiddir með tilreiðara máltækniáætlunarinnar (verkþáttur I3, https://clarin.is/en/resources/tokenizer/), markaðir með ABLTagger (S. Steingrímsson ofl., 2019) sem var þjálfaður á samruna uppfærðu málheildanna tveggja, MIM-GULL og OTB, og lemmaðir með útgáfu af Nefni (S. L. Ingólfsdóttir ofl., 2019) sem hafði verið aðlöguð að nýrri markaskrá.

Miðill	Orð
eidfaxi.is	5.685.420
eyjafrettir.is	4.920.429
eyjar.net	3.458.761
fiskifrettir.is	2.374.368
fjardarfrettir.is	450.914
huni.is	3.150.024
kaffid.is	1.049.636
kopavogsbladid.is	397.549
mannlif.is	3.148.790
siglfirdingur.is	487.545
sunnlenska.is	3.437.029
trolli.is	777.441
vb.is	23.619.651
vikudagur.is	3.569.208
viljinn.is	520.718
Samtals	**57.047.483**

Tafla 1: *Nýir vefmiðlar*

Risamálheildin er gefin út árlega, svo að textar úr útgefnum bókum sem við höfum nú leyfi til að nota munu bætast við í næstu útgáfu Risamálheildarinnar, sem verður gefin út í byrjun næsta árs.

Matssett

Matssettið var tekið saman sem hluti af vörðu 1 og skipt í níu undirmálheildir. Endanlegt markmið var að eiga prófunarsett 100.000 orða, eða u.þ.b. 11.000 úr hverri undirmálheild.

Flöggunaraðferð og handvirk yfirferð:

Við notuðum fjórar samfallandi aðferðir til að flagga (e. flag) mörk sem skal yfirfara handvirkt.

1) Við byrjuðum á því að marka matssettið með þrem mörkurum: ABLTagger, IceStagger (H. Loftsson og R. Östling, 2013) og TriTagger, endurútfærslu tölfræðilega markarans TnT eftir Brants (2000). Skrifta flaggaði öll tilvik þar sem markararnir þrír voru ekki sammála.

2) Við notuðum Decca hugbúnaðinn (http://decca.osu.edu/) til að finna afbrigði n-stæða (5-15) í málheildinni og flagga tókana sem um ræðir.

3) Við notuðum IceParser (H. Loftsson og E. Rögnvaldsson, 2007) til að framkvæma grunnþáttun á málheildinni og finna villur í fallasamræmi innan nafnliða.

4) Að lokum skoðuðum við alla tóka markaða með **e** (erlend orð), **x** (ógreind orð) og **n----s** (sérnöfn án nánari greiningar), sem vilja vera til vandræða, og þar sem markið er **kt** (stutt form nafnorða) og orðin byrja með hástaf.

Samtals voru 19.474 mörk (19,23% af settinu) handvirkt yfirfarin. 4.652 tókar, eða 4,59% af öllu settinu, voru leiðréttir, eða næstum 24% þeirra tóka sem voru flaggaðir og yfirfarnir.

Þegar vitlaust var markað vegna rangrar tilreiðingar leiðréttum við bæði tilreiðinguna og markið. Í flestum tilfellum voru þetta styttingar í enda setninga þar sem punkturinn hafði færst frá styttingunni.

Nákvæmni

Settið var markað með ABLTagger sem hafði verið þjálfaður og sérútbúinn sem hluti af vörðu fyrir I4 og nákvæmnin mæld í mismunandi sviðum. Þar sem tilreiðing settsins hafði verið leiðrétt var settið í sumum tilfellum málfræðilega markað á ný með TriTagger og IceStagger. Nákvæmni markaranna þriggja er sýnd í Töflu 2. Þó röð nákvæmni milli markaranna þriggja sé ekki sú sama (ABLTagger gengur til dæmis ekki mjög vel með bækur á meðan TriTagger gengur best á því sviði) nær ABLTagger alltaf hærri einkunn en hinir tveir markararnir, á öllum sviðum.

		Nákvæmni málfræðilegrar mörkunar(%)		
	Tókar	**ABL**	**IceStagger**	**TriTagger**
Dómsúrskurðir	12.442	95,09 (5)	92,61 (9)	89,11 (9)
Bækur	10.353	94,23 (9)	93,83 (4)	92,98 (1)
Fræðsluvefir	10.772	94,86 (8)	93,88 (3)	91,18 (7)
Lög	12.177	95,55 (4)	94,51 (2)	91,67 (5)
Fréttaefni	11.459	96,07 (2)	93,55 (7)	91,38 (6)
Skoðanir	11.662	95,76 (3)	93,59 (6)	91,15 (8)
Þingræður	12.358	94,88 (7)	93,64 (5)	92,16 (3)
Íþróttir	9.272	94,87 (6)	92,67 (8)	92,33 (2)
Sjónvarp og útvarp	10.766	96,69 (1)	94,58 (1)	91,90 (4)
Samtals	**101.261**	**95,35**	**93,60**	**91,53**

Tafla 2: *Nákvæmni mismunandi undirmálheilda þegar þær eru markaðar með ABLTagger, TriTagger og IceStagger. Tölurnar í svigunum sýna röðina þar sem sviðið með töluna 1 hefur mestu nákvæmni fyrir gefinn markara.*

Tafla 3 sýnir nákvæmni ABLTagger fyrir bæði þekkt og óþekkt orð (orð sem finnast eða finnast ekki í þjálfunarsetti). Í heild er nákvæmnin 95,35%. Hæsta einkunn er 96,69% fyrir Sjónvarp og útvarp og lægsta einkunn er 94,23% fyrir bækur.

	Tókar	Samtals (%)	Óþekkt (%)	Þekkt (%)	Hlutfall óþekkt
Sjónvarp og útvarp	10.766	96,69	91,77	97,06	6,99%
Fréttaefni	11.459	96,07	85,22	97,25	9,80%
Skoðanir	11.662	95,76	85,19	96,77	8,69%
Lög	12.177	95,55	84,85	96,54	8,46%
Dómsúrskurðir	12.442	95,09	82,38	96,21	8,12%
Þingræður	12.358	94,88	78,61	96,09	6,77%
Íþróttir	9.272	94,87	79,63	96,55	9,95%
Fræðsluvefir	10.772	94,86	80,29	96,52	10,27%
Bækur	10.353	94,23	78,78	95,28	6,33%
Samtals	**101.261**	**95,35**	**83,01**	**96,47**	**8,34%**

Tafla 3: Nákvæmni ABLTagger. „Óþekktir" eru tókar sem voru ekki til staðar í þjálfunarsettum. „Þekktir" eru tókar sem eru til staðar í þjálfunarsettum (en mögulega í öðrum skilningi/samhengi).

Viðræður við leyfishafa

Viðræður hafa átt sér stað við Rithöfundasamband Íslands, Félag íslenskra bókaútgefenda og mennta- og menningarmálaráðuneytið.

Rithöfundasamband Íslands

Viðræður við Rithöfundasamband Íslands (RSÍ) hófust í janúar 2020. Málheildin var kynnt stjórn RSÍ á fundi í febrúar og mögulegar niðurstöður samkomulags voru ræddar. Viðbrögð RSÍ voru jákvæð og fengum við eftirfarandi ályktanir sendar frá þeim eftir fundinn:

1) Stjórn RSÍ samþykkir að Stofnun Árna Magnússonar (SÁM) hafi samband við rithöfunda til þess að safna textum fyrir Risamálheildina. RSÍ er einnig tilbúin til að senda meðlimum sínum ályktun þess efnis að RSÍ mæli með því að rithöfundar samþykki þessa beiðni.
2) RSÍ samþykkir að aðstoða við skrif formlegra skilaboða til höfunda þar sem þeir geta staðfest þátttöku sína í verkefninu.
3) RSÍ er fylgjandi þeirri hugmynd að samhliða undirskriftum útgáfusamninga geti höfundar gefið útgefanda rétt til að láta af hendi rafræna útgáfu verka þeirra, sem bætast í Risamálheildina.
4) RSÍ samþykkir hugmynd um upplýst samþykki höfunda varðandi notkun texta þeirra í þessum greiða. RSÍ mun ekki senda bréf vegna þessa í eigin nafni, en mun gjarnan senda slíkt fyrir hönd SÁM til meðlima RSÍ.

Í framhaldi þessarar ályktunar samþykktu SÁM og RSÍ að RSÍ myndi senda bréf til meðlima sinna fyrir hönd Stofnunar Árna Magnússonar. Í bréfi þessu voru höfundarnir upplýstir um að ef þeir mótmæltu ekki fyrir 1. júní 2020 myndi SÁM telja þá samþykka söfnun texta þeirra til að bæta í Risamálheildina. Bréfið var sent í apríl og 1. júní höfðu ekki borist nein mótmæli. Niðurstaðan varð því sú að allir meðlimir RSÍ samþykktu að áður útgefnu efni þeirra yrði bætt í málheildina. Þetta þýðir að við

munum ekki þurfa að fá leyfi frá hverjum og einum leyfishafa varðandi áður útgefna texta heldur getum við lagt áherslu á hvernig efninu verður safnað.

Hagþenkir: félag höfunda fræðirita og kennslugagna
Viðræður við Hagþenki hófust í vor. Markmiðið var að gera sams konar samkomulag við félagið eins og við gerðum við RSÍ. Stjórn Hagþenkis gat ekki brugðist við beiðni okkar strax vegna sumarleyfa. Viðræðurnar standa nú yfir og vonumst við til að skrifa brátt undir samkomulag við Hagþenki.

Félag íslenskra bókaútgefenda
Viðræður við Félag íslenskra bókaútgefenda héldu áfram og var lokið í lok september. Markmiðið var að félagið myndi bæta valkvæðu ákvæði í staðlaðan samning milli útgefenda og höfundarréttarhafa, þar sem kæmi fram að höfundarréttarhafar leyfðu notkun á textum sínum í málheild undir sams konar skilmálum eins og samþykktir voru fyrir Markaða íslenska málheild (MÍM), útgefna 2012.

Félag íslenskra bókaútgefenda var upplýst um niðurstöður viðræðna við RSÍ. Stjórnin fór yfir samkomulag þar sem kemur fram að félagið myndi hvetja meðlimi sína til að afhenda texta endurgjaldslaust til SÁM, til að bæta þeim í Risamálheildina. Stjórnin gat ekki komist að samkomulagi vegna ólíkra sjónarmiða stjórnarmeðlima varðandi möguleika þriðja aðila til að nota gögnin í hagnaðarskyni. Stjórnin komst að þeirri niðurstöðu að hvert útgáfufyrirtæki ætti að taka ákvörðun um þátttöku í þessu verkefni á eigin forsendum. Því munum við þurfa að hafa samband við hvern útgefenda fyrir sig og kynna þeim verkefnið til að reyna að sannfæra þá um að afhenda okkur gögnin þeirra. Þetta kann að tefja vinnuna en vonandi mun það ekki hafa mikil áhrif á niðurstöðurnar. Við erum nú að undirbúa viðræður við útgáfufyrirtækin og skoða mögulegar leiðir til að kynna þeim hvernig við ætlum að meðhöndla gögn þeirra og bæta í Risamálheildina. Markmiðið er einnig að kynna þeim hvernig þeir gætu haft ávinning af afleiddum afurðum við eigin útgáfu.

Menntamálastofnun
Viðræður við Menntamálastofnun hafa einnig staðið yfir síðan í febrúar 2020. Stofnunin er stjórnsýslustofnun á menntasviði og helsti útgefandi kennsluefnis fyrir grunnskóla á Íslandi. Slíkt efni yrði alveg nýr flokkur texta í Risamálheildinni. Stofnunin hefur verið mjög jákvæð og viljug að afla texta fyrir málheildina. Margir þeirra höfunda sem skrifa efni fyrir stofnunina eru meðlimir RSÍ svo samþykki þeirra er þegar til staðar; sumir höfundanna eru einnig meðlimir Hagþenkis sem við erum að vinna að sams konar samkomulagi við. Önnur höfundarréttarmál hafa verið leyst svo við munum safna textum frá Menntamálastofnun í síðustu vikum september. Stofnunin hefur einnig tekið jákvætt í að bæta ákvæði í útgáfusamninga sína við höfunda, sem veitir SÁM aðgang að textum þeirra til að bæta í Risamálheildina. Samningarnir gilda í 10 ár og eru síðan undirritaðir aftur til endurútgáfu. Þannig er mögulegt að bæta ákvæði í endurútgáfusamninga seinna. Þessu er ekki lokið en við áætlum að svo verði fyrir lok þessa árs.

Heimildir

Brants, Thorsten. 2000. TnT: A statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, Seattle, WA,USA.

Loftsson, Hrafn, og Eiríkur Rögnvaldsson. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. In J. Nivre, H-J. Kaalep, K. Muischnek and M. Koit (eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA-2007)*. Tartu, Estonia.

Loftsson, Hrafn. 2009. Correcting a POS-Tagged Corpus Using Three Complementary Methods. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. s. 523-531. Athens, Greece.

Loftsson, Hrafn, og Robert Östling. 2013. Tagging a morphologically complex language using an averaged perceptron tagger: The case of Icelandic. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA-2013)*, NEALT Proceedings Series 16. Oslo, Norway.

Ingólfsdóttir, Svanhvít Lilja, Hrafn Loftsson, Jón Friðrik Daðason, Kristín Bjarnadóttir. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, s. 310-315. Turku, Finland.

Steingrímsson, Steinþór, Örvar Kárason og Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. Í Proceedings of RANLP 2019, Varna, Bulgaria.

G4 - Beygingarlýsing íslensks nútímamáls fyrir máltækni (BÍN/DMII)

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

Varða 3 – lýsing

M2 - Forritaskil og vefþjónusta tilbúin og gefin út samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins. Gögn eru gefin út og aðgengileg á textaformi og/eða sem pakki á tvíundaformi (e. packaged binary format) undir opnu leyfi. Skýrsla um orðhlutafræðilega greiningu orðaforða.

M3 - Orðtökutól lagað að BÍN. Skýrsla um yfirferð á margræðum beygingarmyndum tilbúin sem og um rökliðagerð (e. valency structures).

Afurðir

- BÍN-kjarni tilbúinn: https://bin.arnastofnun.is/api/
- Vefþjónusta máltækni tilbúin: https://bin.arnastofnun.is/DMII/LTData/data/mimisbrunnur/
- Orðtökutól: Sjá skýrslu I2.
- Samanburður við aðra hluta neðar.

Öllum markmiðum var náð. Verkþátturinn heldur áfram á öðru ári.

Skýrsla

1. Forritaskil: BÍN-kjarni

Staða

- Aðgengi: https://bin.arnastofnun.is/api/
- Upplýsingar: https://bin.arnastofnun.is/DMII/dmii-core/
- Leyfi: CC BY-SA 4.0
- CLARIN: http://hdl.handle.net.20.2500.12537/12

BÍN-kjarninn er undirheild BÍN-gagnasafnsins (DMII), sem var búinn til til að uppfylla kröfu um vísandi útgáfu BÍN. Hann er hannaður til notkunar í skólum og fyrir þá sem eru að læra tungumálið. BÍN-kjarninn er aðgengilegur í gegnum RESTful forritaskil og er ætlaður til notkunar af þriðja aðila. Forritaskilin gera notendum kleift að senda einfaldar fyrirspurnir og taka við öllum beygingarmyndum á JSON-formi sem svari. Kjarninn inniheldur nú (15.9.2020) 58.000 uppflettimyndir.

Vefsíða BÍN-kjarnans inniheldur lýsingu á afmörkun orðaforðans í kjarnanum, uppruna, og forsendur til innleiðingar beygingarafbrigða. Hlekkur á fulla lýsingu allra flokka beyginga í íslensku í stærri útgáfu BÍN, með dæmum um öll beygingarmynstur, er einnig aðgengileg. BÍN-kjarninn var upprunalega gefinn út 25. september 2019. Vinna við þróun kjarnans stendur yfir, í samstarfi við ritstjóra *Íslenskrar nútímamálsorðabókar* sem er meginheimild orðaforða kjarnans.

2. Vefþjónusta: Gögn fyrir máltækni

Staða

- Aðgengi: https://bin.arnastofnun.is/DMII/LTData/data/mimisbrunnur/
- Upplýsingar: https://bin.arnastofnun.is/DMII/LTdata/
- Leyfi: CC BY-SA 4.0 : https://bin.arnastofnun.is/DMII/LTdata/conditions/
- CLARIN: http://hdl.handle.net/20.500/12537/5

Máltæknigögn BÍN eru sem stendur aðgengileg í fjórum útgáfum, með ítarlegum skýringum hverrar þeirra á vefsíðunni.

1. Uppflettimyndir: https://bin.arnastofnun.is/DMII/LTdata/word-list/
2. Orðmyndir (beygingarmyndir): https://bin.arnastofnun.is/DMII/LTdata/wordforms-list/
3. Sigrúnarsnið: https://bin.arnastofnun.is/DMII/LTdata/s-format/
4. Kristínarsnið: https://bin.arnastofnun.is/DMII/LTdata/k-format/

Listi uppflettimynda er nýleg viðbót að beiðni notenda. Hin nýja viðbótin er Kristínarsnið, sem inniheldur (að hluta til) vísandi útgáfu BÍN. Þeirri merkingarfræðilegu flokkun sem notuð er í þessari útgáfu er lýst nákvæmlega á vefsíðunni, og einnig í Bjarnadóttir, Hlynsdóttir & Steingrímsson 2019. Vinna við þessa útgáfu stendur yfir og ný útgáfa verður gerð aðgengileg með nákvæmum tilvísunum í íslenskan málstaðal þegar þeirri vinnu er lokið, með flokkun gagna, borið saman við vitnun í vísandi útgáfu í hluta 4.

3. Orðhlutafræðileg greining orðaforðans

Staða:

- Gagnagrunnur er í smíðum og verður tilbúinn 2020. (Frumgerð með dæmagögnum er í notkun).
- Gagnagreining:
    - 25.806 grunnorð og afbrigði tilbúin til innflutnings
    - U.þ.b. 132.000 ákvæðisorð tilbúin til innflutnings
    - 58.000 samsett orð greind, formi þeirra verður umbreytt fyrir innflutning.

Verkefnið sem snýr að orðhlutafræðilegri greiningu orðaforðans gengur undir nafninu *Orðföng* (*MorphIce*). Stutta lýsingu má finna í ritinu sem vitnað er til að ofan, Bjarnadóttir ofl. 2019. Heildaruppbygging orðhluta er innifalin, með lemmuðum orðhlutum, orðflokk fyrir orðhluta og heildarlýsingu á uppbyggingu samsettra mynda (eins og í nafnorðum, þar sem afbrigði mynda í fyrsta hluta samsetningar er stofn (með eða án hljóðvarps), beygingarmyndir (ef.et., ef.ft., (sjaldgæfari) þgf.et., þgf.ft.) og orðmyndir með bandstöfum. Fyrir stutta lýsingu á myndun samsettra íslenskra orða, sjá Bjarnadóttir 2002. Greiningin gerir ráð fyrir tvígildu hrísluformi (e. binary branching structures), eins og lýst er í Bjarnadóttir 2002, 2005. Orð er einfaldlega skráð sem „A = B + C“, þar sem A er samsett orð, B ákvæði og C höfuð samsetta orðsins. Forsenda gagnasafnsins er að B og C verði að vera til staðar áður en hægt er að bæta A við. Greiningin nær til bæði afleiðslu og samsetningar.

Gagnasafnið inniheldur afmarkaðar töflur fyrir ákvæði og höfuð, þ.e., B & C. Öll grunnorð eru skráð sem möguleg höfuð (C) í lista orða, sem og öll viðskeyti og bundnir (samsetningar)liðir sem mynda höfuð samsettra orða. Þar sem orðmyndunarreglur eru endurkvæmar, eru öll greinanleg orð einnig möguleg fyrir C. Í dæmunum fyrir neðan er A orðið sem er greint, B er ákvæði, og C er höfuð.

3.1 Tafla fyrir orð (höfuð)
Grunnorð: **blár**, lo., aðskeytt: **blána** 'verða blár' v., samsett orð: **Bláalónið**.

A	*Word*	*Greining*	*Orðflokkur*	*Uppbygging*	*BIN.ID*
	blár	blá,r	lo.	grunnform	408186
A	blána	blá=n,a	so.	aðskeytt	414465
B	1st part	blá --> blár	lo..		
C	2nd part	-n,a	so.	aðskeyti	
A	Bláalónið	Bláa+lón,ið	no. hk.	samsetning	508812
B	1st part	blá,a --> blár	lo.		
C	2nd part	lónið	no. hk..	grunnform	

Mynd 1. *Dæmi úr töflunni um orð (A) („=“ aðskeyting, „+“ samsetning, „,“ beygingarending)*

Staða fullbúinna gagna 15.9.2020:
Grunnorð: 15.956 & afleidd: 9.713, þ.e. 25.806 af 26.866 orðum skráð sem almennt mál í beygingarhluta BÍN. Gögnum u.þ.b. 58.000 samsettra orða þarf að umbreyta yfir á rétt form, en greiningu er lokið. Endurkvæma aðgerðin í gagnagrunninum verður notuð til að ljúka greiningu fyrsta hluta 150.000 samsettra orða til viðbótar; öðrum þáttum greiningarinnar hefur verið lokið. Endanlegt markmið er greining á öllum orðaforðanum í BÍN, sem inniheldur 296.451 uppflettimyndir (15.09.2020).

3.2 Tafla ákvæða

Íslensk orðmyndun er flókin, sem sést í lemmun og greiningu ákvæða (til að nefna dæmi geta verið allt að fimm mismunandi samsetningarform af nafnorði sem ákvæði). Taflan fyrir ákvæði inniheldur upplýsingar um endingar, málfræðileg mörk og formgerð, eins og sést fyrir neðan í Mynd 2.

1. hluti	*Greining*	*Ending*	*Mark*	*Orð*	*Orðflokkur*	*formgerð.*	*BIN.ID*
blá	blá,	-0	stofn	blár	lo.	grunnorð	424465
bláa	blá,a	-a	ákv.hk.	blár	lo.	grunnorð	424565

Mynd 2. *Færslurnar fyrir fyrstu hlutana í Mynd 1*

Tafla ákvæða inniheldur einnig upplýsingar um breytingar í stofni (ekki sýnt í Mynd 2), eins og í sögninni **auka**: aus/eys/jós/jus/ys/jys og samsvarandi mynstur au-ey-jó-ju>y-jy>j, með viðeigandi merkingu fyrir hvert afbrigði stofns (í þessu tilfelli „yk“ með „>“ sem markar viðeigandi hluta keðjunnar).

Fyrsta útgáfa MorphIce gagnasafnsins inniheldur 132.186 ákvæði, bæði grunnorð og samsetningar (fyrir endurkvæma samsetningu). Verið er að endursmíða gagnasafnið til að tryggja að endurkvæmni orðmyndunarreglna keyri núningslaust.

4. Skýrsla um margræðar beygingarmyndir (og orð)

Margræðar beygingarmyndirnar í meginútgáfu BÍN með flokkun má finna í Kristínarsniði (https://bin.arnastofnun.is/DMII/LTdata/k-format/) og nýjar aðferðir greiningar eru innleiddar í nýrri útgáfu sem hefur vinnuheitið Leiðbeiningarsnið sem er lýst neðar (*Leiðbeiningarsnið*, væntanlegt). Þessi margræðu form má staðsetja innan beygingardæma í BÍN. Nýr hluti kerfisins, sem nefnist **Ritmyndir**, er hirsla aukagagna, þ.e. gagna sem eiga sér ekki stað í beygingarfræðilegum hluta BÍN, eins og villur og eldri orðmyndir. Þær eru geymdar í aðskildum gagnagrunni, en vinna við hann er á lokastigi. Báðum þessum hlutum er lýst neðar.

4.1 BÍN og samanburður við íslenska staðla

4.1.1 Margræðar beygingarmyndir

Yfirferð margræðra beygingarmynda var upphaflega hluti af vinnu við BÍN-kjarnann, sem telst vísandi. Hin stærri BÍN telst lýsandi og inniheldur málgögn sem samræmast ekki staðlaðri málnotkun. Einkunnagjöf og greining beygingarmynda nýtir flokkunarkerfið sem lýst er í útskýringum fyrir Kristínarsnið máltæknigagna á vefsíðunni. Margræðar beygingarmyndir geta verið jafngildar, eins og markaðar/markaðs í eignarfalli eintölu karlkynsnafnorðsins **markaður**. Þetta gildir ekki fyrir beygingarmyndir sagnarinnar **ráða** í þátíð, þar sem rétt fyrsta persóna eintölu er **réð**, en hin ranga mynd **réði** er mjög algeng. Einkunnagjafakerfið sem notað er í BÍN-gögnum endurspeglar þetta.

4.1.2 Margræð orð

Upplýsingar um margræð orð (uppflettimyndir) eru einnig gefnar í einkunnagjöf uppflettimynda sem lýst er á vefsíðu máltækni (eins og þau koma fram á Kristínarsniði). Þetta nær bæði til samhljóma orða (e. homophones) og samtákna orða (e. homographs). Nýlegar viðbætur verða innleiddar í nýrri Leiðbeiningarútgáfu sem er sérstaklega ætluð til notkunar í málmyndun og leiðréttingu og miðast að ákveðnum eiginleikum stafsetningar. Gögnin í BÍN hafa verið borin saman við ritreglur (https://ritreglur.arnastofnun.is/) með það að markmiði að tengja þau saman. Munurinn á samhljóma orðum sem innihalda **i/y** og **í/ý** í stafsetningu (en ekki í framburði) er dæmi um þetta, eins og í orðunum **hýði** og **híði**. Öll lágmarkspör **i/y** og **í/ý** í uppflettimyndum í BÍN hafa verið yfirfarin og flokkuð (937 orð). Samtákna orðin **mæla** 'gera mælingu, taka mál af e-u' (þt. **mældi**) og **mæla** 'tala, segja e-ð' (þt. **mælti**) eru auðkennd með BIN.ID auðkennum þeirra, og merktar sem samtákna orð þar sem samhengi þarf til aðgreiningar (útskýringum á bæði samhljóma og samtákna orðum hefur verið bætt við vefsíðu fyrir almenning, með dæmum til nánari skýringa og hlekkjum á *Íslenska nútímamálsorðabók*, þar sem við á, https://islenskordabok.arnastofnun.is).

Sérstaka áherslu skyldi leggja á mikilvægi sérmerktra leiðbeiningargagna. Íslenskur málstaðall, sem finna má í Reglum um stafsetningu og greinarmerkjasetningu og í Stafsetningarorðabókinni (báðar aðgengilegar á arnastofnun.is), er skylda í íslenskum skólum og fyrir ríkisstarfsmenn. Hins vegar er ekki algilt að þeim sé fylgt og í sumum tilfellum eru fleiri dæmi um villur en rétta stafsetningu í málheildum. Þetta þýðir að ekki er alltaf hægt að nota tölfræðilegar aðferðir til að leiðrétta stafsetningarvillur og fá „réttar“ niðurstöður.

4.2 Ritmyndir: Hirsla fyrir gögn áður utan BÍN

Hirslan *Ritmyndir* inniheldur óstaðlaðar ritmyndir einstakra beygingarmynda sem eru tengdar uppflettimyndum í BÍN. Það eru fjórar megingerðir gagna: eldri orðmyndir, stafsetningarvillur, innsláttarvillur og tilbrigði í beygingu (e. inflectional variants). Hver færsla inniheldur hina upprunalegu rituðu mynd (eins og hún birtist í heimild), samræmd/rétt nútímamynd (eins og hún birtist í meginútgáfu BÍN), uppruna, aldur, einkunn, tegund/málsnið, tegund villu og málfræðilegt mark.

Gögnin á fyrsta stigi verkefnisins byggja á lista orðmynda úr eldri textum frá 15. til 19. aldar, og lista villna úr nútímaíslensku, þar á meðal lista íslenskra villna í leitarbeiðnum (e. Icelandic Search Query Errors (IceSQuEr)) úr leitartóli BÍN-vefsíðunnar (IceSQuEr er aðgengileg hjá CLARIN: http://hdl.handle.net/20.500.12537/78). Hirslan inniheldur nú u.þ.b. 3.900 villur og 20.000 eldri orðmyndir (15.09.2020).

Ritmynd	*Stöðluð m.*	*Uppruni*	*Aldur*	*Einkunn*	Tegund	Stafs.v.	*Innsl.v.*	*Uppflettimynd*	*BIN.ID*
ríkisábygð	ríkisábyrgð	RIS19	2019	5	Villa		Staf vantar	ríkisábyrgð	470256
alvarleger	alvarlegir	KLIM	1745	0	Úrelt			alvarlegur	476993
alvarleger	alvarlegar	MILTON	1828	0	Úrelt			alvarlegur	476993
eittvað	eitthvað	FRETTAMIDLAR	2012	5	Villa		Staf vantar	einhver	478799
orðstýr	orðstír	FRETTAMIDLAR	2012	4	Villa	Ý/Í		orðstír	79011
Frackland	Frakkland	VIKTORS_S	1401-1500	0	Úrelt			Frakkland	431533

Mynd 3. *Dæmi um inntaksgögn fyrir hinn nýja gagnagrunn Ritmyndir.*

Fáein dæmi um færslur eru sýndar í Mynd 3. Í fyrsta dálki (ritmynd) er orðið í þeirri mynd sem það birtist í upprunalegum texta og í öðrum dálki er stöðluð eða leiðrétt ritmynd. Uppruni er skráður í þriðja dálk og aldur er skráður sem ár eða árabil í þeim fjórða. Í fimmta dálki er réttmætiseinkunn, notuð á sama hátt og í meginútgáfu BÍN, með einkunn 0 fyrir eldri ritmyndir þar sem einkunnagjöf byggð á nútímaíslensku á ekki við, en 4 og 5 merkja mismunandi villuflokka. Í tegund/málsnið eru stafsetningar- og innsláttarvillur merktar sem villur og eldri form merkt sem úrelt. Af stærðarástæðum eru ekki sýndir allir reitir fyrir hverja færslu í Mynd 3.

Stafsetningarvillur, innsláttarvillur og beygingarafbrigði eru flokkuð ítarlegar, eftir þörfum, eins og sýnt er í Mynd 4:

Tegund innsláttarvillu	*Dæmi*
Ásláttarvilla	e´g → ég
Bil vantar milli orða	aðmargir → að margir
Auka broddur	vílla → villa
Brodd vantar	júni → júní
Staf vantar	ábygð → ábyrgð
Aukastafur	yffir → yfir
Stafagerð	bródir, bróðir, bròðir → bróðir
Stafavíxl	yfri → yfir

Mynd 4. *Flokkun innsláttarvillna í Ritmyndum.*

Eldri afbrigði eru ekki flokkuð sem stafsetningar- eða innsláttarvillur, en aðrir liðir flokkunar geta verið við hæfi, eins og t.d. fyrir eldri beygingarmyndir eða úreltar orðmyndir. Málfræðileg mörk eru aðeins höfð með þegar beygingarform er skýrt, þ.e., þegar beygingardæmi inniheldur ekki samtákna beygingarmyndir (með mörgum mörkum). Vinna við gagnagrunninn er á lokastigi.

5. Yfirferð á rökliðagerðum (e. valency structures)

Staða:

- Greining u.þ.b. 13.000 röklíða (e. verbal valency structures) er tilbúin. Þessar gerðir innihalda enga bundna röklíði (e. fixed arguments) þ.e., eru ekki bundnar ákveðnum orðum (að undanskildri sagnarinnar sjálfrar).
- Röklíðir háðir sérstökum orðum (e. word dependent valency structures) með öðrum lykilorðum en sögnum, u.þ.b. 4.000, eru á lokastigi yfirferðar.
- Gagnaskema er tilbúin fyrir gerð gagnagrunnsins.

Röklíðagerðir (e. Valency Structures)
Gögnin um röklíði eru tvískipt, þ.e., lesóháðar gerðir (e. lexeme independent structures) (án bundinna röklíða)), og lesháðar röklíðagerðir (e. lexeme dependent valency structures) með að minnsta kosti einn bundinn röklið). Orðaröð ræðst af orðfræðilegum atriðum. Allir nafnyrðarökliðir (e. nominal arguments), t.d., frumlög, andlög og stofnhlutar í forsetningarliðum, eru markaðir með falli og hvort þeir standa fyrir lifandi veru eða ekki (+/- lifandi, e. animacy). Eintala er hin ómarkaða tala, en fleirtalan er mörkuð þegar við á. Agnir, atviksorð og neitanir eru hafðar með, sem og aðrir orðflokkar þegar við á. Aðalatriðin í greiningunni eru eftirfarandi, í stafrófsröð [ensku fagheitanna, aths. þýð.]: lýsingarorð (adj.), atviksorð (adv.), tíðbreyting (conj.), nafnháttur (inf.), neitun (neg.), andlag (obj.), ögn (part.), lýsingarháttur þátíðar (ppart.), forsetningarliður (PP), spurnarform (quest.), afturbeygt fornafn (sig), málsgrein (S), frumlag (subj.), sagnbót (sup.), sagnorð (v.).

Gögnin um lesóháðar gerðir innihalda 239 aðgreindar gerðir röklíða 4.709 sagna. Þær algengustu eru: 1. Subj. v. PP, 2. Subj. v. obj., 3. Sub. v., 4. Subj. v. obj. PP, 5. Sub. v. obj. part.
(1. frl. so. fl., 2. frl. so. andl., 3. frl. so., 4. frl. so. andl. fl., 5. frl. so. andl. ögn.)

Greiningin er sýnd í dæminu um röklíðagerðir sagnarinnar **'þekkja'**, með fimm lesóháðum gerðum (af 15) og tveimur lesháðum gerðum (af níu, sjá feitletrað í dæmi). Íðorðin í dæminu hafa verið þýdd yfir á ensku; íslenska er lýsimál upprunalegu gagnanna.

þekkja|Subj:e-r$_{<nom.+anim>}$ |Verb:þekkir|[á e-ð$_{<acc.-anim>}$]<PP>|

þekkja |Subj:e-r$_{<nom.+anim>}$ |Verb:þekkir |Obj:e-a$_{<acc.pl.+anim>}$ |að<Part>

þekkja |Subj:e-r$_{<nom.+anim>}$ |Verb:þekkir |Obj:e-ð$_{<acc.-anim>}$ |[frá e-u$_{<dat.-anim>}$]<PP>

þekkja |Subj:e-r$_{<nom.+anim>}$ |Verb:þekkir |Obj:e-n$_{<.acc.+anim>}$ |[að e-u$_{<dat.-anim>}$]<PP>

þekkjast |Subj:e-m$_{<dat.+anim>}$ |Verb:þekkist |Obj:e-ð$_{<nom.-anim>}$

þekkja |Subj:e-r$_{<nom.+anim>}$ |Verb:þekkir |Obj:e-n$_{<acc.+anim>}$ |[í **sjón**$_{<dat.-anim.word-dep>}$]<PP>|

þekkja |Sb.:e-r$_{<nom.+anim>}$ |Verb:þekkir |Neg:ekki |Obj:[**haus né sporð**]$_{<þf.-anim.osamb>}$ |[á e-m $_{<dat. +anim>}$]<PP>

Athugið: Orðhlutafræðileg greining(*MorphIce/Orðföng*) verður hlekkjuð við viðeigandi röklíðagerðir setningarliðarsamsetninga og agnarsagna, sem að jafnaði birtast sem samsetningar í orðmyndun, eins og í karlkynsnafnorðinu [hvorugkynsnafnorðinu, aths. þýð.] *fyrirkall* og agnarsögninni *kalla fyrir.*

Valency: kalla |Subj:e-r$_{<nom.+anim>}$ |Verb:kallar |Obj:e-n$_{<acc.+anim>}$ |Adv:fyrir
Compound: [[fyrir]$_{<adv>}$[kall]$_{<n.neut>}$]$_{<n.neut>}$ --> [kalla_fyrir] + id.

Reykjavík 15.09.2020

Kristín Bjarnadóttir, ritstjóri BÍN/DMII, kristinb@hi.is
Kristín Ingibjörg Hlynsdóttir, aðstoðarritstjóri BÍN/DMII, kih4@hi.is
Samúel Þórisson, hugbúnaðarhönnuður, samuel.thorisson@arnastofnun.is

Heimildir

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir & Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, bls. 146-154. Turku, Finland. https://www.aclweb.org/anthology/W19-6116.pdf

Kristín Bjarnadóttir. 2002. *A Short Description of Icelandic Compounds.* Vefsíða Orðabókar Háskólans, apríl 2002, http://notendur.hi.is/~kristinb/comp-short.pdf.

Kristín Bjarnadóttir. 2005*. Afleiðsla og samsetning í generatífri málfræði og greining á íslenskum gögnum.* Orðabók Háskólans, Reykjavík.

Icelandic Search Query Errors (IceSQuEr). 2020. Þórunn Arnardóttir, Þórdís Dröfn Andrésdóttir, Þórður Arnar Árnason, Hildur Hafsteinsdóttir, Samúel Þórisson, Einar Freyr Sigurðsson, og Kristín Bjarnadóttir. http://hdl.handle.net/20.500.12537/78.

G5 – Orðskiptingar

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

Varða 3 – lýsing

Nýr listi handvirkt yfirfarinna orðskiptinga gefinn út. Niðurstöður prófana úr fyrsta mati bornar saman við endanlegt mat, marktækar endurbætur í nákvæmni. Lokaútgáfa orðskiptingatóls gefin út samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins.

Afurðir

Öllum markmiðum var náð. Ný útgáfa orðskiptingalista er tilbúin og ný mynstur orðskiptinga fyrir TeX, Hunspell o.s.frv. hafa verið sett saman úr þeim lista. Skipanalínutól til orðskiptinga eftir þessum mynstrum hefur verið hannað ásamt forritaskilum á vef og vefgátt (http://malvinnsla.arnastofnun.is/ordskipting) til aðgangs á netinu.
Orðskiptingagögnin, þ.e. uppfærður orðskiptingalisti og orðskiptingamynstur, hafa verið gefin út á GitHub (https://github.com/krunars/hyphenation-is) og á CLARIN-IS (http://hdl.handle.net/20.500.12537/86).
Skipanalínutólið hefur einnig verið gefið út á báðum stöðum (https://github.com/krunars/skiptir, http://hdl.handle.net/20.500.12537/87).

Skýrsla

Sjálfvirkar orðskiptingar í íslensku hafa varla þróast síðan á níunda áratug seinustu aldar. Árið 1985 gáfu Magnús Gíslason frá Reiknistofnun Háskóla Íslands og Baldur Jónsson frá Íslenskri málstöð út orðskiptingalista og mynstur orðskiptinga fyrir íslensku (Skipta) byggt á málfræði Íslensk-danskrar orðabókar Sigfúsar Blöndals auk fjölda handvirkt orðskiptra texta. Aðskilið sett orðskiptimynstra byggt á orðskiptigögnum þróuðum fyrir IBM á Íslandi var þróað af Jörgen Pind frá Orðabók Háskólans árið 1987 (uppfærð árið 1988), en þau mynstur eru ekki eins víðtæk, og orðskiptingalistinn sem notaður var við gerð settsins er ekki lengur aðgengilegur. Á þeim 32 árum sem liðin eru síðan hefur þessum tveimur settum mynstra verið dreift og eingöngu umpakkað eða kóðun þeirra breytt, en engar breytingar hafa verið gerðar á mynstrunum sjálfum. Ekki er vitað til þess að annað sett orðskiptingamynstra eða annar stór orðskiptingalisti hafi verið gefinn út og íslenskar orðabækur hafa aldrei veitt ítarleg viðmið til orðskiptinga. Því er orðið löngu tímabært að endurskoða mál sjálfvirkra orðskiptinga fyrir íslensku og íslenska orðskiptingu almennt.

