Samstarf um íslenska máltækni
SÍM

Máltækniáætlun fyrir íslensku

Varða 6 – Lokaskýrsla

1. október 2021

Efnisyfirlit

Formáli

Þessi skýrsla dregur saman við vörðu 6 (M6) afurðir fyrir hvern verkþátt sem skilgreindur var í samningnum milli Almannaróms og SÍM (Samstarf um íslenska máltækni) frá júní 2020, í Endurskoðuðum vörðum (e. Revised Milestones) fyrir annað verkefnisár (Y2) frá nóvember 2020, og í Endurskoðuðum vörðum frá mars 2021.

Á öðru ári er aukin áhersla lögð á undirbúning afurða fyrir hugbúnað. Innan SÍM nota verkefnahópar afurðir annarra hópa, og málefni tengd bæði þörfum og þoli/auðveldleika í notkun (e. robustness/ease-of-use) eru og verða rædd og leyst milli hópanna. Miðeind er að gera tvær beinar notendatilraunir með notendum þriðja aðila: málrýni í ritstjórnarumhverfi dagblaðs (sjá L7-skýrslu), og vélþýðingar í faglegu stjórnsýsluumhverfi (sjá V4-skýrslu). Þessir hlutar tengjast tæknilegri nægð afurðanna, auk fyrstu notendaupplifana.

Í öllum verkþáttum, utan sjö, var markmiðum M6 skilað að fullu eða með smávægilegum frávikum. Áætlanir hafa verið gerðar til að ná almennum markmiðum þessara verkefna fyrir M6 fyrir enda ársins 2021, sjá töflu að neðan og viðkomandi skýrslur. Sú vinna sem vantar upp á hefur verið áætluð þannig að hún hafi ekki áhrif á tilföng þriðja árs.

Verkþáttur	**Seinkuð skiladagsetning**
V4e – Þýðingarvél fyrir sérsvið	2021-11-15
L10 – Aðlögun að máltæknihugbúnaði	2021-12-31
H1 – Upptökur með Samrómi	2021-12-31
H3 – Umritun samræðna og fyrirspurna	2021-12-31
H4 – Umritun fyrirlestra	2021-12-31
T2 – Gögn fyrir raddblöndun	2021-10-31
T13 – Stikaðir talgervlar	2021-12-31

Þessi skýrsla inniheldur lýsingar, umræður og niðurstöður vinnu við 40 verkþætti. Vegna umfangs efnisins miðum við að því að gefa skýra yfirsýn yfir stöðu og afurðir hvers verkþáttar á fyrstu síðu viðkomandi kafla.

G1 – Risamálheildin

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

Markmið

Að halda áfram að stækka RMH, bæði með því að uppfæra hana með nýjum gögnum frá fyrri gagngjöfum og með því að bæta við nýjum gagnagjöfum og textagerðum. Að aðgreina undirmálheildir betur með því að gefa þær út hverja fyrir sig með viðeigandi lýsigögnum. Þetta hefur í för með sér að bæta þarf lýsigögnum við nokkrar undirmálheildir. Allar málheildir verða gefnar út á TEI-samhæfu sniði.

Varða 5 – lýsing

Síðustu fjórar málheildir sérsviða skilgreindar (málheild bóka, málheild vísinda- og fræðitexta, málheild dómsúrskurða, málheild samfélagsmiðla). 2020 útgáfa fyrstu fimm undirmálheilda RMH gefnar út. Skýrsla um tiltæka fræðitexta.

Varða 6 – lýsing

Textar úr 750-1000 bókum, safnaðir frá bókaútgefendum, unnir. Textum úr vísinda- og fræðiritum safnað og þeir unnir. Málheild bóka og málheild vísinda- og fræðitexta gefnar út.

Afurðir

Öllum markmiðum vörðunnar var náð, en málheild bóka er minni en við stefndum á, með texta úr 351 bók. Við höldum áfram vinnu við stækkun málheildar bóka á þriðja ári, sjá umræðu og áætlanir varðandi það hér að neðan í fjórða kafla. Afurðir eru eftirfarandi:

- Skilgreiningar síðustu fjögurra málheilda sérsviða, sjá fyrsta kafla í skýrslunni hér að neðan
- Skýrsla um tiltæka fræðitexta, sjá þriðja kafla í skýrslunni hér að neðan
- Útgáfa átta RMH-undirmálheilda:
    - IGC-Parla: http://hdl.handle.net/20.500.12537/111
    - IGC-Laws: http://hdl.handle.net/20.500.12537/116
    - IGC-Adjud: http://hdl.handle.net/20.500.12537/101
    - IGC-Books: http://hdl.handle.net/20.500.12537/126
    - IGC-Social: http://hdl.handle.net/20.500.12537/138
    - IGC-Journals: http://hdl.handle.net/20.500.12537/139
    - IGC-News1: http://hdl.handle.net/20.500.12537/141
    - IGC-News2: http://hdl.handle.net/20.500.12537/142

Verkþáttur G1 heldur áfram á þriðja ári.

Skýrsla

Skilgreiningar átta undirmálheilda

Þrátt fyrir að mismunandi málheildir hafi mismunandi lýsigögn stefnum við að því að halda samræmdri uppbyggingu fyrir þær allar. Þeim er öllum skipt í tvær gerðir: fyrri gerðin er fyrir fullmerktar málheildir, en með hreinum texta fyrir ræðurnar, á meðan seinni gerðin er alveg eins, nema með viðbættum málfræðilegum merkingum við texta ræðanna. Rót xml-skránna er alltaf <teiCorpus>, vistað í einstaka skrá, sem inniheldur lýsigögn allrar málheildarinnar og hlekki í skrárnar sem innihalda TEI-tögin, eða, ef málheildirnar hafa undirmálheildir, í skrár sem innihalda annað teiCorpus-tag. Þegar málheild inniheldur undirmálheild er hver undirmálheild gerð þannig að hún geti staðið sjálfstætt, aðskilin aðalmálheildinni – öll lýsigögnin eru því geymd í teiCorpus-tagi undirmálheildarinnar og TEI-skrárnar vísa aðeins í stak sem er listað í því teiCorpus-tagi.