Núverandi orðskiptingamynstur fyrir íslensku voru upphaflega búin til með TeX, á formi sem var þróað fyrir TeX af Frank M. Liang[1], en síðan á níunda áratug síðustu aldar hafa orðskiptingaraðferðir TeX verið teknar upp af mörgum forritum, meðal annarra Hunspell, sem er notað í mörgum öðrum forritum eins og InDesign og OpenOffice. Raunar hefur það orðið eiginlegur staðall fyrir sjálfvirkar orðskiptingar. Því var talið nauðsynlegt að halda áfram stuðningi við þessa gerð orðskiptinga og gefa út ný orðskiptingamynstur á TeX forminu. Hins vegar munum við einnig gefa út skipanalínutól og veflæga lausn til að styðja við þau sem nota ekki TeX kerfið beint. Við höfum einnig skoðað möguleika orðskiptinga með notkun Kvists (https://github.com/jonfd/kvistur), sem er skiptir samsetninga knúinn af tauganeti og hannaður af Jóni Friðriki Daðasyni. Tilraunaútgáfa veftóls okkar sem notar Kvist mun einnig verða gefið út, en megináherslan er í dag á mynsturbyggðar orðskiptingar. Ný gögn samsetninga sem eru í þróun fyrir BÍN sem hluti af Máltækniáætlun fyrir

[1] Fyrir meiri upplýsingar um þetta orðskiptingamynstur er bent á vísindagrein Liang um efnið: https://tug.org/docs/liang/liang-thesis.pdf

íslensku munu gera okkur kleift að safna meira magni gagna sem verður hægt að nýta fyrir báðar aðferðir þegar þau eru tilbúin.

Við byggjum núverandi vinnu á Skiptu. Algengi mynsturbyggðra skiptinga í TeX stíl gefur Skiptu mikla notkunarmöguleika og er hún með einhverju móti notuð af mörgum aðilum í útgáfugeiranum í dag. Hins vegar hefur hún sínar takmarkanir; mynstrin eru aðeins byggð á þeim gögnum sem notuð eru til að mynda þau. Þau gögn eru áðurnefndur orðskiptingalisti. Þó orðabók Blöndals, og þess vegna Skiptu, innihaldi vítt svið orðaforðans, skortir hana þó stóran hluta nauðsynlegs orðaforða. Mörg tökuorð voru talin framandi í tungumálinu þegar orðabókin var gefin út árið 1920–24 og eru ekki talin með í henni, þó sum þeirra hafi verið í almennri notkun á þeim tíma og séu enn. Að auki hafa þær miklu þjóðfélagslegu og tæknilegu breytingar síðustu 100 ára valdið því að fjölmörg ný orð hafa bæst í tungumál heimsins og er íslenska ekki undanskilin þeirri þróun. Í lista Skiptu reyndi Baldur Jónsson að fanga orðaforða samtímans með dæmum úr dagblöðum, útgefnum bókum o.s.frv., en þau dæmi voru einfaldlega ekki nógu víðtæk til að ná yfir stóran hluta mikilvægs orðaforða. Síðustu þrjá og hálfan áratug frá því Skipta var búin til hafa tæknilegar og þjóðfélagslegar breytingar einnig haft áhrif á málnotkun, sérstaklega tölvubyltingin og tilkoma internetsins.

Hér eru nokkur dæmi orða sem vantar í lista Skiptu: *nælon*, *mínus*, *klósett*, *sinnep*, *internet*, *nettenging*, *músík*, *malaría*, *Kalifornía*, *rósmarín*, *formúla*, *sjónskertur*, *sviðsetja*, *oxun*, *helíum*, *kólibrífugl*, *debetkort*, *sinfónía*, *spjaldtölva*, *snjalltæki*, *tengiltvinnbíll*, *máltækni*.

Eins og hvert annað kerfi byggt á algrími getur orðskiptingakerfið aðeins unnið með þau gögn sem það fær samkvæmt sértækum aðferðum þess: það vinnur eingöngu á nákvæmum röðum bókstafa og getur t.d. ekki getið sér til um uppruna eða framburð orða. Ef ákveðin röð bókstafa er ekki til staðar í þjálfunargögnum gæti styttri mynstrum verið beitt vitlaust. Á sama hátt, ef ákveðinni röð bókstafa er oftast, en ekki alltaf, skipt á ákveðinn hátt, er mikilvægt að sjaldgæfari mótdæmi séu höfð í þjálfunargögnum til að koma í veg fyrir að algrímið alhæfi ranglega. T.d. gætu orðin *ó-bók-lærð-ur* og *ó-bók-vís* orðið til þess að algrímið dragi þá ályktun að orðum sem hefjast á *óbó-* ætti alltaf að skipta eftir fyrsta *ó-*, ef ekki væri fyrir mótdæmið *óbó-leik-ari*.

Vegna þessarar næmni og sökum þess að samsettum orðum skuli skipta á orðhlutaskilum, er mikilvægt að orðskiptingalistinn innihaldi margar mismunandi gerðir samsettra orða. Æskilegt væri að hafa minnst eitt orð fyrir hvern mögulegan fyrsta bókstaf seinni samsetningarliðar fyrir hvern mögulegan fyrri lið; og jafnvel þá geta ákveðnar algengar endingar orða og önnur síður augljós mynstur komið fram og framkallað óvæntar niðurstöður. Einnig er mikilvægt að hafa mikið af beygingarmyndum, sérstaklega þar sem beygingar geta orsakað fyrirbæri eins og brottfall bókstafa og/eða hljóða og stofnbreytingar, sem geta framkallað miklar breytingar á uppröðun bókstafa í orði. Villur og ósamræmi þarf einnig að þurrka út, þar sem röng mynstur byggð á einu orði í listanum geta haft áhrif á endanlega orðskiptingu margra annarra orða.

Orðskiptingareglur fyrir íslensku valda ákveðnu vandamáli öðru en því sem það veldur í fjölbreyttum samsetningum, nefnilega því að bera kennsl á það sem í reglunum er nefnt sýndarsamsetningar (e. pseudocompounds), þ.e. að orðum sem þrátt fyrir að vera ekki búin til með samsetningu í íslensku er engu að síður skipt (með bandstriki) eins og þau séu það, eða hljóðkerfislega eða réttritunarlega breyttar samsetningar sem er skipt á óreglulegum stöðum. Oftast eru þetta tökuorð, oft fleiri en tvö atkvæði. Í gegnum verkefnið höfum við borið kennsl á þónokkra nothæfa flokka sýndarsamsetninga, t.d. þeirra sem halda erlendum orðhlutaskilum, eins og *kon-súll* < lat. *consul*, *Wolf-gang* (þýskt nafn), *dán-lóda* < e. *download* og öfugt, tilfelli þar sem orð skulu ekki meðhöndluð sem sýndarsamsetningar heldur þess í stað skipt á undan sérhljóði, líkt og í *Bolt-on*. Við höfum þróað lausleg viðmið fyrir þessi tilvik sem geta hjálpað með ný orð, en enn þarf að ákvarða hvert tilvik handvirkt. Kristján Rúnarsson hefur unnið mest í þessu verkefni í samstarfi við Jóhannes B.

Sigtryggsson, ritstjóra Stafsetningarorðabókarinnar og Málfarsbankans, leiðbeininga um íslenska málnotkun útgefnar af Stofnun Árna Magnússonar, við að taka þessar ákvarðanir.

Orðskiptingalistinn frá Skiptu hefur nú að fullu verið yfirfarinn handvirkt. Fjölmargar villur og ósamræmi hefur verið lagað og vandamálið við sýndarsamsetningar á móti skiptingu á undan sérhljóða (sem líta má á sem meginregluna þegar orðhlutaskil eru óljós) hefur verið skoðað nánar fyrir mörg orð, sem olli einhverjum breytingum. Að auki hefur þeirri venju að leyfa eins bókstafs forskeyti að skiptast frá orðhlutum inni í orði verið hætt (t.d. *ban-ó-færð* > *ban-ófærð*, *all-á-sjá-leg-ur* > *all-ásjá-leg-ur*). 1785 orðum af samtals 203.964 (0,88%) hefur verið breytt eða eytt. Bætt hefur verið við úr fjölda heimilda: 2554 tökuorð úr BÍN, 1053 aðrar viðbætur, meðal annars orð sem vantaði í orðabók Blöndals, önnur skörð sem fundust við skoðun listans og fjölmörg orð sem fundust við lestur, o.s.frv. Unnið er að ferkari viðbótum úr orðalista (aukaviðbætur) búnum til úr Risamálheildinni, sem inniheldur mörg tökuorð og erlend nöfn, ásamt fleiri sérsniðnum viðbótum (10.885 orð hafa þegar verið meðhöndluð).

Við búum til orðskiptingamynstur með PATGEN, útgáfu 2.4, úr TeX Live. Til að mæla nákvæmni orðskiptingamynstra notum við eftirfarandi gögn:

- **Almennt úrtak:** 10.001sta–11.000asta algengasta orðið í Risamálheildinni, þar á meðal orð sem koma fram í listanum – handahófskennd dæmi úr tíðnilista til að fá hugmynd um heildardreifingu: Efstu orðin á þessum tíðnilista væru ekki áhugaverð dæmi vegna þess að þau eru líklega öll í þjálfunargögnunum.
- **Úrtak óþekktra orða:** 1000 algengustu orðin í Risamálheildinni, að undanskildum eins og tveggja stafa orðum, sem koma ekki fram í upprunalegum lista eða eru ekki meðal 14.492 aukaorða – þ.e. algengra orða sem koma ekki fram í þjálfunargögnum fyrir gömlu Skiptumynstrin eða nýju mynstrin.

Fyrir bæði gagnasettin mælum við svo árangur:

- Gömlu Skiptumynstrin frá 1985 (grunnlínan).
- Nýtt sett búið til úr gamla Skiptulistanum, óbreytt. Þetta var gert vegna þess að sú tiltekna útgáfa PATGEN sem var notuð fyrir upprunalegu mynstrin, sem og stillingarnar, eru ekki þekktar og gæti sá munur valdið talsverðum mun á niðurstöðunum, sem myndi gera áhrif þess að leiðrétta og bæta í þjálfunargögnin óskýr; þetta hefur einmitt verið sýnt fram á.
- Mynstur búin til úr leiðréttum lista án viðbóta.
- Mynstur búin til úr leiðréttum lista með 3607 viðbótum.
- Mynstur búin til úr leiðréttum lista með áðurnefndum 3607 viðbótum og 10885 aukaviðbótum úr Risamálheildinni, o.s.frv.
- Mynstur J. Pind frá 1988 eru einnig mæld til samanburðar.

Eftirfarandi eru niðurstöður fyrir almennt sýni:

Mynstur	**Fullkomin orð**	**Bærileg orð**	**Slæm orð**	**Góðar orðskipt.**	**Slæmar orðskipt.**
Mynstur J. Pind frá 1988	87,8%	6,4%	5,8%	92,6%	4,5%
Upprunal. mynstur Skiptu (1985)	92,7%	4,5%	2,8%	95,0%	2,1%

Ný mynstur úr lista frá 1985	94,3%	1,4%	4,3%	96,8%	3,5%
Leiðrétt	95,1%	1,4%	3,5%	97,5%	2,7%
Leiðrétt með viðbótum	95,5%	1,6%	2,9%	97,7%	2,4%
Leiðrétt með fleiri viðbótum	96,0%	1,5%	2,5%	97,9%	2,0%

Og fyrir óþekktu orðin:

Mynstur	**Fullkomin orð**	**Bærileg orð**	**Slæm orð**	**Góðar orðskipt.**	**Slæmar orðskipt.**
Mynstur J. Pind frá 1988	68,9%	16,6%	14,5%	78,5%	12,3%
Upprunal. mynstur Skiptu (1985)	76,0%	14,3%	9,7%	83,0%	8,1%
Ný mynstur úr lista frá 1985	78,7%	7,3%	14,0%	87,8%	12,9%
Leiðrétt	79,2%	7,1%	13,7%	88,3%	12,4%
Leiðrétt með viðbótum	80,0%	7,4%	12,6%	89,1%	11,2%
Leiðrétt með fleiri viðbótum	82,9%	5,1%	12,0%	91,3%	10,2%

„Fullkomin orð“ er þau þar sem öll orðaskipti eru rétt og engin röng fundust. „Bærileg orð“ hafa engar vitlausar skiptingar en sýna ekki öll möguleg orðaskipti. „Slæm orð“ hafa eina eða fleiri rangar skiptingar. „Góðar orðskiptingar“ og „Slæmar orðskiptingar„ tákna hver um sig heildarfjölda réttra og rangra orðskiptinga sem fundust.

Nýju mynstrin hafa öll verið búin til með sömu stillingum PATGEN, sem sagt með fjórum stigum orðskiptinga og mynsturslengd 2–6, 3–6, 4–6 og 2–8 fyrir hvert stig í röð. Alltaf er notuð „góð vigt“ sem 1, „slæm vigt“ sem 2 og þröskuldur settur við 1. Þau sýna almennt fjölgun fullkominna orða og fækkun bærilegra orða, þ.e. þau fanga betur alla möguleika til réttrar orðskiptingar, en þau búa einnig fleiri slæmar samsetningar og slæm orð. Með viðbótunum er þetta lagfært fyrir almenna úrtakið og bætt umfram grunnlínu þegar fleiri viðbætur eru hafðar með, en það er enn marktæk aukning slæmra orðskiptinga og slæmra orða umfram grunnlínu fyrir sýni óþekktra orða. Þetta gæti verið vegna ofmátunar þjálfunargagnanna, þ.e. almennrar notkunar sjaldgæfra mynstra. Við þurfum enn að fínstilla stillingar fyrir PATGEN til að ná eftirsóknarverðum árangri. Hvað sem því líður ná leiðréttingar og viðbætur fram jákvæðum niðurstöðum og hugsanlega gæti stóraukið þjálfunarsett gert okkur kleift að ná niðurstöðum yfir grunnlínu á öllum mælikvörðum án þess að fórna réttum orðaskiptum.

Það er því brýnt að reglulega verði bætt við í framtíðinni, bæði til að listinn verði uppfærður og til að komandi endurbætur takist. Það er markmið okkar að halda áfram að bæta við hann í tengslum við verkefni framtíðarinnar.

G6 – Framburðarorðabók

Skýrsla: Grammatek
Aðilar: Grammatek

Varða 3 – lýsing

Fyrsta útgáfa framburðarorðabókar sem nær yfir kerfisbundin tilbrigði í framburði, hver færsla einnig málfræðilega mörkuð. Mat á sjálfvirkum hljóðritunum byggt á þessari útgáfu. Ritstýringartól fyrir framburðarorðabók gefið út samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins.

Afurðir

1. Íslensk framburðarorðabók sem inniheldur að auki upplýsingar um mállýsku, orðflokk, samsetningu, forskeyti og tungumál.
2. Reglubyggt rittákn-til-fónem (g2p) tól.
3. Tauganetsbyggt g2p tól.
4. Ritstýringar- og umsjónartól orðabókar.

Öllum markmiðum var náð. Hins vegar hófst handvirk vinnsla orðabókarinnar ekki fyrr en í byrjun 2020 og við áformum því að nota þann forða sem eftir er til að framlengja M3-útgáfu orðabókarinnar til loka nóvember 2020. Verkþátturinn mun halda áfram á öðru ári í nánu samstarfi við undirverkefni taltækni, sérstaklega T10 (sjálfvirk hljóðritun).

Íslenskar framburðarreglur: http://hdl.handle.net/20.500.12537/82
Reglubyggt g2p: http://hdl.handle.net/20.500.12537/83
Tauganetsbyggt g2p: http://hdl.handle.net/20.500.12537/84
Vefforrit orðabókar: http://hdl.handle.net/20.500.12537/85

Skýrsla

Þessi fyrsta útgáfa íslenskrar framburðarorðabókar inniheldur 25.000 færslur, hver merkt með orðflokki, tilbrigðum í framburði, hvort færslan er samsett orð og/eða getur verið ákvæði eða höfuð samsetningar, hvort hún inniheldur forskeyti, og tungumálamerki (e. language label). Enn fremur er mögulegt að geyma athugasemd við hverja færslu, t.d. þegar eitthvað er óljóst.

Listi yfir 100 erlend nöfn frá íslenskri fréttasíðu var settur saman og þau umrituð samkvæmt íslenskum framburðarvenjum, þ.e. eins og búast mætti við að fréttaþulur bæri þau fram í íslenskum fréttatíma. Erlend nöfn eru algeng hindrun fyrir talgervla og er stefnan að skoða meðhöndlun þeirra nánar á öðru ári.

Vegna þess að g2p vörpun í íslensku er almennt regluleg, og þeirra reglna sem safnað var saman við þróun handvirkrar endurskoðunar orðabókarinnar, ákváðum við að þróa reglubyggt g2p fyrir hvert framburðartilbrigði. Reglurnar eru skrifaðar á formi Thrax-málfræði[2] og safnað saman í stöðuferjöld (e. finite-state-transducers) til vinnslu. Skipanalínutólið fyrir keyrslu sjálfvirkrar g2p er forritað í C++. Hægt er að velja framburðartilbrigði, en sjálfgefið val er hefðbundinn skýr framburður.

Nokkur fjöldi tilrauna var gerður með vélrænt nám fyrir sjálfvirka g2p vörpun. Endanlegt líkan er byggt á kóðunar-afkóðunar (e. encoder-decoder) LSTM tauganeti eins og notað var í opnu verkefni

[2] http://www.openfst.org/twiki/bin/view/GRM/Thrax

fyrir margmála g2p á Sigmorphon ráðstefnunni í maí 2020[3]. Íslenska var eitt af 15 tungumálum í verkefninu, en frumkóðinn fyrir öll grunnlínulíkön er opinn (e. open-source) sem og þjálfunargögn og prófunargögn. Fyrst endursköpuðum við niðurstöðurnar sem lýst er í ritinu með áherslu á það líkan sem skilar bestum niðurstöðum fyrir íslensku. Gögnin sem notuð voru við verkið voru dregin úr íslensku útgáfu Wiktionary sem inniheldur hljóðritanir. Við þjálfuðum fyrst og lögðum mat á líkanið með gögnunum óbreyttum, 3600 orðum fyrir þjálfunarsettið og 450 orðum hvert fyrir þróunar- og prófunarsett. Gagnasettið var næst umritað samkvæmt þeim reglum sem þróaðar voru í fyrirliggjandi verkefni og líkanið þjálfað og prófað aftur, sem gaf sams konar niðurstöður.

Í byrjun verkefnisins sömdum við handyfirfarin prófunar- og þróunarsett fyrir sjálfvirka g2p með það að markmiði að hafa breiða dreifingu samstæðra rittákna. Þessi sett eru því erfiðari viðureignar fyrir sjálfvirka umritun en prófunarsett byggt á slembiúrtaki, þar sem líklegt er að sjaldgæfar samstæður rittákna séu hlutfallslega algengari en í slembiúrtak orða.

Fjöldatölur dreifingar n-stæða rittákna og orðalengda í bókstöfum milli prófunarsetta:

Prófunarsett	fjöldi orða	fjöldi tvístæða	fjöldi þrístæða	Meðallengd orða
Sigmorphon-sett	450	429	947	6,11
Núverandi prófunarsett	1000	549	2418	9,67

Eins og við var að búast skiluðu því prófanir á sama líkani á prófunarsetti okkar verri niðurstöðum en á prófunarsetti úr Sigmorphon-verkefninu.

Þjálfunarsettið sem var þróað í fyrirliggjandi verkefni var sett saman á sama hátt og prófunarsettin, með áherslu á að ná vel yfir samstæður rittákna og samtals ~5.700 orð.

Niðurstöður prófana allra líkana/aðferða/þjálfunar- og prófunarsetta:

g2p nálgun	Þjálfunargögn	Prófunarsett	WER	PER
Sequitur	Okkar sett, hefðbundinn framburður	Okkar sett, hefðbundinn framburður	34,40%	7,15%
Sequitur	Okkar sett, hefðbundinn framburður	Sigmorphon prófunarsett	38,00%	9,46%
Thrax (reglubyggt)	-	Okkar sett, hefðbundinn framburður	35,90%	6,49%
Thrax (reglubyggt)	-	Sigmorphon prófunarsett	7,33%	1,58%
LSTM	Okkar sett, hefðbundinn framburður	Okkar sett, hefðbundinn framburður	26,10%	4,07%
LSTM	Okkar sett, hefðbundinn framburður	Sigmorphon prófunarsett	7,33%	1,54**%**
LSTM	Okkar sett, harðmæli	Okkar sett, harðmæli	29,20%	4,57%

[3] Gorman, Kyle o.fl. (2020): The SIGMORPHON 2020 Shared Task on Multilingual Grapheme-to-Phoneme Conversion. In: *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*

Bæði reglubyggðu og tauganetsbyggðu nálganirnar skila almennt góðum árangri. Hins vegar nær hvorug nálgunin til þeirra ólíku reglna sem gilda um samsett orð, t.d. er almenna reglan að rittáknið ‘f’ milli sérhljóða er borið fram [v] eins og í *gefa*, en þegar höfuð samsetningar byrjar á ‘f’, gildir þessi regla ekki, heldur ‘f’ í byrjun orðs: *fjallaferð – fjalla+ferð.* Reglubyggða nálgunin mun alltaf umrita ‘f’ sem [v] og í seinna dæminu, þar sem [f] væri rétt, kemur í ljós að tauganetsbyggðu líkönin gera þetta í flestum tilfellum líka, jafnvel þó þjálfunargögnin innihaldi samsett orð. Fyrir nákvæma g2p munum við því alltaf þurfa greiningu samsettra orða. Harðmæli er tekið með í töflunni að ofan, almennt voru niðurstöður fyrir mismunandi mállýskur sambærilegar hefðbundnum framburði. Allar niðurstöður má sjá undir hlekk á verkefnið.

Ritstýringartól framburðarorðabókarinnar (PEDI)

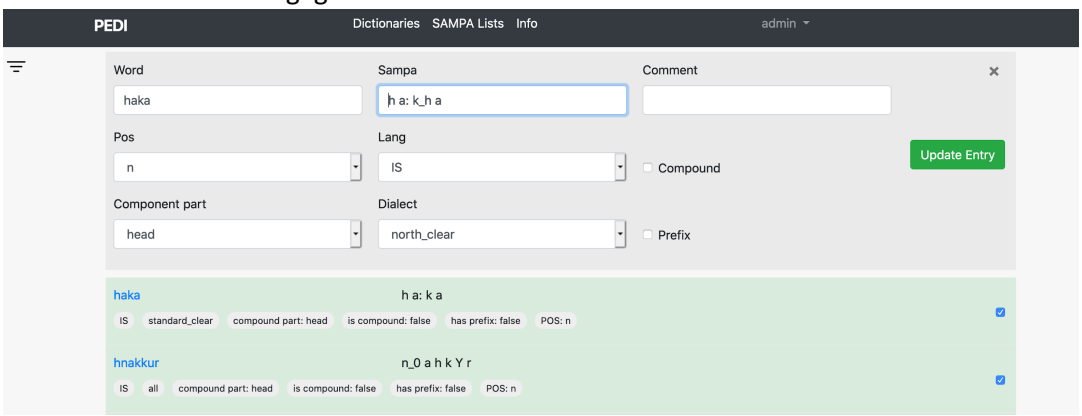
Handvirk vinna við framburðarorðabókina var unnin með ritstýringartóli sem þróað var innan verkefnisins.

Orðalista gömlu framburðarorðabókarinnar var skipt í lista 250-350 orða, hver með góða dekkun n-stæða. Þumalputtareglur fyrir sjálfvirkar merkingar eiginda hvers orðs voru settar saman, t.d. ef orð er líklegra til að eiga sér tilbrigði í framburði, er samsett eða getur tilheyrt mismunandi orðflokkum. Í tilfellum mismunandi orðflokka eða mismunandi mállýska voru búnar til aukafærslur fyrir orðin. Listarnir voru svo færðir í PEDI fyrir starfsfólk orðabókarinnar til að vinna úr.

Í opinni orðabók sýnir PEDI með litakóðum hvaða færslur hafa þegar verið yfirfarnar og merktar „réttar“ og hverjar ekki. Færslur sem innihalda ógildar umritanir eru merktar rauðar. Í PEDI er boðið upp á ýmsar aðferðir til síunar fyrir þá sem vinna með orðalistana til að stjórna ákveðnum hlutum eða vinna í sambærilegum eiginleikum í orðalistum.

Þegar PEDI er opnað eru allir orðabókarlistar sýndir sem eru nú í gagnagrunninum. Þá má velja alla saman eða einn í einu til útflutnings. Form úttaks er einfalt .csv form, sem er það form sem talgervilssteymið vinnur með, aðrir útflutningsmöguleikar verða þróaðir eftir þörfum í samstarfi við undirverkefni T10 á öðru ári.
Tólið innleiðir margmálastuðning (e. internationalization) og notendaviðmótið er í boði á íslensku og ensku.

Að auki hefur skrifta til umbreytingar útfluttrar orðabókar á TEI-form verið innleidd, til að undirbúa samvinnu við önnur orðfræðigagnasöfn.



Skjáskot af ritstýringartóli framburðarorðabókarinnar.

G7 – Íslenskt orðanet

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

Varða 3 – lýsing

M2 – Umbreyting orðanetsins yfir á form sem hæfir máltækni er lokið. Skýrsla um aðferðir sjálfvirkrar greiningar skarða.

M3 - Fyrsta útgáfa íslensks orðanets gefin út á völdu formi. Áframhaldandi vinna við að fylla í skörð samkvæmt áætlun. Greiningu skarða lokið. Vinna við uppfyllingu skarða samkvæmt greiningu er á áætlun.

Afurðir

Öllum markmiðum verkþáttarins var náð. Vinna við verkþáttinn mun ekki halda áfram á öðru ári verkefnisins, en gæti haldið áfram á seinni árum verkefnisins, eftir þörfum.

Skýrsla

1. Inngangur

Íslenskt orðanet er djúpt og víðtækt kerfi sem kortleggur mismikil merkingarfræðileg vensl milli eins og tveggja orða uppflettimynda. Þessi dýpt felur í sér ákveðið flækjustig: Í sinni upprunalegu mynd felur uppbygging orðanetsins í sér orðavensl sem byggja á flókinni blöndu merkingar, málfræði, og samhengis; auðugt samansafn gagna hefur verið skapað yfir nokkurra ára tímabili af fjölda notenda og úr innfluttum gagnasöfnum; víðtækt flokkunarkerfi þess hefur verið handsmíðað með tímanum til að halda utan um nákvæmt innihald þess; og geymsla gagnasafnsins og vefviðmótið er sérhannað fyrir þessa tilteknu málheild.

Hvað varðar umbreytingu yfir á viðeigandi form fyrir máltækni skapar þessi mikla sérsníði tvíþætt vandamál. Í fyrsta lagi, óháð því hvaða gagnaform er valið, er einfaldlega ekki hægt að umbreyta orðanetinu yfir á það. Eina leiðin til að staðla orðanetið og jafnframt viðhalda hinu einstaka notagildi þess er að endurhanna og innleiða það á hinu nýja formi nánast frá grunni. Þetta þýðir að á meðan á endurhönnun stóð var ríkari áhersla lögð á að endurskapa að fullu núverandi virkni, fremur en að bæta við þeim nýju eiginleikum sem nýja formið styður hugsanlega. Í öðru lagi mun áherslan á greiningu á gloppum óhjákvæmilega verða á endurbætur á núverandi gögnum fremur en að stækka þau eða bæta við fleirum. Nýja formið, sem er heldur staðlaðra, er miklu strangara gagnvart villum og óregluleika. (Það skal taka fram að þrátt fyrir að nafn kerfisins hafi áður verið þýtt á ensku sem „Wordnet“ notum við nú heitið „Wordweb“ að beiðni upprunalegs höfundar, til að rugla því ekki saman við orðanet (e. wordnet) á formi Princeton WordNet).

2. Uppbygging orðanetsins

Orðanetið samanstendur í raun af tveimur sjálfstæðum en samtengdum kerfum: Flettum (e. Lemmas) og Hugtökum (e. Concepts). Hið fyrra, Flettur, er nokkurn veginn sambærilegt orðasafni. Efni þess er dregið beint úr hinum ýmsu textaheimildum Orðanetsins, ásamt merkingarfræðilegum venslum byggðum á þeim heimildum. Þessar Flettur innihalda einnig fleiryrtar flettur og orðasambönd (e. phrasemes), en mikill fjöldi þeirra er kóðaður í orðanetið. Seinna kerfið, Hugtök, líkist meira verufræði (e. ontology): Eins lags flokkunarkerfi, búið til af stjórnendum Orðanetsins, sem margar Flettnanna teljast til.

Mikill fjöldi venslanna í Flettukerfinu eru byggð á eða leidd af ákveðinni gerð samhliða uppbyggingar (titluð „pör“ eða „parasambönd“ í Orðanetinu). Þetta eru einfaldlega tvíhliða tengsl (e. binary relations) sem gefa til kynna að tvær Flettur, X og Y, hafi á einhverjum tímapunkti birst í upprunalegum texta með samtengingunni „og“ á milli þeirra. Pör eru óröðuð, þar sem „X og Y" er talið jafngilt „Y og X" (þessi óraðaði eiginleiki er ein af mörgum ástæðum þess að pör er ekki hægt að meðhöndla einfaldlega sem orðastæður).

Aðrar gerðir vensla í Flettukerfinu hafa tilhneigingu til að byggjast á pörum, en þó er til aðskilinn (nokkuð lítill) flokkur vensla - sem Orðanetið merkir sérstaklega sem samheiti, andheiti, og „stiklur“ - sem eru handgerð frekar en fengin beint úr textaheimild.

Vensl milli eininga í Hugtakahluta Orðanetsins eru hins vegar ekki handgerð þó einingarnar sjálfar hafi verið valin af höfundum Orðanetsins. Þess í stað eru þau dregin af pörum sem hvert Hugtak er tengt. Hugtak X og Hugtak Y teljast þannig aðeins vensluð ef einhverjar Flettur sem þær standa fyrir eru sjálfar pör. Víxl af þessu tagi, þar sem einingar í einum hluta Orðanetsins eru tengdar gegnum einingar í öðrum hluta, koma oft fyrir í hönnun nýja kerfisins.

3. Nýja formið

Þessi tvö kerfi Orðanetsins eru nægjanlega ólík til þess að hvort um sig þarfnist eigin líkans.

Fyrir Flettur notum við OntoLex (áður þekkt sem „lemon“ stytting á ensku fyrir „The Lexicon Model for Ontologies“). Það styður flókin málfræðilíkön og er eina gagnalíkanið af sínu tagi sem hægt er að beita á jafn orðhlutafræðilega flókið tungumál og íslensku. Að auki gerir þetta okkur ekki aðeins kleift að endurskapa Flettukerfi gamla Orðanetsins, heldur einnig að kóða ákveðin gögn sem voru til staðar í töflum gamla gagnasafns Orðanetsins en voru ekki beinlínis aðgengileg notendum þess.

Til dæmis fylgja nú hverri einyrtri Flettu orðhluta- og setningafræðileg gögn sem dregin eru úr BÍN. Gögnin ná ekki aðeins yfir uppflettimynd Flettunnar, heldur öll afbrigði hennar að auki, sem veitir möguleika á bæði samhengisbundinni leit að óhefðbundnum orðmyndum og leit byggða á orðhluta- og setningafræðilegum eiginleikunum: kyni, beygingu, veikri/sterkri beygingu, tölu, stigi, mynd, hætti, tíð, persónu og greini. Þessi flokkun nær lengra en verksvið BÍN, þar sem flettur (bæði ein- og fjölyrtar) eru merktar sem upphrópanir, samtengingar, forsetningar, töluorð, bundnir liðir/föst orðasambönd og sérnöfn, eftir því sem gögn Orðanetsins leyfa.

Fjölyrtar Flettur eru merktar á annan hátt. Þetta er að hluta til vegna ófullnægjandi orðhluta- og setningafræðilegra gagna í gagnagrunni Orðanetsins og að hluta vegna þess að nýja formið er vistað í einu XML-skjali í stað fjölda taflna í gagnagrunni. Fjöldi afbrigða sem við þurfum að geyma fyrir einyrtar færslur þýðir að skrá Orðanetsins verður mjög stór; endurtekning þeirrar vinnu margsinnis fyrir fjölyrtar Flettur skilaði sér í gríðarstóru skjali. Þess í stað höfum við, þar sem mögulegt er, bætt við hlekk á viðeigandi einyrta færslu fyrir hvert orð fjölyrtrar Flettu (ásamt orðhluta- og setningafræðilegum eiginleikum þeirrar færslu, að sjálfsögðu). Þetta er ný virkni fyrir Orðanetið og vonir standa til að þetta muni bæta aðgengi og innbyrðis vensl gagnanna.

Þó við höfum flutt meirihluta upplýsinga Orðanetsins þörfnuðust nokkrir þættir þess endurrömmunar (e. reframing). Einn af eiginleikum Orðanetsins er að það getur endurspeglað flókin vensl milli færslna, en umfang sumra þessara vensla geta verið óljós og ekki auðvelt að kóða. Til dæmis eru orð og setningarliðir í <oddum> og [hornklofum] notaðir sem nokkurs konar algildistákn, standa fyrir hlutmengi annarra orða sem deila ákveðnum eiginleikum. ≤Oddar≥ eru vanalega notaðir til að tákna málfræðilegan flokk og ákveðna orðhluta- og setningarfræðilega eiginleika, oftast fornöfn; á meðan hornklofar eru oftar notaðir í tengslum við við merkingu, og gegna gjarnan (mjög óformlega) hlutverki skilgreiningar á eiginleikum hugtaka. Til að koma þessum mikilvægu

upplýsingum til skila þurfti að sýna aðgát og fyrirhyggju og úrdráttur þeirra úr sérútbúnu gagnasafni og kóðun þeirra á tölvulæsilegt málfræðibyggt líkan reyndist ekki auðvelt, en við fundum bestu nálgun sem við gátum. Við höfum að auki hannað Fletturnar í hinu nýja Orðaneti vandlega (og Python-umbreytingarkóða sem notaður er í ferlinu) á slíkan hátt að viðbótarupplýsingar setningarfræði fjölyrtra Flettna, til dæmis - megi kóða án mikilla vandræða í framtíðinni.

Til að kóða Hugtökin notum við SKOS (Simple Knowledge Organization System), vinsælt verufræðilíkan byggt á RDF. Virkni OntoLex var hönnuð með samþættingu við SKOS í huga og líkönin tvö passa vel saman þegar þeim er beitt á Orðanetið. Líkanagerð Hugtakanna sjálfra hefur verið einfalt ferli og þarfnast lítilla skýringa. Eiginleikar SKOS koma helst að notum þar sem líkönin tvö skerast.

Þó að OntoLex sé án efa besti kosturinn til að sýna flókna málfræði Orðanetsins nær það ekki yfir ákveðin atriði hönnunar Orðanetsins. Þar ber helst að nefna að OntoLex getur ekki með góðu móti sýnt merkingarfræðileg vensl eins og pör úr hluta 2, eða kóðað merkingu Flettna ef þær eru ekki tengdar hugtökum (eins og er stundum tilfellið). SKOS, sem er nokkuð sveigjanlegt, hentar fullkomlega til að taka á þessu vandamáli. Við höfum hannað aðskilið hugtakalag milli líkananna tveggja þar sem þessir eiginleikar eru innleiddir í SKOS en má meðhöndla eins og þeir teljist til OntoLex. Með þessum hætti komumst við hjá því að búa til óstöðluð uppflettiorð sem gætu annars flækt notkun Orðanetsins og aðskiljum alla ódæmigerða notkun líkananna í greinilega afmarkaðan hluta kerfisins á sama tíma og grunnvirkni hins upprunalega Orðanets er viðhaldið. Ef nýrri gerð merkingarvensla væri bætt í Orðanetið, væri auðvelt að láta þau passa í þetta lag.

4. Afurðir

Meginafurð þessa verkefnis er RDF-skrá sem má hlaða inn í hvaða hugbúnað sem styður SKOS og OntoLex. Við mælum með VocBench 3, þróuðu af gervigreindarteyminu (the Artificial Intelligence Research Group) við Tor Vergata Háskólann í Róm. Auk þess að veita báðum líkönum þolinn stuðning (e. robust support) býður VocBench 3 upp á sérstaka möguleika fyrir mikið magn gagna. Þar sem RDF-skráin er nálægt 2 GB að stærð og inniheldur milljónir lína eru þessir möguleikar hugsanlega skilyrði fyrir því að vinna með Orðanetið.

Ásamt RDF-skránni sjálfri höfum við búið til handbók sem útskýrir hönnunina, lýsir innihaldinu og fyrirhugaðri notkun. Orðanetið er flókið tól og RDF-skráin sem inniheldur það líka. Að okkar mati er nauðsynlegt að fara nánar út í smáatriði varðandi suma þætti til að tryggja að notendur geti kannað notkunarmöguleikana til fullnustu.

Lykildæmi um þessa notkunarmöguleika, og nauðsyn þess að skjala þá, er hvernig við köllum fram hinar ýmsu gerðir orðavensla. Í upprunalega Orðanetinu eru þessi vensl föst. Megingerðirnar eru sjö og fyrir utan fáeina endurröðunarmöguleika er ekki hægt að breyta þeim. Nýja Orðanetið hefur engar slíkar takmarkanir: Svo lengi sem notandi getur mótað ásetning sinn í gilda SPARQL fyrirspurn má beita henni á gögn Orðanetsins. Áskorunin felst nú í því að skrifa fyrirspurnir sem draga réttar ályktanir um uppbyggingu Orðanetsins. Handbókin veitir ekki aðeins nauðsynlegar upplýsingar um slíkt, heldur gefur hún einnig lista SPARQL fyrirspurna sem samsvara öllum venslum sem finnast í gamla Orðanetinu. Þessar fyrirspurnir ættu að veita hverjum sem vill kynna sér nýja Orðanetið góðan grunn til að byrja á.

Greiningarskýrsla um gloppur var aukaafurð. Eins og hefur komið fram var lögðum við áherslu á að bæta núverandi gögn Orðanetsins. Sumar þessara breytinga voru gerðar vegna upplýsinga úr villuleit sem var sett saman áður. Afgangurinn tilheyrði almennt einum þessara þriggja flokka. Í fyrsta flokknum voru Flettur sem innihéldu ofgnótt orða - sem þýddi oftast að sum þessara orða höfðu verið ætluð sem víxlanlegar tillögur – sem olli oft ósamræmi við málfræðilegu skýringar Flettnanna. Í

öðrum flokknum voru Flettur sem voru einfaldlega vitlaust merktar, stundum vegna þess að kyn, tala, fall eða einhver annar málfræðilegur eiginleiki var vitlaust skráður. Í þriðja flokknum voru Flettur þar sem textinn eða merkingarnar innihéldu óþekkt tákn, eða vantaði mikilvægar upplýsingar. Í heildina inniheldur skýrslan u.þ.b. 8.000 breytingar gerðar á Flettum í gagnasafninu.