Lýsigögn sem allar málheildir eiga sameiginleg eru: a) titill málheildarinnar, og b) nöfn þeirra sem bera ábyrgð á stofnun málheildarinnar, staðsett í <titleStmt>, c) útgefandi málheildarinnar, d) aðgengileiki (leyfi), og e) dagsetning útgáfu málheildarinnar í <publicationStmt>, f) titill og g) dagsetningar uppruna <sourceDesc>, og h) lýsing verkefnis og i) skilgreining taga sem notuð er í <encodingDesc>. Sumar málheildir hafa engin nánari lýsigögn en aðrar taka fram t.d. fólk og stofnanir í <profileDesc> eða aðra flokkun í <clasaDecl>.

1.1 IGC-News1 og IGC-News2

Tvær málheildir með texta frá fréttamiðlum þar sem aðeins hluta þeirra má gefa út undir CC-BY-leyfi. IGC-News1 inniheldur texta með CC-BY-leyfi og ICG-News2 inniheldur texta með takmörkuðu leyfi (MIM).

Hver málheild inniheldur margar undirmálheildir þar sem textar frá mismunandi miðlum eru vistaðir sem einstakar málheildir. Aðal teiCorpus-skrárnar fyrir hverja málheild (IGC-News1 og IGC-News2) innihalda lýsigögn um alla málheildina og hlekki í teiCorpus-skrá fyrir hvern miðil.

1.2 IGC-Parla

IGC-Parla inniheldur allar þingræður sem eru aðgengilegar á www.althingi.is.

Við fylgdum ParlaMint-skemanu[1] fyrir þingræðurnar. Aðal teiCorpus-skráin inniheldur upplýsingar um tegund mælanda, tegund fundar, sal, flokk, efni, kjördæmi, ráðuneyti, ríkisstjórn, stjórnmálaflokka og mælendur (nafn, kyn, fæðingardag og flokksaðild).

1.3 IGC-Laws

IGC-Laws inniheldur þrjár mismunandi undirmálheildir: 1) Lagasafn Íslendinga, 2) greinargerðir og athugasemdir úr alþingisfrumvörpum, og 3) ályktanir og tillögur lagðar fram á Alþingi. Aðal teiCorpus-skráin inniheldur hlekk á þessar þrjár undirmálheildir.

[1] https://github.com/clarin-eric/ParlaMint/tree/main/Schema

*Proposals and resolutions (*ályktanir og tillögur)*:* Undirmálheildin IGC-Laws1 inniheldur ályktanir og tillögur lagðar fram til Alþingis, dagsettar frá nóvember 1988 til enda 2020.

Bills (frumvörp): Undirmálheildin IGC-Laws2 inniheldur greinargerðir og athugasemdir úr alþingisfrumvörpum sem hafa verið lögð fram á Alþingi síðan í október 1988 til enda 2020, og eru aðgengilegar á www.althingi.is.

Laws (lög): Undirmálheildin IGC-Laws3 inniheldur lagasafn Íslendinga sem eru aðgengileg á www.althingi.is, (útgáfa 150b, útgefin í maí 2020).

1.4 IGC-Books

IGC-Books inniheldur texta úr 351 bók, gefnar út frá 1968 til 2020. teiCorpus-skráin inniheldur lista yfir alla höfunda, þýðendur og ritstjóra (nöfn, kyn og fæðingarár). Hver TEI-skrá inniheldur eina bók. Hausinn inniheldur upplýsingar um útgefandann, höfund/ritstjóra/þýðanda og titilinn. Ef bókinni er skipt í nokkra hluta með mismunandi höfundum er hver hluti inni í og titill hvers hluta og nafn höfunda(r) eru í <bibl>.

1.5 IGC-Journals

IGC-Journals inniheldur 53.191 greinar frá vísinda- og fræðiritum og 4 vefsíðum með vísinda- og/eða fræðigreinar, frá 1979 til 2020.

1.6 IGC-Social

IGC-Social inniheldur þrjár undirmálheildir: IGC-Social1 inniheldur texta frá tveimur netspjallborðum, IGC-Social2 inniheldur texta frá fjórum bloggum og IGC-Social3 inniheldur tíst (e. tweets). Aðal teiCorpus-skráin inniheldur hlekki á teiCorpus-skrár þessara undirmálheilda.

Forums: IGC-Social1 inniheldur tvær undirmálheildir, eina fyrir hvert spjallborð. Aðal teiCorpus-skrá IGC-Social1 inniheldur hlekki á teiCorpus-skrár hverrar undirmálheildar. Báðar þeirra innihalda hlekki á TEI-skrár þar sem hver TEI-skrá inniheldur texta frá eins mánaðar tímabili. Setningunum hefur verið stokkað upp og það eru engar upplýsingar um höfunda eða nákvæmar dagsetningar.

Blogs: IGC-Social2 inniheldur fjórar undirmálheildir, eina fyrir hvert svið.

Tweets: teiCorpus-skráin inniheldur hlekki á TEI-skrárnar, en hver TEI-skrá inniheldur tíst frá eins mánaðar tímabili. Setningunum frá öllum tístunum var stokkað upp og það eru engar upplýsingar um notendur eða nákvæmar dagsetningar. Vegna leyfisskilmála Twitter hafa skrárnar verið „þurrkaðar“, þ.e. allir textar hafa verið fjarlægðir, en með einfaldri skriftu sem verður gefin út með málheildinni geta notendur sótt öll tíst frá twitter.com og þannig „endurvökvað“ málheildina.

Ætlunin var að hafa fjórðu undirmálheildina sem innihéldi texta frá kommentakerfi vefmiðla en þar sem næstum öll íslensk fjölmiðlafyrirtæki nota Facebook fyrir athugasemdir reyndist söfnun gagnanna vera hæg og erfið.

Tölfræði

Málheildirnar átta innihalda samtals 1.862.139.277 lesmálsorð eða 2.447.029.470 tóka (sjá töflu 1), samanborið við u.þ.b. 1.555 milljón lesmálsorð við síðustu útgáfu RMH.

IGC-Social er stærsti þátturinn í þessari aukningu. Fréttamálheildirnar tvær fóru í gegnum hreinsunarferli þar sem fréttir á erlendum tungumálum (aðallega ensku og pólsku) voru teknar út auk endurtekninga. IGC-Adjud minnkaði einnig í stærð þar sem tilvísanir í skjölum hæstaréttardómstóla í skjöl héraðsdómstóla voru fjarlægðar til að sleppa við endurtekningar.

Málheildir	**Setningar**	**Orð**	**Tókar**
GC-News1	20.952.712	354.459.688	390.196.047
IGC-News2	49.246.907	855.480.334	952.174.584
IGC-Parla	10.283.804	212.873.555	234.062.485
ICG-Adjud	2.579.022	56.595.797	61.798.812
IGC-Laws	2.200.328	40.612.997	44.629.736
IGC-Social	29.820.072	319.959.250	360.307.303
IGC-Books	949.169	13.340.865	15.155.246
IGC-Journals	1.450.193	25.918.973	28.967.810
Samtals	**117.482.207**	**1.879.241.459**	**2.087.292.023**

Skýrsla um tiltæka fræðitexta

Söfnun fræðitexta hefur gengið mjög vel. Ritstjórar og útgefendur fræðirita hafa svarað fyrirspurnum okkar mjög jákvætt. Mörg þessara fræðirita eru aðgengileg á netinu svo við gátum auðveldlega skoðað hvort þau gætu nýst í þetta verkefni. Þetta hefur einnig í för með sér að auðvelt er fyrir ritstjóra og útgefendur þessara rita að gefa leyfi til notkunar þessara texta þar sem einhverjir þeirra voru þegar aðgengilegir öllum. Þrátt fyrir það ákváðum við að textum þessarar undirmálheildar skyldi skipta í setningar og stokka upp innan hvers skjals. Þessi ákvörðun var aðallega til þess að við þyrftum ekki að hafa samband við hvern og einn höfund til þess að fá leyfi til að nota texta þeirra, þar sem það hefði verið flókið og mjög tímafrekt. Þess í stað vörðum við tíma okkar í samskipti við ritstjóra og útgefendur, söfnun skráa og vinnslu textanna.

Við höfum fengið leyfi til að nota texta 42 fræðirita yfir vítt svið fræðigreina. Magn gagna sem við náðum að safna var í raun of mikið til að við gætum unnið þau fyrir þessa vörðu. Ástæðan er aðallega sú að mörg þessara rita geyma skrár aðeins á PDF-formi og hafa því ekki getað veitt okkur skrár á öðrum formum til vinnslu. Þetta hefur gert söfnun skráa mun auðveldari, en vinnslu

þeirra erfiðari og tímafrekari. Því höfum við enn mikið magn efnis til að vinna sem verður bætt í málheildina á undan næstu útgáfu. Útgáfan sem við gefum út núna inniheldur texta 20 fræðirita.

Málheild bókatexta

Málheild bókatexta inniheldur texta úr 351 bók. Við stefnum á útgáfu málheildar með textum úr 750-1000 bókum og náðum sannarlega að safna nægilega til að ná því markmiði.
Nokkur vandamál komu upp við söfnun bókatexta fyrir málheildina, sem er lýst að neðan, sem orsakaði hægari framgang og vinnslu bóka fyrir þessa málheild sem olli því að fjöldi bóka í þessari útgáfu er lægri en við höfðum vonað.

Þeir bókaútgefendur sem hafa svarað fyrirspurnum okkar hafa almennt verið jákvæðir varðandi útvegun texta fyrir málheildina. Hins vegar sendu mjög fáir þeirra skrár um leið og við þurftum að ítreka fyrirspurnir okkar oft við næstum alla útgefendur sem hafa útvegað texta fyrir málheildina. Því hefur söfnunarferli textanna tekið mikið lengri tíma en búist var við. Eins og búast mátti við þurftu sumir útgefendur að ræða málið innanhúss áður en ákvörðun var tekin, og þar sem þetta er ekki forgangsmál fyrir þeim var stundum töf mánuðum saman. Í flestum tilfellum þar sem fyrirspurn okkar var samþykkt tók enn lengri tíma fyrir útgefendurna að safna skjölum og afhenda þau. Samskiptahluti þessa verkefnis hefur því verið mjög tímafrekur.

Í fyrstu einblíndum við á stærri útgefendur í stað minni. Hins vegar kom svo í ljós að minni útgefendur voru gjarnan liðlegri, á meðan sá stærsti ákvað á endanum að koma ekki að verkefninu. Þar sem við áætluðum fjölda bóka út frá fjölda útgefinna bóka frá hverjum útgefanda höfum við ekki getað safnað eins mörgum bókum og við bjuggumst við. Í vörðuskýrslu M4 töldum við upp samstarfsfúsa útgefendur og fjölda bóka sem þeir höfðu gefið út, samkvæmt upplýsingum af vefsíðum þeirra. Á þeim tímapunkti töldum við að staðfestir samstarfsfúsir útgefendur myndu vera með um 3500 (3568) bækur í boði, þó að við gerðum ekki ráð fyrir að geta safnað textum úr þeim öllum. Sumir þessara útgefenda breyttu ákvörðun sinni seinna og ákváðu að útvega ekki neina texta, eftir skil vörðuskýrslu M4. Á jákvæðari nótum náðum við að fá fleiri útgefendur til að vinna með okkur og sumir þeirra útveguðu jafnvel fleiri bækur en við gerðum ráð fyrir.

Eitt vandamál var nokkuð óvænt í þessu ferli. Eins og kom fram í vörðuskýrslu M4 náðum við samþykki allra meðlima Rithöfundasambands Íslands og Hagþenkis, félags höfunda fræðirita og kennslugagna á Íslandi, sem þýddi að við gætum safnað textum úr öllum bókum eftir þessa meðlimi. Við gerðum ráð fyrir að þetta gæfi okkur aðgang að textum úr næstum öllum bókum sem við gætum safnað. Hins vegar kom í ljós að aðeins um þriðjungur þeirra höfunda sem við höfðum safnað bókum frá voru meðlimir þessara samtaka, sem þýddi að við höfðum ekki leyfi til að nota texta úr um tveimur þriðju bókanna sem við söfnuðum. Þetta þýðir að við höfum mikið magn unnins texta sem við getum ekki enn bætt í málheildina. Á þriðja ári munum við reyna að hafa samband við höfunda þessara bóka til að fá leyfi til að nota textana, það verður einfalt að bæta þeim í málheildina svo lengi sem við fáum leyfi til þess. Eftir að við fengum meðlimaskrá þessara tveggja samtaka höfum við lagt áherslu á söfnun bóka eftir höfunda þeirra. Þetta krefst þess að við gerum lista bóka sem eru gefnar út af ákveðnum útgefanda sem eru skrifaðar af meðlimum þessara samtaka. Svo þarf útgefandinn að leita eftir skrám þeirra bóka, sem orsakast í því að við fáum vanalega örfáar bækur í hvert skipti með mislangri bið á milli afhendinga.