Notendahandbók sem útskýrir hönnun Orðanetsins og hugsanlega notkun má finna á https://github.com/HjaltiDan/Ordanet/raw/master/Handbook%20-%20Wordweb.pdf

Ítarlegar skýrslu um greiningu skarða má finna á https://github.com/HjaltiDan/Ordanet/raw/master/Report%20-%20Gap%20Analysis.pdf

G10 – Gullstaðall fyrir málfræðilega mörkun

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

Varða 2 - lýsing

Uppfærður Gullstaðall gefinn út. Öðrum aðgengilegum málheildum breytt á valið form með endurskoðaðri markaskrá.

Afurðir

Öllum markmiðum var náð. Gullstaðallinn (MIM-Gold málheildin) og Íslenska orðtíðnibókin (IFD) eru aðgengilegar hjá CLARIN. Skjal sem útlistar þær breytingar sem gerðar hafa verið á markaskránni og hvernig margræð tilfelli við mörkun voru leyst er innifalið í MIM-Gold 20.05.

- MIM-Gold 20.05:
  https://repository.clarin.is/repository/xmlui/handle/20.500.12537/39

- MIM-Gold 20.05 - þjálfun/prófun:
  https://repository.clarin.is/repository/xmlui/handle/20.500.12537/40

- IFD 2020.05 - þjálfun/prófun:
  https://repository.clarin.is/repository/xmlui/handle/20.500.12537/38

Verkþátturinn var kláraður við M2.

Skýrsla

Inngangur

Við höfum aðlagað tvær helstu málheildirnar sem eru helst notaðar til að prófa og þjálfa málfræðilega markara á íslenskum textum, MIM-Gold og Íslensku orðtíðnibókina (IFD - gamla gullstaðalinn), yfir í nýja orðhluta- og setningafræðilega markaskrá fyrir íslensku. MIM-Gold var einnig tilreidd aftur, með tilreiðaranum úr Máltækniáætlun.

Vélræn umbreyting

Gullstaðlinum hafði þegar verið varpað með skriftu yfir í nýja markaskrá sem hluta af vörðu 1. Íslensku orðtíðnibókinni var varpað á sams konar máta. Skriftan var ekki í nokkrum vafa hvert ætti að vera nýtt mark sumra marka. Forsetningum (ao, aþ, ae) var til að mynda alltaf varpað í **a**f, **sþghen** í **ssg** og **sþmhen** í **ssm**. Greinarmerkjum var einnig almennt hægt að varpa með nýjum mörkum án nánari skoðunar. Önnur mörk, til dæmis erlend orð, þörfnuðust nánari skoðunar. Nokkrar reglur og listar voru búnir til svo skriftan gæti varpað sumum þessara marka frekar en í flestum tilfellum voru þau flögguð til nánari skoðunar.

Handvirk leiðrétting

Eftir umbreytingu voru Gullstaðallinn og IFD yfirfarin og leiðrétt handvirkt. Eftir leiðréttingu var Gullstaðallinn tilreiddur aftur, með nýja tilreiðara Máltækniáætlunar (Tilreiðari frá Miðeind, sjá undirverkefni I3) með ákveðnum takmörkunum til að koma í veg fyrir að villum yrði bætt við Gullstaðalinn. Eftir tilreiðslu las skrifta í gegnum Gullstaðalinn og leiðrétti eða flaggaði línur sem höfðu breyst í tilreiðingunni. Flögguðu línurnar voru svo yfirfarnar handvirkt.

Útgáfa

Gullstaðallinn (MIM-Gold) 20.05 inniheldur þær 13 undirmálheildir sem málheildin samanstendur af, auk eins skjals sem inniheldur alla málheildina. PDF skrá lýsir nýju markaskránni, breytingunum sem gerðar voru og hvernig margræð tilfelli við mörkun voru leyst.

Prófunar- og þjálfunarsett fyrir Gullstaðalinn og IFD innihalda 10 skrár með þjálfunarsettunum og 10 skrár með prófunarsettunum sem má nota til að gera tífalda krossprófun.

I2 – Orðtökutól

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

Varða 3 – lýsing

Skýrsla um þarfir, form úttaks og skilgreiningar á aðferðum sem notaðar eru. Stopporðaskrár skilgreindar og samsettar. Orðtökutól er tilbúið.

Afurðir

1) Orðtökutól sem nota skal með Beygingarlýsingu íslensks nútímamáls og ISLEX en má laga að öðrum orðabókum eða orðalistum á vefnum.
2) Stopporðaskrá, sem inniheldur m.a. ítarlega stopporðaskrá sem sótt var handvirkt í Risamálheildina.
3) Dæmi um okkar úttaksskrár.

Allar afurðir má finna hjá CLARIN-IS:
https://repository.clarin.is/repository/xmlui/handle/20.500.12537/79
og á Github:
https://github.com/steinunnfridriks/ordtaka

Verkþátturinn klárast í M3.

Skýrsla

Markmið með orðtökutólinu er að styðja við þróun og stækkun orðabóka og orðalista á vefnum, einna helst Beygingarlýsingu íslensks nútímamáls (BÍN) og ISLEX.

Tólið er hannað utan um Risamálheildina (IGC) og upplýsingarnar sem eru í skjölum hennar á TEI-formi. Það er að segja, besta útkoman fæst með því að nota tiltæk málfræðileg mörk, uppflettimyndir og orðmyndir sem eru skilgreindar í Risamálheildinni. Orðtökutólið má hins vegar nota á hvaða málheild sem er, sem inntak sem notar annaðhvort sama TEI-form og Risamálheildin eða venjulegt textaform, eftir því hvað notandi vill.

Orðtökutólið styðst við nokkra mismunandi hluti. Eftir þörfum þarf notandinn að hafa aðgang að viðeigandi gagnaskrám BÍN (sem má sækja með skriftu úr pakkanum okkar), gagnaskrá ISLEX og málheildinni sem á að nota (í flestum tilfellum Risamálheildin). Notandinn þarf einnig að hafa aðgang að skránni all_filters.txt sem er innifalin í pakkanum okkar, til að fá aðgang að sérútbúnum stopporðalista Risamálheildarinnar. Uppsetningarskráin býr þá til nauðsynleg gagnasöfn til að keyra forritið. Í stuttu máli ber forritið orðaforða úr viðeigandi gagnasöfnum saman við orðaforða málheildarinnar og sækir lista orða sem væri hægt að bæta í gagnasafnið. Virkni forritsins má lauslega skipta í tvennt, eftir því hvort notandi vill nota uppflettimyndir eða orðmyndir sem grunn fyrir úttaksskrár.

Úttaksskrár, sem má skoða á GitHub síðu okkar, eru eftirfarandi:

1) Tíðni orðmynda án aukalegra upplýsinga.
2) Tíðni uppflettimynda án aukalegra upplýsinga.
3) Tíðni orðmynda með upplýsingum um allar mögulegar uppflettimyndir fyrir gefnar orðmyndir.

4) Tíðni orðmynda með upplýsingum um allar mögulegar orðmyndir fyrir gefnar uppflettimyndir.
5) Tíðni uppflettimynda ásamt upplýsingum um gerðir texta þar sem sú uppflettimynd birtist. Tíðni einstakra textagerða má einnig skoða í lækkandi röð.

Fyrstu tveir valkostir úttaksskráa eru einföldustu myndir þeirra. Þær innihalda ekkert nema orðin sjálf og tíðni þeirra, sem er nytsamlegt til að meta hvort þau eiga heima í orðabókum og orðalistum á vefnum eða ekki. Ef tiltekið orð kemur aðeins fyrir í fáein skipti í málheildinni gæti það verið síður hentugt fyrir stækkun orðasafnsins af ýmsum ástæðum, meðal annars vegna þess hvað það er sjaldgæft í tungumálinu, mögulega vitlaust stafsett eða hluti af sérhæfðum orðaforða sem er ekki notaður af almennum málnotendum. Aðrir valkostir úttaks innihalda aukaupplýsingar sem gætu hæft ákveðnum rannsóknum. Þriðji valkosturinn gefur upplýsingar um hvaða uppflettimyndir koma fyrir í málheildinni fyrir gefna orðmynd. Þetta veitir upplýsingar um hvort orðmyndin geti talist til fleiri en eins orðflokks, auk þess hvort sjálfvirk lemmun virkar. Á hinn bóginn gefur fjórði valkosturinn upplýsingar um allar mögulegar orðmyndir fyrir ákveðna uppflettimynd, sem gerir mögulegt að kanna hvort ákveðin orðmynd er mikið algengari eða sjaldgæfari en aðrar og hvort orðmyndin sé aðeins notuð sem hluti af ákveðnu orðatiltæki. Seinasti valkosturinn gefur upplýsingar um þær gerðir texta sem ákveðið orð birtist í, sem gerir það mögulegt að útbúa sérhæfðan orðalista (t.d. orðalista íþróttatengdra orða).

Þrátt fyrir að Risamálheildin hafi bylt þróun íslenskra máltæknitóla er hún ekki gallalaus. Samkvæmt skilgreiningu sýnir málheildin raunverulega notkun tungumálsins og því innsláttar- og stafsetningarvillur og erlend orð sem teljast ekki nothæf til orðtöku. Þar sem hún er sjálfvirkt mörkuð og lemmuð, geta beygingar ákveðinna orða einnig verið rangar. Af þessari ástæðu höfum við handvirkt útbúið víðtækan lista orða sem skal hunsa þegar orðtökutólið er notað. Þessi listi inniheldur erlend orð, innsláttar- og stafsetningarvillur, villur við lemmun og skammstafanir. Flestir fyrrnefndra flokka útskýra sig sjálfir (t.d. að innsláttarvillur eiga ekki heima í orðabók). Sérnöfn er einnig hægt að hunsa ef þurfa þykir, aðallega vegna þess að þau eru sérstaklega algeng í tungumálinu svo þau geta skyggt á úttaksskrár orðalista ef þau eru höfð með.

Þennan stopporðalista má nota utan tólsins. Í fyrsta lagi gæti hver sem notar Risamálheildina í sinni vinnu notað listann til að undanskilja ógagnleg orð frá niðurstöðunum. Þar sem listinn er í venjulegri textaskrá er mjög auðvelt að aðlaga hann hvaða formi sem er til að nota í öðrum verkefnum. Í öðru lagi geta hönnuðir sjálfvirka lemmarans vonandi fengið innsýn í hvað þarf að laga til að bæta virkni hans., með því að skoða hvaða villur eru gerðar við lemmun í Risamálheildinni. Í þriðja lagi mætti bæta skammstöfunum í listanum í safn skammstafana og þýðinga þeirra sem þegar eru aðgengilegar, eða sem grunn að nýju enn aðgengilegra safni. Að lokum gætu innsláttar- og stafsetningarvillurnar, ásamt tíðni þeirra, nýst við þróun málrýnikerfis fyrir íslensku og við að meta hvers konar villur er líklegast að almennir notendur geri.

Við þróun orðtökutólsins voru sumir hlutar sem ekki skiluðu sér í hinni endanlegu afurð. Þar á meðal forrit sem fer ítrekað í gegnum Risamálheildina og leitar að ákveðnum málfræðilegum mynstrum í leit að föstum orðasamböndum/orðatiltækjum, sem nýtist aðallega við önnur verkefni eins og Íslenskt orðanet. Annað forrit leitar af samsettum orðum sem ekki er að finna í BÍN og myndar sjálfkrafa beygingarmyndir þeirra. Þó það sé hugsanlega nytsamlegt, myndar það gjarnan óþarfar beygingar (eins og t.d. *banka* sem getur bæði verið no. *banki* og so. *banka*). Svo er forrit sem fer ítrekað í gegnum Risamálheildina, sækir upplýsingar um tíðni gefins orðs á ársgrundvelli og úttakið er línurit, en það er frekar hægvirkt. Við höfðum einnig með forritaskil fyrir bæði Risamálheildina og BÍN, sem geta verið nytsamleg fyrir smáar beiðnir en ætti ekki að nota fyrir stórar beiðnir þar sem það getur hugsanlega orsakað kerfishrun netlægra gangagrunna þeirra. Að lokum höfðum við með forrit sem sækir allar uppflettimyndir úr Risamálheildinni sem veita nytsamlegar upplýsingar en nýtast hugsanlega ekki við almenna notkun.

I3 – Textatilreiðari

Skýrsla: Miðeind
Aðilar: Miðeind

Varða 2 - lýsing

Tilgangur
Að veita öðrum NLP tólum tilreiddan texta.
Tilreiddur texti sýnir skil setninga og sýnir hvern tóka aðskilinn með hvítbili.

M2

- Skýrsla um lista jaðartilvika tilbúin. Tilreiðarinn mun hafa innleitt reglur til að taka á öllum þekktum tilvikum. Sjaldgæf tilvik og/eða mjög margræð mynstur sem ekki er tekið á, verða skjöluð.
- Stikaðar stillingar fyrir öll skilgreind tilvik hafa verið innleiddar.
- Tilreiðarinn verður fínstilltur til að taka á algengum mynstrum sem koma fram í fréttatextum og verða öll prófuð á því sviði með öllum stikuðum stillingum. Við stefnum á nákvæmni á bilinu 96%-98% fyrir allar stillingar í slíkum prófunum.
- Tilreiðarinn er nú að fullu útfærður samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins.

Verkefninu var lokið við vörðu M2 og er haft með hér fyrir heildarútlit.

Afurðir

Öllum markmiðum vörðunnar var náð.

Innleiðing og útgáfa
Tilreiðarinn hefur verið afhentur samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins og er aðgengilegur almenningi á GitHub og CLARIN. Textakóðun er UTF-8 í alla staði. Kóði og athugasemdir eru á ensku. Hugbúnaðurinn er undir MIT leyfi, í samræmi við þær lágmarkskröfur í leyfamálum sem skilgreindar eru fyrir allt verkefnið.

Tilkynningar höfundarréttar og leyfa eru innifaldar í höfuðathugasemd í öllum frumskrám.
Travis CI er notað fyrir stöðuga innleiðingu og prófanir.

Hugbúnaðurinn virkar á bæði Python 2.7 og Python >= 3.5, og hefur verið prófaður á Linux, MacOS og Windows. Hægt er að setja hann upp sem Python pakka og skipanalínutól með pip, og hann er hýstur í Python Package Index (PyPi) sem tokenizer. Skjölun fyrir uppsetningu og notkun er á GitHub.

Tilreiðarinn er í stöðugri þróun, með nokkrum nýjum útgáfum síðan í vörðu M2.

Greining og mælingar
Stór prófunarsvíta/prófunarsafn var þróað fyrir tilreiðarann. Það er aðgengilegt á GitHub.
Innbyggðar prófanir eru í skránni test/test_tokenizer.py og er hægt að keyra með pytest-tólinu.

Skráin test/toktest_large.txt inniheldur prófunarsett 13.075 lína. Línurnar prófa greiningu setninga, greiningu tóka og flokkun tóka. Til greiningar inniheldur test/toktest_large_gold_perfect.txt viðbúið úttak grunntilreiðslu (sjá eftirfarandi hluta) og test/toktest_large_gold_acceptable.txt inniheldur

núverandi úttak grunntilreiðslu. Samanburður skránna tveggja (fullkominn á móti ásættanlegum) skjalfestir jaðartilvik, ásamt skýrslu úr vörðu M1.

Skráin test/toktest_edgecases.txt inniheldur runutexta úr nýlegum fréttagreinum, með 50 setningum og 957 tókum. Stafsetningarvillur voru leiðréttar. Skráin var notuð til að prófa grunntilreiðslu, bæði greiningu setninga og greiningu tóka. Gullstaðal fyrir þá skrá má finna í skránni test/toktest_edgecases_gold_expected.txt.
Tilreiðarinn nær 100% nákvæmni fyrir greiningu setninga í prófunarsettinu og 99,7% nákvæmni fyrir greiningu tóka (3 villur).

Stikaðar stillingar
Eftir því sem verkefnið þróaðist og þarfirnar urðu skýrari, voru tveir hamir og stillingar tilreiðslu innleiddar: fulltilreiðsla (e. *deep* tokenization) og grunntilreiðsla (e. *shallow* tokenization). Þessir tveir hamir tilreiðslu voru metnir fullnægjandi fyrir flest verkefni.

Grunntilreiðsla skilar hverri setningu úr streymi textaheimilda sem streng (eða sem línu texta í úttaksskrá) þar sem einstakir tókar eru aðgreindir með bilum. Þetta hentar t.d. við vélþýðingar og mörkun. Undirverkefni I4 – Málfræðilegur markari – í Máltækniverkefninu notar grunntilreiðslu sem inntak, ásamt fulltilreiðslu til meðhöndlunar flóknari tóka.

Fulltilreiðsla skilar hverjum einstökum hlut tóka sem hefur verið markaður með tegund tóka og öðrum upplýsingum fengnum úr tókanum. Til dæmis er færslan (e. tuple) *(ár, mánuður, dagur)* unnin upp úr dagsetningum og skilgreiningum skammstafana er flett upp og þær geymdar með orðatókum sem tákna skammstafanir. Þetta hentar t.d. við þáttun og leiðréttingu málfræði. Verkefni I5 – Þáttari – og L6 og L7 – Málrýnir – í Máltækniverkefninu, reiða sig á fulltilreiðslu.

Fyrir hvern ham virkni styður tilreiðarinn fínstillingu með skipanalínu eða stikum falla í Python. Þessu er lýst í skjölun verkefnisins (á ensku).
Frekari þróun tilreiðarans er nú talin hluti af undirverkefnum þáttunar og málrýnis.

I4 –Málfræðilegur markari

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík, Stofnun Árna Magnússonar í íslenskum fræðum

Varða 3 – lýsing

Tilgangur
Að þróa fyrsta flokks málfræðilegan markara (e. PoS tagger) fyrir íslensku. Þessi verkþáttur felur í sér greiningar villna sem gerðar eru af tiltækum íslenskum mörkurum fyrir mismunandi gerðir texta, auk rannsóknar á því hversu mikilli nákvæmni má ná við mörkun íslenskra texta.

M2

- Prófunarsett sviða hafa verið undirbúin samkvæmt endurskoðaðri íslenskri markaskrá.
- Prófanir á prófunarsetti Gullstaðals hafa verið gerðar.
- Villur hafa verið greindar og flokkaðar; skýrsla um einkenni hvers prófaðs markara tilbúin.
- Fínstillingar besta/u markara/líkans/a og prófanir á öllum sviðum eru vel á veg komnar.

M3

- Besti markari/líkan hefur verið afhent samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins.
- Skýrslum um niðurstöðu prófana á öllum völdum sviðum texta er lokið.
- Markmiðið fyrir afhentan markara er minnst 95% á prófunarsetti Gullstaðals.

Afurðir

Öllum markmiðum vörðunnar var náð. Verkþátturinn heldur áfram á öðru ári verkefnisins.

- Prófunarsett sviða hafa verið undirbúin, sjá skýrslu um matssett í G1 (Risamálheildin).
- Prófanir hafa verið gerðar á prófunarsettum sviða, sjá skýrslu um matssett í G1 (í skýrslunni vísar ABLTagger til ABLTagger 1.0).
- Prófanir á Gullstaðli hafa verið gerðar og nákvæmni er viðunandi, sjá hlutann „Niðurstöður Gullstaðals“ fyrir neðan.
- Villur greindar og flokkaðar, sjá hlutana „Villugreining“ og „Villuflokkun” fyrir neðan.
- Endurskoðaður markari hefur verið gefinn út, sjá hlekk fyrir neðan.

Málfræðilegur markari: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/53

Skýrsla

Líkanið (ABLTagger) sem kynnt var í Steingrímsson ofl. (2019)[4] er BiLSTM málfræðilegur markari með samþættuðum orða- og bókstafagreypingum (e. combined word and character embeddings), aukinn með orðhlutafræðilegu orðasafni og greiningu orðasafnsflokks. Líkanið var innleitt með notkun

[4] Steingrímsson, Steinþór, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria.

DyNet[5,6]. Líkanið var þróað enn frekar fyrir vörður 2 og 3 og endurinnleitt með notkun PyTorch[7,8]. Þetta var gert til að sameina notkun innan máltækniverkefnisins og lágmarka áhættu, þar sem PyTorch er vel viðhaldið. Eldri útgáfa markarans eru nú nefnd útgáfa 0.9, og nýrri útgáfan 1.0. Þessi útgáfunúmer eru valin þar sem ABLTagger var aldrei skilað til CLARIN og útgáfunúmer 1.0 er gjarnan notað til að tákna fyrstu almennu útgáfu. Útgáfa 1.0 er aðgengileg á CLARIN[9].

Meirihluti vinnunnar fyrir vörðurnar tvær var tæknilegs eðlis og fól í sér árangursgreiningu mismunandi hluta líkanskins og framkvæmd tilrauna. Líkanið samanstendur af fjórum hlutum. Fyrstu þrír tákna orð/tóka á mismunandi hátt: líkan stafa-til-orðs (e. characters-to-word), orðagreypingar og upplýsingar frá Beygingarlýsingu íslensks nútímamáls (BÍN). Fjórði hlutinn sameinar hinar þrjár birtingarmyndirnar á runubundinn hátt yfir alla setninguna og ályktar um málfræðileg mörk.

Greining sýndi að upprunalega líkanið náði að mörgu leyti slökum árangri. Líkan stafa-til-orðs var ekki þjálfað nægilega með viðunandi hætti. Orðagreypingarnar voru „auðveldlega“ ofmátaðar við þjálfunargögnin sem olli því að þjálfun stöðvaðist snemma. BÍN-hlutinn var uppfærður á meðan á þjálfun stóð sem olli slökum árangri á óséðum tókum. Tekið var á öllum þessum vandamálum í nýju útgáfunni. Fleiri fínstillingar voru gerðar á líkaninu, meðal annars var orðflokkagreining í sérstöku skrefi fjarlægð, sem hægði verulega á mörkuninni og þjálfunarferlinu[10], forþjálfaðar orðagreypingar (þjálfaðar á Risamálheildinni) voru innleiddar, flækjustig líkansins hækkaði og *regularization* var notað til að draga úr ofmátun. Þetta bætti heildarskilvirkni og árangur líkansins, en almenn uppbygging þess er óbreytt.

Niðurstöður fyrir Gullstaðal

Fyrir neðan má sjá niðurstöður nákvæmniprófana á markaranum/mörkurunum (í fleirtölu í tilfelli krossprófana með notkun mismunandi þjálfunar- og prófunarsetta. Fyrsta gagnasettið, sem vitnað er til sem Gullstaðalsins, er MIM-Gold[11] og inniheldur u.þ.b. 1 milljón handmarkaðra tóka. Íslenska orðtíðnibókin (IFD)[12] er eldri staðall, byggð á eldri textum, sem inniheldur u.þ.b. 590.000 handmarkaða tóka. Bæði gagnasettin nota endurskoðuðu markaskrána (sjá G10 – Gullstaðall fyrir málfræðilega mörkun).

Fyrir neðan er sett fram nákvæmni ABLTagger v1.0 þegar hann var þjálfaður og prófaður með MIM-Gold og/eða IFD. Síðan berum við heildarnákvæmnina saman við ABLTagger 0.9,. Báðum málheildum hefur verið skipt í 10 hluta, þar sem tíundi hlutinn er notaður fyrir stillingar á forstilltum breytum (e. hyperparameter tuning), og fyrstu 9 brotin fyrir þjálfun og prófanir (níföld krossprófun).

ABLTagger 1.0	**Alls (% / #tóka)**	**Óþekktir (% / #tóka)**	**Þekktir (% / #tóka)**
MIM-Gold	95,59 / 100095	84,90 / 8511	96,59 / 91584
IFD	96,40 / 59027	90,33 / 4044	96,85 / 54983

[5] https://github.com/clab/dynet

[6] https://github.com/steinst/ABLTagger

[7] https://pytorch.org/

[8] https://github.com/cadia-lvl/POS

[9] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/53

[10] Þörf er á nánari greining varðandi hvort skref sem greinir orðflokk bætir árangur. Þessi greining er hluti af áætlun næsta árs.

[11] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/40

[12] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/38

MIM-Gold+IFD	95,78 / 100095	84,13 / 7435	96,72 / 92660

„Alls“ er meðalnákvæmni yfir 9 hluta málheildarinnar, ásamt fjölda tóka innan þess flokks, í þessu tilfelli, allir tókar. „Óþekktir“ er meðalfjöldi tóka sem voru ekki til staðar í þjálfunargögnunum. „Þekktir“ er meðalfjöldi tóka sem voru til staðar í þjálfunargögnunum (en hugsanlega í annarri merkingu/samhengi).

ABLTagger 1.0	**Þekktar greyp. (e. WordEmb) (% / # tóka)**	**Þekkar greyp.+BÍN (% / # tóka)**	**Þekkt BÍN (% / # tóka)**	**Þekkt WordsAsChars (% / # tóka)**
MIM-Gold	97,74 / 16367	96,30 / 73176	88,84 / 9	97,79 / 2032
IFD	99,40 / 7653	96,46 / 46882	96,36 / 67	93,13 / 382
MIM-Gold+IFD	97,53 / 16408	96,52 / 74195	86,31 / 10	97,34 / 2047

ABLTagger 1.0	**Óþekktar greyp. (% / # tóka)**	**Óþekktar greyp.+BÍN (% / # tóka)**	**Óþekkt BÍN (% / # tóka)**	**Óþekkt WordsAsChars (% / # tóka)**
MIM-Gold	77,91 / 2053	92,33 / 4512	89,99 / 167	73,11 / 1779
IFD	66,53 / 213	93,08 / 3112	94,54 / 476	67,71 / 243
MIM-Gold+IFD	76,91 / 2013	92,92 / 3493	92,10 / 165	73,76 / 1764

Flokkarnir „Þekktir“ og „Óþekktir“ eru sundirliðaðir frekar í mismunandi flokka. Þar sem við notuðum forþjálfaðar orðagreypingar og gögn frá BÍN verður hugtakið „þekktir tókar“ brenglað. Til dæmis: Tóki sem er ekki í þjálfunargögnunum („Óþekktir“), en er til staðar í forþjálfuðum orðagreypingum („WordEmb“) og BÍN, er því flokkaður sem „Óþekktar greyp.+BÍN“. „*WordAsChars*“ eru tókar sem eru hvorki í forþjálfuðum orðagreypingum né BÍN. Summan af tókum í „Þekkt x“ og „Óþekkt x“ er fjöldi tóka í „Þekkt“ og „Óþekkt“, hvert um sig.

Fyrir MIM-Gold, er meðal nákvæmni með notkun nífaldrar krossprófunar **95,59%**. Þetta er yfir 95% eins og ákvæði segja um. Borið saman við 94,47% nákvæmni í útgáfu 0.9, hefur nákvæmni aukist um 1,12 prósentustig, sem fækkar villum um 20,3%.

Með því að nota allt gagnasett IFD sem aukaleg þjálfunargögn (MIM-Gold+IFD) hækkar nákvæmni nífaldrar krossprófunar í **95,78%**, sem er hækkun um 0,69 prósentustig frá 95,09% í útgáfu 0.9.

Fyrir IFD er nákvæmni nífaldri krossprófunar **96,40%**, eða hækkun um 0,85 prósentustig frá 95,54% í útgáfu 0.9.

Villugreining

Þónokkrar tífaldar[13] krossprófanir voru keyrðar með mismunandi stillingum málfræðilegu markaranna þriggja, ABLTagger 1.0, ABLTagger 0.9, IceStagger og TriTagger með notkun

[13] Nífaldar fyrir ABLTagger 1.0.

ofangreindra þjálfunargagna. Sökum lengdar skýrum við aðeins frá ABLTagger útgáfu 1.0 og 0.9 hér. Fyrir nánari upplýsingar má sjá skýrslu í fullri lengd.[14]

ABLTagger 1.0	**Alls %**	**Þekkt orð %**	**Óþekkt orð %**	**Hlutf. óþekktra orða %**
MIM-Gold	95,59	96,59	84,90	8,50
IFD	96,40	96,85	90,33	6,85
MIM-Gold+IFD	95,78	96,72	84,13	7,43

Athugið að í ABLTagger 1.0 skilgreinum við „Þekkta“ og „Óþekkta“ tóka eftir því hvort þeir eru í þjálfunarsettinu.[15] Þetta er ekki síst vegna þess að í framtíðinni munum við nota orðflísatilreiðingu, sem mun flækja hugtakið „Þekktir tókar“ enn frekar. Fyrir neðan eru tölur sem byggja á sömu niðurstöðum en þar sem „Þekkt“ og „Óþekkt“ er skilgreint á sama máta og ABLTagger 0.9, það er að segja IceStagger og TriTagger. Við skilgreinum „Þekkta tóka“ sem tóka í þjálfunarsettinu, í forþjálfuðum orðagreypingum og í BÍN. Með öðrum orðum, nákvæmni „Óþekktra“ er nákvæmni „Óþekkt *WordsAsChars*“.

ABLTagger 1.0 modified “Known”	**Alls %**	**Þekkt orð %**	**Óþekkt orð %**	**Hlutf. óþekktra orða %**
MIM-Gold	95,59	96,00	73,11	1,78
IFD	96,40	96,52	67,71	0,41
MIM-Gold+IFD	95,78	96,18	73,76	1,76

ABLTagger 0.9	**Alls %**	**Þekkt orð %**	**Óþekkt orð %**	**Hlutf. óþekktra orða %**
MIM-Gold	94,47	95,58	66,64	3,84
IFD	95,54	95,86	53,95	0,76
MIM-Gold+IFD	95,09	95,93	63,96	2,63

Villuflokkun

Fyrir neðan eru tíu algengustu villurnar fyrir ABLTagger 0.9 og sæti þeirra fyrir ABLTagger 1.0 við mörkun með **MIM-IFD**. MIM-IFD er ekki sama gagnasett og MIM+IFD og er notað hér fyrir afturhæfi. Þetta gagnasett samanstendur af þjálfunargögnum frá MIM og IFD, í hverju broti, auk prófunargagnanna. Þetta orsakar smærra þjálfunarsett og stærra prófunarsett samanborið við MIM+IFD. Fyrir enn fleiri villur og niðurstöður sem innihalda IceStagger og TriTagger sjá skýrslu í fullri lengd.

[14] Skýrslan er einnig geymd í GitHub hirslunni: https://github.com/cadia-lvl/POS/blob/master/I4%20-%20POS%20tagger%20-%20M2%20%26%20M3.odt

[15] Við tökum meðaltal tóka í hverjum flokki yfir settin.

	ABLTagger 1.0			ABLTagger 0.9		
markari>rétt	**nr.**	**fjöldi**	**hlutfall**	**nr.**	**fjöldi**	**hlutfall**
af > aa	1	1653	2,65	1	1989	2,57
aa > af	2	1421	2,28	2	1694	2,19
nveþ > nveo	4	979	1,57	3	1373	1,78
n----s > e	5	886	1,42	4	1269	1,64
nveo > nveþ	3	988	1,58	5	1243	1,61
nheo > nhfo	7	561	0,90	6	684	0,88
e > n----s	6	770	1,23	7	677	0,88
nkeo > nkeþ	9	497	0,80	8	650	0,84
nkeþ > nkeo	8	510	0,82	9	584	0,76
nhen > nheo	13	435	0,70	10	567	0,73

Á milli ABLTagger 0.9 og 1.0 sjáum við að flestir flokkar hafa færri villur (u.þ.b. 20% færri villur) og dreifingin helst sú sama. Tíunda villan í útgáfu 1.0 er „“ct > c“, með 491 tilfelli.

I5 – Þáttari

Skýrsla: Miðeind
Aðilar: Miðeind

Varða 3 – lýsing

Tilgangur
Að afhenda fyrsta flokks þáttara fyrir íslensku sem þjóna þörfum ólíkra máltækniverkefna.

Ákvæði
Tilreiðari (I3)
Málfræðilegur markari (I4)

M3: Ný vandamál greind og leyst.

Afurðir

Markmiðum var náð. Verkþátturinn mun halda áfram á öðru ári verkefnisins.

1. Yfirlit

Þetta undirverkefni felur í sér mat og endurbætur á þeim þátturum sem til eru fyrir íslensku. Til að gera það var almennt þáttunarskema skilgreint og prófunarsett sem notar það skema búið til. Prófunarsettið hefur verið notað til að árangursmæla þáttarana og niðurstöðurnar hafa komið að góðum notum við að stýra endurbótum og forgangsraða þeim.

Þáttunarhluti IceNLP[16], *IceParser*, er hlutaþáttari sem hefur þegar verið notaður við fjölda verkefna, t.d. MerkOr (1, GitHub), gagnagrunn merkingarfræði fyrir íslensku og í tilraun í sjálfvirka upplýsingaöflun (1, GitHub). Hann var einnig notaður til þáttunar upprunalegrar útgáfu grunnþáttuðu prófunargagnanna fyrir þetta undirverkefni.

Greynir[17] er (full-)þáttari sem hefur meðal annars verið notaður til að leiðrétta málfræðivillur (undirverkefni L7 af máltækniverkefninu), fyrir þáttun upprunalegrar útgáfu fullþáttaðra prófunargagna innan þessa undirverkefnis, í textavinnslu fyrir vélþýðingar (sjá undirverkefni V3b), sem vélin bakvið raddstutt app/gerviþjón (e. voice assistant app), í hugbúnaði fyrir sjálfvirkan útdrátt hugtaka og einnig fyrir þáttað málheildarverkefni (GreynirCorpus) utan máltækniverkefnisins.

Meðal annarra nota þáttaranna er merkingarfræðileg greining, samvísun og greining á endurvísunum, upplýsingaheimt, spurningasvörun, vélþýðingar, talþjónar og rannsóknir. Þáttarana má einnig nota til að aðstoða við myndun margvíslegra tegunda þjálfunarmálheilda fyrir tauganet.

2. Prófanir og mat

2.1 Prófunarmálheild

[16] Á GitHub: https://github.com/hrafnl/icenlp
[17] Á GitHub: https://github.com/mideind/GreynirPackage

Handvirkt þáttuð og mörkuð prófunarsett fyrir bæði djúp- og fullþáttun voru búin til sem hluti af þessu undirverkefni. Prófunarmálheildin inniheldur 450 setningar af handahófi úr íslenskum textum frá síðastliðnum tíu árum. Setningarnar voru valdar af handahófi úr fjórum flokkum texta: 300 setningar úr fréttatextum og 50 setningar hver um sig frá þingtextum, Vísindavefnum og bókmenntatextum úr útgefnum bókum. Að auki eru 250 setningar fyrir hvern þeirra þriggja síðarnefndu flokka tilbúnar til þáttunar og viðbótar í málheildina þegar og ef þarf.

Hver skrá í prófunarmálheildinni var tvisvar þáttuð og mörkuð handvirkt; einu sinni samkvæmt almennu *djúp*þáttunarskema (e. general *deep* parsing scheme) og einu sinni samkvæmt almennu *grunn*þáttunarskema (e. general *partial* parsing scheme). Þessum skemum er lýst nánar í kafla 2.2 hér fyrir neðan.

Verkefni utan máltækniáætlunar, *GreynirCorpus*, stefnir á að handþátta allt að 5.000 setningar úr fréttatextum samkvæmt skema Greynis fyrir nóvember 2020. Þegar þetta er skrifað inniheldur málheildin 4.530 handvirkt merktar setningar. Fréttatextarnir fyrir prófunarmálheildina voru sóttir úr þessu verkefni. Þingstextarnir, vísindatextarnir og bókmenntatextarnir voru sóttir úr Risamálheildinni (sjá undirverkefni G1).

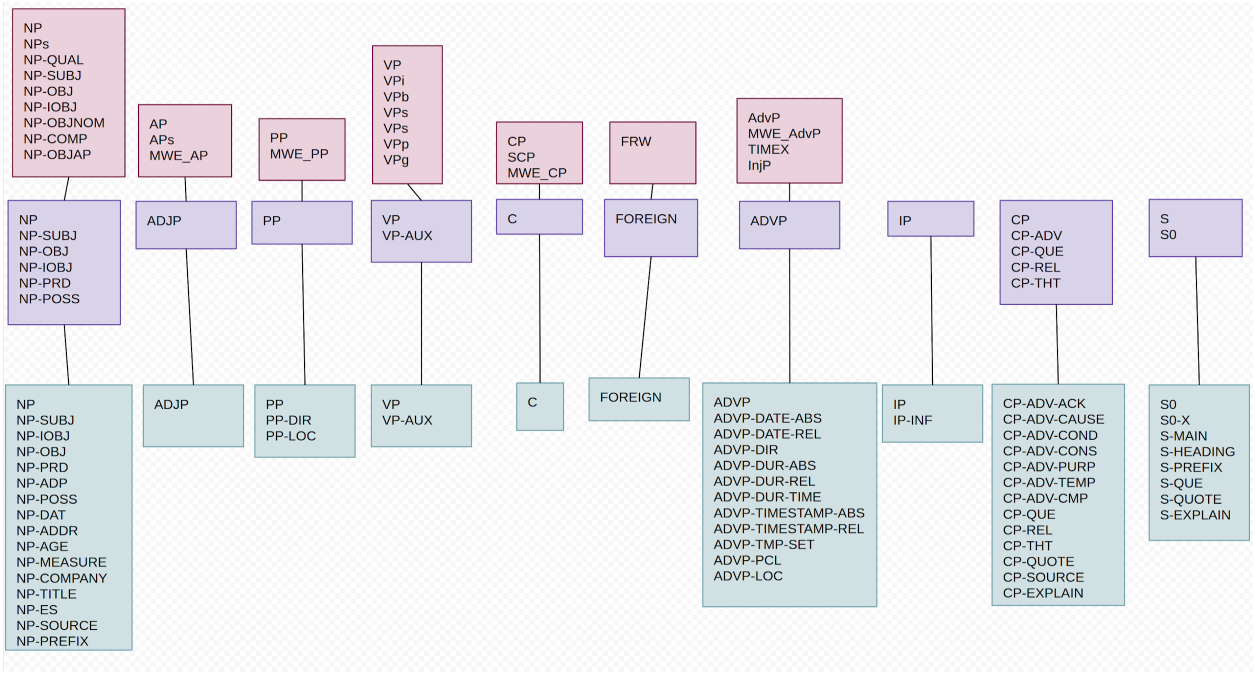
2.2 Þáttunarskemu
Prófunarsettið fylgir staðlaða almenna skemanu sem skilgreint var í vörðu M1, með þeim frávikum að *nonterminals* S, CP, og IP voru aðeins tekin með í djúpþáttunarskemanu, og *nonterminals* NP-ADP og FOREIGN var bætt við í bæði skemu, þar sem bæði Greynir og IceParser veita nauðsynlegar upplýsingar til auðkenningar þeirra.

2.3 Merkingarferli (e. annotation process)
Til að undibúa handvirka merkingu og til að meta úttak þáttara var búin til pípa. Skjölum innan hvers flokks var skipt í einfaldar textaskrár sem hver inniheldur 10 setningar. Hver skrá var þáttuð sjálfvirkt með Greyni. Skrárnar voru svo leiðréttar með því að nota betrumbætta kvísl (e. fork) handvirka mörkunartólsins Annotald[18], lagaða að þáttunarskema Greynis.