Á næstu nokkrum mánuðum munum við vinna við að safna fleiri bókum frá völdum útgefendum og að hafa samband við höfunda þeirra bóka sem við höfum þegar unnið.

G4 – Beygingarlýsing íslensks nútímamáls fyrir máltækni

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

Markmið

Að afhenda gögn fyrir næstu kynslóð BÍN-gagnasafnsins, reglulega uppfært og opið (e. openly distributed). BÍN fyrir máltækni samanstendur af fjórum hlutum: **Beygingarlýsingu íslensks nútímamáls** (BÍN, The Database of Modern Icelandic Inflection), **Orðföngum** (orðhlutafræðilegri greiningu, MorphIce), **Ritmyndum** (Non-Standard Word Forms) og **Rökliðum sagna** (Valency Structures). BÍN-gögnin verða greypt í Python-pakka með stöðluðum forritaskilum. Áfram verður hægt að hlaða gögnunum niður af vef BÍN (bin.arnastofnun.is) á ýmsum formum, með víðtækum lýsigögnum. Snið niðurhalanlegra máltæknigagna úr öðrum hlutum BÍN-gagnasafnsins verður skilgreint í samráði við notendur og þau gerð aðgengileg til niðurhals á vefsíðunni. BÍN-gögnin verða einnig aðgengileg á CLARIN.

Vörður 5-6 – lýsing

- **Beygingarlýsing íslensks nútímamáls í heild**
  - Gagnagreining, uppbygging gagnagrunns, viðbót gagna, og skilgreining úttaks (með lýsigögnum). Allir þættir vinnunnar eru samtengdir, og munu orsaka viðbætur í orðaforða og orðmyndir í BÍN, þ.á.m. víðtæka einkunnagjöf og millivísanir.
- **BÍN-gagnasafnið (DMII)**
  - Vinnu við BÍN-gagnasafnið er ekki skipt í vörður hér, þar sem vinnan fer að mestu eftir öðrum hlutum BÍN sem lýst er annars staðar. Einkunnagjöf orða og orðmynda er dregin frá öðrum hlutum vinnu við BÍN. Búist er við víðtækum viðbótum í orðaforða, með nýjum eiginleikum greiningar, og víðtækum millivísunum.
- **MorphIce**
  - M5: Endurkvæmar samsetningar, og endurkvæmni í afleiðslu og samsetningu. Meirihluti orðaforðans í BÍN verður greindur, og stefnt er á um 300.000 uppflettimyndir (e. lemmas).
  - M6: Setningarliðarsamsetningum lokið með hlekkjum í rökliðagerðir. Skilgreiningar úttakssniðs fyrir máltækni, hönnun vefsíðu, útgáfa lýsigagna.
- **Rökliðir sagna**
  - M5: Þróun gagnagrunns, hlekkir í agnarsagnir í MorphIce.
  - M6: Öllum hlekkjum lokið í MorphIce.
- **Ritmyndir**
  - M5: Fínstilling greiningar og gagnasafnið að mestu tilbúið.

- M6: Úttaksgögn skilgreind fyrir mismunandi notkun, þar á meðal notkun innan DIM.

Afurðir

Rökliðir sagna: Úttak gagnagrunns sent til CLARIN: http://hdl.handle.net/20.500.12537/163
BÍN: Nýtt snið (Stórasnið, „The Comprehensive Format“) gefið út með viðbótarupplýsingum varðandi óstaðlaðar orðmyndir og sent til CLARIN: http://hdl.handle.net/20.500.12537/162
Listi íslenskra skammstafana: Aukaafurð MorphIce, send til CLARIN: http://hdl.handle.net/20.500.12537/164
MorphIce: MorphIce er ekki tilbúin til útgáfu.

Verkþætti G4 lýkur við M6.

Skýrsla

BÍN og ritmyndir

Gögn úr **BÍN**, þeirra á meðal **Ritmyndir**, eru í sniðinu **Stórasniði** (titill á ensku „The Comprehensive Format“), sem er gefið út á vefsíðu BÍN og sent til CLARIN. Ítarlegar skýringar á íslensku eru aðgengilegar á vefsíðunni bin.arnastofnun.is.

Stórasniði er skipt í þrjár CSV-skrár, þ.e. orð (ord.csv), beygingarmyndir (beygmynd.csv), og ritmyndir/óstaðlaðar orðmyndir (ritmynd.csv). Skrárnar eru tengdar með BIN.ID uppflettimynda. Upp á gegnsæi og til að gera skrárnar auðveldari í lestri fyrir fólk eru sumir reitir úr ord.csv endurteknir í skránum beygmynd.csv og ritmynd.csv. Skráin ord.csv inniheldur 303.339 uppflettimyndir, skráin beygmynd.csv inniheldur 6.559.797 beygingarmyndir, og skráin ritmyndir.csv inniheldur 48.517 orðmyndir (15. sept. 2021).