Við þetta skref var upphaflega áætlað að varpa þáttuðum skrám frá skema Greynis yfir á almenn þáttunarskemu full- og grunnþáttunar. Prófunarmálheildin myndi síðan samanstanda af þessu. Þetta virkaði vel með skema fullþáttunar, af augljósum ástæðum, en þegar kom að skema grunnþáttunar ollu skil liðanna vandræðum. Sem dæmi, ef við höfum nafnlið sem inniheldur tilvísunarsetningu birtist endi nafnliðarins á eftir tilvísunarsetningunni í tréi fullþáttunar. Grunnþáttunarskemað merkir ekki undirskipaðar setningar (e. subclauses), svo þar birtist endi nafnliðarins á undan tilvísunarsetningunni. Þetta tiltekna dæmi gæti virst einfalt að laga í eftirvinnslu, en þar sem dæmin voru mun fleiri og mörg flóknari og lúmskari, ákváðum við að það yrði minni vinna að fara aðra leið.

Hver skrá var þáttuð sjálfvirkt með IceParser. Þáttuðum skrám voru svo varpað yfir á almennt grunnþáttunarskema. Þetta gerði okkur kleift að nota Annotald-tólið til að leiðrétta þær handvirkt, þar sem almennu skemun eru undirmengi af skema Greynis (og grunnþáttunarskemun eru undirmengi IceParser skemans); hins vegar inniheldur IceParser-skemað formgerðir sem ekki eru til staðar í skema Greynis og því ekki í Annotald. Rauðu reitirnir tákna IceParser (grunn)skemað, fjólubláu reitirnir almenna skemað, og bláu reitirnir (fullþáttað) skema Greynis. Fjólubláu boxin sem aðeins tengjast bláum boxum tákna liði sem birtast aðeins í almenna fullþáttaða skemanu (e. general deep schema).

[18] Sjá https://github.com/mideind/Annotald



1. Vörpun IceParser og Greynis yfir á almennu skemun.

Endanleg niðurstaða var málheild með 90 skrám fjögurra textaflokka, hver með 10 setningum. Fyrir hverjar 45 upprunalegar textaskrár eru til tvær handvirkt þáttaðar og merktar útgáfur; ein fullþáttuð og ein grunnþáttuð. Þáttuðu skrárnar fylgja formi Penn-trjábankans.

```
( (META (ID-CORPUS 0a3a6244-0e88-11e8-8126-04014c605401.5)
        (ID-LOCAL reynir_corpus_00120.psd,.7)
        (URL http://www.mbl.is/frettir/innlent/2018/02/10/threngslin_opnud_fyrir_umferd/)
        (COMMENT ))
  (S0 (S-MAIN (IP (ADVP (ao Þá (lemma þá)))
                  (VP (VP-AUX (so_0_fh_p3_ft_nt_gm hafa (lemma hafa)))
                      (ADVP (ao einnig (lemma einnig)))
                      (VP (VP (so_0_sagnb_mm borist (lemma bera)))
                          (NP-SUBJ (no_ft_nf_kvk tilkynningar (lemma tilkynning))
                                   (PP (P (fs_þf um (lemma um)))
                                       (NP (lo_ft_þf_kvk_sb fastar (lemma fastur))
                                           (no_ft_þf_kvk bifreiðar (lemma bifreið))
                                           (PP (P (fs_þgf á (lemma á)))
                                               (NP (sérnafn_et_þgf_kk Laugarvatnsvegi (lemma Laugarvatnsvegur)
                                                                                      (exp_seg Laugar-vatnsvegur))
                                                   (C (st og (lemma og)))
                                                   (sérnafn_et_þgf_kvk Biskupstungnabraut (lemma Biskupstungnabraut)
                                                                                          (exp_seg Biskupstungna-braut)))))))))))
      (grm ,)
      (C (st en (lemma en)))
      (S-MAIN (IP (NP-SUBJ (fn_ft_þgf_kk þessum (lemma þessi))
                           (no_ft_þgf_kk vegum (lemma vegur)))
                  (VP (VP-AUX (VP (so_0_op_subj_þgf_fh_p3_et_nt_gm hefur (lemma hafa)))
                              (ADVP (ao ekki (lemma ekki)))
                              (VP (so_0_sagnb_gm verið (lemma vera))))
                      (VP (so_0_lhþt_et_nf_hk_sb lokað (lemma loka))))))
      (grm .)))
```

2. Dæmi um fullþáttaða setningu í málheildinni, sem fylgir formi Penn-trjábankans.

2.4 Pípa prófana

Pípa prófana var innleidd fyrir innri mælingar. Pípan varpar þáttunarúttaki hvers þáttara yfir á viðeigandi almennt skema áður en það er prófað með viðeigandi gullstaðli. Pípa prófana notar hefðbundna matsforritið *evalb*[19] til að mæla heimt út frá svigum, nákvæmni út frá svigum, F-gildi út frá svigum, meðaltal sviga sem skarast, og nákvæmni marka. Mæling á nákvæmni marka nær aðeins til málfræðilegrar eins og er; ítarlegri mælingar verða innleiddar síðar. Sama gildir um málfræðileg hlutverk sem hefur verið flaggað. Mælingarnar eru sjálfgefnar fyrir hvern flokk texta og fyrir hvern

[19] Sjá https://nlp.cs.nyu.edu/evalb/

þáttara allrar prófunarmálheildarinnar. Pípan er aðgengileg á GitHub (https://github.com/mideind/ParsingTestPipe).

3. Forgangsröðun

Eftir greiningu villna sem hver þáttari olli bjuggum við til (sem hluta af vinnu við M2) forgangsraðaðan lista endurbóta fyrir hvern þáttara. Eftirfarandi endurbótum var forgangsraðað fyrir IceParser:

- Styðja ætti fleiri margra orða segðir (e. MWEs) (og í mörgum tilfellum sameina í einn tóka).
    - Forsetningar (þvert á, fram hjá)
    - Atviksliðir (hins vegar, hér á landi)
    - Tímaliðir (klukkan 19.10, í ár, á morgun, árið 2020, desember 2013)
    - Nafneiningar (e. named entities) (Kristín Björgvinsdóttir, Finsbury Park)
    - Prósentur og aðrir liðir mælinga (80 prósent, 20 kíló)
- Í sumum tilfellum er punkturinn sem lýkur setningu ekki úttekinn sem sér tóki, heldur geymdur í enda síðasta tóka.
- Heildstæðara safn greinarmerkja myndi koma í veg fyrir að þau séu greind sem óþekkt nafnorð.
- Marga orða samsetningar/orðasambönd með bandstriki (land- og sjávardýr, sölusamningur og -laun, B.Ed.-próf, 1473-1543) ætti að sameina í einstaka tóka.
- Í fáeinum tilfellum virðist markarinn ekki skila marki fyrir tóka.
- Betri markari myndi bæta niðurstöður.
- Heildstæðari orðaforði og betra mállíkan myndi koma í veg fyrir margar mörkunarvillur (sem orsaka þáttunarvillur). Í mörgum tilfellum myndi tíðni mismunandi marka bæta niðurstöður.
- Markaskrána þarf að uppfæra í samræmi við undirverkefni G10.
- Gagnasett sem inniheldur forsetningarsagnir og agnarsagnir (bjóða í e-ð, lauma e-u inn) myndi bæta greiningu.
- Skoða ætti hvort lýsingarorðssagnfylling ætti að varpa (e. project) nafnlið.
- Bæta ætti við ákveðniliðum fyrir aðra orðflokka en lýsingarorð, til dæmis töluorð og fornöfn.
- Þörf er á betri greiningu erlendra orða.
- Vefföngum þarf að bæta við; eins og er greinast þau sem nafnorð eða erlend orð, eftir endingu þeirra.

Eftirfarandi endurbætur voru forgangsraðaðar fyrir fullþáttun í Greyni:

- Nöfn persóna og aðrar nafneiningar ætti að setja saman í einn tóka.
- Í sumum tilfellum er tókum ranglega sameinað við margra orða segðir. Þetta þarf að fínstilla.
- Þáttarinn er of ágengur í því að skilgreina nafnliði sem andlög, frekar en að taka forsetningarliði, tengiliði, eða í sumum tilfellum atviksliði, til greina sem andlag. Þetta má bæta með betri skilgreiningu sagnaramma – sérstaklega forsetningarsögnum (sögnum sem taka með sér forsetningu) og agnarsögnum – og betra mállíkani.
  PP attachment er vel þekkt vandamál. Betri þekking á forsetningarsögnum myndi bæta niðurstöðurnar.
- Meðhöndlun setninga sem innihalda spor og tóma liði ætti að bæta.
- Bæta ætti yfirborðsgerðum setninga (e. uncovered sentence structures) við málfræðireglurnar. Aðaláherslu ætti að leggja á setningar sem innihalda markaða orðaröð þar sem liðir hafa verið færðir fremst, eða í þolmynd.
- Atvikssetningar líta oft út eins og CP-THT inni í PP og sumar eru frekar skilgreindar eftir merkingu en setningarfræðilegri hegðun. Þetta þarf að samræma og skjala.
- Mikil vinna hefur verið lögð í skilgreiningu tímaliða. Það eru nokkur óalgeng tilfelli sem þarf að skoða betur, sérstaklega formgerð liðanna.
- Villur við ákvörðun orðflokka leiða alltaf til (a.m.k. sumra) þáttunarvillna. Lagfæring slíkra villna ætti að hafa forgang.

4. Endurbætur

Í vinnu við fullþáttarann í aðdraganda M3 höfum við byrjað að bæta samhengisfrjálsa málfræði hans og leitarnám (e. heuristics) þáttarans í samræmi við forgangsröðunina sem lýst er ofar. Að auki höfum við lagt áherslu á að endurbæta kortlagningu milli þáttunarskema, vegna fjölmargra auðleystra vandamála á því sviði.

Sameining nafna í einn tóka: Greining margorða nafna sem margra aðskildra tóka leiddi af sér óþarfa villur. Þar sem markmiðið er að auðvelda útdrátt merkingar, sameinuðum við nöfn í einn tóka.

Rangar margra orða segðir (e. multi-word expressions (MWEs)): Við fjarlægðum rangar skilgreiningar margra orða segja sem við rákumst á, þar sem of áköf sameining í margra orða segðir kom í veg fyrir rétta þáttun.

Sagnarammar og fjölbreyttari andlög: Vinna er þegar hafin við ítarlegri skilgreining forms fyrir sagnaramma. Við munum halda þeirri vinnu áfram á öðru ári og bæta *actions on verb frame data* sem hefur þegar verið safnað í stillingarskrár, til dæmis agnir sem almennt teljast með sögnum.

PP attachment: Bættir sagnarammar ættu að hjálpa með stóran hluta. Við höfum einnig takmörkuð gögn um algenga forsetningarliði innan nafnliða (*karlinn í tunglinu*, *kötturinn í sekknum*) og lýsingarorðsliða og vinna er hafin við hugbúnað sem notar gögnin sem þumalputtareglur fyrir einkunnagjöf.

Spor og tómir liðir: Skema Greynis inniheldur ekki spor og tóma liði sem sérstaklega skilgreindar setningagerðir. Verið er að bæta sérhæfðum reglum í samhengisfrjálsu málfræðina til að taka á algengustu setningagerðunum.

Yfirborðsgerð setninga: Við bættum við málfræðireglum fyrir setningagerðir sem ekki þáttuðust, til dæmis frá falsk-jákvæðri greiningu úr málrýni (L7 undirverkefnið). Þetta fækkaði óþáttanlegum setningum.

Atvikssetningar: Við skjöluðum viðbúna hegðun undirskipaðra tenginga og álíka setningarliða. Allar viðeigandi málfræðireglur voru uppfærðar, ásamt gullstaðlinum.

Aðrir hlutir sem hefur verið forgangsraðað verða teknir fyrir á öðru ári. Auk þess má nefna að upplýsingarnar sem fengust frá verkefni handvirkrar mörkunar GreynirCorpus hafa reynst mjög dýrmætar og hafa leitt til endurbóta bæði innan og utan forgangsraðaðra sviða.

IceParser verður endurbættur fyrir M4. Niðurstöðurnar úr matinu og forgangsröðunin sem sagt er frá hér ætti að vera mikilvæg til að leiða endurbætur.

5. Mat

Eftirfarandi grunnlínumælingar voru gerðar.

IceParser var prófaður á grunnþáttuðu prófunarsetti (450 setningar), sem gaf þessar niðurstöður:

`Heimt út frá svigum: 77,37%`	`Meðalskörun: 0,96`
`Nákvæmni út frá svigum: 79%`	`Nákvæmni marka: 93,48%`

`F-gildi út frá svigum: 78,13%`

Greynir var prófaður á fullþáttuðu prófunarsetti (450 setningar), sem gaf þessar niðurstöður:

Heimt út frá svigum: 74,29%

Nákvæmni út frá svigum: 77,85%

F-gildi út frá svigum: 75,97%

Meðalskörun: 2.50

Nákvæmni marka: 93,95%

Þess ber að geta að prófunarsettin innihalda gallaðar og óformlegar setningar, setningar með stafsetningar- og málfræðivillum og margar erlendar setningar. Engar villur voru leiðréttar fyrir fram í textunum. Þessar óreglulegu setningar hafa verið markaðar og teknar með í gullstaðli, en það má deila um hversu vel má búast við að þáttararnir taki á þeim. Raunhæft markmið um F-gildi fyrir prófunarsettið er því nokkuð neðar en 100%.

Bæði IceParser og Greynir geta skilað setningagerðum sem eru viðbætur við almenna skemað.

Niðurstöðurnar sem birtast hér eru fyrir alla málheildina, en tilhneigingin er að fréttatextar nái bestu einkunn og bókmenntatextar verstu, sem kemur ekki á óvart.

Varðandi málefni utanaðkomandi mælinga má nefna að undirverkefni L7 - Kerfisbundin þróun málrýnis – notar þáttarann Greynir. Þegar ekki er hægt að þátta setningu í málrýnitóli er hún flögguð sem villa. Með því að skoða þessi falsk-jákvæðu tilfelli, þ.e. setningar markaðar sem villur af málrýnitóli en ekki í villumálheild gullstaðalsins, fundum við þónokkrar setningagerðir sem þurfti að dekka með samhengisfrjálsri málfræði Greynis. Þetta er dæmi um gagnkvæman ávinning SÍM samstarfsins, þar sem málrýnitólið hefur ávinning af endurbótum þáttara og öfugt.

Við hlökkum til að sjá fleiri flókin verkefni sem nýta þáttarann, til að fá meiri utanaðkomandi mælingar.

6. Útgáfa

Greynisverkefnið hefur verið þróað áfram með ítrunaraðferðum og hafa millistigsútgáfur verið gefnar út á GitHub og í Python Package Index (PyPI). Við lok M3 var þáttari Greynis gefinn út á GitHub, Python Package Index (PyPI) og á CLARIN (http://hdl.handle.net/20.500.12537/76) undir MIT leyfinu, sem Python pakki (GreynirPackage).

Eins og nefnt er í 4. hluta verður IceParser endurbættur og gefinn út á öðru ári.

V2 – Samhliða málheildir

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

Vörður 2 & 3 – lýsing

M2: Fyrsta flokks aðferðir við samröðun og síun prófaðar og aðlagaðar málheildinni. Handvirkt yfirfarin prófunar-/þróunarsett útgefin.

M3: Skýrsla um síunar- og samröðunaraðferðir. Uppfærður kóði fyrir vinnslupípu útgefinn.

Afurðir

Handvirkt yfirfarin prófunar-/þróunarsett hafa verið gefin út á CLARIN:
- https://repository.clarin.is/repository/xmlui/handle/20.500.12537/24.

Verkþátturinn heldur áfram á öðru ári verkefnisins.

Skýrsla

Prófunar-/þróunarsett

OpenSubtitles og EMA
Sem hluti af vörðu 1 voru bútar valdir til mats úr undirmálheildunum Open Subtitles og Lyfjastofnun Evrópu (EMA). Bútarnir voru prófaðir með KEOPS (https://keops.prompsit.com), þangað til við höfðum náð um 5.000 bútum sem voru gildir sem réttar þýðingar úr hverri undirmálheild. Við fjarlægðum þá búta þar sem enski textinn innihélt þrjú eða færri orð og geymdum þá í aðskildri skrá (test_short_sentences.tmx). Afgangnum var skipt í tvennt, prófunar- og þróunarsett.
Þeir bútar sem voru valdir fyrir prófunar-/þróunarsett en töldust gallaðir á einhvern hátt voru vistaðir í sér skrám.

Þjálfunarsettið samanstendur af afgangi bútanna.

Skrifta var útbúin fyrir OpenSubtitles til að endurbyggja prófunarsettið. Textarnir sem við samröðuðum komu frá www.opus.nlp (sem eru upprunalega frá http://opensubtitles.org) og við getum ekki dreift allri málheildinni.

EES
Nokkur skjöl voru valin af handahófi úr gagnasafni reglugerða Evrópska Efnahagssvæðisins (EES, e. EEA), sem hluti af vörðu 1. Af þeim voru 22 leiðrétt handvirkt svo að samröðun og þýðing væri rétt. Löngum setningum var einnig skipt í smærri hluta þar sem mögulegt var.

Við fjarlægðum ekki stuttar setningar eða endurtekningar úr EES settinu þar sem við vildum halda samhengi og skrám heilum, þar sem mögulegt var.

Skrárnar ees_test.tmx og ees_dev innihalda prófunar- og þróunarsettin. Afgangur bútanna myndar þjálfunarsettið.

	Þýðingareiningar
EEA - dev:	2.035
EEA - test:	1.991
EEA - test_short_sentences:	662
EEA - train:	1.697.927
EMA - dev:	2.254
EMA - rest:	240
EMA - test:	2.255
EMA - test_short_sentences:	491
EMA - train:	399.093
Open Subtitles - dev:	1.531
Open Subtitles - rest:	777
Open Subtitles - test:	1.532
Open Subtitles - test_short_sentences:	2.277
Open Subtitles - train:	1.298.511

Tafla 1: Prófunar-, þróunar- og þjálfunarsett

Skýrsla um síunar- og samröðunaraðferðir. Uppfærður kóði fyrir pípuvinnslu útgefinn

Áður var íslensku samhliðagögnunum samraðað með LF Aligner og þau síuð með algrími sem gefur setningum einkunn, byggðu á orðapokaaðferð tvímála orðasafns (e. bilingual lexicon bag-of-words method). Síunarferlið gaf 3,5 milljón þýðingabúta, af alls 4,3 milljónum búta í gögnunum. Handvirkt mat á hluta bútanna gefur til kynna að u.þ.b. 5% samröðuðu bútanna í endanlegu málheildinni séu gallaðir, en við metum svo að næstum 50% fjarlægðra búta séu ekki gallaðir og hugsanlega nothæfir.

Til að gera þau samröðuðu gögn nothæfari sem við höfum, höfum við samraðað þeim aftur og síað þau aftur með fínstilltari aðferðum.

Samröðunaraðferð okkar notar margmála (e. cross-lingual) greypingar, þjálfaðar á mjög stórum einmála málheildum, til að bootstrappa (e. bootstrap) SMT-kerfi með Monoses. Monoses, upphaflega þjálfað á gögnum sem voru aðeins búin til með margmála greypingunum, er notað til að bakþýða ítrekað setningar frá einmála málheildunum til að búa til frasatöflu fyrir SMT-kerfið.

Þetta SMT-kerfi er því næst notað til að þýða samhliða gögnin okkar, skipta í setningar með tilreiðara máltækniverkefnisins (undirverkefni I3), í stað þess að LF Aligner giski á setningamörk. Þýddu bútarnir eru notaðir til að giska á hvaða samraðanir eru nákvæmastar, með því að nota Bleualign. Setningarnar eru loks síaðar með Bicleaner. Með þessari aðferð metum við nákvæmni samraðaðra para úr runutextum sem 91,6% í stað aðeins 79,1%. Taka skal fram að sumum málheildunum er auðveldara að samraða en runutextum, til dæmis skjátextum og fjölda smærri undirmálheilda.

Enn sem komið er hefur stærsti hluti textanna verið sóttur í Opus (KDE4, OpenSubtitles, Ubuntu, KDE4), Tilde (EMA) og ELRC (Hagstofa Íslands). Einu textasöfnin sem við söfnuðum sjálf, með skröpun (e. scraping) vefsíðna, eru EES (gagnagrunnur.ees.is), ESO (eso.org), Tatoeba (tatoeba.org) og íslenska útgáfa Biblíunnar (biblian.is). Við höfum búið til Python-skriftur sem safna nýjum gögnum frá ESO, EES og Tatoeba með skröpun, en vegna þess hve kóðinn er óstöðugur (minniháttar breytingar á

vefsíðunum verða til þess að kóðinn virkar ekki) og í sumum tilfellum flókinnar uppbyggingar gagnasafnanna sem við notum til að fylgjast með hvað hefur þegar verið sótt, teljum við ekki gerlegt að gefa út kóðann fyrir söfnun texta. Pípa fyrir samröðun og síun er aðgengileg á GitHub: https://github.com/stofnun-arna-magnussonar/align-filter.

V2b – Val og síun gagna fyrir bakþýðingu

Skýrsla: Miðeind
Aðilar: Miðeind

Varða 3 – lýsing

Tilgangur
Að afhenda stóra samraðaða tilbúna samhliða málheild fyrir beggja átta þýðingar milli íslensku og ensku, til notkunar í viðbótarþjálfun vélrænna þýðingarkerfa. Þessi málheild verður byggð á bakþýddum einmála textum, til að vega á móti skorti á samhliða ensk-íslenskum málheildum.

M2
- Bráðabirgðalíkan tvíátta þýðinga þjálfað á síuðum bakþýddum textum.
- Bráðabirgðarannsókn á blöndunarhlutföllum valinna sviða þýðinga í þýðingarverkefninu.

M3
- Skýrsla um síunar- og valaðferðir.
- Tilbúin samhliða málheild útgefin samkvæmt stöðlum CLARIN.
- Kóði útgefinn samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins.

Afurðir

Öllum markmiðum var náð.
Tilbúna samhliða málheildin sem var búin til innan þessa undirverkefnis er aðgengileg á CLARIN http://hdl.handle.net/20.500.12537/70.

Kóði fyrir þjálfun og smíði er hluti af GreynirT2T búntinu, sjá hirslu https://github.com/mideind/GreynirT2T og http://hdl.handle.net/20.500.12537/71 á CLARIN.
Verkþátturinn heldur áfram á öðru ári verkefnisins.

Skýrsla

Aðferðafræði
Sem hluti af undirverkefni V3 var þýðingarkerfi grunnlínu valið og þjálfað á takmarkaðri samhliða málheild. Edunov o.fl. (2018)[20] náðu bestu niðurstöðum með því að nota tiltækt kerfi við þýðingu einmála málheilda og innleiða svo vélþýddu gögnin sem *gervi*-þjálfunargögn (e. *synthetic* training data). Við beitum sömu aðferðafræði og þýðum íslenskar og enskar málheildir í þessum tilgangi, sem taldar eru upp hér fyrir neðan.

Að bæta gervigögnum gögnum við hin raunverulegu samhliða þjálfunargögn víkkar séða málfræði og róf formgerða tungumáls, sem skilar betra þýðingarlíkani. Í þjálfunarferli er mikið stærri tilbúinni málheild blandað við raunverulegu málheildina, en hlutfallinu er lýst neðar. Hins vegar, þó að heildarhlutfall raunverulegra gagna sé lægra en tilbúinna gagna, sést hvert setningapar oftar af netinu en hvert tilbúið par. Þetta er vegna þess að oftar er rennt í gegnum raunverulegu málheildina við þjálfun en þá tilbúnu, vegna þess að sú síðarnefnda er mikið stærri en sú fyrri.

[20] Understanding Back-Translation at Scale

Ferli íslensk-ensku bakþýðingarinnar hefur verið ítrað fjölmörgum sinnum, svo að tilbúna samhliða málheildin sem þetta leiðir af sér hefur verið sköpuð af líkönum sem voru einnig þjálfuð með tilbúnum samhliða gögnum. Við vitnum til líkans sem hefur ekki verið þjálfað með tilbúnum gögnum sem ítrun 0. Tilbúnu ensku gögnin eru afurð líkans ítrunar 1 og tilbúnu íslensku gögnin eru afurð líkans ítrunar 2[21]. Augljóslega minnkar árangur með frekari ítrunum.

Þýðing og síun

Einmála málheildirnar sem voru þýddar fyrir þetta undirverkefni eru taldar upp í töflunni fyrir neðan. Íslensku málheildirnar voru fengnar frá Risamálheildinni (sjá undirverkefni G1), Europarl var fengin frá CLARIN, og Newscrawl var fengin frá WMT[22].

Upprunatungumál	**Nafn**	**Stærð** í setningum (notuðum)
Íslenska	Dómsúrskurðir	1,8m
Íslenska	Hæstaréttarúrskurðir	2m
Íslenska	Lög (Alþingislög)	814k
Íslenska	Vísindavefurinn	268k
Íslenska	Wikipedia	405k
Íslenska	Alþingisræður	6,2m
Íslenska	Ýmislegt	350k
Íslenska	Dagblað (Morgunblaðið)	13,6m
Íslenska	Dagblað (Vísir)	4,9m
Íslenska	Útvarpsviðtöl og útvarpsfréttir (Rás 1 og Rás 2)	1m
Íslenska	**Samtals**	**31,3m**
Enska	Newscrawl	33,4m
Enska	Wikipedia	9,3m
Enska	Europarl	2,0m
Enska	**Samtals**	**44,7m**
	Fjöldi setningapara í bakþýddri tilbúinni málheild (eftir síun)	**76m**

Síun gagnanna, eftir þýðingu, er gerð í samræmi við (Pinnis et. al 2018)[23] eins og fyrir raunveruleg gögn (sjá einnig skýrslu M3 fyrir undirverkefni V3).
Í síunarferli voru um 20% íslensku gagnanna fjarlægð, og 10% ensku frumgagnanna.

---

[21] Við gerðum 4 ítranir af *þjálfa-svo-skapa*, þ.e. við bjuggum til EN_0, IS_1, EN_1, IS_2. Meginmarkmiðið var að hvor endanlegra málheildanna tveggja væru búnar til af líkani með þjálfunargögnum blönduðum við *úttak líkans sem hefur séð tilbúin gögn*. IS_1 tilbúnu gögnin voru ekki sköpuð með slíku líkani og eru því ekki notuð sem endanlega afurð.

[22] Workshop on Machine Translation

[23] Pinnis, M.: Tilde's Parallel Corpus Filtering Methods for WMT 2018. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers.* Brussels, Belgium (2018)

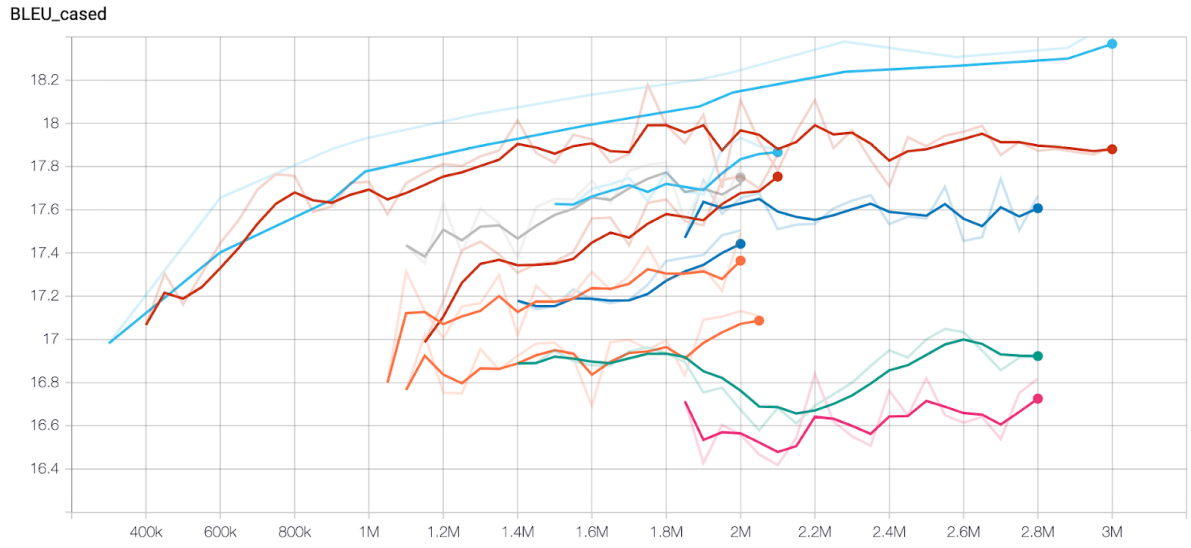
Niðurstöður

Frumheimildir fyrir bakþýðingu voru valdar með það að markmiði að bæta gæði þýðinga fyrir fréttir og reglugerðir Evrópska efnahagssvæðisins (EES). Þær eru valdar meðal annars til að passa við fyrirhugað samvinnuverkefni við Þýðingamiðstöð utanríkisráðuneytisins sem hluti af vörðum annars árs.
Rannsókn á bestu blöndunarhlutföllum var gerð og fjölmörg líkön þjálfuð til að meta hversu viðkvæm uppbyggingin er fyrir mismunandi stig tilbúinna þjálfunargagna.

Við mátum eftirfarandi:

a) Hvað blöndunarhlutfall er best að nota?[24]
b) Bætir það árangur kerfis ef líkaninu er leyft að vita hvaða pör eru tilbúin eða raunveruleg á meðan þjálfun stendur? (mörkuð eða ómörkuð gervigögn).
c) Hver er besta aðferðin við útdrátt þýðinga í neti þjálfuðu með blöndu tilbúinna og raunverulegra gagna? (úrtak, geislaleit (e. beam search) eða geislaleit með suði (e. beam search with noise)).
d) Afköst þýðinga, í BLEU-einkunnum, bæði á gögnum innan sviðs og utan sviðs (prófunargögn utan sviðs eru fengin úr ágripum ritgerða og frá textum Votta Jehóva, útgefnum á mörgum tungumálum).

Bestu niðurstöður, eins og mælast með BLEU á gögnum utan sviðs, fengust með 33% raunverulegum + 67% mörkuðum tilbúnum setningapörum, með notkun geislaleitar með suði. Þetta er efsta, ljósbláa línan í línuritinu fyrir neðan. Allar keyrslur eru keyrðar á Tensor2Tensor í Tensorflow, með Transformer grunn-arkitektúr.
Árangur í þýðingum ritgerðarágripanna er talinn mest lýsandi fyrir *downstream* niðurstöður vegna almenns og flókins eðlis þeirra. Málheild Votta Jehóva er svipuð Biblíunni að sumu leyti, sem er innan sviðs, en er mun frjálslegri í eðli sínu og inniheldur texta ýmissa „mjúkra“ (létt-trúarlegra) umræðuefna.



[24] Prófuð hlutföll voru 0,15, 0,33, 0,5, og 0,67 sem hlutföll af raunverulegum-til-tilbúinna setningapara.

Þróun BLEU-einkunna netsins við mismunandi þjálfunarkeyrslur í áttina en→is, reiknað á held-out utan óðals samhliða prófunarmálheild.[25]

V3 – Grunnlína í vélrænum þýðingum

Skýrsla: Miðeind**,** Háskólinn í Reykjavík
Aðilar: Miðeind, Háskólinn í Reykjavík

Varða 3 – lýsing

Tilgangur
Að afhenda vélþýðingarkerfi sem getur verið grunnur að tóli fyrir þýðendur til að þýða milli ensku og íslensku (í báðar áttir). Kerfið verður hannað til að þýða texta á ákveðnu sviði, eftir því hvaða málheildarþjálfun er í boði, auk þeirra sem verða til á meðan máltækniverkefnið stendur yfir.

M2

- Prófunar- og þróunarsett eru leiðrétt og mörkuð.
- Skýrsla um skref for- og eftirvinnslu texta fyrir hvert kerfi afhent.

M3

- Öll kerfi hafa verið prófuð á endanlegum prófunarsettum, með notkun for- og eftirvinnsluskrefa sem eiga við hvert kerfi.
- Skýrsla um styrki og veikleika hvers kerfis afhend, ásamt tillögu um áframhaldandi þróun.
- Áætlun um prófunarsvítu fyrir vélþýðingar fyrir íslensku.

Afurðir

Markmiðum hefur verið náð, með örfáum minniháttar undantekningum sem frekar er skýrt frá fyrir neðan. Þýðingarkerfi milli íslensku og ensku sem virkar vel er aðgengilegt og í notkun sem vefþjónn með REST API á www.velthyding.is, sem styður einnig notendaviðmótið sem þróað var í undirverkefni V5. Í núverandi mynd er kerfið oft samkeppnishæft við t.d. Google Translate, sérstaklega þegar þýtt er úr ensku yfir á íslensku.

Kerfið sem hefur verið afhent hefur verið stækkað með bakþýddum tilbúnum málheildum í samræmi við niðurstöður undirverkefnis V2b, sem bætir afköst þýðinga umtalsvert og einnig orðaforða og málfimi (e. fluency) kerfisins.

Þau þrjú þýðingarkerfi sem komu til greina fyrir M1 hafa þegar verið metin með prófunarsetti sem var búið til fyrir grein sem var gefin út á árinu (Jónsson o.fl., 2020)[26]. ParIce prófunar- og þróunarsettin, sem gefin voru út fyrir vörðu M2 í þessu undirverkefni, þörfnuðust breytinga til að

[25] Prófunarmálheildin fyrir þetta línurit var fengin úr samröðuðum ágripum ritgerða sem sóttar voru á Skemmuna. https://skemman.is, sem er sameiginlegt verkefni íslenskra háskólastofnanna þar sem ritgerðir og rannsóknir eru gefnar út.
[26] Experimenting with Different Machine Translation Models in Medium-Resource Settings, *23rd International Conference on Text, Speech and Dialogue* 2020.

hámarka nákvæmni mælinga[27], sem hindra samanburð héðan í frá við núverandi útgáfu. Við töldum ótímabært að halda öllum kerfum aðeins til að ná áreiðanlegum BLEU-mælingum á nýlega útgefnu prófunarsetti, og höldum því áfram að greina frá BLEU-einkunnum á prófunarsettinu sem notað var í greininni hér. Grunnlínur fyrir endanlega ParIce prófunarsettið verða skilgreindar fyrir M4 vörðuna á öðru ári.

Samkvæmt endurskoðuðum vörðum fyrir verkefnið var áætlað að greina handvirkt 500 setningar þýddar af með þrem mismunandi þýðingarkerfum: tölfræðilíkani Moses), BiLSTM ítruðu endurkvæmnistauganeti (OpenNMT), og *attention-based* tauganeti (Transformer). Eftir nokkra umhugsun og í ljósi augljósra yfirburða athyglis-líkansins (Transformer) var ákveðið, í samráði við verkefnisstjóra, að takmarka handvirka greiningu við 200 setningar.

Að undanskildum ofangreindum undantekningum, teljum við að öllum markmiðum hafi verið náð. Kóði og tenglar á Docker-gáma og þýðingarlíkan grunnlínu eru aðgengilegir á CLARIN http://hdl.handle.net/20.500.12537/72 og á GitHub https://github.com/mideind/nnserver.

Verkþættinum er lokið við M3 og grunnur lagður að meginþróun opins vélþýðingakerfis fyrir íslensku á komandi verkefnaárum.

Skýrsla

Við bárum saman kerfin þrjú með notkun bæði sjálfvirkra mælinga og mannlegs mats. Handvirk villugreining var einnig gerð á sömu gögnum. Við lýsum uppsetningu síunarpípu hér fyrir neðan. Nafnakennsl voru gerð á hreinsuðu samhliða gögnum. Við komumst að þeirri niðurstöðu að bestu aðferðirnar fyrir ensku-þýsku (staðallinn í NMT rannsóknum í reynd) hæfa að mestu okkar þörfum. Við lýsum mögulegum leiðum fyrir prófunarsvítu fyrir þýðingarlíkön til og frá íslensku, sérstaklega fyrir tungumálaparið íslenska-enska.

Niðurstöður á samanburði kerfa

Sem hluti af vörðu M2 voru þrjú vélþýðingarkerfi metin. Transformer-kerfi sem var þjálfað með notkun Tensor2Tensor og Tensorflow[28], BiLSTM-líkan sem notar OpenNMT, og Moses, tölfræðikerfi aðlagað fyrir íslensku. Samanburð á BLEU-einkunnum kerfanna má sjá í töflunni hér fyrir neðan. Taka skal fram að eftirfarandi tölfræði er fyrir líkön sem nota ekki bakþýdd gögn við þjálfun. Prófunarsettið sem var notað var bráðabirgðahreinsað sett búið til fyrir (Jónsson, 2020) en ekki það sem gefið var út á CLARIN sem hluti af vörðu M2 fyrir undirverkefni V2 og V3.

Líkan	EES[29]	EMA[30]	OpenSubtitles[31]	Örmeðaltal
BiLSTM	38,68	41,60	23,32	38,12
Moses	49,70	54,93	26,11	48,60
Transformer	**56,31**	**58,37**	**34,71**	**54,71**

Í samræmi við almenna reynslu af meðalstórum og stórum þýðingarlíkönum, fást bestar niðurstöður (BLEU-einkunn) fyrir Transformer-líkanið yfir önnur prófunarsett.

[27] Við höfum meðal annars fundið lítilsháttar smit setningapara úr þróunarsettinu yfir í prófunarsettið og prófunarsettið inniheldur nokkur pör sem er ekki hægt að búast við vélþýðingarkerfi—eða manneskja—myndi.

[28] Við höfum síðan fært okkur yfir í Fairseq og Pytorch, sem hafa tekið yfir sem staðlar í greininni í reynd.

[29] Reglugerðir Evrópska Efnahagssvæðisins.

[30] Gagnasafn Lyfjastofnunar Evrópu (e. European Medicines Agency) fyrir lyfjaupplýsingar.

[31] Stórt lýðvirkjunardrifið safn skjátexta frá kvikmyndum og sjónvarpsþáttum.

Eins og er lýst nánar í (Jónsson, 2020), voru málfimi og nægð einnig metin af manneskjum. Niðurstöður, sem sjá má í töflunni fyrir neðan, voru í samræmi við ofangreint. Google Translate náði (sem kemur ef til vill ekki á óvart) betri árangri á prófunarsetti utan óðals, en á prófunarsetti innan óðals náði Transformer-líkan okkar bestum árangri.

Prófunarsett	Líkan	Málfimi	Nægð
Innan sviðs	BiLSTM	2,49	2,01
	Moses	3,64	3,84
	Transformer	**4,30**	**4,33**
	Google Translate	3,80	4,16
Utan sviðs	BiLSTM	1,85	1,30
	Moses	2,54	2,32
	Transformer	3,20	2,86
	Google Translate	**3,40**	**3,80**

Fyrir ítarlegri skýrslu um niðurstöðurnar að ofan og skýrslu um styrki og veikleika hvers kerfis, sjá (Jónsson, 2020).

Einnig greindum við allar tilteknar þýðingarvillur hvers kerfis, byggt á villuskema frá Lommel o.fl.[32]. Við vitnum í það rit fyrir flokkun og lýsingar villuflokka.

Við notuðum flokkana *nákvæmni (e. accuracy)* og *málfimi (e. fluency)* á setningarstigi þegar enginn annar flokkur var við hæfi eða þegar matið var óvisst. Flokkurinn óskiljanlegt (e. *unintelligible)* var einnig notaður á setningarstigi ef engir aðrir flokkar pössuðu við tiltekna setningu, sem þýðir að setningin var annaðhvort óskýr hvað varðar merkingar- og setningafræði, eða að setningin hefði þurft of margar leiðréttingar (15+) til að verða samþykkt. Allir aðrir flokkar voru notaðir á orðastigi.