- **I. Orð**, ord.csv. 1 = aðalorð; 2 = BIN.id; 3 = orðflokkur/kyn nafnorða; 4 = orðmyndunarflokkur (ósamsett/samsett), 5 = hluti BÍN; 6 = einkunn orðs; 7 = tegund; 8 = málfræði og notkun; 9 = sýnileiki (þ.e. vísun í mikilvægi í nútímaíslensku)
  - **A**: 10 = framburður
  - **B**: Orðhlutafræðileg bygging ósamsettra orða; 11 = tegund A (t.d. nefnifallsendingar); 12= tegund B (bakstöðugreining á stofni); 13 = fjöldi atkvæða (í orðstofni); 14 = fjöldi sérhljóða og samhljóða
  - **C. Vísanir**: 15 = millivísanir milli mismunandi tilbrigða orðs (t.d. milli tveggja uppflettimynda sama orðs, vegna stafsetningar o.fl.); 16 = ruglingsmengi orða, t.d. samhljóma orða o.fl.; 17 = vísun frá tökuorði til íslensks jafngildis; 18 = vísun í stafsetningarreglu; 19 = vísun í Nútímamálsorðabók; 20 = önnur mynd aðalorðs; 21 = málfræðilegt mark annarrar myndar aðalorðs.
- **II. Beygingarmynd**, beygmynd.csv: 1 = aðalorð, 2 = BIN.id, 3 = orðflokkur, 4 = beygingarmynd, 5 = málfræðilegt mark; 6 = einkunn beygingarmyndar; 7 = tegund/stíll beygingarmyndar; 8 = gildi beygingarmyndar (notað í vali á tilbrigðum o.fl.).
- **III. Óstaðlaðar orðmyndir**, ritmynd.csv: 1 = aðalorð, 2 = BIN.id, 3 = orðflokkur, 4 = orðmynd; 5 = stöðluð orðmynd; 6 = málfræðilegt mark; 7 = einkunn orðmyndar; 8 = tegund/stíll orðmyndar, 9 = greining á stafsetningartilbrigðum; 10 = greining á innsláttarvillum; 11 = greining á beygingartilbrigðum, 12 = dagsetning; 13 = uppruni

Dæmi

ord.csv:

```
allsnægt;127071;kvk;s;alm;1;;TALA;K;;;;;;;;;;3010;allsnægtir;NFFT
dudda;514764;kvk;g;alm;2;BARN;;V;;X.a;;1;cvC.a;;;;;;;
skrítinn;165191;lo;g;alm;1;;Í-Ý,MERK,STAFS;K;;X.n;Xin.n;2;cccvcvc.n;165192;;;;37158;;
skrýtinn;165192;lo;g;alm;1;;Í-Ý,MERK,STAFS;K;;X.n;Xin.n;2;cccvcvc.n;165191;;;;37233;;
```

beygmynd.csv:

```
hönd;426053;kvk;handar;EFET;1;;;
hönd;426053;kvk;höndur;NFFT2;0;URE;VIK;
hóll; 485361;kk;hóli;ÞGFET2;2;SJALD;VIK;
```

ritmynd.csv:

```
áætlun;138223;kvk;áætlunnar;áætlunar;EFET;4;VILLA;NN4N-END;;;;RIS19
Kristín;362296;kvk;krístin;Kristín;NFET;5;VILLA;LAG4HA;ASL;VIXLBRODD;BIN_LEIT
bóndi;10162;kk.bóndum;bændum;ÞGFTgr;4;VILLA;BEYGVILLA;BIN_LEIT
```

Skammstafanaskrá BÍN

[Þessi hluti BÍN var ekki á áætlun **annars árs**. Skráin er aukaafurð Morphlce, tilbúin til útgáfu.]

Greining á íslenskum skammstöfunum, með nokkrum erlendum skammstöfunum sem eru algengar í íslenskum textum. Þetta verkefni inniheldur nú 4.194 færslur og er aðgengilegt á Vefsíðu BÍN og sent til CLARIN, með ítarlegum útskýringum á íslensku. Ensk útgáfa útskýringa kemur bráðum.

Útgefið úttak hefur eftirfarandi hluta: form færslu; staðlað form færslu; réttmætiseinkunn; undirliggjandi texti; þýðing erlends texta; frekari útskýringar (eftir þörfum); flokkun skammstafana, skammstafanir, og stutt form; svið notkunar; innri millivísun; og gögn um mögulega beygingu undirliggjandi texta þegar talgervill les. Þetta á alltaf við um skammstafanir og um sum hánefni. Nauðsynlegir reitir greiningar fyrir beygingargögn eru +/-margorða, +/-beygt, +/-innri beyging, með upplýsingar um beygingarfræðilega hausa og vísanir í beygingarfræðileg viðmið í BÍN. Dæmi um úttak, með þýddum hugtökum og brottfalli auðra reita:

- `IMF;International Monetary Fund;Alþjóðagjaldeyrissjóðurinn;;+;e;acron-II;FIN;+;-;-;;AGS`
- `AGS;Alþjóðagjaldeyrissjóðurinn;;International Monetary Fund;;;acron-II;FIN;+;-;-;;IMF`
- `SÞ;Sameinuðu þjóðirnar;;United Nations;;;acron-II;INST;+;+;+;[Sameinuðu]<int-infl:adj-def.sameinaður>[þjóðirnar]<infl-head:fem.pl-def.þjóð>;UN`
- `UN;United Nations;Sameinuðu þjóðirnar;;+;e;acron-II;INST;+;-;-;;SÞ`

BÍN rökliðir

Form úttaks sem verður sent til CLARIN fljótlega. DVS.NR verður notað fyrir hlekki í Morphlce. Hausar eru innifaldir í hverjum reit til glöggvunar. Gögn og útskýringar munu birtast á vefsíðu BÍN.

- `DVS.NR:1;LEMMA:abbast;FRL-HAUS:e-r;FRL-FALL:nf;;FRL-ANIM:+;;SO:abbast;FL-HAUS:upp á e-n;FS:upp á;FL-FALL:þf;;FL-ANIM:+`
- `DVS.NR:2;LEMMA:aðhlýða;FRL-HAUS:e-r;FRL-FALL:nf;;FRL-ANIM:+;;SO:aðhlýðir;ANDL-HAUS:e-u;ANDL-FALL:þgf;;ANDL-ANIM:-`
- `DVS.NR:3;LEMMA:aðla;FRL-HAUS:e-r;FRL-FALL:nf;;FRL-ANIM:+;;SO:aðlar;ANDL-HAUS:e-n;ANDL-FALL:þf;;ANDL-ANIM:+`

- DVS.NR:4;LEMMA:aðlaga;FRL-HAUS:e-r;FRL-FALL:nf;;FRL-ANIM:+;;SO:aðlagar;ANDL-HAUS:e-ð;ANDL-FALL:þf;;ANDL-ANIM:-;;ANDL-HAUS:e-u;ANDL-FALL:þgf;;ANDL-ANIM:-
- DVS.NR:5;LEMMA:aðlaga;FRL-HAUS:e-r;FRL-FALL:nf;;FRL-ANIM:+;;SO:aðlagar;ANDL-HAUS:e-ð;ANDL-FALL:þf;;ANDL-ANIM:-;;FL-HAUS:að e-u;FS:að;FL-FALL:þgf;;FL-ANIM:-

MorphIce

MorphIce er ekki tilbúin til útgáfu. Okkur fannst skilgreining G4-verkþáttarins nokkuð stór miðað við fjölda úthlutaðra mannmánuða á öðru ári. Þetta olli töfum á forritun fyrir þennan þátt og því er útgáfu frestað. Þegar gagnasafninu verður lokið og er komið í gang verður stór hluti orðaforðans tilbúinn til útgáfu með minniháttar fyrirhöfn. Kristín Bjarnadóttir, ritstjóri BÍN, mun sjá til þess að MorphIce ljúki.

G4b – BÍN í Python-pakka

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

Tilgangur

(Frá undirverkefni G4): Afhenda næstu kynslóð BÍN-gagnasafnsins, greypt í Python-pakka.

Markmið

Til að einfalda og auðvelda notkun fyrir forritara og vísindafólk verður BÍN pakkað í Python-pakka með stöðluðum forritaskilum. Allt sett uppflettimynda og beygingardæma verður sett í forrit í þjöppuðu formi (e. binary compressed format) (áætluð stærð þess er ~100 MB) og minnisvarpað (e. memory-mapped) á keyrslutíma fyrir mjög hraðvirkan aðgang (~míkrósekúndur fyrir hverja leit). Ekkert utanaðkomandi gagnasafn eða önnur ákvæði verða þörf til að vinna með BÍN.

Afurðir

M5

- Tilbúinn Python-pakki er gefinn út á PyPI og á CLARIN.

Skýrsla

Afurðum var skilað.

BÍN (*Beygingarlýsing íslensks nútímamáls*)-pakkinn er aðgengilegur á Python Package Index (https://pypi.org/project/islenska/) og má setja upp með skipuninni:

```
pip install islenska
```

Engin frekari skref eru þörf til að byrja að nota BÍN í kóða, þar sem allt gagnasafnið er innifalið í Python-pakkanum í þjöppuðu formi. Gagnaleit er keyrð innvært með C++ einingu fyrir hámarkshraða. Pakkinn hefur verið prófaður á Linux (x86 og ARM), macOS og Windows.

Frumkóðinn, ásamt ítarlegri skjölun, er aðgengilegur á GitHub á https://github.com/mideind/BinPackage. Hann er með MIT-leyfinu. Nýjustu útgáfuna er einnig hægt að sækja frá hirslu CLARIN á http://hdl.handle.net/20.500.12537/137.

Pakkinn styður leit eftir hvaða færslu BÍN sem er á „Kristínarsniði“ og skilar 15 eigindum fyrir hverja færslu. Hann styður einnig leit að beygingarmynd hvaða færslu sem er, til dæmis

fleirtölumynd nafnorðs ef eintölumynd er gefin, eða viðtengingarhátt sagnar ef framsöguháttur er gefinn.

G6 – Framburðarorðabók

Skýrsla: Grammatek
Aðilar: Grammatek

Markmið

Framburðarorðabók með merktum tilbrigðum í framburði sem nota má til að þjálfa líkan rittákns-til-fónems (g2p) fyrir hvert tilbrigði. Tól til að fara yfir orðabókina og útbúa g2p-líkan.

Varða 6 – lýsing

M5: Hljóðritunarreglur fyrir erlend orð tilbúnar og lista erlendra orða og skammstafana úr fréttatextum hefur verið bætt í orðabókina, ásamt fleiri íslenskum orðum, þannig að samtals 15.000 orðum hefur verið bætt í orðabókina.

M6: Lokaútgáfa íslensku framburðarorðabókarinnar gefin út, sem inniheldur u.þ.b. 60.000 færslur. Lokaútgáfa ritstýringartóls orðabókarinnar gefin út.

Afurðir

1. Ný íslensk framburðarorðabók sem inniheldur 58.222 færslur
2. Hljóðritunarreglur fyrir erlend orð á málfræðiformi Thrax.
3. Ritstýringartól orðabókarinnar „PEDI“

Öllum markmiðum vörðunnar var náð.

Íslensk framburðarorðabók á GitHub: https://github.com/grammatek/iceprondict
Thrax málfræði, íslensk og ensk með „íslenskum“ framburði:
https://github.com/grammatek/g2p-thrax
CLARIN: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/154
Ritstýringartól orðabókarinnar „PEDI“ á GitHub: https://github.com/grammatek/pedi

Verkþætti G6 lýkur við M6.

Skýrsla

Á fyrsta ári var megináhersla lögð á hljóðritun orðalista með góðri dreifingu mögulegra samsetninga rittákna í íslensku, til þess að búa til þjálfunarsett fyrir sjálfvirkar einingar rittákns-til-fónems (g2p). Á öðru ári, samhliða vinnu við tíðnilista frá Risamálheildinni (RMH) (verkþáttur G1), var lögð sérstök áhersla á algeng erlend orð og orð sem fylgja ekki g2p-reglum á íslensku.

Erlend orð

Erlend orð voru valin með hjálp tíðniupplýsinga frá Risamálheildinni, en einnig með því að skoða nýlega fréttatexta og handvelja nokkur mikilvæg orð (t.d. *COVID*). Við bjuggum einnig til lista mikilvægra orða fyrir Android-viðmótið, orð úr stillingum og nöfn helstu appa, sem fylgir kröfum úr verkþætti T5. Alls eru næstum 4.000 enskar (eða aðrar erlendar) færslur í orðabókinni. Orðin voru umrituð samkvæmt reglum skilgreindum fyrir M4, og reglunum var einnig varpað yfir á Thrax-málfræðireglur fyrir sjálfvirka umritun erlendra orða (sjá verkþátt T10).