Hver dálkur sýnir niðurstöðurnar úr hinum 200 greindu setningum. Tf=Transformer, Mo=Moses, biL=biLSTM.

[32] Lommel, Arle, o.fl. 2014. „Using a new analytic measure for the annotation and analysis of MT errors on real data.“ *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*.

	EN → IS Innan sviðs			EN → IS Utan sviðs			IS → EN Innan sviðs			IS → EN Utan sviðs		
	Tf	Mo	biL	Tf	Mo	biL	Tf	Mo	biL	Tf	Mo	biL
Nákvæmni	1	1	0	0	1	4	45	1	2	2	1	0
- Orði/orðum sleppt	10	21	15	29	37	23	10	18	21	23	33	27
- Röng þýðing	19	27	30	98	100	60	22	42	60	149	75	114
- Viðbót	17	32	231	8	21	153	16	13	146	18	33	98
- Engin þýðing	1	4	7	5	23	1	0	35	0	0	65	1
Málfimi	1	5	2	5	4	3	2	6	0	7	4	1
- Innsláttur	28	11	21	7	7	8	0	6	23	12	11	3
- Málfræði	0	2	3	2	2	4	0	7	3	7	18	9
-- Kerfisorð	5	7	2	21	9	6	5	5	10	15	12	13
-- Orðaröð	3	4	1	14	24	6	2	16	3	14	41	3
-- Orðmynd	3	36	11	27	68	31	2	12	6	12	8	15
- Stafsetning	2	2	2	0	2	1	1	2	0	0	0	0
Óskilgreint	2	9	56	7	18	57	1	9	29	12	34	55
Samtals	**92**	**161**	**381**	**223**	**316**	**357**	**106**	**172**	**303**	**271**	**335**	**339**

Við vitnum í Lommel (2014) fyrir lýsingu á merkingarskema (e. labeling scheme). Þessar niðurstöður eru í samræmi við BLEU-einkunnir og við niðurstöður um mannlega málfimi og nægð (e. adequacy). Við tökum fram að Transformer villurnar snúa aðallega að varðveislu merkingar, ekki að setningafræði eða málfimi. Einnig var Transformer-líkanið sem notað var til villugreiningar ekki (enn) þjálfað með tilbúinni, bakþýddri málheild (sjá undirverkefni V2b) sem virðist bæta orðaforða og sérstaklega málfimi.

Niðurstöður alls mats (BLEU-einkunnir, mannlegrar málfimi (e. human fluency) og nægðar, og villugreiningar) benda sterklega til þess að áframhaldandi þróun ætti að vera byggð á Transformer-kerfinu. Þessi niðurstaða er í samræmi við það sem fjallað hefur verið um í nýlegum fræðigreinum.

For- og eftirvinnsla

Fyrsta forvinnsla tiltækra samhliða málheilda var gerð í samræmi við Pinnis o.fl. (2018)[33]. Þetta fjarlægði u.þ.b. helming upprunalegu 3,6 milljón setningaparanna í ParIce samhliða málheildinni.

a) Tómar setningar fjarlægðar
b) Nákvæmlega eins eða nánast eins setningar fjarlægðar
c) Pör síuð eftir hlutfalli setningalengdar
d) Hámarks- og lágmarkslengdarsíun
e) Hámarkslengd tóka
f) Lágmarks meðallengd tóka
g) Hvítlisti bókstafa
h) Misræmi talna (uppruni og takmark ætti að innihalda sama fjölda raða)
i) Einstök pör eftir að tákn utan stafrófs voru fjarlægð

[33] Pinnis, M.: Tilde's Parallel Corpus Filtering Methods for WMT 2018. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers.* Brussels, Belgium (2018)

j) *Full case mismatching pairs removed*
k) Pör með brengluð tákn fjarlægð (kóðunarvillur eða aðgreiningarmerki vantar)
l) Sértæk íslensk OCR vandamál löguð, þ.e. *b* og *þ ruglað saman*
m) Stöðlun gæsalappa, áherslumerkja, bandstrika o.s.frv.
n) Línuskiptingar lagaðar (frá orðskiptingum með bandstriki við enda lína)

Við handvirkt mat þýðinga frá líkaninu reyndist tíðasta vandamálið vera beinþýðing margra nafneininga. Kerfið þarf mikið magn þjálfunardæma til að geta þýtt (eða hleypt í gegn, eins og stundum þarf) nöfnunum í nafneiningum. Nöfn Evrópskra stofnana þýðast vel (vegna málheildar reglugerða EES sem var notuð), en meirihluti íslenskra nafna þýðist ekki rétt frá íslenska yfir á ensku, þar sem þau verða gjarnan gölluð eða breytast í orðasalat.

Við keyrðum for- og eftirvinnslupípu (sjá undirverkefni **V3b**) með nafnakennslum, fínstilltum fyrir íslensk og ensk gögn í réttri röð með notkun BERT mállíkana[34]. Tölfræði nafneininga fyrir persónur má sjá í töflunni fyrir neðan. Tölfræðin nær aðeins til hreinsaðra gagna.

Uppruni	**Samtals bútar[35]**	**Bútar þar sem mannanöfn voru merkt**
EMA	276.282	688
Open subtitles	498.845	55.449
Biblían	25.832	7.782
EES	682.969	394
Tatoeba	4.987	433

Flest nöfnin sem fundust í samhliða gögnum eru algeng ensk nöfn eins og John eða Paul (jafnvel þegar Biblían er undanskilin), með fáum fullum nöfnum og nær engum íslenskum nöfnum – sem við var að búast úr þessum heimildum.

Gögn með merktum nafneiningum (e. named entity annotated data) gera okkur kleift að draga út aðskilda samhliða málheild með staðgenglum eininga (e. entity placeholders) sem passa bæði við upprunamál og markmál, og auðvelda sjálfvirka myndun aukinna þjálfunargagna með því að skipta út íslenskum nöfnum og samsvarandi ensku nöfnum fyrir staðgengla eininganna. Við gerum þessi gögn aðgengileg á CLARIN[36].

Áframhaldandi þróun

Á seinastliðnum árum hafa attention-byggð þýðingarlíkön sem innleiða einnig mállíkön (sérstaklega mynstruðu mállíkön og BERT-afleiður) skilað verulegum endurbótum á þeim þýðingarlíkönum sem miðað er við í dag (e. state-of-the-art), annaðhvort með því að innleiða markmið mállíkana í þjálfuninni eða með því að nota greypingar úr tiltæku líkani, með þýðingu sem eina markmiðið.[37]

[34] Fyrir ensku notuðum við stórt BERT-líkan frá Hugging Face, fínstillt á CONLL NER-mörkun, sem er hluti af Transformer-pakka þeirra. Fyrir íslensku var nýstárlegur BERT-byggður NER-markari notaður, sem verður gefinn út á næstu mánuðum.

[35] Bútar eru almennt setningar en geta verið styttri eða lengri eftir samröðun.

[36] http://hdl.handle.net/20.500.12537/74

[37] Sjá t.d. *MASS: Masked Sequence to Sequence Pre-training for Language Generation* (Song o.fl. 2019) og *Incorporating BERT into Neural Machine Translation* (Zhu o.fl. 2020)

Þar sem raunverulegar samhliða þýðingarmálheildir eru af skornum skammti fyrir íslensku en meira en nóg er til af einmála gögnum til að þjálfa betri mállíkön, lofa þessar aðferðir góðu fyrir betri þýðingar, sérstaklega þegar bakþýðingum verður bætt við (sjá undirverkefni V2b) á öðru ári máltækniverkefnisins, eins og hefur verið ákveðið.

Áætlun um prófunarsvítu fyrir íslenska vélþýðingu

Við vitnum í hlutann að ofan varðandi handvirkt mat.
Við mælum ekki með matsþáttum sem ná aðeins til sértækra eiginleika íslensk máls, frekar að almennum matsverkefnum verði beitt á íslensku, líkt og þeim sem þegar eru til fyrir ensku. Að fá undirstöðueiginleika málfræðinnar rétta er ekki aðalvandamálið við nútíma þýðingarkerfi (sem nýta meðalstór- eða stór málföng); en varðveisla merkingar getur reynst erfið. Merkingafræðileg greining og samanburður slíkra greininga þvert á tungumál er því nauðsynlegur. Í þessu sambandi leggjum við áherslu á skort á gögnum og kerfum sem þarf til að meta réttmæti þýðinga. Eftirfarandi eru tillögur okkar að viðmiðunarmælingum á gæðum og afköstum þýðingarkerfa milli ensku og íslensku sem birtast ekki nógu skýrt í BLEU-einkunnum. Tillögurnar eru ekki í neinni sérstakri röð, fyrir utan það að orðmyndunin er talin síður mikilvægari en hinar.

Tillögur að aðferðum til að meta gæði þýðinga

Lausn samvísana, bæði innan og milli setninga. ***It*** í setningunni *The trophy does not fit in the brown suitcase because **it** is too small* ætti, , að þýðast í kvenkynsorðið *hún* á íslensku (*taska*), þar sem ***it*** vísar til töskunnar (*the suitcase*), á meðan karlkynsorðið *hann* myndi vísa til bikarsins (*the trophy*). Lágmark nokkurra þúsunda viðeigandi þjálfunardæma þarf fyrir þetta verk.

Einræðing orðamerkingar. Setningin *Maðurinn gaf í þegar hann sá grænt ljós* myndi beinþýðast sem *The man **gave in** when he saw a green light*, en rétt þýðing er *The man **accelerated*** [or ***stepped on the gas***] *when he saw the green light*. Þetta krefst réttrar einræðingar orðamerkingar til að greina merkinguna í samhenginu *gefa* og *grænt ljós*. Best væri að geta metið færni þýðingakerfis við einræðingu orðamerkinga gagngert. Þetta myndi kveða á um mörkun íslenskra texta með orðamerkingum, helst með sömu skrá orðamerkinga og er í boði fyrir mörkun ensku. Til dæmis nokkur þúsund línur íslensks texta mörkuð með orðamerkingum með notkun orðamerkingaskrár frá Princeton WordNet[38]. Þessar setningar ættu að innihalda margræða orðasafnshluta sem fólk getur ráðið í eftir í samhengi.

Orðtök/Frasar. Setningin *Þegar lögreglumennina bar að garði var slökkt á ljósunum*, yrði bókstaflega þýdd *When the policemen came to the yard, the lights were off*, ætti rétt að þýðast sem *When the policemen arrived, the lights were off*. Þetta er nokkuð sambærilegt hlutanum að ofan, nema hér viljum við þjálfunarsetningar þar sem orðtök og föst orðasambönd eru notuð. Við leggjum til sköpun tveggja samhliða málheilda setninga; eina þar sem enskar setningar innihalda orðtök eða frasa þýdda yfir á íslensku, og aðra þar sem íslenska hliðin inniheldur orðtök. Slík prófun gerir gagngert mat á hæfni þýðingarkerfa þegar kemur að orðtökum og föstum orðasamböndum.

Þol nafna (e. name robustness). Þetta felur í sér tól og málheildir til að meta þýðingarhæfni fyrir margar gerðir nafneininga eins og nafna persóna, vara, samtaka, hópa, staða, eða önnur sérnöfn, o.s.frv. Þetta kveður á um gerð samhliða málheildar merktra nafneininga (e. NER annotated) þar sem hverri gerð eininga eru gerð skil, helst með nokkur hundruð dæmum fyrir hverja gerð.[39]

---

[38] WordNet – Merkingarfræðilegur gagnagrunnur fyrir ensku. Sjá https://wordnet.princeton.edu/

[39] Athuga skal að sú sem við veitum sem hluta af þessari vörðu er langt frá því að vera ákjósanleg í þessum tilgangi þar sem hún er tölvugerð og úr þjálfunargögnunum, svo ekki sé minnst á skort nafna í gögnunum.

Kyn orða fyrir undirskilyrtar þýðingar á íslensku. Venjuleg nafnorð eiga sér málfræðilegt kyn í íslensku, en ekki er alltaf einfalt að greina kyn fornafns sem skal nota til að vísa til þeirra nafnorða. Fornöfn skulu valin skv. samlögun nema þau séu beinlínis bundin. Stundum koma einnig upp aðstæður við þýðingu á ensku þar sem kynjuð fornöfn halda ranglega kyni sínu í gegnum þýðingu. Sjá til dæmis: *Hann er nokkuð frábrugðinn amerískum fótbolta, þó að hann minni mjög á hann* sem Google Translate skilar sem ***He** is quite different from American football, although he is very reminiscent of him*. En karlkyns fornafnið *hann* vísar til íþróttar og ætti því að vera ***it*** á ensku. Til að leysa þetta mælum við með gerð samhliða málheildar með merktum fornöfnum og viðeigandi þýðingum. Nokkur hundruð setningar fyrir hvora átt þýðingar ættu að nægja til að meta færni kerfisins.

Gæðamat. Sjálfvirkt vélrænt mat þýðingargæða þarfnast frekari mannmarkaðra gagna fyrir málfimi og nægð. Slíkt gangasett hefur einkum tvenns konar notagildi: 1) Vélþýðingarkerfi ætti að geta gefið gilda tölfræði þýðinga sem endurspegla val mennskra merkjara. 2) Þjálfun kerfis sem metur gæði þýðinga gagngert. Við mælum með gerð tvímálamálheildar þar sem hver upprunasetning hefur minnst tvær mögulegar þýðingar og ein þeirra er frekar valin af mennskum merkjara en hin(ar).

Orðmyndun. Ef gefið er að líkan skilji sett róta, getur það skilið samsett orð mynduð úr þeim rótum? Dæmi um þetta er setningin: *In the news today, it was stated that there are now 10 **domestic infections***, þar sem *domestic infections* ætti helst að þýða með samsetta orðinu *innanlandssmit*. Fyrir hina áttina: ***Kíkisglerið** var ógagnsætt svo ég sá ekki neitt* ætti að þýðast sem *The lens on the binoculars was opaque so I didn't see anything,* frekar en *The binoculars were opaque [...]*. Við mælum með gerð málheildar þar sem íslenska hliðin inniheldur samsett orð, með leyfðum mögulegum þýðingum þess orðs á hlið markmálsins. Markmiðshliðin ætti á þýðingarstigi að innihalda einn þessara möguleika. Við mælum einnig með gagnstæðri málheild þar sem nokkuð sterk hefð eða tilhneiging fyrir því að þýða ákveðið enskt samsett orð eða frasa sem eitt orð í íslensku.

Við viljum einnig benda á að utan þessara viðmiðunarmælinga að ofan, væru **meiri prófunargögn** frá víðari sviðum en voru aðgengileg við þjálfun mjög nytsamleg, sérstaklega hvað varðar formfasta hluti eins og samræður.

V3b – Textavinnsla (for- og eftirvinnsla)

Skýrsla: Miðeind
Aðilar: Miðeind

Varða 3 – lýsing

Tilgangur
Að hafa sett aðferða fyrir vinnslu nafneininga (NE). Nafneiningar eru fundnar og paraðar innan uppruna- og markmiðssetningaparanna í samhliða málheildum og skipt út fyrir staðgengla með upplýsingum um fall og tölu. Skiptin NE-fyrir-staðgengil eru innleidd í inntakið og skiptin staðgengill-fyrir-NE í úttakspípum þýðingarkerfisins.

M3

- Tól fyrir for- og eftirvinnslu nafneininga í inn- og úttaki vélþýðinga.

Afurðir

Öllum markmiðum var náð.

Nafnakennslamerkt samhliða þjálfunarmálheild er aðgengileg hjá CLARIN: http://hdl.handle.net/20.500.12537/74.[40]

Skýrsla

Tólið samanstendur af eftirfarandi meginhlutum:

Forvinnsla þjálfunarmálheilda fyrir uppruna- og markmiðstungumála
a) Nafnakennsl
b) Skipti á staðgenglum fyrir nafneiningar, með upplýsingum um fall og tölu

Eftirvinnsla paraðra gagna
a) Nafneiningar mátaðar saman
b) Síun para og tölfræði útbúin

Tólið er forritað til að geta notað einingahluta, en reglubyggt nafnkennslaforrit (e. NER processor) og málfræðilegur markari (með notkun þáttara Greynis, sjá undirverkefni I5) eru sjálfgefnir. Sýslari fyrir þolnari/harðgerðari tauganafnkennsl (e. neural NER) og líkön málfræðilegrar mörkunar eru til staðar og nothæf ef þeim er veitt aðgengi að þjálfuðum *model checkpoints* og *architecture classes*.

Mælikvarðinn sem valinn er til að tengja nafneiningar milli tungumála er Jaro-Winkler fjarlægðin[41].

[40] Kóðinn (ásamt skjölun) er aðgengilegur á GitHub í hirslunni https://github.com/mideind/GreynirSeq/tree/develop/src/greynirseq/ner
[41] https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

Dæmi og flæði pípu

Eftirfarandi er dæmi um pörun með notkun íslensks BERT-líkans, sérstilltu sem NE-markara og BERT-líkan frá Hugging Face[42] sérstillt fyrir ensku.

1. Eftirfarandi setningapör eru gefin:
    - *At the public library reading Thus Spake Zarathustra by Friedrich Nietzsche.*
    - *Á bókasafni að lesa Svo mælti Zaraþústra eftir Friedrich Nietzsche.*
2. Setningarnar eru markaðar með nafneiningum (e. NER-tagged)
    - At the public library reading **[**$_{\text{MISC}}$ Thus Spake Zarathustra**]** by **[**$_{\text{PERSON}}$ Friedrich Nietzsche**]**.
    - Á bókasafni að lesa **[**$_{\text{MISC}}$ Svo mælti Zaraþústra**]** eftir **[**$_{\text{PERSON}}$ Friedrich Nietzsche**]**.
3. *Friedrich Nietzsche* passar fullkomlega í báðum setningum og hefur Jaro-Winkler fjarlægð 0.
4. Við fáum út að tákngreindu spannirnar tvær, **9:11:Person** og **8:10:Person,** séu samstæðar.
5. Setningarnar eru málfræðilega markaðar til að draga út upplýsingar um fall og tölu, og samstæðu nafnkennsla-spannirnar (e. NER spans) eru merktar.
    - At the public library reading Thus Spake Zarathustra by **[**$_{\text{PERSON.Noun.Sing.}}$ Friedrich Nietzsche**]**.
    - Á bókasafni að lesa Svo mælti Zaraþústra eftir **[**$_{\text{PERSON.Noun.Sing.Acc}}$ Friedrich Nietzsche**]**.

Gögnin sem fást má svo nota til að mynda tilbúið setningapar sjálfkrafa, með réttu falli og tölu fyrir hverja nafneiningu, til að auðga þjálfunargögn með fjölbreyttara nafnasetti. T.d.:

- At the public library reading Thus Spake Zarathustra by **[**$_{\text{PERSON.Noun.Sing.}}$ Helga Jónsdóttir**]**.
- Á bókasafni að lesa Svo mælti Zaraþústra eftir **[**$_{\text{PERSON.Noun.Sing.Acc}}$ Helgu Jónsdóttur**]**.[43]

Aðrar einingar eins og *Thus spake Zarathustra*, merktar sem *MISC* í dæminu að ofan, mætti einnig innleiða í textavinnslupípu í vörðu seinna í verkefninu.

[42] Sjá https://huggingface.co/

[43] Athugið að nafnið hefur verið fallbeygt rétt í þolfalli.

V5 – Viðmót þýðingarvélar

Skýrsla: Miðeind
Aðilar: Miðeind

Tilgangur

Að afhenda vefþjón og notendaviðmót á netinu sem gerir þýðendum og öðrum notendum kleift að slá inn texta (beint, eða með því að senda inn hreinan texta (UTF-8) eða Microsoft Word skjöl, eða með því að kalla á HTTP/JSON byggð forritaskil) og fá hann þýddan með notkun bakenda í undirverkefni V3. Notendaviðmótið ætti að veita möguleika á samsíða samanburði úttaks kerfanna þriggja og ætti að geta sýnt þá upprunalegu setninguna sem samsvarar hverri markmiðssetningu, í flettiglugga og/eða með auðkenningu (í seinna verkefni (V5b) verður þetta vefviðmót bætt til að veita notendum möguleika til þess að velja og/eða slá inn aðrar/betri þýðingar fyrir hverja markmiðssetningu).

M2

- Veframendi vélþýðingar (vefviðmót notenda og hugbúnaður á vefþjóni) afhentur samkvæmt viðmiðunarreglum hugbúnaðarþróunar innan verkefnisins.

Afurðir

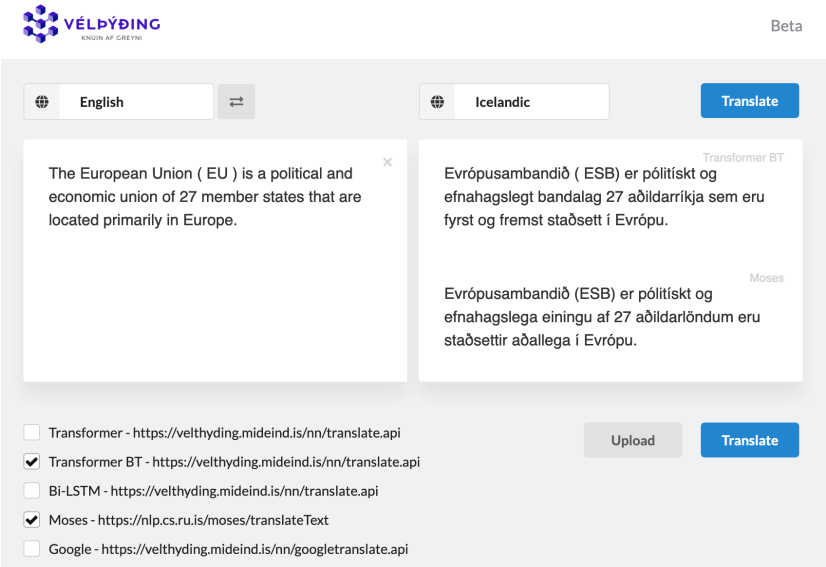
Öllum markmiðum var náð við vörðu M2 með opnum sýningarvef á velthyding.is. Athugið að þessi vefur hefur breyst síðan vörðu M2 var skilað, þar sem hann hefur verið uppfærður til að útiloka (nú) úrelta þýðingarbakenda sem nota Moses og BiLSTM líkönin. Fyrir vörðu M3 varð þetta undirverkefni grunnurinn að undirverkefni V5 (Lýðvirkjun).

Kóði vefsíðunnar er aðgengilegur á CLARIN https://repository.clarin.is/repository/xmlui/handle/20.500.12537/23 og á Github https://github.com/mideind/Velthyding.

Kóðinn er forritaður í React Javascript umhverfinu með notkun krækja og ES6, en fluttur yfir í venjulegt Javascript fyrir þjónustu. Engan sérstakan bakendaþjón þarf, aðeins vefþjón sem getur framreitt statískar skrár (e. serve static files). Hins vegar verður þýðingarbakendi að vera í gangi og aðgengilegur biðlara í samskipanlegri vefslóð.
Verkþátturinn heldur áfram á öðru ári verkefnisins.

Skýrsla

Þó að sýningarvefurinn sé að einhverju leyti merktur „Miðeind“, er sjálfgefin stilling í kóðanum að hafa slökkt á slíkum merkjum. Ef þurfa þykir má auðveldlega skipta út merkinu og síðufót, ásamt þýðingarbakenda vefslóðanna, í stillingarskrá biðlara.



Skjáskot af vefbiðlara í keyrslu á velthyding.is við vörðu M2.
Athugið að reitirnir til að velja þýðingarkerfi til samanburðar hafa verið fjarlægðir af vefsíðunni fyrir vörðu M3 í öðrum undirverkefnum vélþýðinga.

V5b – Lýðvirkjun

Skýrsla: Miðeind
Aðilar: Miðeind

Varða 3 – lýsing

Tilgangur
Að hanna þjónustu sem gerir notendum kleift að leiðrétta þýðingar sem sýndar eru í prófunarumhverfi notendaviðmóts og senda endurbætur beint inn í kerfið, sem verða að lokum innleiddar í þjálfunarmálheild.

M3

- Útfærsla endurgjafarkerfis
    - Þýðingar geymdar í bakenda
    - Hægt að leggja til betri þýðingar
    - Tillögur að öðrum þýðingum úr tauganeti sýnilegar

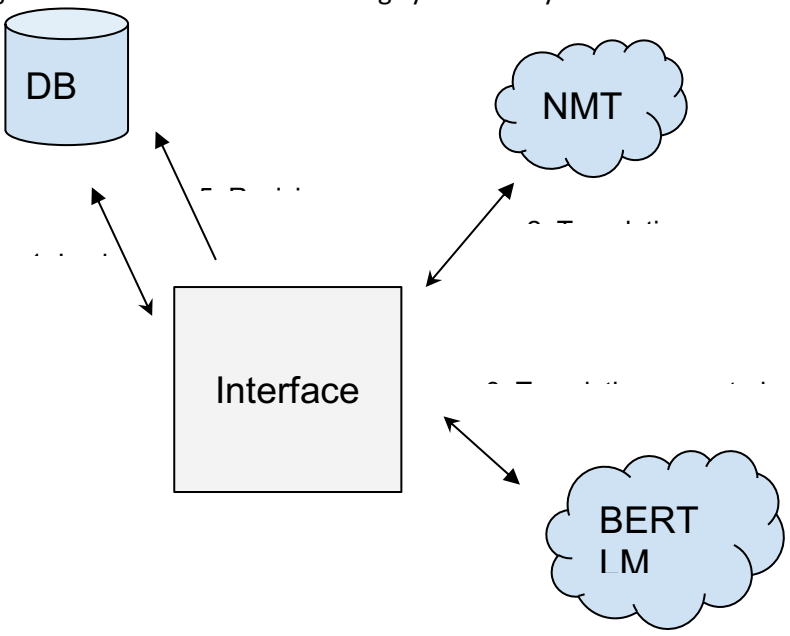
Afurðir

Öllum markmiðum var náð og sýningarkerfi er í gangi á www.velthyding.is. Taka skal fram að þetta kerfi tengir saman margvíslega flókna hluta og er aðeins mögulegt að afhenda sem vefsíðu að svo stöddu. Verkþátturinn heldur áfram á öðru ári verkefnisins.

Skýrsla

Kerfið er byggt ofan á tiltækt notendaviðmót úr undirverkefni V5, sem lauk sem hluta af M2. Viðbætur og breytingar á viðmótinu lúta aðallega að stuðningi fyrir tillögur og endurgjöf.

Kerfið tengir nú saman nokkra hluta eins og sýnt er hér fyrir neðan.



1. Notendur skrá sig (valkvætt) til að geyma, meðhöndla og rekja eigin þýðingar og breytingar.

2. Notandinn slær inn upprunatexta á annaðhvort íslensku eða ensku og smellir á takkann „Translate“ (þýða). Tauganetið sýnir notandanum grunnlínuþýðingu [t.d. *Let's go to the cinema tonight and watch the film*].

3. Notandinn getur smellt á orð í grunnlínuþýðingunni og fengið aðrar tillögur sem dregnar eru úr íslensku eða ensku mállíkani (BERT), eða breytt þýðingunni handvirkt innan setningarinnar. [Dæmi um tillögu: *theatre* í stað *cinema*] [Dæmi um leiðréttingu: *Let's go to the theatre...*]. Þetta er sýnt frekar í skjáskotum í lok kaflans.

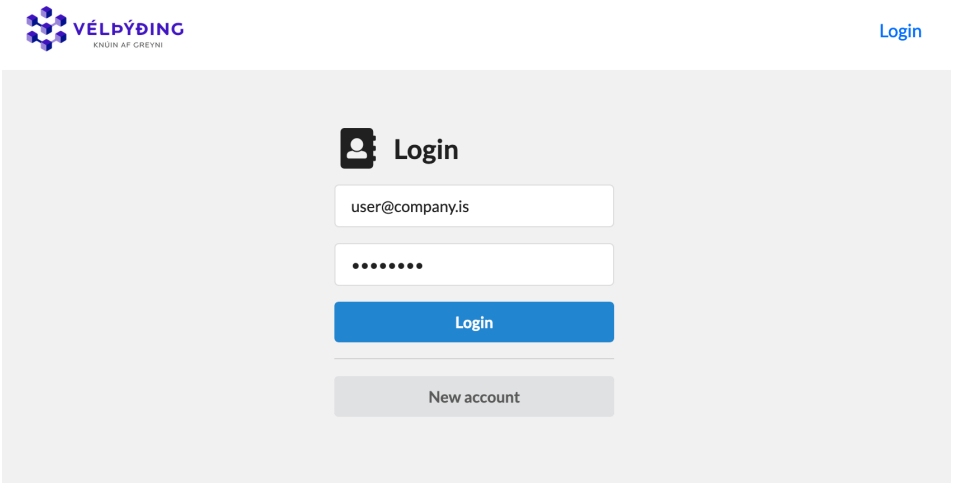
4. Ný þýðing er búin til, byggð á uppfærðu setningunni, með gefinni leiðréttingu/forskeyti. Tauganetið mun ekki breyta úttakinu fram að staðsetningu leiðréttingarinnar og að henni meðtalinni, en býr til nýjan enda á þýðinguna. [Dæmi um úttak: *Let's go to the theatre tonight and watch the* ***show*** – þar sem upprunalega orðinu *film* hefur verið skipt út fyrir *show* til að passa við samhengi *theatre* á móti *cinema*].

5. Upprunalegi textinn, grunnlínuþýðingin og leiðréttingin er geymd sem lína í gagnagrunninum og tengd við skráðan notanda, ef hann hefur skráð sig inn, til frekari athugunar í framtíðinni.

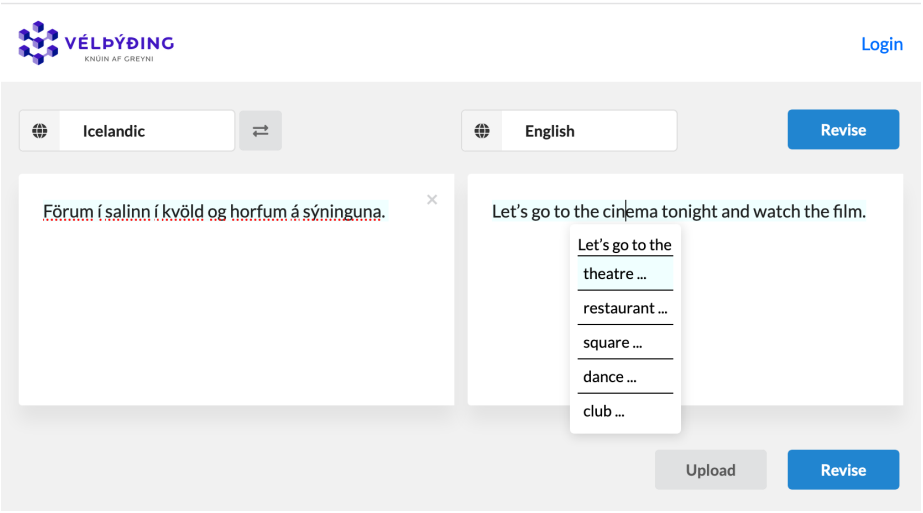
Nýting leiðréttinga þýðinga

Geymdar þýðingar og samsvarandi leiðréttingar má yfirfara handvirkt vegna algengra vandamála með úttak þýðinganna. Með þessum hætti má nota þá vitneskju sem fæst úr yfirferð leiðréttinga til að fínstilla þjálfunarstillingar eða sýna fram á skort á ákveðnum gerðum þjálfunargagna. Gögnin geta einnig verið grunnur að sniðmátum fyrir tilbúin gögn sem búa skal til fyrir þjálfun. Að lokum má einfaldlega bæta leiðréttum þýðingum í þjálfunarmálheild kerfisins fyrir næstu keyrslu þjálfunar; möguleiki á millistigi væri að bæta við slíkum þýðingum sem tilbúnum gögnum frekar en raunverulegum.

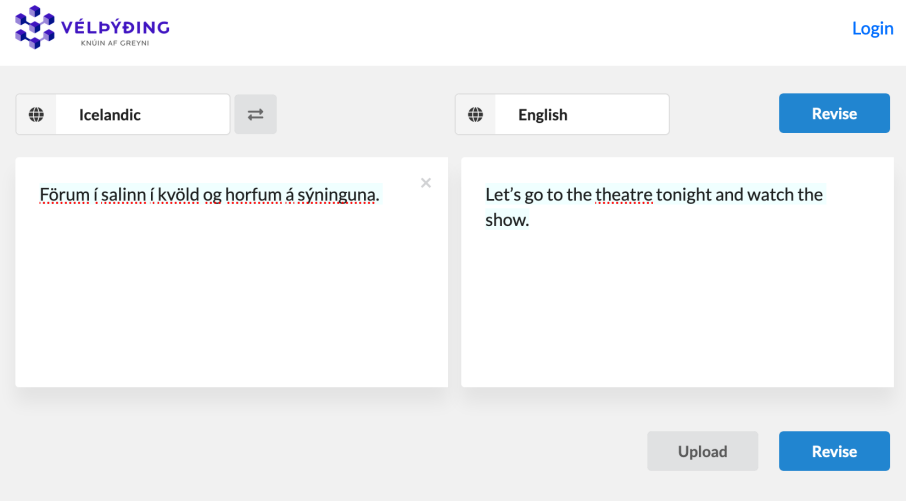
Hugsanlega mætti í framtíðinni þjálfa kerfi, sem er þekkt sem endurröðunarlíkan, til að velja betri kost tveggja þýðinga, þar sem ein væri hin upprunalega og önnur sú endurbætta. Þetta myndi hins vegar krefjast þúsunda dæma til viðbótar (með lýðvirkjun) í upphafi. Kerfið sem út úr því kæmi mætti þá nota til að meta gæði þýðinga á einkunnakvarða.
Þetta gæti leitt til þróunar tauganetsvélþýðingarkerfis sem er fínstillt með því að nota styrkingarnám (e. reinforcement learning), þar sem gefin einkunn úr einkunnakvarða er notuð til að bæta gæði upprunalega vélþýðingarkerfisins.



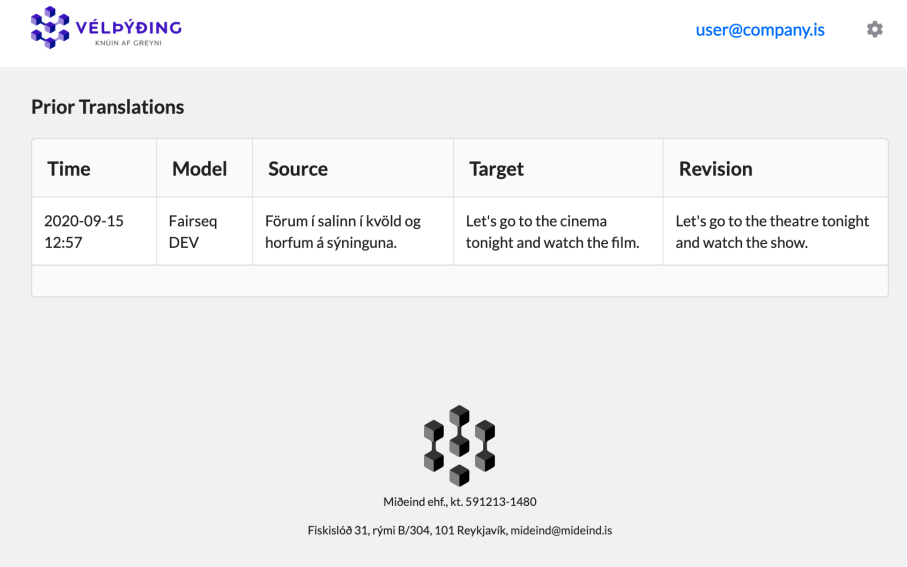
Skjáskot frá `velthyding.is` þar sem notandinn getur (valkvætt) skráð sig inn í þýðingarkerfið.



Skjáskot sem sýnir þýðingu setningar á íslensku yfir á ensku.
*Notandinn hefur smellt á orðið **cinema** og kerfið stingur upp á öðrum möguleikum fyrir orðið.*



*Notandinn hefur valið **theatre** úr lista annarra möguleika,*
og kerfið uppfærir þýðinguna í framhaldi af því.
*Athugið hvernig orðið **film** hefur breyst í **show**, til að endurspegla nýja samhengið.*



Skjáskot sem sýnir hvernig innskráður notandinn getur
farið yfir leiðréttingar úr gagnagrunni kerfisins.

L1 – Almenn villumálheild

Skýrsla: Háskóli Íslands
Aðilar: Háskóli Íslands, Miðeind

Varða 3 – lýsing

Tilgangur

Að þróa villumálheild, sem er nauðsynleg fyrir þróun/þjálfun og mat á málrýnikerfum.

M2: Merkt villumálheild útgefin samkvæmt stöðlum CLARIN. Áætluð stærð við byrjun verkefnisins er 5.000 setningar, sem innihalda að minnsta kosti eina villu hver.

M3: Merkt villumálheild útgefin.

Afurðir

Öllum markmiðum vörðunnar var náð og endanleg málheild er mikið stærri en upprunalega var stefnt að í þessum verkþætti, eins og lýst er neðar. Útgáfa 1.0 iceErrorCorpus hefur verið gefin út á GitHub á https://github.com/antonkarl/iceErrorCorpus og á CLARIN á http://hdl.handle.net/20.500.12537/73. Verkþættinum er lokið við M3.

Skýrsla

Málheild merktra íslenskra texta hefur verið samsett og útgefin með opnum leyfum. Málheildin er á TEI XML-formi. Stafsetningar- og málfræðivillur eru markaðar með leiðréttingum og flokkaðar, og markanir veita einnig tillögur að betra orðalagi og betri stíl.

Uppruni gagna

Textarnir í villumálheildinni eru þrenns konar: nemendaritgerðir, fréttatextar af netinu og Wikipedia-greinar á íslensku. Hluta textanna hafði þegar verið safnað og þeir gefnir út (án merkinga) sem hluti af Risamálheildinni (IGC).

Stærð málheildar

- 158 ritgerðir nemenda með 4.743 endurskoðunarspannir (e. revision spans) og 5.842 flokkuðum villum.
- 2.638 fréttagreinar af netinu með 15.852 endurskoðunarspannir og 18.839 villum.
- 883 Wikipedia greinar með 20.249 endurskoðunarspannir og 26.149 villum.

Merkingaferli

Merkingaferli okkar notar lagskipta aðferð sem endar í safni XML skjala á TEI-formi með endanlegum villumerkingum. Ferlið hafði einnig birtingarmyndir í miðju ferli.

Fyrsta skrefið í ferlinu fól í sér handvirkan yfirlestur og leiðréttingu með hvers konar tóli sem gat leiðrétt villur og einnig viðhaldið upprunalegum texta (snið 1). Við notuðum Track Changes eiginleikann í Microsoft Word vegna þess að merkjararnir okkar kunna vel á þetta viðmót og það gerir þeim kleift að vinna nokkuð hratt. Það er hins vegar engin þörf fyrir Word, því það nota má hvaða textaritil sem er fyrir þetta skref.

Í framhaldinu notuðum við Python-skriftu til samröðunar og samskeytingar réttra og rangra útgáfa í xml-grind sem inniheldur kortlagningu villna til leiðréttinga.