Óhefðbundinn framburður

BÍN (verkþáttur G4) inniheldur upplýsingar um uppflettimyndir sem fylgja ekki hefðbundnum g2p-reglum íslensku. Til dæmis er hefðbundna reglan fyrir *f* milli sérhljóða (V) : *V-f-V : [V-v-V].* Þetta gildir fyrir orð eins og: *sofa : [s O: v a]* og *afi : [a: v I].* Hins vegar hafa tökuorð tilhneigingu til að brjóta þessar reglur þar sem samhljóðinn er borinn fram eins og 1:1 yfirborðsgerð þeirra, þ.e. [f] fyfir *f*: *sófi : [s ou: f I]* og *grafík : [k r a: f i k].* Það er almennt ekki mögulegt að greina þessar undantekningar frá orðmyndunum einum, og því þurfa þær að vera til staðar í orðabókinni frekar en að láta g2p-kerfi meðhöndla þær.

Algengasta mynstrið í undantekningalistanum, og á sama tíma það vandasamasta, er mynstrið V-ll-V, þ.e. tvöfalt ‘l’ milli sérhljóða. Almenna reglan er að bera tvöfalt *l* á „íslenska mátann“, þ.e. sem [t l]. Helstu undantekningarnar frá þessari reglu eru nöfn, og gælunöfn, og – aftur – tökuorð. Við þurfum ekki bara að taka á þessum tilbrigðum, heldur eru til þónokkur samhljóma orð þar sem annað hefur framburðinn [t l] og hitt [l]:

villa (‘skekkja, mistök’) : v I t l a og *villa (‘hús’) : v I l a*
galli (‘veila’) : k a t l I og *galli (‘samfestingur’) : k a l I*

Þetta er alvarlegt vandamál fyrir talgervilskerfi, þar sem þessi orð eru nokkuð algeng, mismunurinn í framburði er töluverður og pirrar því notendur.

Við fengum lista uppflettimynda frá BÍN með flokkuðum undantekningum frá framburðarreglum. Eftir að við drógum beygingardæmin úr gagnagrunninum settum við saman lista yfir 16.000 orð sem fylgja ekki hefðbundnum g2p-reglum. Við byrjuðum valferlið með því að velja þær orðmyndir sem finnast í lista 30.000 algengustu orðmynda í Risamálheildinni, því næst var V-ll-V-mynstrið valið og öll grunnorð og samsetningar með V-ll-V samhljóma orðum, og að lokum var tíðnilisti búinn til úr frétta- og dómstólagögnum frá Risamálheildinni. Alls inniheldur listi óreglulegra orðmynda 4.377 orð. Sjálfvirka g2p-kerfið, sem er keyrt áður en listarnir eru handyfirfarnir, var töluvert verra en fyrir listana sem innihalda að mestu regluleg orð, og bjó oft til aðrar færslur fyrir framburðarmállýskur þar sem ætti ekki að vera neinn munur. Kerfið er augljóslega eitthvað „ráðavillt“ og getur ekki sagt t.d. hvaða framburður er réttur fyrir V-ll-V.

Þegar samsetningagreiningu verður bætt við g2p-ferlið á þriðja ári (sjá verkþátt T10), með notkun framburðarorðabókarinnar til að aðstoða við val réttra umritana fyrir óregluleg samsett

orð sem standa eftir á listanum, ætti að vera hægt að bæta eftirstandandi orðmyndum í orðabókina. Athugið að orðin sem eftir standa á listanum eru óalgeng eða ekki til staðar í gögnum okkar (Risamálheildinni) og því teljum við ekki þess virði að umrita þau öll handvirkt.

I4 – Málfræðilegur markari/Lemmari

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

Markmið

Að þróa málfræðilegan markara og lemmara fyrir íslensku samkvæmt nýjustu aðferðum. Þessi verkþáttur felur í sér greiningu á villum sem fyrri markarar og lemmarar hafa gert, sem og rannsókn á því hversu mikilli nákvæmni er hægt að ná við mörkun texta á íslensku.

Varða 6 – lýsing

- M5: Tilraunir með þrepaskipt mörkunarlíkan, velja það BERT-líka líkan sem virkar best og nota gögn frá BÍN. Seinni helmingur MIM-GOLD er auðgaður með uppflettimyndum.
- M6: Uppfært líkan fyrir sameiginlega mörkun og lemmun tilbúið. Markmiðið er að nákvæmni lemmunar sé >= 97% með því að nota (gold) markað inntak. Endanleg villugreiningarskýrsla um lemmun og mörkun.

Afurðir

Markmiðum vörðunnar hefur verið skilað og þessum verkþætti lýkur við M6. Það eru örfá minniháttar frávik og aðlaganir:

- Líkan fyrir sameiginlega mörkun og lemmun var ekki gefið út vegna skertrar nákvæmni við málfræðilega mörkun þegar bæði verkin eru keyrð á sama tíma. Þess í stað er hvoru líkani skilað sér, og má keyra í röð fyrir lemmun. Þetta frávik kemur ekki að sök fyrir markmið verkþáttarins þar sem bæði tólin verða afhent og má keyra sem eina pípu.
- Engar tilraunir voru gerðar til að þróa þrepaskipt mörkunarlíkan. Markmið þessa verks var að endurbæta frekar málfræðilega mörkun en þar sem þróun lemmarans reyndist tímafrekari var tíma frekar úthlutað til að ná kjarnamarkmiðum verkefnisins.

Öðrum markmiðum vörðunnar var skilað eins og upprunalega var skilgreint:

- Taugalemmarinn nær nákvæmni upp á 98,3% sem fer fram úr markmiði þessarar vörðu.
- Líkanaskrár fyrir málfræðilegu markarana má finna hér: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/115 og fyrir lemmarann hér: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/134. Leiðbeiningar fyrir uppsetningu má finna hér: https://github.com/cadia-lvl/POS.
- MIM-GOLD var auðguð með uppflettimyndum og notuð til að þjálfa taugalemmarann.

Verkþætti I4 er lokið við M6.

Skýrsla

Markmið þessa verkþáttar er að eiga málfræðilegan markara (e. PoS-tagger) og lemmara fyrir íslensku. Þessi tól skulu vera auðveldlega aðgengileg og hraðvirk, þar sem þau eru gjarnan kjarnatól fyrir frekari málvinnslu. Þessi tól hafa verið þróuð og gefin út með auðvelda notkun í huga. Notandi getur valið að forgangsraða hraða eða nákvæmni við val á málfræðilegum markara. Taugalemmarinn reyndist erfiðari í þróun en búist var við og nær ekki gæðum umfram reglubyggða lemmarann. Sjá neðar fyrir nánari greiningu lemmara.