Lokaskrefið í ferlinu var svo handvirk merking villna og endanlegt TEI-formað úttak var vistað til útgáfu í málheildinni (snið 2).

Til að bæta málheildina er verið að bæta leiðréttingum til viðbótar við hana til að tryggja mestu mögulegu nákvæmni. Þessar leiðréttingar fundust ekki við handvirkan yfirlestur en fundust með notkun málrýnisins GreynirCorrect sem er í þróun í undirverkefni L7. Vélrænu merkingarnar eru yfirfarnar og rétt greindar villur eru leiðréttar í málheildinni. Þessi vinna hefur einnig verið gerð á prófunarmálheildinni í undirverkefni L3. Þessar viðbótarleiðréttingar hafa þegar bætt árangursmælingar á GreynirCorrect í undirverkefni L7.

Merkingarskema

Merkingarskemað felur í sér 212 villukóða, sem endurspegla allar gerðir leiðréttinga sem gerðar eru á textum málheildarinnar. Hver villukóði tilheyrir yfirflokki og eru þeir 19. Listi villukóðanna, lýsingar þeirra og yfirflokka er að finna í GitHub hirslunni sem sagt er frá að ofan. Listinn inniheldur einnig dæmi fyrir hvern villukóða.

Hver villa í TEI XML-skránum er birt með endurskoðunarspönn, sýnt hér fyrir neðan. Hver endurskoðunarspönn inniheldur raðfellt heiltölu(auð)kenni. Villunni og leiðréttingu hennar er lýst ásamt villukóðanum (nefnt *xtype*).

```
<revision id="3">
  <original><c>"</c></original>
  <corrected></corrected>
    <errors>
      <error xtype="extra-quot" eid="0" />
    </errors>
</revision>
```

Flækjustig villna er mismunandi, frá einföldum villum eins og hér að ofan til villunnar hér á eftir. Villan sem um ræðir inniheldur tvær endurskoðunarspannir, með auðkennin 11 og 12. Villukóði hennar er „wording“ og til að sýna að villukóðinn á einnig við endurskoðunarspönn 12, hefur sú spönn villukóðann „dep“ og „depID“. „DepID“ er vísir á fyrri hluta villunnar, vísinn sem sýndur er sem endurskoðunarspönn 11.

```
<revision id="11">
  <original><w>í</w><w>dag</w></original>
  <corrected/>
    <errors>
      <error xtype="wording" idx="1b-1" eid="0"/>
    </errors>
</revision>
<c>,</c>
<w>einkum</w>
<w>vestræn</w>
<w>samfélög</w>
<c>,</c>
<revision id="12">
  <original><w>í</w><w>dag</w></original>
  <corrected/>
    <errors>
      <error xtype="dep" depId="1b-1" eid="0"/>
    </errors>
</revision>
```

Villa getur einnig falið í sér fleiri en einn villukóða, eins og sýnt er í dæminu fyrir neðan. Til að sýna að villukóðarnir tveir eigi við sama orð og villu er vísir þeirra sá sami, í þessu tilviki „10-1“.

```
<revision id="1">
  <original><w>að</w></original>
  <corrected/>
    <errors>
      <error xtype="extra-word" idx="10-1" eid="0"/>
      <error xtype="style" idx="10-1" eid="0"/>
    </errors>
</revision>
```

Villugerðir

Villumálheildin hefur verið greind og tíðni hverrar gerðar villu er aðgengileg á https://github.com/antonkarl/iceErrorCorpus/tree/master/freqInfo. Meðfylgjandi eru upplýsingar um tíðni villna í almennri villumálheild í heild sinni í ritgerðum nemenda, fréttum á netinu og Wikipedia-greinum.

L2 – Sérhæfðar villumálheildir

Skýrsla: Háskóli Íslands
Aðilar: Háskóli Íslands, Miðeind

Varða 3 – lýsing

Tilgangur
Að þróa sérhæfðar villumálheildir sem samanstanda af textum frá börnum og unglingum, textum frá lesblindum og textum frá fólki sem notar íslensku sem annað mál.

M2: Uppruni gagna skráður, söfnun hafin.

M3: Gagnasöfnun lokið. Merkingarskema uppfært í samræmi við algengar villur fyrir hverja villumálheild. Merkingar vel á veg komnar.

Afurðir

Á meðan sum undirverk í L2 eru á undan áætlun, hefur markmiðum sumra undirverka vörðunnar ekki verið náð að fullu. Undirverkefni L1 tók lengri tíma en áætlað var sem orsakaði tafir við vinnu sérhæfðra villumálheilda. Áhersla var lögð á að ljúka undirverkefni L1 áður en vinna við L2 hæfist af fullum krafti, svo að fullklárað merkingarskema til að byggja á væri tilbúið.

Textum frá fólki sem notar íslensku sem annað mál hefur verið safnað og því markmiði M4 að merkja og gefa út L2 villumálheild með 3000 villum til CLARIN hefur verið náð. Málheildin hefur verið gefin út á GitHub á https://github.com/antonkarl/iceErrorCorpusL2 og CLARIN á http://hdl.handle.net/20.500.12537/67. Merkingarskemað frá almennri villumálheild í undirverkefni L1 hefur verið uppfært til að taka á villum sem fundust í textum L2.

Undirbúningur við söfnun texta frá lesblindum er hafinn og textarnir verða lesnir yfir og merktir. Komið hefur verið á samskiptum við Félag lesblindra á Íslandi. Varðandi villumálheildina með textum frá unglingum og börnum, er hlutanum með textum frá unglinum lokið og hann útgefinn (M4) og vinna er hafin við texta frá börnum. Einn af flokkunum í almennu villumálheildinni í undirverkefni L1 eru ritgerðir nemenda, skrifaðar af framhaldsskólanemum á aldrinum 15-20 ára og eru þeir textar nógu líkir textum fullorðinna til að þeim sé best lýst sem skrifum fullorðinna frekar en sérhæfðri villumálheild. Notendur geta að sjálfsögðu skoðað þennan flokk til að rannsaka þennan hóp. Þó við höfum ekki gefið út villumálheild barna mun uppbyggingin og aðferðafræðin sem þegar er til staðar frá hinum undirverkunum flýta vinnu við og lok þessa undirverks.

Í stuttu máli er vinna við flestar villumálheildirnar annaðhvort á áætlun eða á undan áætlun, annaðhvort við vörður M3 eða M4. Undirhlutar töfðust þar sem við álitum skilvirkast að ljúka fyrst merkingarskema og allri aðferðafræði sem beitt er við gerð málheildanna. Í heild sinni er verkþátturinn á áætlun og erum við örugg um að öllum afurðum vörðu M4 verði lokið á áætlun. Verkþátturinn heldur áfram á öðru ári verkefnisins.

Skýrsla

Uppruni gagna

Textarnir frá fólki sem notar íslensku sem annað mál samanstanda eins og er aðeins af ritgerðum nemenda við Háskóla Íslands sem læra íslensku sem annað mál. Leitað verður til annarra hópa útlendinga á Íslandi, til dæmis skiptinema sem dvelja á Íslandi í stuttan tíma og læra íslensku.

Textum frá lesblindum verður safnað í gegnum námsráðgjafa hjá Háskóla Íslands. Textarnir þurfa að koma frá fólki sem hefur verið greint með lesblindu, sem námsráðgjafar hafa upplýsingar um.

Aðrar leiðir fyrir báðar gerðir texta eru til skoðunar, með því að hafa samband við fólk sem tengist útlendingum á Íslandi og lesblindum.

Núverandi stærð málheildar

Málheildin samanstendur eins og er af 14 mörkuðum ritgerðum nemenda, 3.028 endurskoðunarspönnum og 3.992 flokkuðum villum. Vinna við frekari viðbætur í málheildina heldur áfram á næsta stigi.

Merkingararferli

Merkingarferli okkar er það sama og fyrir almenna villumálheild í undirverkefni L1. Það felst í notkun lagskiptrar aðferðar sem endar með safni XML-skjala á TEI-formi með endanlegum villumerkingum. Ferlið hefur einnig birtingarmyndir í miðju ferli. Þar sem sama ferli var notað í undirverkefni L1, hafa merkjararnir reynslu af því og allar handvirkar merkingar eru því nokkuð fljótunnar.

Fyrsta skrefið í ferlinu felur í sér handvirkan prófarkalesturlestur og leiðréttingu með hvers konar sem sem getur leiðrétt villur og einnig viðhaldið upprunalegum texta. Við notum Track Changes eiginleikann í Microsoft Word fyrir þetta vegna þess að merkjarar okkar kunna vel á þetta viðmót og það gerir þeim kleift að vinna nokkuð hratt. Það er hins vegar engin þörf fyrir Word, það nota má hvaða textaritil sem er fyrir þetta skref.

Í framhaldinu notuðum við Python skriftu til samröðunar og samskeytingar réttra og rangra útgáfa í xml-grind sem inniheldur kortlagningu villna til leiðréttinga.

Lokaskrefið í ferlinu er svo handvirk merking villna og endanlegt TEI-formað úttak hefur verið vistað til útgáfu í málheildinni.

Merkingarskema

Merkingarskemað felur í sér 233 villukóða sem endurspegla allar gerðir leiðréttinga sem gerðar eru á textum málheildarinnar. Hver villukóði tilheyrir yfirflokki, og eru þeir 22. Þremur flokkum var bætt við þá 19 yfirflokka sem eru í merkingarskema fyrir almennu villumálheildina í undirverkefni L1, orðasafn, setningafræði og tal. Listi villukóðanna, lýsingar þeirra og yfirflokka er að finna í GitHub hirslunni sem er gefin að ofan. Listinn inniheldur einnig dæmi fyrir hvern villukóða.

L3 – Tölfræðileg úrvinnsla og prófunarmálheildir

Skýrsla: Háskóli Íslands, Miðeind
Aðilar: Háskóli Íslands, Miðeind

Varða 3 – lýsing

Tilgangur: Að fá tölfræði um algengustu villur og búa til aðgreindar prófunarheildir. Þörf er á þeim til að skilgreina hvaða villur og tegundir villna er mikilvægast fyrir málrýni að meðhöndla.

M2: Borin hafa verið kennsl á algengar villur í villuheildum og þær greindar.
M3: Prófunarmálheildir með mismunandi tegundum villna tilbúnar og skipt í lokaða og opna hluta.

Afurðir

Öllum markmiðum vörðunnar var náð. Prófunarmálheildin hefur verið gefin út og er aðgengileg á https://github.com/antonkarl/iceErrorCorpus/tree/master/testCorpus.
Verkþættinum er lokið við M3.

Skýrsla

Yfirlit

Handvirkt merktu villumálheildinni frá undirverkefni L1 hefur verið skipt með handahófskenndu úrtaki í þróunarmálheild (90% af heildinni) og prófunarmálheild (10%).

Prófunarmálheildin er þegar í notkun við prófanir í undirverkefni L7 og hefur reynst vel til að mæla árangur. Dæmi um merkingar má sjá í skýrslunni fyrir undirverkefni L1 og niðurstöður má sjá í skýrslunni fyrir undirverkefni L7.

Uppruni gagna

Prófunarmálheildin samanstendur af þremur mismunandi gerðum texta: nemendaritgerðum, fréttagreinum af netinu, og Wikipedia-greinum.

Stærð málheildar

Prófunarmálheildin samanstendur af samtals 367 textum:

- 18 ritgerðir nemenda með 646 endurskoðunarspönnum og 800 flokkuðum villum.
- 267 fréttagreinar af netinu með 1.346 endurskoðunarspönnum og 1.661 flokkuðum villum.
- 82 Wikipedia greinar með 1.381 endurskoðunarspönnum og 1.957 flokkuðum villum.

Eftirvinnsla

Auk upphaflegrar handvirkrar mörkunar var sjálfvirka málrýni GreynirCorrect, sem er í þróun í undirverkefni L7, beitt á prófunarmálheildina til að finna afgangsvillur. Vélrænu merkingarnar voru yfirfarnar og rétt greindar villur leiðréttar í prófunarmálheildinni. Þetta var gert til að tryggja að prófunarmálheildin væri eins nákvæm og mögulegt er og til að lágmarka fjölda rangt flokkaðra falskjákvæðra villna sem rekja má til sjálfvirkrar villugreiningar í undirverkefni L7.

Vélrænu markanirnar eru yfirfarnar og rétt greindar villur eru leiðréttar í málheildinni. Þessi vinna hefur einnig verið gerð á prófunarmálheildinni í undirverkefni L3. Þessar viðbótarleiðréttingar hafa þegar bætt árangursmælingar á GreynirCorrect í undirverkefni L7.

Villugerðir

Villumálheildin úr undirverkefni L1 hefur verið greind og tíðni fyrir hverja gerð villu er aðgengileg á https://github.com/antonkarl/iceErrorCorpus/blob/master/freqInfo/errorTypeFreq_data.tsv.

Fyrir prófunarmálheildina eingöngu er tíðni hverja gerða villu aðgengileg á https://github.com/antonkarl/iceErrorCorpus/blob/master/freqInfo/errorTypeFreq_testCorpus.tsv.

Eftirfarandi 10 villugerðir eru algengastar í prófunarmálheildinni:

Villugerð	Tíðni
orðalag (e. wording)	19
aukakomma (e. extra-comma)	16
orð vantar (e. missing-word)	15
kommu vantar (e. missing-comma)	11
óorð (e. nonword)	11
bandstrik vantar (e. missing-dash)	11
svigi fyrir kommu (e. bracket4comma)	9
punkt vantar (e. missing-period)	8
nafnorðabeying (e. nominal-inflection)	6
fleiri en eitt orð vantar (e. missing-words)	6

Eins og sjá má er skipting tíðninnar að mestu í samræmi við tíðnirnar í þróunarmálheildinni, sem staðfestir að handahófskennda úrtakið fyrir prófunarmálheildina er tölfræðilega áreiðanlegt til mælingar.

L4 – Orðalistar og mállíkön

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

Varða 3 – lýsing

Tilgangur

Að útbúa nauðsynleg gögn fyrir málrýni, eins og orðalista með gildum íslenskum orðmyndum, villuorðabók með algengum stafsetningarvillum og leiðréttingum þeirra, og mállíkan sem byggt er með hjálp Risamálheildarinnar til að sjá um tölfræðilegar villuleiðréttingar.

M3

- Orðalistar tilbúnir.
- Þjálfun mállíkana lokið.

Afurðir

Öllum markmiðum vörðunnar var náð. Verkþættinum er lokið við M3 og gögn og líkön fyrir áframhaldandi þróun undirverkefnis L7 eru tilbúin.

Orðalistar tilbúnir

Eftirfarandi orðalistar voru búnir til:

1. Bannorð
Búinn var til listi íslenskra orða sem geta á einhvern hátt talist óviðeigandi, bannorð og/eða orð sem eru hlaðin í merkingu eða notkun. Listinn inniheldur einnig orð sem eru ekki endilega óviðeigandi en gætu talist óheppilegt umræðuefni fyrir börn eða óviðeigandi í ákveðnu samhengi. Orðunum er skipt upp í flokka eftir annaðhvort merkingu, formi eða notkun.. Hvert orð hefur því verið merkt með stuttri útskýringu (á íslensku) á því hvernig það getur verið óviðeigandi og þá í hvaða samhengi.

Listinn er aðgengilegur á Github (https://github.com/antonkarl/iceTaboo) og í hirslu CLARIN (http://hdl.handle.net/20.500.12537/64).

2. Algengar stafsetningarvillur og leiðréttingar þeirra
Búinn var til listi yfir algengar stafsetningarvillur ásamt leiðréttingum. Orðmyndirnar koma úr þróunarhluta almennu villumálheildarinnar í undirverkefni L1 og samanstanda af orðmyndum merktum sem ‘óorð (e. nonwords)’ í handvirkri merkingu villna.

Listinn er tilbúinn og aðgengilegur hjá CLARIN (http://hdl.handle.net/20.500.12537/63) á `.tsv` sniði.

3. Kerfisbundnar villur
Búinn var til listi listi sjálfvirkt útbúinna orðmynda sem innihalda kerfisbundnar villur, ásamt leiðréttingum þeirra. Listinn samanstendur af eftirfarandi kerfisbundnum villum:

- Orð með bókstaf sem vantar stafmerki
- Orð þar sem stafmerki er bætt við bókstaf
- Orð sem ættu að innihalda ‘y’ en innihalda þess í stað ‘i’
- Orð sem ættu að innihalda ‘i’ en innihalda þess í stað ‘y’

- Orð sem ættu að innihalda ‘ý’ en innihalda þess í stað ‘í’
- Orð sem ættu að innihalda ‘í’ en innihalda þess í stað ‘ý’
- Orð sem ættu að innihalda ‘hv’ en innihalda þess í stað ‘kv’
- Orð sem ættu að innihalda ‘kv’ en innihalda þess í stað ‘hv’

Listinn er tilbúinn og aðgengilegur hjá CLARIN (http://hdl.handle.net/20.500.12537/50) á `.tsv` formi.

Listinn var búinn til með því að nota orðalista frá BÍN (Beygingarlýsing íslensks nútímamáls fyrir máltækni, sjá undirverkefni G4) sem inniheldur ólík íslensk orð og beygingar þeirra. Listi kerfisbundinna villna samanstendur af tveim dálkum skipt með dálkstaf, þar sem fyrri dálkurinn sýnir ranga orðmynd og seinni dálkurinn sýnir leiðréttingu hennar. Dæmi úr orðalistanum er sýnt fyrir neðan.

```
sjálfsýmynd     sjálfsímynd
sjálfsýmynda    sjálfsímynda
sjálfsýmyndanna sjálfsímyndanna
sjálfsýmyndar   sjálfsímyndar
sjálfsýmyndarinnar      sjálfsímyndarinnar
sjálfsýmyndin   sjálfsímyndin
sjálfsýmyndina  sjálfsímyndina
sjálfsýmyndinni sjálfsímyndinni
sjálfsýmyndir   sjálfsímyndir
```

4. Landaheiti og örnefni
Þrjú gagnasett tengd villum í örnefnum hafa verið búin til. Gagnasettin eru á JSON-formi og kóðuð í UTF-8.

- *isprep4isloc* inniheldur gild íslensk forskeyti fyrir ýmis íslensk örnefni. Gagnasettið er nytsamlegt við greiningu rangra forskeyta fyrir íslensk örnefni.[44]
- *isprep4cc* inniheldur íslensk forskeyti fyrir ýmis lönd og sjálfstjórnarsvæði. Gagnasettið verður nytsamlegt vil greiningu rangra forskeyta örnefna annarra en íslenskra.
- *cities_is2en* tengir borgarnöfn á íslensku við samsvarandi nöfn á ensku.[45] Gagnasettið er nytsamlegt til að greina þegar ensk borgarnöfn eru notuð, og má leiðrétta yfir á samsvarandi íslenskt nöfn; eða til að para íslensk borgarnöfn við ensk í samhliða málheildum fyrir þýðingar.

5. Kerfisbundnar beygingarvillur
Við höfum einnig safnað saman og skilgreint algengar, rangar beygingarreglur. Þessum reglum hefur verið beitt á færslur í BÍN til að búa til kerfisbundnar beygingarvillur sem innihalda upplýsingar úr BÍN ásamt leiðréttingum, auk hlekks (með auðkennisnúmeri (e. ID number)) á rétt beygingardæmi, þar sem það á við.

6. Gildar orðmyndir
Auk þessara orðalista skal taka fram að BÍN (Beygingarlýsing íslensks nútímamáls fyrir máltækni, undirverkefni G4) er notuð sem grunnur orðalista með gildar íslenskar orðmyndir sem notaðar eru af málrýni. Einnig inniheldur þáttari Greynis (sjá undirverkefni I5), sem málrýnir byggir á, lista gildra orðmynda sem finnast (enn) ekki í BÍN. Þar sem hlutverk BÍN er að vera lýsandi, ekki vísandi (e. prescriptive), höfum við merkt færslur sem innihalda rangar orðmyndir með leiðréttingum og réttu marki. Listinn er aðgengilegur undir hirslu GreynirCorrect á GitHub.

[44] Dæmi: *Ég bý í Vík* og *Ég bý á Akranesi* eru rétt, en *Ég bý á Vík* and *Ég bý í Akranesi* eru röng.

[45] Til dæmis: Lundúnir - London, Stokkhólmur - Stockholm

Að lokum skal nefna að á öðru ári verður samþjöppuð útgáfa BÍN búin til og gefin út sem Python-pakki (undirverkefni G4). Önnur gögn úr undirverkefni G4 munu einnig vera nytsamleg, sérstaklega *Ritmyndir*, sem er hirsla aukagagna, og BÍN-kjarninn. Við höfum deilt öllum viðeigandi gögnum úr þessu kjarnaverkefni með undirverkefni G4 til að tryggja fulla nýtingu þeirra.

Þjálfun mállíkana lokið
Fyrirliggjandi Python-pakki og mállíkanið *Icegrams*[46], þrístæðulíkan sem var upprunalega eingöngu þjálfað á fréttatextum, var endurþjálfað, fínstillt og stækkað til að innihalda víðara svið texta úr Risamálheildinni. Ítarlega lýsingu á sköpunarferlinu og tölfræði má finna á GitHub[47].

Útgefni *Icegrams*-pakkinn inniheldur Python hjúp (e. wrapper) (API), með afkastamiðuðum hlutum forrituðum í C++, sem sendir fyrirspurn í samþjappaða gagnasafnið á auðveldan og fljótlegan hátt (þ.e. um ~10 míkrósekúndur fyrir hverja uppflettingu).

Icegrams-pakkinn er aðgengilegur á GitHub (https://github.com/mideind/Icegrams) og á CLARIN (http://hdl.handle.net/20.500.12537/80) undir MIT-leyfi. Hugbúnaðinn má setja upp sem Python pakka með `pip install icegrams`, og er hýstur í Python Package Index (PyPi) sem 'icegrams'.

[46] Á GitHub: https://github.com/mideind/Icegrams; í Python Package Index (PyPI): https://pypi.org/project/icegrams/

[47] Sjá https://github.com/mideind/Icegrams/blob/master/doc/overview.md

L6 – Málrýnir fyrir stakorðavillur

Skýrsla: Miðeind
Aðilar: Miðeind, Háskóli Íslands

Varða 2 - lýsing

Tilgangur
Að afhenda sjálfstæðan hugbúnaðarpakka og vefþjónustu sem getur þekkt og stungið upp á leiðréttingum einfaldra (samhengisfrjálsra) stafsetningarvillna í hefðbundnum textum, skrifuðum af ritfærum textahöfundum á íslensku. Á seinni stigum máltækniverkefnisins mun þessi eining verða hluti af fullgildum málrýni fyrir íslensku.

Ákvæði
Tilreiðari (I3)

M2

- Endanleg útgáfa samhengisfrjálsa málrýnisins fyrir stakorðavillur er afhent.
- Leiðréttingar virka að fullu.
- Vefþjónusta innleidd.
- Einfalt vefforrit komið í gagnið.

Afurðir

Öllum markmiðum vörðunnar var náð.

Yfirlit
Verkþættinum er lokið við M2. Þessi skýrsla lýsir stöðunni við enda verkefnisins og er höfð með til að gefa heildarmyndina. Áframhaldandi þróun undirverkefnis L6 er talin hluti af L7 frá M3 og áfram.

Útgáfa
Endanleg útgáfa málrýnisins hefur verið afhend og má finna í hirslu CLARIN á http://hdl.handle.net/20.500.12537/22 sem og á Github á https://github.com/mideind/GreynirCorrect. Skjölun er aðgengileg á https://yfirlestur.is/doc/.

Leiðréttingar
Virkni leiðréttinga hefur verið innleidd. Við prófun málrýnisins á villugreiningu og leiðréttingu með bráðabirgðaprófunarsetti markaðra samhengisfrjálsra villna fengust eftirfarandi niðurstöður:

Villugreining (grunnlína við enda M2)
```
Hittni: 99,48%
Nákvæmni: 98,06%
Heimt: 88,63%
F-gildi: 93,11%
```

Villuleiðrétting
```
Nákvæmni: 88,16%
```

Þegar þessar niðurstöður eru bornar saman við niðurstöður úr sama prófunarsetti við enda M1 sést að ógreindum villum fækkaði úr 66 í 39 (af 343 heildarvillum). Rétt leiðréttum villum fjölgaði úr 240 í 268. Þessar niðurstöður lofa góðu og F-gildið var mun hærra en 60% markmiðið sem sett var fram.

Vefþjónusta
Vefþjónustan hefur verið innleidd og er komin í gangið á https://yfirlestur.is/correct.api. Skjölun má finna á https://github.com/mideind/Yfirlestur#https-api.

Vefforrit
Vefforrit er í gangi á https://yfirlestur.is/.

L7 – Kerfisbundin þróun málrýnis

Skýrsla: Miðeind
Aðilar: Miðeind, Háskóli Íslands

Varða 3 – lýsing

Tilgangur
Að hafa málrýnikerfi fyrir íslensku sem nær yfir ákveðið safn samhengisháðra villna, sem eru greindar og þeim forgangsraðað á meðan verkefninu stendur. Markmið verða skilgreind á ítraðan hátt eftir forgangsröðun villuflokka sem skilgreindir eru í undirverkefni L3 með gögnum frá undirverkefnum L1 og L2. Að greina þessar villur og stinga upp á leiðréttingum.

Ákvæði
Tilreiðari (I3)
Reglubyggður þáttari (I5)
Almenn villumálheild (L1)
Tölfræðileg úrvinnsla og prófunarmálheildir (L3)
Orðalistar og mállíkön (L4)
Málrýnir fyrir stakorðavillur (L6)

M3

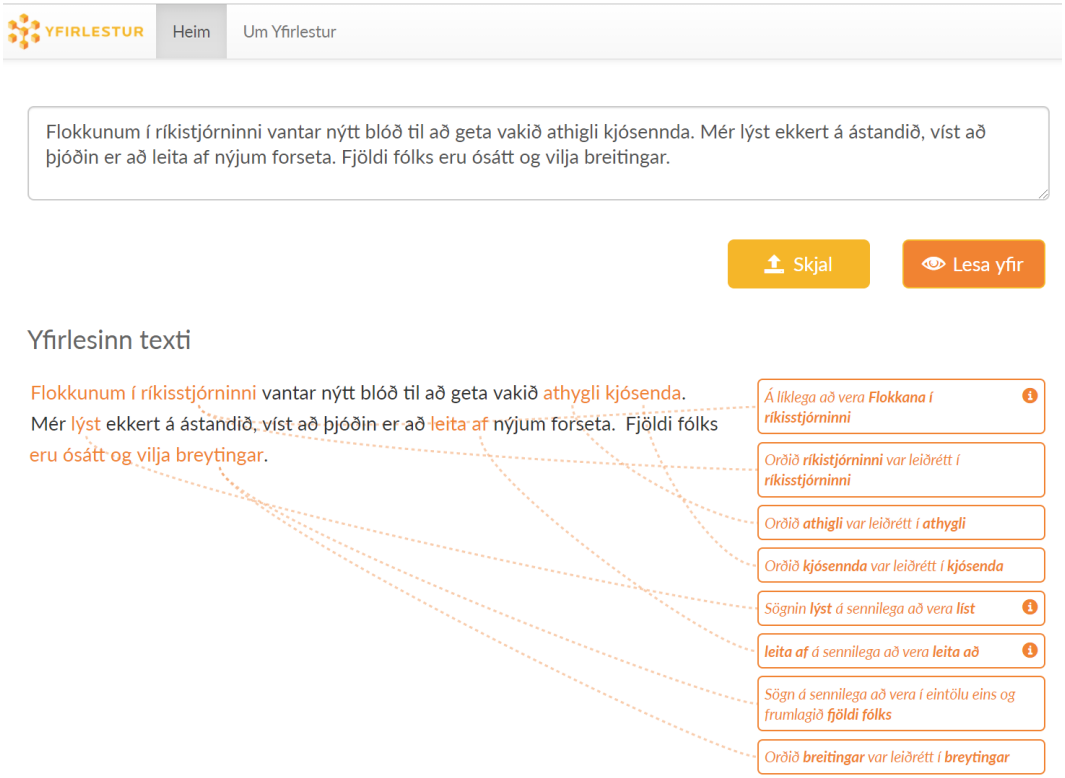
- Leiðrétting villna byggð á forgangsröðun villna úr villugreiningu í L3 innleidd.
- Nýjum vandamálum forgangsraðað og verið er að innleiða meðhöndlun þeirra.

Afurðir

Öllum markmiðum vörðunnar var náð. Verkþátturinn heldur áfram á öðru ári verkefnisins.

Yfirlit
Þessi verkþáttur felur í sér þróun málrýnieiningar í Python, auk vefviðmóts til sýnis þar sem notendur geta slegið inn texta og fengið hann sjálfkrafa prófarkalesinn og athugasemdir við stafsetningar- og málfræðivillur. Verkefnið byggir á gögnum sem hefur verið safnað í öðrum undirverkefnum, einna helst mörkuðu villumálheildinni *iceErrorCorpus* úr L1 og orðalista og mállíkön úr L4 og L6.

Verkefnið hefur verið unnið á ítraðan hátt, með reglulegum útgáfum á GitHub og á Python Package Index (PyPI). Vefviðmótið er uppi á https://yfirlestur.is og er uppfært reglulega með nýjustu hugbúnaðarútgáfum.



Skjáskot af `yfirlestur.is`, sem sýnir textann sem notandi sló inn merktan með leiðréttingum og tillögum fyrir stafsetningu og málfræði.

Ný útgáfa pakkans GreynirCorrect er aðgengileg á CLARIN á (http://hdl.handle.net/20.500.12537/77), sem og á GitHub á (https://github.com/mideind/GreynirCorrect).

Eftir því sem villu- og prófunarmálheildir undirverkefna L1/L3 urðu aðgengilegar, milli vörðu M2 og M3, var mælingadrifin nálgun tekin upp. Með notkun sjálfvirks matstóls hefur *iceErrorCorpus* þróunarsettið (úr L1) verið notað til að greina og meta veika hlekki í hugbúnaðinum og forgangsraða þróun, og prófunarsettið (úr L3) hefur verið notað til að mæla árangur.

Villur í forgangi og leiðrétting þeirra

Undirverkefni L1 sér fyrir lista villuflokka sem inniheldur tíðni hvers flokks. Þessi gögn hafa upplýst forgangsröðun meðhöndlunar fyrir villuflokkana.

Villuflokkar voru fyrst greindir til að meta hvort þeir ættu að teljast innan sviðs eða utan sviðs fyrir sjálfvirkan málrýni. Sumar merkingarnar taka á vandamálum tengdum orðalagi eða stíl með útskýringum, sem er eins og er mjög erfitt að greina eða leiðrétta á sjálfvirkan hátt og ekki er hægt að búast við slíkum leiðréttingum af tiltækum tólum. Þessa flokkun má endurskoða eftir því sem nýjum möguleikum til meðhöndlunar er bætt við.

60 algengustu villuflokkarnir voru því greindir. Tíðnigögn eru tiltæk í lýsigögnum undirverkefnis L1. Af þessum 60 villuflokkum voru 42 taldir innan sviðs. Þeir voru síðan greindir enn frekar. Það felur í sér að skilgreina hvaða gögn þurfti og hvernig þau skyldi meðhöndla og á hvaða stigi, þ.e. á tókastigi (e. token-level), á setningarstigi (e. sentence-level), sem venjulega málfræðivillu sem greina má eftir samhengisfrjálsum reglum, o.s.frv.
Villuflokkunum var síðan skipt í hópa eftir þessum eiginleikum.

Hér fyrir neðan er listi villuflokka og lýsing á meðhöndlun hvers og eins. Fyrir hvern villuflokk tilgreinum við hvaða forgangsröðuðu villukóða þeir innihalda.

1. Villur á tókastigi (e. token-level errors)
1a. Villur tengdar há- og lágstöfum (upper4lower-common, lower4upper-proper, upper4lower-proper)
Meðhöndlun rangra há-/lágstafa í orðum hefur verið innleidd. Bráðabirgðagögn eru notuð sem stendur. Frekari gagnaöflun byggist á vinnu í undirverkefni G4.

1b. Tvítekning orða (extra-word, extra-words, extra-conjunction)
Algengustu orðmyndum sem geta réttilega komið fyrir tvisvar (eða oftar) í röð hefur verið safnað saman og þær skjalaðar. Leiðréttingar fyrir önnur tilfelli hafa verið innleiddar.

1c. Orð ranglega rituð sem eitt orð (merged-words, missing-space)
Algengustu tilfellum samruna hefur verið safnað saman og þau skjöluð. Leiðréttingar byggðar á þeim hafa verið innleiddar. Algengustu tilfellin snúast aðallega um þekkt sambönd orða sem eru ranglega túlkuð, og rituð, sem aðskeyti. Villur þar sem handahófskennd orð eru rituð sem eitt orð er erfiðara að meðhöndla, þar sem þau geta í mörgum tilfellum talist samsett orð.

1d. Stök orð ranglega rituð sem tvö orð eða fleiri (split-word, split-compound)
Algengustu tilfellunum hefur verið safnað saman og þau skjöluð, og leiðréttingar byggðar á þeim hafa verið innleiddar. Þessar villur eru skyldar (1c) þar sem þær snúast um þekkt aðskeyti sem eru ranglega túlkuð, og rituð, sem einstök orð. Annað algengt tilfelli er þegar samsett orð eru skrifuð sem tvö (eða fleiri) einstök orð. Þetta getur í mörgum tilfellum verið túlkað sem nafnliður með nafnliði í eignarfalli að framan og er því ekki endilega villa. Við tökum ekki á þessu tiltekna atriði sem slíku að svo stöddu, en tökum á skýrum villum af þessari gerð sem við höfum rekist á.

1e. Beygingarvillur (nominal-inflection, n4nn, nn4n, verb-inflection)
Eins og nefnt er í skýrslu fyrir undirverkefni L4 hefur verið settur saman listi yfir beygingarvillur ásamt leiðréttingum. Þennan lista þarf að sameina öðrum minni lista frá fyrri tilraun. Meðhöndlun listanna byggir á vinnu við BÍN (undirverkefni G4).

1f. Samsetningarvillur *(compound-collocation)*
(1c) og (1d) væri hægt að skilgreina sem samsetningarvillur, en áherslan hér er á samsett orð sem innihalda stafsetningarvillu, eða til dæmis fyrri hluta í fleirtölu þar sem hefð er fyrir notkun eintölu. Við höfum skilgreint lista slíkra fyrri hluta sem lýsa skýrum tilfellum sem við höfum rekist á og tökum á þeim. Stækkaða mállíkanið úr undirverkefni L3 hjálpar þar sem það eykur líkur á því að finna samsett orð í þrístæðunum.

1g. Bannorð (taboo-word, gendered)
Við höfum innleitt meðhöndlun bannorða. Til að byrja með voru gögnin okkar af skornum skammti, en gögnunum sem undirbúin voru í undirverkefni L4, sem innihalda víðtæka lista bannorða og fínstillta flokkun þeirra, verður bætt við á öðru ári og ættu að auka dekkun umtalsvert.

1h. Icegrams (missing-letter, extra-letter, missing-accent, nn3n, letter-rep, nonword)
Eins og kemur fram í skýrslu fyrir L4 hefur undirsett Risamálheildarinnar, sem samanstendur mestmegnis af réttum textum, verið notað til að þjálfa nýja útgáfu *Icegrams*-þrístæðumálalíkansins. Þessi stækkaða útgáfa felur í sér víðtækara og réttara tölfræðilíkan af nútímaíslensku. Talið er að þetta líkan muni bæta leiðréttingartillögur algengra stafsetningarvillna, eins og þar sem vantar staf, vantar kommu, o.s.frv.

2. Villur á setningarstigi (e. Sentence-level errors)
2a. Hliðskipun innan nafnliða; hliðskipun milli sagna og röklíða; rangar forsetningar með sögnum (agreement-pred, wrong-prep, agreement-concord, agreement)
Þegar tengslin milli undirverkefna L6 og L7 voru skoðuð, var ákveðið að leggja áherslu á að ljúka við og gefa út meðhöndlunarvirkni fyrir L6 (og bæta við gögnum þar sem hægt er), áður en vinna við L7 hæfist. Þessar villur eru útskýrðar í hluta 5 fyrir neðan.

3. Kóði. Þessi verk tilgreina hvað þarf að innleiða eða lagfæra í kóðanum.
3a. Tillögur
Hugmyndin um að beita öllum tillögum (þ.e. líka „viðvörunum“, ekki bara villum) sem leiðréttingum í skipanalínutóli var byggð á takmörkuðum gögnum. Við hlökkum til að prófa þessa hugmynd á þróunarsetti villumálheildarinnar úr L1.

3b. Þrístæðuleit
Þegar þrístæðan (t_1, t_2, t_3) er notuð í málrýni til að meta líkindin á t_3, tókum við eftir því að upprunalegar útgáfur tókanna t_1 og t_2 voru notaðar í leitinni jafnvel þó t_1 og/eða t_2 hefðu verið leiðréttar í millitíðinni. Þetta hefur verið lagfært.

Niðurstöður
Við höfum innleitt sjálfvirka matssvítu til að nota með þróunar- og prófunarmálheildum úr L1/L3. Eins og er mælir hún nákvæmni, heimt og F1-skor á setningarstigi, þ.e. byggt á því hvort setningar eru sagðar innihalda villu eða ekki, réttilega eða ranglega (rangt/rétt neikvætt/jákvætt). Greiningin nær enn sem komið er ekki til staðsetningar eða flokkunar villna eða merkinga innan hverrar setningar.

Fleiri mæligildum verður bætt í prófunarsvítuna eftir vörðu M3, sem og niðurstöðum fyrir hvern (yfir)flokk villna og textaflokk.

Fyrstu grunnlínuniðurstöður úr iceErrorCorpus prófunarsettinu eru eftirfarandi:

```
Setningar unnar:              5230
   essays       :              888
   onlineNews   :             2569
   wikipedia    :             1773

Heimt:                      0,6116
   essays       :           0,5694
   onlineNews   :           0,5676
   wikipedia    :           0,6695

Nákvæmni:                   0,4833
   essays       :           0,5885
   onlineNews   :           0,3756
   wikipedia    :           0,5632

F1 skor:                     0,5399
   essays       :           0,5788
   onlineNews   :           0,4521
   wikipedia    :           0,6118
```

Þessar niðurstöður eru fyrir heilar setningar, þ.e. ef villa er merkt í prófunarmálheild og villa er fundin af forritinu er setningin dæmd rétt jákvæð (e. true positive), þó það sé strangt til tekið ekki sama villan. Tilraunir með frekari, nákvæmari mælingar eru hafnar en niðurstöður liggja ekki fyrir. Hluti þeirrar vinnu felur í sér samröðun villukóða og afmörkun villuspannar milli sjálfvirka tólsins og handvirks villumörkunarskema.

Einnig skal taka taka fram að greining á setningum sem búa til rangar jákvæður (e. false positives) (þ.e. setningar sem eru ekki merktar í prófunarmálheild en GreynirCorrect skilar þeim sem röngum) hafa nýst sem utanaðkomandi mæling fyrir þáttara Greynis í undirverkefni I5. Í mörgum tilfellum innihalda slíkar rangar jákvæðar setningaform sem vantar í samhengisfrjálsa málfræði þáttarans. Þetta er augljóst dæmi um gagnkvæman ávinning SÍM verkefnisins í víðari skilningi, þar sem endurbætur á þáttaranum leiða til endurbóta málrýnisins.

Forgangsröðun nýrra villna og innleiðing meðhöndlunar þeirra

Við höfum greint algengustu villurnar sem urðu eftir í þróunarmálheildum við fyrri vinnu. Þegar þessari greiningu var lokið var listanum umbreytt í verkefnalista, röðuðum saman eftir hvort þau fela í sér söfnun gagna eða innleiðingu kóða. Kóðunarverkunum var síðan raðað saman eftir því í hvaða hluta kerfisins skal innleiða þau.
Vinna við þessi verk er hafin og villumeðhöndlun fyrir sum þeirra hefur verið innleidd.
fyrir neðan er verkalisti sem skýrir forgangsraðaðar villur. Fyrir hvert verk tilgreinum við hvaða forgangsraðaða villukóða það skal taka á.

4a. Viðbót, samruni og útvíkkun gagna; meðhöndlun algengra, kerfisbundinna beygingarvillna (wrong-prep, nominal-inflection, missing-letter, extra-letter, agreement errors, …)
Þetta verður að mestu tekið fyrir á öðru ári og byggist á þriðju vörðu undirverkefnis L4.

4b. Hástafavillur (upper4lower-common, lower4upper-proper, upper4lower-proper)
Meðhöndlunaraðferðir fyrir hástafavillur eru tiltækar en nota takmörkuð gögn. Þar sem þessar villur virðast nokkuð algengar í villumálheildunum leggjum við áherslu á söfnun gagna fyrir þær á öðru ári.

4c. Sértækum tilvikum bætt við stafskiptafjöldaalgrím (e. edit distance algorithm) (missing-letter, extra-letter, missing-accent)
Á öðru ári munum við byrja á því að meta hvernig stækkaða míllíkanið reynist á ein- eða tvístafa villum og stafmerkjavillum innan orða. Gamla líkanið átti oft erfitt með leiðréttingu þeirra, en hið nýja gæti gert lista sértækra tilvika óþarfan. Ef bæta þarf slíkum lista við verður það gert á öðru ári.

4d. Reglum bætt við fyrir hætti í undirskipuðum setningum (ind4sub)
Reglur sem ákvarða réttan hátt fyrir algengustu og augljósustu tilvik innan undirskipaðra setninga verður bætt við og þau skjöluð.

4e. Reglum bætt við fyrir samræmi við umsögn (agreement-pred)
Málfræðireglur tölutengdra (ein-/fleirtölu) villna hafa verið innleiddar.
Varðandi fallbeygingarvillur, þá er vinna hafin við skilgreiningu ítarlegri forma sagnflokka fyrir þáttara Greynis í undirverkefni I5. Sú vinna mun halda áfram á öðru ári og gagna verður aflað. Þegar því er lokið munum við bæta algengum röngum tilvikum við gögnin.
Meðhöndlun beggja gerða villna sem tengjast samræmi við umsögn (e. predicate agreement error) hefur verið innleidd en kann að þurfa aðlögun að gagnaformi sem valið er fyrir sagnflokkana. Samræður eru hafnar við eigendur mögulegra gagna. Þegar þau gögn hafa verið móttekin er kóðun þegar til staðar fyrir algengustu setningarföllin.

4f. Reglum bætt við fyrir samræmi innan nafnliða (agreement-concord)
Villureglum þegar samræmi vantar innan nafnliða verður bætt við.

H1 – Upptökur með Samrómi

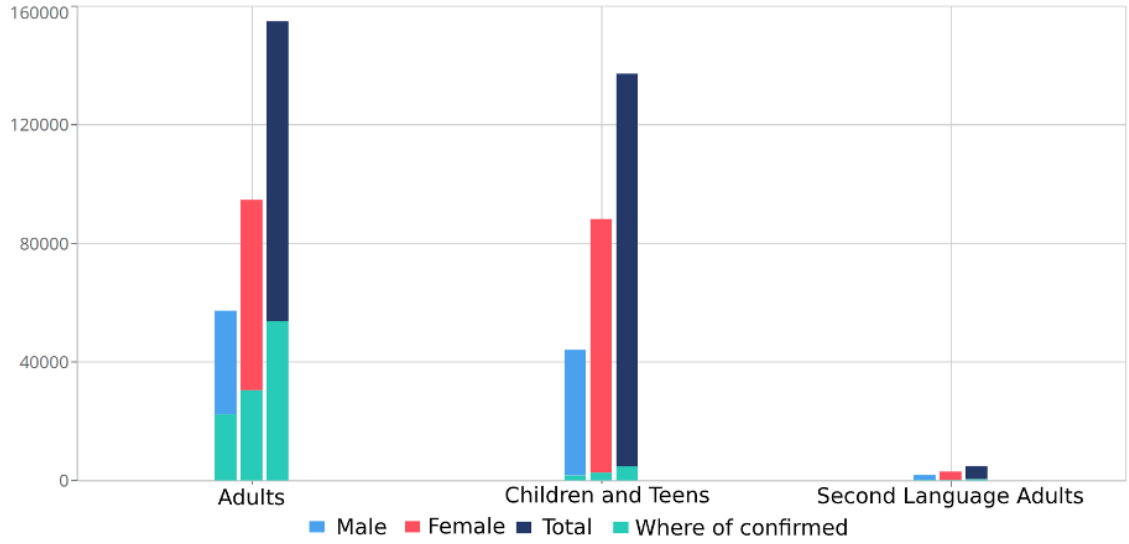
Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Varða 3 – lýsing

M3: Vegið gagnasafn 150,000 segða fullorðinna málhafa með íslensku að móðurmáli gefið út í samræmi við staðla CLARIN. Gagnasafn í kynjajafnvægi með 120,000 segðum unglinga gefið út í samræmi við staðla Clarin. 10,000 segðir fullorðinna málhafa sem hafa íslensku sem annað mál gefið út í samræmi við staðla Clarin (áætlað er að bæta við þetta gagnasafn á öðru ári verkefnisins).

Afurðir

Upphaflega var áætlað að gefa allar segðir sem safnað var út fyrir lok fyrsta árs verkefnisins (M3). Hins vegar höfum við nú ákveðið að gefa aðeins út yfirfarnar upptökur sem þýðir að fjöldi útgefinna segða er minni en fjöldi þeirra segða sem safnað hefur verið. Við höfum safnað meiri gögnum en lýst var fyrir M3, nema þegar kemur að þeim sem hafa íslensku sem annað mál, en við leggjum okkur sérstaklega fram um það núna. Þess vegna lítum við svo á að markmiðum vörðunnar hafi verið náð. Þegar þessi orð eru rituð höfum við safnað samtals 300 þúsund segðum, en rauntímatölfræði (e. live statistics) er aðgengileg hér og hér. Afurðir vörðunnar eru sem hér segir: við höfum safnað 155 þúsund segðum fullorðinna málhafa, 137 þúsund segðum frá börnum og unglingum og 5 þúsund segðum frá fólki sem hefur íslensku sem annað mál. Kvenkyns þátttakendur eru fleiri í öllum þremur flokkunum, eða um 62% í hverjum flokki. Þetta má sjá í línuriti 1.



Græni liturinn í súlunum vísar til fjölda hljóðdæma sem hafa verið yfirfarnar af sjálfboðaliðum á Samromur.is. Alls hafa verið greidd 170 þúsund atkvæði sem leiðir til þess að 60 þúsund hljóðdæmi hafa verið metin fullgild. Til að hljóðdæmi sé metið fullgilt með þessari aðferð verður tveir aðskildir kjósendur að meta það jákvætt. Með því að nota annað umhverfi til bráðabirgða höfum við farið yfir 80 þúsund hljóðdæmi með aðstoð nokkurra sumarstarfsmanna. Samtals höfum við nú staðfest um 140 þúsund hljóðdæmi og af þeim hafa 85% verið merkt fullgild.
Við erum í því ferli að gefa út fyrstu útgáfu Samróms í samræmi við aðra vörðu. *100,000 segðir fullorðinna málhafa hafa verið útgefnar í samræmi við staðla CLARIN*. Sú málheild verður gefin út á CLARIN.

Samrómur hefur vakið mikla athygli meðal almennings og þátttaka hefur verið mjög góð, við áætlum að allt að tíu þúsund einstaklingar hafi tekið þátt (nákvæm tala er á reiki því hver manneskja er talin aftur ef hún hefur notað mörg tæki). Núverandi kynjahalli málheildarinnar hefur merkilegt nokk haldist frekar stöðugur frá upphafi, við vonumst til að jafna hann út á öðru ári með því að nota markmiðaðar auglýsingar (e. targeted advertising), meðal annars. Við erum að vinna að því að sannprófa (e. verify) málheildina með fljótvirkari hætti, en skrefum til að ná því markmiði er lýst í skýrslunni. Fyrsta árið hefur verið mjög jákvætt en jafnframt bent okkur á hvar áherslur okkar þurfa að liggja á öðru ári.

Skýrsla

Endurbætur á Samrómur.is

Uppfærð útgáfa verður gefin út í lok september/byrjun október. Við höfum endurgert kóðann að hluta sem mun gera okkur betur kleift að kemba fyrir villum, endurgreiða tækniskuld (e. technical debt) og bæta ýmsa þætti. Í ljósi þess að áhersla hefur verið lögð á að yfirfara hljóðdæmi erum við að bæta það ferli með þrennum hætti.
Í fyrsta lagi verður upplýsingagjöf til sjálfboðaliða stórbætt til að fækka villum, innskráningarkerfi hefur verið komið fyrir svo notendur geti betur fylgst með eigin ferli bæði þegar kemur að því að gefa raddsýni og fara yfir hljóðdæmi. Þetta mun gera ferlið meira hvetjandi og vonandi halda notendum virkari. Í öðru lagi sjáum við fram á að framlag sjálfboðaliða mun eitt og sér ekki nægja til að yfirfara allt safnið þannig að við áformum að bjóða nokkrum greiðslu til að fara yfir hljóðdæmin. Í þriðja lagi munum við bæta við sjálfvirku gæðaprófunarkerfi sem kallast Maraosjio[48].

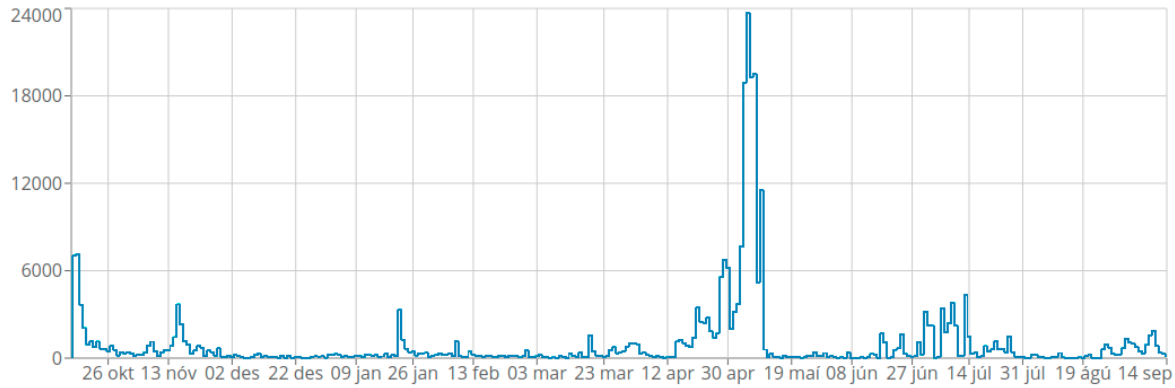
Það kerfi hefur reynst gagnlegt við að greina, með nákvæmum hætti, upptökur sem eru í góðum gæðum sem og upptökur í mjög slæmum gæðum. Þetta opnar möguleikann á því fyrir okkur að krefjast aðeins minnst eins atkvæðis frá sjálfboðaliðum fyrir sumar upptökurnar og leyfir okkur þannig að nýta hvert atkvæði með skilvirkari hætti án þess að það komi niður á gæðum.

Aðgerðir til að ná til fólks

Lestrarkeppni grunnskólanna

Í vor stóðum við fyrir lestrarkeppni meðal allra grunnskóla í landinu. Við gerðum stutt myndband með Guðna Th. Jóhannessyni, forseta Íslands, þar sem hann hleypti keppninni af stokkunum. Keppnin stóð í viku og 1,430 nemendur og starfsfólk tók þátt og tók upp 144 þúsund segðir. Rauntímatölfræði var birt á vefsíðunni og keppnin var nokkuð spennandi undir lokin. Það má glögglega sjá á línuriti 2, sem sýnir heildarþátttöku á dag frá því söfnun hófst til dagsins í dag. Toppurinn í miðjunni er keppni grunnskólanna. Við áformum að blása til sams konar keppni á komandi önn.

[48] Eyra - Speech Data Acquisition System for Many Languages. M Petursson, S Klüpfel, J Gudnason - Procedia Computer Science, 2016.

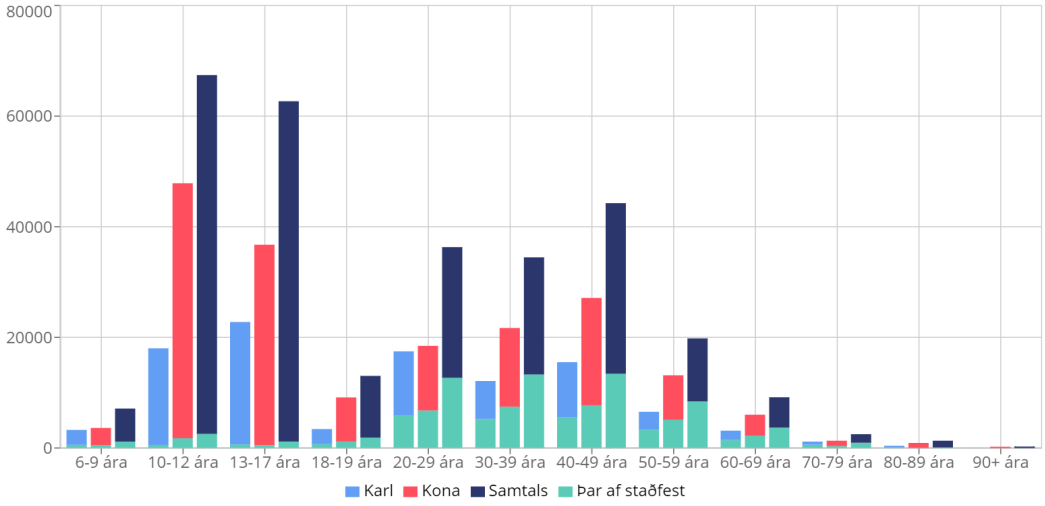


Íslenska sem annað mál

Eins og sjá má á línuriti 1 hafa þeir sem hafa íslensku sem annað mál ekki tekið mikið þátt með núverandi markaðsátaki. Við höfum nokkrar tilgátur um það, einkum teljum við að þeir haldi að sá hreimur sem heyrst getur í tali þeirra geri þá vanhæfa til þátttöku. Til að taka á þessum misskilningi gerðum við kynningarmyndband með Elizu Reid forsetafrú, þar sem hún kynnti herferðina „Íslenska er allskonar“. Hún var fullkomin talskona þessarar herferðar þar sem hún talar íslensku með enskum hreim. Okkur tókst að vekja máls á því hvaða hlutverk tungumálið skipar í sambandi við það að innflytjendur verði og upplifi sig sem hluta af samfélaginu. Myndbandið leiddi til þónokkuð mikilla umræðna á samfélagsmiðlum og hefur fengið meira en 50 þúsund áhorf. Þjóðfélagshópurinn sem ætlunin var að ná til brást vel við kallinu og við tvöfölduðum fjölda upptaka þess hóps. Við vonumst til þess að þetta markaðsátak muni leiða til frekari þátttöku þessa hóps á öðru ári.

Ýmsar aðgerðir

Til að bregðast við fækkun starfa í boði vegna Covid-19 kynnti ríkisstjórnin Aðgerðir stjórnvalda vegna sumarnáms og sumarstarfa. Nokkrir nemendur störfuðu hjá Samrómi í sumar. Þeir fóru yfir hljóðdæmi eins og sagt var frá að ofan og hjálpuðu okkur með örfá sértæk gagnasöfn. Eins og sjá má á línuriti 3 minnkar þátttaka eftir því sem aldur þátttakenda hækkar. Þetta kemur ekki á óvart en til að safna saman fleiri radddæmum frá eldra fólki heimsóttum við tíu dvalarheimili aldraða í sumar. Hver heimsókn stóð yfir í um það bil tvær klukkustundir og skilaði um 1.500 radddæmum. Þetta er mjög tímafrekt þar sem þátttakendur þurfa oftast hjálp við að komast um vefsíðuna.



H2 – Umritun efnis úr sjónvarpi og útvarpi

Skýrsla: Háskólinn í Reykjavík, Creditinfo
Aðilar: Háskólinn í Reykjavík, Creditinfo, RÚV

Varða 3 – lýsing

Markmið: Að umrita efni úr sjónvarpi og útvarpi til að þjálfa talgreina til notkunar á sviði fjölmiðla.

M3: 125 klukkustundir af umrituðu útvarps- og sjónvarpsefni gefið út samkvæmt stöðlum CLARIN.

Afurðir

Við höfum umritað og samraðað samtals 98 klukkustundum af efni úr sjónvarpi og útvarpi. Ekkert af því hefur verið gefið út enn, það verður gert þegar heildarmagn ganga hefur verið undirbúið eins og áætlað er fyrir M3. Góðu verklagi við handvirka umritun hefur verið komið á og samkvæmt skipulagi er áætlað að M3 verði skilað í síðasta lagi 1. desember 2020. Verkþættinum verður haldið áfram á öðru ári verkefnisins.

Skýrsla

Útvarpsefni

Handvirk umritun gagna var unnin af Creditinfo, Fjölmiðlavaktinni, af reynslumiklu teymi í handvirkri umritun hljóðefnis fyrir fjölmiðlavöktun. Hins vegar, þar sem kröfurnar fyrir talgreinisgögnin eru nokkuð strangari (tímamerkingar, nákvæm umritun talaðra orða, umritun annarra hljóða eins og hláturs, hósta o.s.frv.), tók nokkurn tíma að skilgreina viðeigandi snið og vinnulag. Fljótt kom í ljós að þar sem Creditinfo hefur ekki (enn) aðgang að sjónvarpsstreymi RÚV reyndist erfitt að fá réttar tímamerkingar. Vinna við handvirka umritun hefur því eingöngu snúið að útvarpsefni.

Þrátt fyrir brösulega byrjun vorum við bjartsýn á að geta haldið áætlun, þar til umritun stöðvaðist algerlega í COVID-19 faraldrinum í mars/apríl vegna ofurálags hjá Creditinfo. Fólki var bætt við verkefnið yfir sumarmánuðina og verður til loka M3. Enn fremur mun hluti þjálfaða teymisins halda áfram að útbúa gögn fyrir næstu vörðu á öðru ári.

Sjónvarpsefni

Sjónvarpsefnið verður útbúið með því að samraða skjátextum frá síðu 888 í textavarpinu.
Til að hefja vinnu við þetta skoðuðum við gögn sem voru undirbúin í undirverkefninu samræðugreind (e. diarization) (H14), þ.e. skipt upp í mælendur (e. speaker turns) með mælendaauðkennum. Mest af þessu eru kvöldfréttir og fréttaskýringaþátturinn Kastljós. Skjátextarnir eru aðgengilegir án endurgjalds á RÚV API, á vtt sniði, t.d.: https://api.ruv.is/api/subtitles/5008564T0.mp4/is. Við rákum okkur á tvö vandamál með þessi gögn: a) tímamerkingar skjátextanna eru mjög ónákvæmar og b) textinn er oft umorðun á því sem sagt er. Til að sigrast á þessum vandamálum hlóðum við niður þeim skjátextagögnum sem áttu við upptökurnar okkar. Við stöðluðum textann og notuðum Talgreini Alþingis til að samraða og skipta gögnunum upp að nýju. Þetta leysti vandann við tímamerkingarnar og fjarlægði mikið af misræminu milli hljóðs og texta (bútunum þar sem textinn endurspeglar illa hljóðið var hent). Við getum ekki fjarlægt allt misræmi milli hljóðs og texta, en að lokum stóðum við uppi með gögn sem við búumst við að verði nothæf í talgreinaþjálfun, á kostnað um það bil 50% af upphaflegu gagnamagni. Við byrjuðum með átta klukkustunda langar upptökur, bæði tónlist og tal. Eftir hreinsunarferlið og auðkenningu mælenda voru fjórar klukkustundir eftir. Við eigum 18 klukkustundir til viðbótar af samræðugreindargögnum sem verða meðhöndlaðar á sama hátt. Þess

vegna stefnum við að því að afhenda um að bil 13 klukkustundir af gögnum sem verða nothæf við talgreinaþjálfun.

Helst ætti að nota talgreini sem þjálfaður er á efni innan óðals fyrir samröðun hljóðs og texta. Þess vegna leggjum við til að sjónvarpsefninu verði endursamraðað þegar talgreinir hefur fengið þjálfun byggða á útvarpsgögnunum.

H3 – Umritun samræðna og fyrirspurna

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Varða 3 – lýsing

Að umrita samræður og fyrirspurnir til að afla meira efnis um talað mál í frjálsu flæði. Magn ganganna sem verður safnað verður skilgreint þegar búið verður að koma upptökuaðstöðu í gagnið og fyrsti fundurinn tekinn upp. Samtöl verða tekin upp á fyrsta og öðru ári en upptaka á fyrirspurnum hefst og lýkur á öðru ári. M2: Aðstaða fyrir fundi og upptökur tilbúin. Fyrsta samtalið tekið upp.

M3: Fyrsta sett af samtölum tekið upp og umritað.

Afurðir

Öllum markmiðum vörðunnar var náð. Samtalasett er tilbúið og hefur verið umritað. Verkþátturinn heldur áfram á öðru ári verkefnisins.

Skýrsla

Við útbjuggum okkar eigin hugbúnað til að taka upp samtöl. Um er að ræða notendavænt VoIP-app sem þátttakendur geta gengið að á netinu og lagt sitt af mörkum til gagnasöfnunarinnar. Fyrirmyndir gagnasafnsins eru AMI og CallHome, ensk gagnasett sem samanstanda af samtölum á fundum og símtölum.

Hvernig?
Með hugbúnaðinum er kynnt til sögunnar notendaviðmót þar sem þátttakendur geta boðið öðrum þátttakendum í raddsamtal. Þegar komið er inn á spjallsvæðið þurfa þátttakendur að samþykkja ákveðna skilmála varðandi söfnunina og veita lýðfræðilegar upplýsingar: aldur og kyn. Umritun samtalsins fylgir sömu viðmiðum og H4 fyrirlestrarnir sem umritaðir voru af nemendum í sumar.

Snið/Form
Við ákváðum að gera eftirfarandi kröfur til gæða gagnanna:

- Hver þátttakandi hefur sinn eigin hljóðnema og raddir þeirra eru teknar upp hver í sínu lagi.
- Hvert samtal skal mynda eina skrá með mismunandi rásum fyrir hvern mælanda.
- Þátttakendur skulu hafa íslensku að móðurmáli og tala tungumálið reiprennandi.
- Lýsigögn (e. metadata) hvers þátttakenda innihalda aldur og kyn.
- Lengd samtals skal vera allt að 35 mínútur.
- Hvert samtal er umritað og svo staðfest af tveimur einstaklingum.

Framvinda
Fyrsta sett samtala sem tekið var upp er um 30 mínútur að lengd og inniheldur einn kvenkyns og þrjá karlkyns þátttakendur sem allir þekkja hvern annan vel. Við umritunina var Gecko notað til að búa til skjátextaskrár. Hvers kyns persónuupplýsingar sem gætu hafa smogið inn á meðan upptökum stóð voru fjarlægðar. Upptökurnar tvær af hverju samtali voru sameinaðar í eina skrá með tveimur rásum, einni fyrir hvorn þátttakenda og lýsigögnin geymd í aðskildri skrá.

Að lokum; hugbúnaðurinn okkar er kominn í gagnið og ferlum við upptökur og umritanir hefur verið komið á. Þannig að við erum tilbúin að takast á við það verkefni að taka upp og umrita afganginn af samtölunum.

Framtíðaráform

Í næstu vörðu er ætlun okkar að bæta nýjum eiginleikum við gagnasöfnunarvettvanginn.

- Sjá þátttakendum fyrir umræðuefnum og ráðum sem hvetja til lengri samtala.
- Hvetja þátttakendur til að gefa ekki upp neinar persónuupplýsingar og forðast málskipti (e. code-switching) í samtölum sínum.
- Gera kleift að taka upp fundi með mörgum (3+) viðmælendum.

H4 – Umritun fyrirlestra

Skýrsla: Tiro ehf.
Aðilar: Tiro ehf., Háskólinn í Reykjavík og Háskóli Íslands

Varða 2 - lýsing

Markmið: Að eiga safn umritaðra fyrirlestra til að þjálfa kerfi talgreina á því sviði. Þessari vinnu verður lokið á tveimur árum. Markmið fyrra ársins er að koma kerfinu upp og framkvæma fyrstu umritanir til staðfestingar á aðferðinni (e. proof of concept).

Afurðir

Öllum markmiðum vörðunnar var náð.
Fyrsti bunki hljóðgagna undirbúinn og 15 klukkustundir af fyrirlestrum umritaðar og aðgengilegar á CLARIN sem Kennslurómur - Icelandic Lectures 20.09 (http://hdl.handle.net/20.500.12537/81).
Vinna við verkþáttinn heldur áfram á öðru ári verkefnisins.

Skýrsla

Fimm fyrirlesarar gáfu samtals 38 fyrirlestra sem voru í upphafi umritaðir sjálfvirkt og síðan leiðréttir handvirkt með fría veflæga talgreininum og leiðréttingakerfinu frá Tiro (https://tal.tiro.is), auk yfirferðar prófarkalesara. Tímasamskipun umrituðu bútanna var leiðrétt handvirkt af sumarstarfsmönnum Háskólans í Reykjavík og síðan staðfestir með Viterby þvingaðri samröðun (e. forced alignment).

Eftirfarandi tafla sýnir dreifingu fyrirlestranna.

Tölvunarfræði	9 fyrirlestrar	2 þátttakendur	6 klukkustundir
Vinnumarkaðshagfræði	13 fyrirlestrar	1 þátttakandi	2 klukkustundir
Viðskiptagreind	1 fyrirlestur	1 þátttakandi	20 mínútur
Íslensk málvísindi	15 fyrirlestrar	1 þátttakandi	7 klukkustundir

Fyrirlestrarnir voru teknir upp af fyrirlesurunum sjálfum í alls konar umhverfi. Sumir voru teknir upp fyrirfram en aðrir eru fyrirlestrar úr kennslustundum sem innihalda samræður í bakgrunni, sem heyrast illa eða ekki. Upptökurnar voru afhentar á alls kyns skráarsniðum svo við drógum aðeins út hljóðið og breyttum kóðanum á öllu í 16 kHz 16 bit PCM WAVE skrár.

Fyrirlestrarnir hafa verið gerðir aðgengilegir á CLARIN sem Kennslurómur - Icelandic Lectures 20.09 (http://hdl.handle.net/20.500.12537/81). Málheildin inniheldur umritunarleiðbeiningar sem þeir sem umrituðu fengu (á íslensku).

H6 – Leyfismál talgreinisgagna

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík, RÚV, Creditinfo

Varða 2

Leyfi fyrir opinn aðgang að fréttaupptökum RÚV frágengið.

Afurðir

Þrír samningar hafa verið gerðir sem ná yfir leyfismál fyrir efni RÚV:

1. Samningur milli RÚV og Háskólans í Reykjavík (HR) varðandi meira en 900 klukkustundir af efni sem RÚV hefur þegar afhent HR. Samkvæmt þessu samkomulagi getur HR gefið út allar gerðir sem þjálfaðar hafa verið á þessum gögnum, sem og aðrar afleiddar niðurstöður, með opnu leyfi.
2. Samningur milli RÚV og Creditinfo (CI) varðandi efni sem CI umritar: CI er heimilt að afhenda HR bútuðu upptökurnar.
3. Samningur milli CI og HR varðandi efnið sem CI afhendir HR: HR verður útgefandi bútuðu, umrituðu hljóðgagnanna frá CI.

Samningur 2 hefur verið undirritaður, samningur 1 er tilbúinn til undirritunar og HR og CI eru að fara yfir lokaútgáfuna af samningi 3 fyrir undirritun.

Upphaflega var áætlað að gera eitt samkomulag við RÚV varðandi þeirra gögn. Uppsetning keðju, þ.e. hver afhendir hverjum hvaða gögn og með hvaða skilyrðum, þurfti hins vegar að vera vönduð og það tók nokkurn tíma að komast að niðurstöðu. Tveir höfundarréttarhafar eru að umritaða hljóðefninu: RÚV er rétthafi hljóðsins en CI textans. Þar af leiðandi voru gerðir samningar milli allra þátttakenda með það að markmiði að HR geti gefið út allar niðurstöður verkefnisins með opnu leyfi.

H12 – Greinarmerkjasetning og endapunktaskynjun

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Tilgangur

Að afhenda stuðningstól sem sem munu sjálfkrafa bæta úttak talgreina með almennum greinamerkjum og endapunktum.

Varða 2 - lýsing

Góð þjálfunar- og prófunarsett fyrir greinarmerkjasetningarlíkön útbúin. Samanburður á mismunandi líkanahögun fyrir innsetningu greinarmerkja framkvæmdur vísindagrein skrifuð. Kóða fyrir mismunandi högun skilað. Skýrandi greining á notkun hljóðeiginleika fyrir endapunktaskynjun.

Varða 3 – lýsing

Sameinuð eining (e. combined module) fyrir sjálfvirka greinamerkjasetningu í úttaki talgreina og endapunktaskynjun afhent afhentar í samræmi við viðmiðunarreglur hugbúnaðarþróunar verkefnisins. Einingin samanstendur af textabyggðu líkani fyrir greinamerkjasetningu og hljóðeiginleikalíkani fyrir endapunktaskynun. Mat á sameinuðu einingunni, niðurstöðurnar ættu að sýna umtalsverðar framfarir frá M1.

Afurðir

Öllum markmiðum vörðunnar var náð. Verkþættinum er lokið við M3.

Aðalafurðin er Python-pakkinn punctuator-isl. Hann má setja upp með skipuninni: `pip install punctuator-isl`.

Notandinn getur sett inn skrá eða streng frá skipanalínunni til að setja inn greinarmerki og textanum verður skilað með greinarmerkjum annaðhvort til skipanalínunnar eða sem úttaksskrá. Notandinn getur einnig keyrt virknina beint úr Python-skel. Sjá dæmin fyrir neðan. Maður getur valið milli tókaflokkunar *transformers* og uppfærðs Punctuator 2 líkans. Þessi líkön eru geymd á CLARIN, https://repository.clarin.is/repository/xmlui/handle/20.500.12537/52, og hlaðið niður af punctuator-isl í fyrsta skipti sem þau eru notuð. Hlekkur á kóðann sem notaður er til að hreinsa gögnin og keyra líkönin er hér: https://github.com/cadia-lvl/punctuation-prediction

Tvær forskriftir voru útbúnar fyrir endapunktaskynjunina. Önnur til notkunar inni í verkfærakistu Kalda og virkar fyrir tilbúnar upptökur. Hin vinnur með eftirfarandi talgreiningartól: https://github.com/Uberi/speech_recognition, sem gerir notandanum mögulegt að setja inn talgreini að eigin vali sem mun undirbúa hljóðið. Skriftan sem við gerðum vinnur með langt tal í rauntíma, þ.e. fyrirlestra, þar sem við viljum taka upp stöðugt á meðan umritun stendur. Kóðinn er aðgengilegur á: https://github.com/cadia-lvl/end-point-detection

Skýrsla

Við könnuðum þrjú mismunandi líkön sem spá fyrir um greinarmerkjasetningu. Við endurskrifuðum *Punctator 2* Theano model eftir Ottokar Tilk í Tensorflow 2, þar sem Theano er ekki lengur viðhaldið með virkum hætti. Við prófuðum líka tvær gerðir Transformer-líkana. Frá því þau komu til sögunnar árið 2017 hafa Transformer-líkön, sem ekki nota lög af afturvirkum tauganetum, hafa náð framúrskarandi árangri við ýmis verk, þar á meðal greinarmerkjasetningu.

Fyrsta líkanið sem við prófuðum er BERT-byggt líkan frá Hugging Face sem tekst á við vandamálið sem tókaflokkunarvanda (parar orð við gerð greinamerkis eða bil). Við notuðum forþjálfað ELECTRA-líkan frá Jóni Friðriki Daðasyni, þjálfuðu á Risamálheildinni og notuðum Hugging Face uppskriftina til að fínstilla það að spá fyrir um greinarmerki (e. punctuation prediction). Seinna Transformer-líkanið var runu-til-runu líkan frá Fairseq, búið til fyrir tauganetsvélþýðingar sem „þýðir“ ógreinarmerkjasetta texta í greinarmerkjasetta. Bæði líkönin nota PyTorch.

Við gerðum umfangsmikinn samanburð á öllum þremur líkönunum þegar þau voru þjálfuð með sömu gögnunum, sem samanstóðu af 55 milljónum tóka enskra gagna úr Europarl-málheildinni og samsvarandi hluti af íslenskum gögnum úr Risamálheildinni. Runu-til-runu Transformer-líkanið gaf góðar niðurstöður með prófunarsettunum en er óstöðugt og breytir stöku sinnum orðum í textanum, bæði skiptir út orðum og býr til ný orð. Prófanir okkar á textum þar sem orðvillutíðni (e. WER; word-error-rate) er hækkuð viljandi í skrefum sýndu líka að það hrynur algerlega með hærri orðvillutíðni. Þess vegna er það ekki innifalið í pakkanum sem afhentur var og við sleppum því í niðurstöðunum hér fyrir neðan.

Gögnin

Endanlegu þjálfunar- og matsgögnin eru blanda af gögnum úr Risamálheildinni (RMH, sjá undirverkefni G1) og áður hreinsuð gögn frá Alþingi. Sett Risamálheildarinnar inniheldur 164 milljónir tóka, en þar af eru í kringum 14 milljón greinarmerki. Alþingissettið inniheldur 65 milljónir tóka, af þeim eru 4,5 milljónir greinarmerki. Þannig að heildarþjálfunarsettið inniheldur í kringum 18,5 milljónir greinarmerkjatóka. Við einbeittum okkur að því að læra að spá fyrir um þrjú mismunandi greinarmerki: punkta, kommur og spurningarmerki. Taflan sem hér fer á eftir sýnir dreifingu þessara þriggja greinarmerkja.

Greinarmerki	Þjálfun		Prófun	
	Atvik	% af texta	Atvik	% af texta
Punktur	10.600.000	4,6	44,000	4,8
Spurningarmerki	380.000	0,2	1,330	0,1
Komma	7.600.000	3,3	31,000	3,4

Niðurstöður

Öllum niðurstöðum fyrir gagnasöfnin er lýst hér fyrir ofan. Eins og sjá má í töflum 2, 3 og 4 skila nýju greinarmerkjalíkönin í öllum tilfellum betri niðurstöðum en upprunalegu líkönin, auk þess að vera

bæði þjálfuð með vélnámsforritasöfnum sem er stöðugt viðhaldið. Tf2 Punctuator2 og fínstillta ELECTRA transformer-líkanið sem sýnd eru í töflunni hér fyrir neðan eru þau sem voru gefin út á CLARIN.

Tafla 2: F1-einkunnir fyrir upprunalega Punctuator 2 líkanið, Tensorflow-útgáfuna af Punctuator 2 og fínstillta ELECTRA transformer-líkanið

F1 → Líkan ↓	**Punktur**	**Komma**	**Spurningarmerki**	**Samtals**
Original Punctuator 2	89,3	66,1	74,0	80,1
Tf2 Punctuator 2	90,0	69,8	82,3	**82,2**
Transformer	89,0	70,2	81,2	81,8

Tafla 3: Nákvæmniseinkunnir upprunalega Punctuator 2 líkansins, Tensorflow-útgáfu Punctuator 2 og fínstillta ELECTRA transformer-líkansins

Nákvæmni Líkan ↓	**Punktur**	**Komma**	**Spurningarmerki**	**Samtals**
Original Punctuator 2	89,4	72,1	68,7	82,7
Tf2 Punctuator 2	90,1	79,2	85,8	**86,2**
Transformer	89,6	80,1	82,4	86,1

Tafla 4: Heimtareinkunnir (e. recall scores) upprunalega Punctuator 2 líkansins, Tensorflow-útgáfu Punctuator 2 og fíntstilltu ELECTRA transformer-líkansins

Heimt → Líkan ↓	**Punktur**	**Komma**	**Spurningarmerki**	**Samtals**
Original Punctuator 2	89,2	61,1	80,2	77,7
Tf2 Punctuator 2	89,9	62,3	79,0	**78,6**
Transformer	88,5	62,5	80,0	77,9

H14 – Framburðarmállýskur, hljóðgreining og samræðugreind

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Vörður 2 & 3 - lýsing

Að búa til þróunarumgjörð (þróa ramma) fyrir breiðari og dýpri greiningu á tali sem getur gert talgreiningu nákvæmari og auðgað úttak talgreinis. Greining með tölfræðiupplýsingum um framburðarmállýskur og hljóðeiginleika má nota sem inntak fyrir talgreini, annaðhvort til að nota beint sem eiginleika fyrir talgreiningu eða fyrir endapunktaskynjun eða aðrar tegundir af bútun fyrir talgreininn. Auðgað úttak getur innihaldið samræðugreind (e. diarization), málsnið (e. register) og tón (e. tone).

M2: Hljóðgreining á löngum upptökum og samræðugreind málnotenda. Að útbúa forskrift fyrir sæmræðugreind mælanda byggða á átta klukkustunda íslensku samræðugreindargagnasafni. Útbúa hljóðgreiningartól.

M3: Mállýskugreining.

Afurðir

Öllum afurðum fyrir hljóðgreiningu og samræðugreind málnotenda var skilað. Hljóðgreininar- og samræðugreindartólin hafa verið útbúin og gefin út. Undirbúningi er lokið og gagna hefur verið aflað fyrir mállýskugreiningu, en endanlegar niðurstöður greiningarinnar verða afhentar fyrir lok október 2020, með afurðum M3 undirverkefnisins G6 (framburðarorðabók). Verkþættinum er lokið fyrir M3 sem útvegar verðmæt tól og greiningar fyrir frekari þróun talgreinakerfa fyrir íslensku.

Skýrsla

Samræðugreind málnotenda

Að útbúa samræðugreindartól íslenskra málnotenda er tvíþætt: 1. að útbúa samræðugreindargagnasafn og 2. að útbúa leiðbeiningar fyrir samræðugreind.

Þar sem engin íslensk samræðugreindargagnasöfn eru til ákváðum við að búa til átta klukkustunda byrjunargagnasafn úr útsendingargögnum RÚV sem kallast Rúv-di. Þetta fól í sér að afmarka mælendur og merkja málnotendur. Nemendur voru ráðnir til starfa í sumar í gegnum áætlun Vinnumálastofnunar um sumarstörf fyrir námsmenn. Þeir tvöfölduðu upphaflega átta klukkustunda samræðugreindargagnasafnið og útbjuggu skriftur til að auðvelda öðrum að búa til samræðugreindargagnasafn með hálfsjálfvirkum hætti. Skrifturnar eru aðgengilegar á Diar-AZ: https://github.com/cadia-lvl/diar-az. Rúv-di merkta gagnasafnið var síðan notað í leiðbeiningum til að þjálfa líkön og meta villutíðni samræðugreindar (e. DER; diarization-error-rate).

Við notuðum Kalda til að útbúa forskriftirnar, sama tól og verður notað fyrir talgreina, til að lágmarka vinnu við samþættingu hugbúnaðarkerfa. Forskriftirnar eru sett af skriftum notaðar á gagnasafn til að búa til samræðugreindarlíkön eða til að afkóða ný gögn, í þessu tilfelli til að skipta niður í

mælendur (e. diarize). Við gerðum fjórar forskriftir. Fyrstu tvær hljóðgreina ný gögn sjálfvirkt. Þær eru byggðar á núverandi líkönum og samræðugreindarforskriftanna fyrir ensku. Seinni tvær forskriftirnar útbúa samræðugreindarlíkön byggð á íslenskum gagnasöfnum: Rúv-di og Alþingisræðum. Seinni tvær forskriftirnar má einnig nota á önnur gögn með því að fylgja sama gagnasniði. Kaldi forskriftirnar má finna í þessari kóðahirslu: https://github.com/cadia-lvl/kaldi-speaker-diarization.git og íslensku líkönin verða gefin út á CLARIN þegar samningur milli RÚV og HR er fullfrágenginn.

Úttak niðurstaðna samræðugreindarforskriftanna eru Rich Transcription Time Marked (rttm) skrár. Þær eru bornar saman við gullstaðal rttm-skráa úr merkta gagnasafninu, Rúv-di til þess að reikna DER (villutíðni samræðugreindar, sjá hér fyrir ofan). Því lægri sem DER er, því betra. Með því að nota sama matssett fyrir mismunandi forskriftir er okkar besta DER 22,58%. Það eru niðurstöður úr íslensku líkönunum okkar, sem er meira en 18% framför frá besta enska líkaninu. Margar af þeim villum sem eftir standa eru til komnar vegna þess að forskriftirnar merkja tal sem skarast ranglega. Þetta hefur enn ekki verið leyst.

Íslensku forskriftirnar eru stórt skref í þá átt að fækka þeim umritunarvillum sem talgreinar gera þegar greina þarf tal frá mörgum mælendum.

Hljóðgreining

Þegar kemur að hljóðgreiningu lögðum við áherslu á að greina einkenni hljómfalls. Þau skref sem við tókum til að búa til okkar eigin útdráttartól voru þríþætt: fyrst kynntum við okkur bestu tólin og aðferðirnar við að greina einkenni hljómfalls, síðan þróuðum við leiðbeiningar úr þessum tólum og gáfum að lokum út eitt slíkt, sniðið að okkar þörfum.

Við könnuðum tól sem greina hljóðeiginleika: Praat, Kaldi og DisVoice. Þegar við prófuðum Kaldi ákváðum við að best væri að nota Kaldi ekki beint. Í þessu tilfelli var áætlunin sú að fylgja forskriftum úr öðru verkefni á GitHub: https://github.com/jcvasquezc/DisVoice. Útdrætti hljóðeinkenna með Praat er verið lokið og inniheldur upplýsingar um málhafa úr RTMM-skrá. Næst þróuðum við forskriftir sem nothæfar eru sem inntak fyrir kerfi talgreina.

Að lokum ákváðum við að gefa út tól okkar til útdráttar einkenna hljómfalls sem byggt er á Praat. Það má nálgast hér: https://github.com/cadia-lvl/kaldi-speaker-diarization.git. Það lýsti hljóðeinkennum málhafa best. Tólið skilar hljómfallseinkennum með málhafamerkjum með nauðsynlegri rttm inntaksskrá. Eftir að hafa fengið inntakið markar tólið auðkenni málhafa með hljómfallseinkennum á 0,01 sekúndu fresti á upptökunni. Hljómfallseinkennin sem dregin eru fram eru tónhæð og *harmonicity*, bæði greind með eiginfylgni (AC) aðferðinni, og styrkur (e. intensity). Fyrir nánari lýsingu, sjá LESIST á GitHub-hirslunni.

Mállýskur

Mállýskur eru ekki mjög áberandi í íslensku og tíðni þeirra hefur lækkað hratt undanfarna áratugi. Engu að síður eru nokkrar mállýskur sem vert er að skoða í samhengi máltækniáætlunar fyrir íslensku. Sú afstaða hefur verið staðfest eftir eftir nokkra óformlega upplýsingaöflun, t.d. með samtölum við einn helsta sérfræðing Íslands, Eirík Rögnvaldsson. Með viðbótarupplýsingum frá vefsíðunni mállýskur.is, eldra átaki í mállýskurannsóknum og enn fremur með staðfestingu frá rannsakendunum Höskuldi Þráinssyni og Bjarka Karlssyni, hefur sú staðhæfing verið staðfest.

Þess vegna var útbúin rannsóknaráætlun fyrir „harðmæli“, algengustu mállýskuna sem töluð er á Norðurlandi. Prófanir verða gerðar með því að nota talgreini fyrir meirihlutaframburð íslensku til að umrita sumar af upptökunum sem safnað var í þessu verkefni og bera saman orðvillutíðni til að dæma um frammistöðuna. Samanburðurinn verður gerður með því að bera upptökur sem

einskorðast við mállýskur saman við upptökur með meirihlutaframburði. Auk þess verður gerður samanburður á þessum niðurstöðum með því að beita framburðarorðabókinni sem þróuð var á fyrsta ári verkefnisins, verkþáttur G6 - *Framburðarorðabók*.

Gagngjafar

Upptökurnar af mallyskur.is voru sóttar og greindar. Í ljós kom að gæði þeirra voru ekki nægilega mikil fyrir umritun með talgreini. Í raun, eins og lýst var af höfundum þeirra, reyndust þær frekar vera sagnfræðileg heimild um mállýskur en verkfæri til umritunar með talgreinum.

Tveir aðrir gagnagjafar fyrir mállýskurannsóknir hafa verið gefnir út í gegnum gagnaöflun verkþáttarins *H1 - Upptökur með Samrómi* og *T1 - Upptaka á einingavalsgögnum*. Fyrir gögn sem aflað var í gegnum H1 var reynt að ná til einstaklinga sem töluðu mállýskur með því að auglýsa í staðbundnum fréttatímum eftir mállýskuhöfum til að taka þátt í gegnum lýðvirkjun (e. crowdsourcing) á samromur.is. Rétt rúmlega 800 upptökur fengust með þessari aðferð og áætlað er að nota þessar upptökur og svipaðan fjölda upptaka úr gagnasafninu á samromur.is sem samanburð. Seinni gagnagjafinn úr T1 er úr einingavalsgögnum þar sem 8 málhafar voru ráðnir til þess að tala inn 40 klukkustundir hver. Tveir þessara málhafa voru valdir vegna norðlensks framburðar þeirra og þess vegna eru til gæðagögn tveggja mállýskuhafa og sex málhafa þess sem álitinn er meirihlutaframburður. Að auki tóku þessir málhafar upp sömu lista af setningum (e. prompt sets) þannig að sömu segðir megi nota til að bera saman orðtíðnivillur.

Niðurstöður þessarar greiningar verða notaðar til að bera kennsl á nauðsyn þess að laga talgreinana að mállýskum í íslensku, til dæmis með framburðarorðabókum. En þar sem þessi úrræði eru hluti af afurðum M3 og um það bil að verða afhent, þá getum við ekki búist við alhliða greiningum á mállýskum, eins og áætlað var, fyrr en M3 er lokið.

T1 – Upptaka á einingavalsgögnum

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík, RÚV

Vörður 2 & 3 - lýsing

M2: Samkomulagi náð við alla átta mælendur. Sex þátttakendur hafa lokið upptökum, 20 klukkustundum tals hver.

M3: Upptökum á átta þátttakendum að fullu lokið. Gögn frá minnst fjórum mælendum útgefin með viðeigandi lýsigögnum samkvæmt stöðlum CLARIN.

Afurðir

Öllum markmiðum vörðunnar var náð.
Átta mælendur hafa lokið upptökum á um 20 klukkustundum tals hver, alls 160 klukkustundum. Það eru fjórar kven- og fjórar karlraddir með tveimur mismunandi mállýskum, sjá töflu 1 fyrir nánari útskýringar.

	kyn	aldur	mállýska	upptökur	# klst (kk:mm)[49]
v0	karl	60	Linmæli	15.622	25:31
v1	kona	60	Linmæli	13.297	15:14
v2	karl	35	Linmæli	19.859	18:41
v3	kona	26	Linmæli	20.060	20:32
v4	karl	50	Linmæli	13.991	18:05
v5	kona	71	Linmæli	13.699	21:03
v6	karl	39	Harðmæli	17.397	19:43
v7	kona	33	Harðmæli	16.895	20:47

Tafla 1: Nákvæmt yfirlit radda

Fjórar raddir, v1-4, eru tilbúnar til útgáfu með viðeigandi lýsigögnum.

Upptökurnar sem eru í málheildinni eru klipptar svo að hvert radddæmi sé byrji og endi á 300 ms þögn. Að auki hafa allar þagnir í miðjum upptökum verið styttar niður í 300 ms. Þetta var gert sjálfkrafa með þvingaðri samröðun með hjálp Montreal Forced Aligner til að finna þagnir.
Allt hljóð var tekið upp í 44kHz og verður gefið út á því formi.
Textaskrár með bæði upprunalegum og stöðluðum textum eru innifaldar. Allar upplýsingar um mælendur hafa verið fjarlægðar fyrir utan aldur, kyn og mállýsku.
Verkþátturinn heldur áfram á öðru ári verkefnisins og lýkur þá.

[49] Klukkustundir eru reiknaðar sem heildarlengd hljóðsins, fyrir klippingu, mínus 1 sekúndu sem ljósmerki í upptökuhugbúnaðinum leggur til að mælendur bíði við hvern enda.

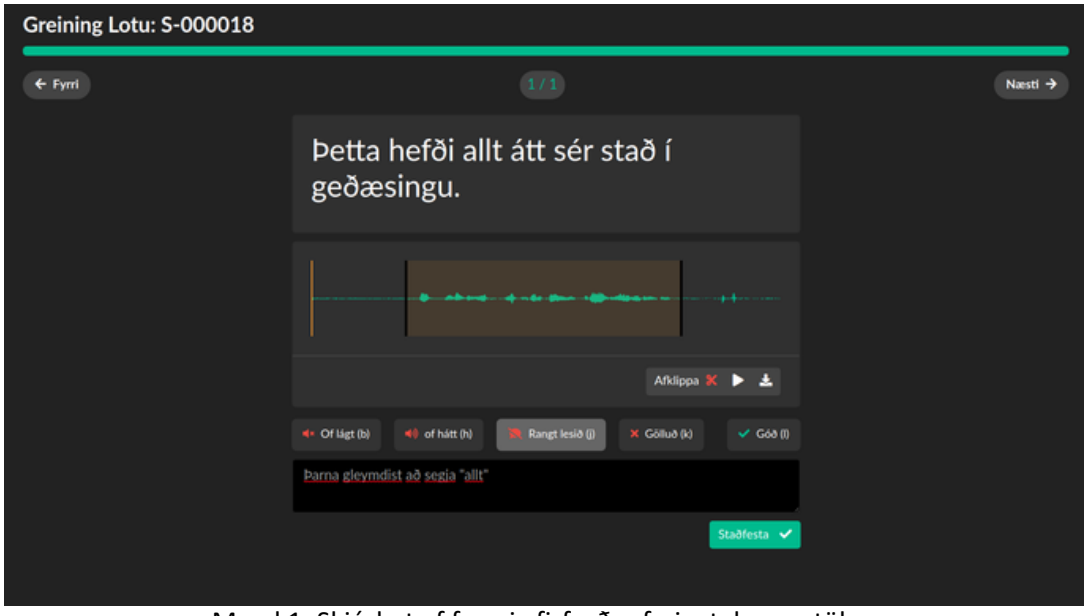
Verkefnið hélt greiðlega áfram eftir fyrstu vörðu, þar sem næsti hópur mælenda var ráðinn fyrir lok mars. RÚV viðhélt ströngum aðgangstakmörkunum vegna COVID-19 og aðeins ómissandi starfsfólki var hleypt inn í bygginguna svo við þurftum að aflýsa öllum upptökum. Augljóslega olli þetta töfum á upptökum.

Upptökur byrjuðu aftur í maí og héldu áfram eftir það.

Annað hljóðver með sömu uppsetningu var sett upp á Akureyri til að taka upp síðustu tvær raddirnar, með staðbundinni mállýsku.

Yfir sumarið bættum við kerfi til yfirferðar allra radda í upptökuhugbúnaðinn.

Þeir sem unnu við yfirferð gátu merkt upptökur sem góðar eða slæmar eftir gæðum þeirra. Slæmum upptökum var safnað saman í einn af fjórum flokkum - ósamræmi hljóðs við texta, of háværar, of lágværar, og hljóðgalla eða önnur vandamál. Hluti af ferli yfirferðar fólst í handvirkri klippingu þagna úr byrjun, miðju og enda upptaka. Samtals hafa rétt undir 16 þúsund upptökur verið yfirfarnar.



Mynd 1: Skjáskot af formi yfirferðar fyrir staka upptöku.

T2 – Gögn fyrir raddblöndun

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík, RÚV

Vörður 2 & 3 - lýsing

M2/M3: Handrit til upplesturs tilbúið. Færanleg upplestursaðstaða uppsett. Samtals 20 þátttakendur teknir upp.

Afurðir

Vörðum er ekki lokið.

Í dag er handritið tilbúið til upplesturs, veflægt upplestursumhverfi hefur verið sett upp til að afla mögulegra þátttakenda, en enginn þátttakandi hefur hafið upptökur.

Nú standa yfir prufur á netinu sem við munum velja þátttakendur úr. Meðlimir þriggja kóra eru að lesa 20 setningar hver í umsóknarferli á netinu. Nokkrir kórar til viðbótar hafa verið beðnir að taka þátt í prufunum og við metum sem svo að við munum hafa úr um 120 mismunandi röddum að velja við lok september.

Val á þátttakendum hefst í byrjun október. Fyrsti hópur þátttakenda mun hefja upptökur um leið og allir þátttakendur hafa verið valdir.

Upptökurnar verða gerðar í Stúdíó 9 hjá RÚV, á sama stað og upptökur í T1, með sömu uppsetningu.

Næstu vörður (4 og 5) kalla á 20 upptökur til viðbótar, samtals 40 fyrir allar vörður. Við áætlum að upptökur allra 40 radda fyrir verkefnið muni taka um 10 vikur. Þar sem upptökuverið verður sett upp og allar 40 raddirnar valdar, er rökrétt að klára allar upptökurnar í einu. Við stefnum á að hafa 20 raddir tilbúnar fyrir 15. desember, og ljúka þar með M3, og hinar 20 fyrir enda vörðu 4 (febrúar 2021).

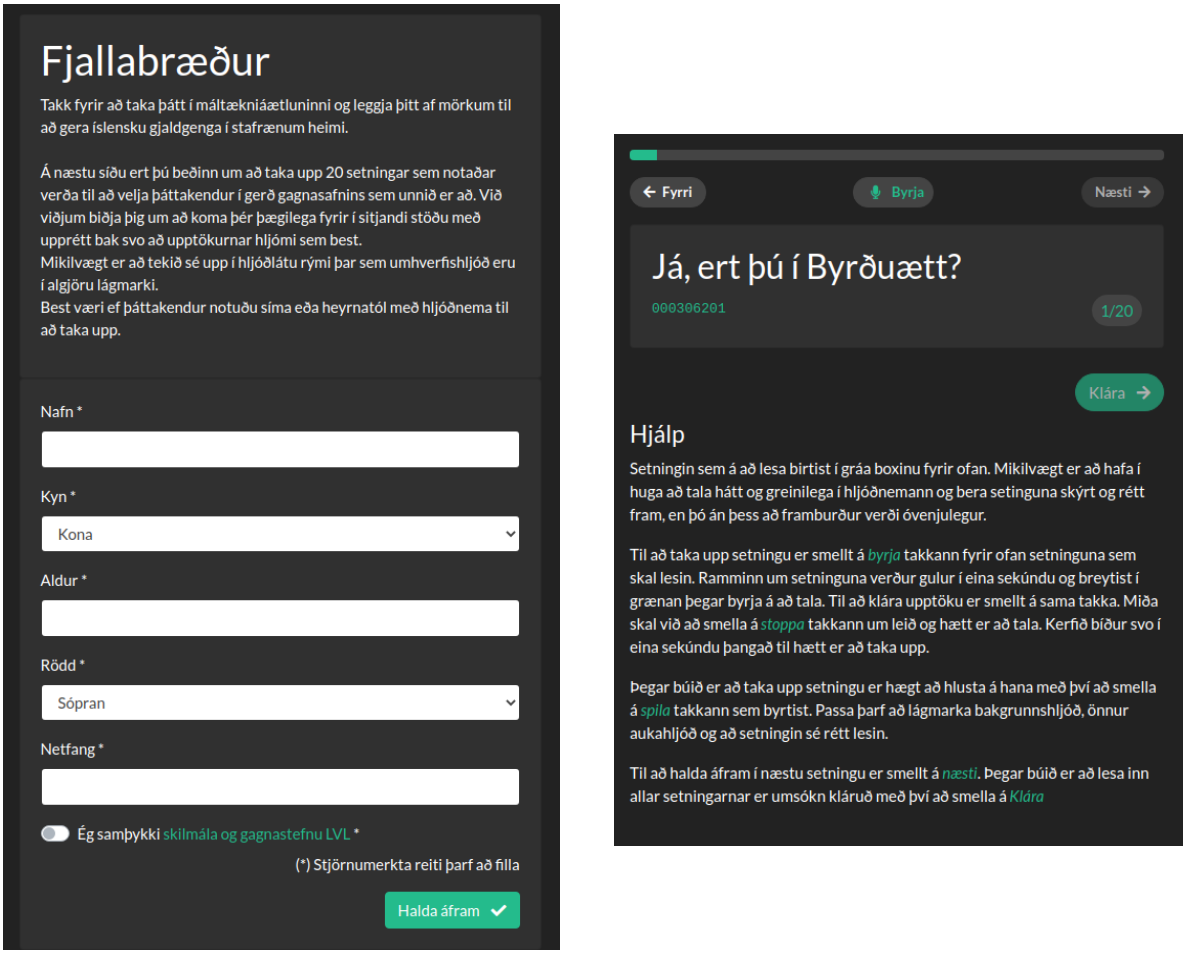
Skýrsla

Meginástæða tafarinnar er lokun hljóðversins vegna COVID-19 og önnur vandamál vegna faraldursins. Áætlunin að ofan miðar að því að ná fyrri áætlun eins fljótt og auðið er, án þess að fórna gæðum lokaafurðarinnar.

Fyrir prufurnar innleiddum við skráningarform og nýtt upptökuviðmót í LOBE, upptökuhugbúnaðinn. Viðmótið er svipað því sem notað er til upptöku í hljóðveri en hefur verið einfaldað eins mikið og hægt er til að auðvelda notkun.

Umsækjendur lesa lista setninga sem þeim er gefinn í umsókninni. Fyrir umsóknina völdum við tuttugu setningar sem hljóma eðlilega[50] efst af leslista sem notaður var fyrir raddir í T1 og verður notaður fyrir þetta verkefni. Umsóknum frá hverjum kór er safnað í aðskilin söfn margra mælenda.

[50] Sumar setninganna efst af listanum er erfitt að lesa eða hljóma óvenjulegar. Til að halda víðu sviði tvístæða höldum við okkur við efsta hluta listans en völdum setningar sem er auðveldara að lesa greiðlega og án vandamála.



Mynd 1: Umsóknarformið og einfaldað upptökuviðmót fyrir umsóknir á netinu.

Eftirfarandi er ítarlegra yfirlit mikilvægra þátta áætlunarinnar sem enn hafa ekki komið til framkvæmda.

Val þátttakenda verður gert hálfsjálfvirkt með notkun einhvers forms klösunar hljóðfræðilegra eiginleika sem safnað verður úr umsóknunum. Eiginleikana má taka saman annaðhvort úr þjálfuðu i-vector líkani eða handvirkt völdum hljóðfræðilegum breytum eins og tónhæð eða blöndu beggja.

Umsóknirnar innihalda aldur og kyn mælenda svo getum fengið jafna dreifingu beggja breyta.

Við ákváðum að nota sama leslista fyrir alla þátttakendur, sama lista og raddirnar átta í T1 notuðu. Með þessu hámörkum við dreifingu tvístæða fyrir alla þátttakendur á kostnað fjölbreytni í sameinuðu setti. Þetta hefur kosti og galla eftir því hvernig gögnin verða notuð.

T6 – Veflesari

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Varða 3 – lýsing

Hugbúnaðarhönnun veflesara tilbúin.

Markmið

Að útbúa talgervil sem getur lesið vefsíður. Meginmarkmiðið er að þróa tól sem gerir talgervingu fyrir íslensku aðgengilega fyrir vefhönnuði til þess að nota á vefsíðum sínum. Tvö aukamarkmið eru að búa til viðbót fyrir vafra og að hvetja til notkunar á opnum íslenskum talgervli meðal vefhönnuða þriðju aðila. Á fyrsta ári þessa verkþáttar verður veflesarinn hannaður og á síðari stigum fullkláraður og bættur, eftir því sem vinnu við talgervil verkefnisins miðar áfram.

Afurðir

Öllum markmiðum vörðunnar var náð. Hugbúnaðarhönnun veflesara er tilbúin. Við höfum undirbúið ýmis hönnunargögn.
Vinna við T6 heldur áfram á öðru ári verkefnisins.

Skýrsla

Áður en við hönnuðum veflesara rannsökuðum við tiltæka veflesara sem eru í notkun um allan heim, en sérstaklega á Íslandi. Samkvæmt okkar heimildum er WebReader frá ReadSpeaker víðtækasti almenni veflesarinn fyrir íslensku. Því væri best að ná fram sömu eiginleikum til að veita notendum nánast hnökralausa upplifun þegar þessir veflesarar eru notaðir á íslenskum vefsíðum. Hönnun okkar gerir ráð fyrir því að veflesarinn verði tengdur við tiltæka vefþjónustu talgervils eins og AWS Polly eða forritaskil talgervils sem gerð verða innan undirverkefnis T4 - Vefgáttir fyrir talgervla.

Eftirfarandi er samantekt hugbúnaðarhönnunar okkar. Veflesarinn er nefndur Web Reader Icelandic (WebRICE). Hugbúnaðarhönnun okkar er tvískipt, þarfir og eiginleikar, og skýringarmyndir.

Þarfir og eiginleikar

Þarfir og eiginleikar okkar eru bæði fyrir WebRICE og undirliggjandi forritaskil talgervils. Heildargátlista má finna á https://languageandvoice.files.wordpress.com/2020/09/checklists.pdf.

WebRICE

Við höfum skipt hugbúnaðarþróuninni í þrjú stig: frumgerð, afurð með lágmarkseiginleika, og endanlega afurð. Frumgerð okkar leggur áherslu á virkni, veflesari sem getur spilað, gert hlé og stöðvað, en ekki mikið meira. Afurð með lágmarkseiginleika leggur áherslu á notendavænan veflesara sem virkar með öllum helstu vöfrum. Fyrir endanlega afurð okkar viljum við veflesara sem er jafn góður og WebReader frá ReadSpeaker, ef ekki betri. Innan hvers stigs eru markmið og gátlistar fyrir þarfir og eiginleika. Gert er ráð fyrir að hvert stig nái yfir allt úr fyrri stigum.

Forritaskil talgervils (e. TTS API)

Í þessari þarfagreiningu teljum við upp lágmarkseiginleika fyrir vefgáttir T4 til að virka með WebRICE. Þau fela í sér að skila talmerkjum til þess að geta litað textann jafnóðum og hann er lesinn og að skila hljóði innan hæfilegs tíma.

Skýringarmyndir

Skýringarmyndir okkar fela í sér eitt klasarit og tvö runurit. Klasaritið útskýrir uppbyggingu klasa í WebRICE á meðan runuritin útskýra samspil milli klasa. Skýringarmyndirnar má finna á https://languageandvoice.files.wordpress.com/2020/09/webrice_diagrams.pdf. Skýringarmyndirnar eru skrifaðar í UML[51],[521,2].

WebRICE klasarit

Þetta rit sýnir hvernig klasar í WebRICE líta út og hvernig þeir eru tengdir. **Í stuttu máli** samanstendur *Reader* af 7 hnöppum. Þessir hnappar erfa allir frá hugræna *MainButton* klasanum. Svo eru hinir klasarnir okkar. *UserHighlightPopUp* gerir notandanum kleift að hlusta á texta sem er litaður. *HighlightTracker*, sem er útbúinn af Reader, sér um litun lesins texta. *SpeechManager* hefur samskipti við forritaskilin og að lokum sér *ClientStoreManager* um geymslu á hlið notandans. Punktalínan gefur til kynna ákvæði. Til dæmis: a --> b þýðir að a er háð b.

WebRICE onPlay runurit

Þetta rit sýnir hvað gerist þegar notandi smellir á aðal spilunarhnapp. **Í stuttu máli**, þegar notandinn smellir á „spila“ athugum við hvort notandinn hefur spilað þennan texta á vefsíðunni áður. Ef svo er sækjum við hljóðbútinn og talmerkingar (e. speech marks) úr geymslu biðlara. Ef ekki sækjum við hann frá forritaskilum. Við athugum svo hvort notandinn vill hafa sjálfvirkt skrun og/eða litun texta aðgengilega. Ef svo er virkjum við þá möguleika.

WebRICE onDOMContentLoaded runurit

Þetta rit sýnir hvað gerist þegar notandi hleður vefsíðu sem inniheldur WebRICE. **Í stuttu máli** athugum við hvort textinn sem notandinn hefur þegar hlustað á er sá sami og hann getur lesið nú. Svo búum við til tvo hnappa. Ef þemum fyrir *HighlightTracker* hefur ekki verið hlaðið í geymslu biðlara gerum við það. Svo búum við til *HighlightTracker*.

[51] UML-klasarit https://www.uml-diagrams.org/class-diagrams-overview.html

[52] UML-runurit https://www.uml-diagrams.org/sequence-diagrams.html

T7 – Forskriftir að röddum

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Vörður 2 & 3 - lýsing

M2: Frumgerð forskriftar einingavals tilbúin.
M3: Fyrsta útgáfa forskriftar tilbúin.

Afurðir

Öllum markmiðum vörðunnar var náð.
Fyrsta útgáfa forskriftar fyrir The Festival Speech Synthesis System er tilbúin fyrir raddirnar fjórar sem gefnar voru út í T1. Forskriftirnar eru aðgengilegar á GitHub síðu LVL: https://github.com/cadia-lvl/unit-selection-festival/
Vinna við T7 heldur áfram á öðru ári verkefnisins.

Skýrsla

Það voru nokkrir mismunandi valkostir í boði til að búa til forskrift fyrir einingaval. Helst ber að nefna Festival/Festvox, MaryTTS og Merlin. Fyrir fyrstu útgáfu ákváðum við að nota Festival sem virðist í mestri notkun sögulega og er enn nokkuð vel haldið við.

Fyrir forskriftina þurftum við að þróa tvö aðföng að auki, utan hljóðgagnanna til þess að hafa fullbúinn talgervil: líkan rittákns-til-fónems (g2p) og ítarlega töflu sérkenna fyrir öll fónem í íslenska tungumálinu.

Fyrir töflu sérkenna fónema var hverju fónemi gefið gildi fyrir 11 mismunandi máleiginleika (e. linguistic attributes). Þessi gildi eru notuð af einingavalslíkani til að velja hvaða hljóð skal nota saman. Eiginleikarnir og öll möguleg gildi eru skjöluð innan kóðans og yfirlit er aðgengilegt í skjöluninni (phonemes.md).

Fyrir g2p-líkanið notuðum við Sequitur G2P líkan þjálfað á framburðarorðabók sem var samsett sérstaklega fyrir talgervil. Þjálfaða g2p-líkanið er ekki hluti af afurðunum en möguleiki þess að þjálfa slíkt líkan er innifalinn í forskriftinni. Við munum þróa þennan hluta verkþáttarins með undirverkefnum G6 (Framburðarorðabók) og T10 (Sjálfvirk hljóðritun) á öðru ári.

T13 – Stikaðir talgervlar

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Varða 3 – lýsing

Skýrsla um enda-til-enda/stikaða nálgun við íslenska talgervingu.

Afurðir

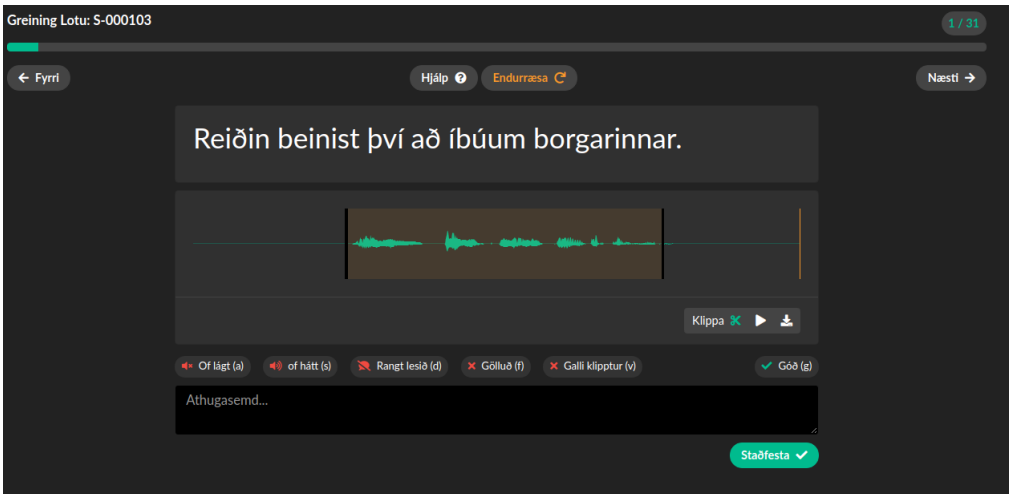
Öllum markmiðum vörðunnar var náð.
Könnun á bestu nálgunum við talgervingu hefur verið framkvæmd og tilraunir með gögnum úr T1 eru hafnar. Endanleg afurð fyrir T13 verða 8 aðgengilegar talgervilsraddir með heildstæðum framenda og hlutdrægar matsskýrslur í formi MOS, MUSHRA eða A/B prófana. Vinna við T13 heldur áfram á öðru ári verkefnisins.

Skýrsla

Val á þátta-tauganeti (e. feature network) fyrir þetta verk var nokkuð einfalt, við völdum að skoða Tacotron 2 (sjá: https://arxiv.org/abs/1712.05884) þar sem það telst fremst í flokki þegar kemur að talgervingu (sjá https://paperswithcode.com/sota/speech-synthesis-on-north-american-english).

Hirslu fyrir talgervingu Mozilla (sjá https://github.com/mozilla/TTS) er vel viðhaldið og sveigjanleg fyrir þróun með notkun Tacotron 2 líkansins. Við völdum þessa hirslu sem grunnstoð. Hún býður upp á bæði Tensorflow og PyTorch útfærslur og við kusum PyTorch.

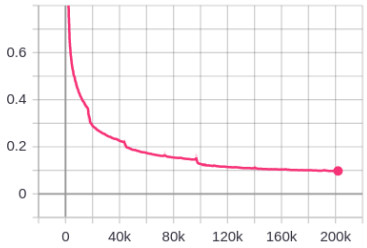
Þar sem gagnaöflun fyrir T1 stóð enn yfir lögðum við áherslu á gögn úr fyrstu tveimur röddunum sem hafði þegar verið aflað. Það er mikilvægt í enda-til-enda (e. end-to-end) talgervingu (og flestum öðrum gerðum talgervingarlíkana) að klippa úr allar þagnir í upphafi og enda. Mismunandi lengd þagnar við hvern enda veldur slæmum árangri tímalengdarlíkans. Þagnir má klippa úr handvirkt eða sjálfvirkt með notkun einfaldrar leitaraðferðar. Vandamálið við seinni aðferðina er að árangur slíkrar leitaraðferðar er aðeins hægt að meta handvirkt, sem tekur langan tíma. Af þessari ástæðu ákváðum við að klippa handvirkt eins margar upptökur og mögulegt var með hjálp sumarnema. Nemendurnir náðu að klippa 15.956 upptökur. Þetta mætti nota sem matssett gagna fyrir leitaraðferð klippinga eða til að þjálfa ákveðið klippingalíkan. Sú vinna er ekki hafin eins og er en mun að öllum líkindum verða hluti af næstu vörðu.



Mynd 1: Nemandi klippir hverja upptöku sem hann samþykkir.

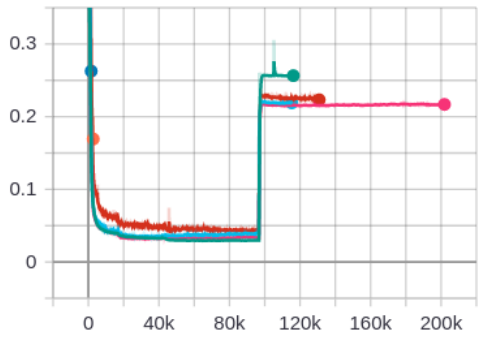
Í stað ítarlegrar lýsingar á Tacotron 2, þar sem upprunalega ritið er mjög ítarlegt, ætti stutt yfirlit að nægja. Tacotron 2 er í grunninn end-to-end, runa-til-runu, sjálfkrafa-afturvirkt (e. auto-regressive) *feature network* hannað sérstaklega fyrir talgervingu. Mikilvægur munur á Tacotron 1 og 2 er að það seinna skilyrðir WaveNet fyrir spár frá *feature network*. Þessi skilyrðing skýrir einstaklega háa MOS-einkunn sem náðist á verki bandarískrar ensku upp á 4,53.

Við gerðum engar stórar breytingar á grunnlíkaninu fyrir tilraunir okkar. Þjálfun var dreift yfir þrjú skjákort og gert að halda áfram þar til ákveðnum tapþröskuldi (e. loss threshold) var náð, en þó ekki lengur en í 10 daga.



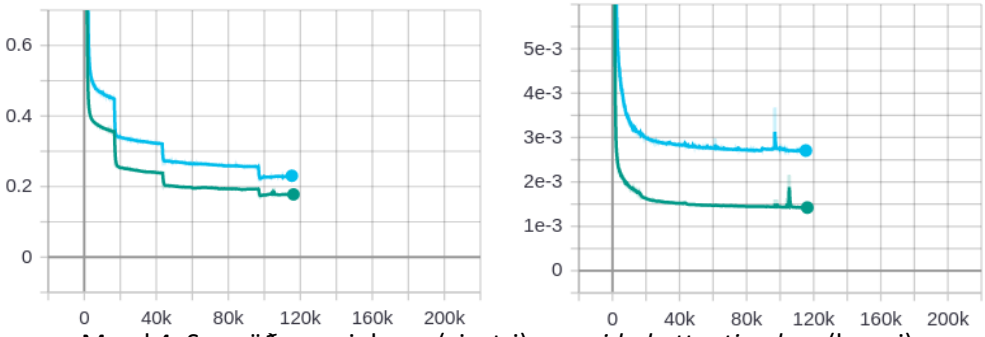
Mynd 2: Tap afkóðara sem fall af altækum skrefum (e. global steps) (lægra = betra)

Tilraunin var að mestu vel heppnuð og marktækt minna tap var greint á flestum sviðum. Mjög óheppilegt var að stoppnetið (e. stopnet), eining sem spáir fyrir um hvenær skal hætta að senda nýja ramma, lenti alltaf í furðulegum vandamálum eftir u.þ.b. 100.000 altæk skref. Þetta gerðist stöðugt fyrir allar mismunandi raddirnar, með mismunandi stillingum, jafnvel með því að nota handahófskennt vægi frumstillingar (e. random weight initialization) og handahófskennda samsetningu gagnabunka. Við höfum ekki fundið skýringu á því hvers vegna þetta gerist en aðrir rannsakendur sem nota Mozilla TTS hirsluna hafa orðið varir við álíka hegðun. Þetta gefur til kynna að stoppnetið sé illa forritað eða grundvallarvandamál sé í gagnasettinu. Hið síðarnefnda er ólíklegt en þó mögulegt. Þar sem gögnin hafa ekki enn verið klippt og sannreynd ákváðum við að klippa einfaldlega enda hverrar upptöku við ákveðinn hljóðstyrk. Þessi mjög svo einfalda aðferð við klippingu þýðir að líklega eru upptökur þar sem klippingin er of ágeng (t.d. ef mælandinn talar mjög lágt) eða ekki nógu ágeng (t.d. ef mælandinn hóstar óvart í byrjun upptöku og bíður áður en hann byrjar að tala). Þetta skýrir hins vegar ekki, a.m.k. ekki að okkar mati, hvers vegna við sjáum þetta gerast ítrekað í kringum 100.000asta altæka skrefið. Engar áætlaðar breytingar í yfirstillingum eiga sér stað í kringum þetta skref og því er þessi hegðun ekki skýrð með því. Þetta lýsir myrku hlið enda-til-enda líkana, þegar eitthvað fer úrskeiðis er oft mjög erfitt að leiðrétta þau. Mikilvægt er að komast að því hvers vegna þetta gerist, með tilliti til vinnu í framtíðinni.



Mynd 3: Stoppnetstap (e.stopnet loss) sem fall af altækum skrefum (lægra = betra)

Inntakið í líkan Tacotron 2 er einfaldlega textinn sem er lesinn. Þetta er kosturinn við enda-til-enda líkön, þar sem engin þörf er á flókinni útfærslu. Hins vegar má aðstoða Tacotron 2 í að ákvarða samröðun milli rittákna og hljóðeiginleika ef gott g2p líkan er til staðar. Við erum með g2p líkan til bráðabirgða, svo við ákváðum—í stað þess að þjálfa á textanum—að þjálfa á viðeigandi hljóðritunarmerkjum (e. phonetic labels) sem g2p líkanið segir til um. Við gerum ráð fyrir að tap samröðunarlíkansins ætti að minnka hraðar með því að nota hljóðritunarmerki sem inntak frekar en rittákn, vegna þess að þau eru skýrari.



Mynd 4: Samröðunareinkunn (vinstri) og *guided attention loss* (hægri) fyrir líkan byggt á rittáknum (ljósblátt) á móti líkani byggðu á hljóðritunarmerkjum (ljósgrænt) (lægra = betra)

Og þetta er einmitt það sem við sjáum. Athyglislíkanið (e. attention model) sem liggur til grundvallar þessa runu-til-runu líkans er mikilvægur þáttur þess að ná fram góðum niðurstöðum með notkun líkansins í heild sinni og við mælum því með notkun hljóðritunarmerkja sem inntaks í runu-til-runu líkön fyrir talgervingarverkefnið.

Niðurstöður tilrauna úr þjálfun Tacotron 2 líkans á fyrstu röddunum eru aðgengilegar hér: https://drive.google.com/drive/folders/1jR_HgsAp0fTFQGclJ6FlRRDitR8b_FHA?usp=sharing.

Viðtal við Ríkisútvarpið um undirverkefni T1 og T12 er aðgengilegt hér: https://lvl.ru.is/2020/07/10/learn-about-our-tts-process-as-described-by-atli-thor/.