Málfræðilegur markari

Fyrir M5 var ABLTagger v3.0.0 gefinn út. Í nýju útgáfunni eru líkön sótt sjálfkrafa, sem auðveldar notkunina frekar. Enn fremur voru tvö PoS-líkön gefin út, lítið og stórt líkan sem vinna með endurskoðaða markaskrá sem ná nákvæmni ~96,2% og ~97,8% á MIM-Gold[2], hvert um sig. Þessi líkön byggja á ELECTRA-small og ELECTRA-base. Ákvörðun var tekin um að hætta að spá fyrir um „x“ (rangur tóki) og „e“ (útlendur tóki) mörk úr endurskoðuðu markaskránni þar sem þessi mörk eru í það minnsta vafasöm sem úttak úr markara. Við gefum því upp tölfræði fyrir nákvæmni en hunsum þessi mörk, og þetta veldur minniháttar aukningu í nákvæmni (um 0,1-0,2).

Nákvæmni málfræðilegrar mörkunar á MIM-GOLD

Líkan	Heildarnákvæmni	Nákvæmni á óséðum tókum[3]	Nákvæmni á séðum tókum
ELECTRA-base	97,84 ±0,07	92,20 ±0,68	98,34 ±0,04
ELECTRA-small	96,22 ±0,16	86,03 ±1,02	97,13 ±0,10

Hraði og stærð málfræðilegra markara

Líkan	Diskrými	Tókar/sek @ CPU	Tókar/sek @ GPU
ELECTRA-base	435MB	10	1100
ELECTRA-small	60MB	360	10000

Lemmari

Við M5 voru fyrstu keyrslur taugalemmarans gerðar á undirbúningsgögnum (óleiðréttum uppflettimyndum). Til þess að bæta árangur, og fella meiri þekkingu á uppflettimyndum inn í lemmarann, voru gögn dregin úr BÍN og þeim breytt í aukaþjálfunargögn. Þessi þjálfunargögn voru án samhengis setninga og gátu því ekki verið tvíræð í þeim skilningi að orðmynd +

[2] níföld krossprófun (e. 9-fold cross-validation), að undanskildum "x" og "e" mörkum.
[3] Óséður tóki er tóki sem ekki er í þjálfunarsettinu.

málfræðilegt mark gæti ekki orsakað tvær mismunandi uppflettimyndir. Öll slík dæmi voru síuð úr. Enn fremur stönguðust BÍN-gögnin oft við MÍM-gögnin, t.d. hefur BÍN uppflettimyndina „Þjóðleikhúskjallarinn“ á meðan MÍM hefur uppflettimyndina „Þjóðleikhúskjallari“, án greinis. Við fjarlægðum allt slíkt ósamræmi með frekari síun BÍN-gagnanna.

Á meðan þessi greining stóð yfir uppgötvuðum við einnig meira ósamræmi í MÍM-gögnunum; innsláttarvillur, ósamræmi uppflettimynda í bæði ein- og fleirtölumynd, vitlausar há- og lágstafanir, og fleira. Við leiðréttum margar þessara villna en reyndum að vera hófsöm, þar sem við vildum ekki vinna með allt annað gagnasett.

Taugalemmarinn er byggður á ELECTRA-small líkaninu og fær að auki málfræðilega markið á tókanum sem skal lemma. Uppflettimyndin er svo mynduð staf fyrir staf með sjálfvirku aðhvarfi (e. autoregressive RNN). Lemmarinn er fyrst forþjálfaður, án samhengis setninga, á BÍN-gögnunum og svo fínstilltur á MÍM-gögnunum.

Í töflunni að neðan berum við saman Nefni, reglubyggðan lemmara, við taugalemmarann.

Nákvæmni uppflettimynda á MIM-GOLD

Líkan	**Heildarnákvæmni**	**Nákvæmni á óséðum uppflettimyndum[4]**	**Nákvæmni á séðum uppflettimyndum**
Nefnir	98,58 ±0,17	94,55 ±1,81	98,78 ±0,11
Lemmatizer	98,33 ±0,16	86,59 ±1,20	98,93 ±0,10

Í töflunni sjáum við að lemmararnir ná svipaðri nákvæmni í heild en reglubyggði lemmarinn nær mikið betri árangri á óséðum uppflettimyndum. Þetta bendir til þess að taugalemmarinn alhæfi ekki umfram þjálfunargögnin og að forþjálfunin með BÍN skili ekki niðurstöðum eins og búist var við. Ef við skoðum vitlausar spár taugalemmarans sjáum við að þær eru alls ekki ásættanlegar: „glæpabókmenntur“ þar sem ætti að vera „glæpabókmenntir“ eða „viðreisnnn“ þar sem ætti að vera „viðreisn“. Orðið „viðreisn“ er sannarlega í BÍN og lemmarinn var þjálfaður á slíkum dæmum en svo virðist sem hann nái ekki að halda þeim upplýsingum í fínstillingu.

Nákvæmni uppflettimynda á MIM-GOLD með málfræðilegum mörkum sem eru ekki gullmörk

Líkan	**Heildarnákvæmni**	**Nákvæmni á óséðum uppflettimyndum**	**Nákvæmni á séðum uppflettimyndum**
Nefnir	97,97 ±0.18	91,32 ±1,58	98,30 ±0,14
Lemmatizer	98,06 ±0.16	85,93 ±1,32	98,69 ±0,10

[4] Óséð uppflettimynd er uppflettimynd sem er ekki í þjálfunarsettinu. Bæði líkönin gætu hafa séð uppflettimyndirnar áður í BÍN-gögnunum.

Í töflunni að ofan berum við saman lemmarana tvo við keyrslu á MIM-GOLD-prófunarsettunum með málfræðilegum mörkum frá ELECTRA-base málfræðilega markaranum, þ.e. sum mörkin eru vitlaus. Eins og við má búast minnkar nákvæmni þeirra að einhverju leyti en helst svipuð.

Hraði og stærð lemmara

Líkan	Diskrými	Tókar/sek @ CPU	Tókar/sek @ GPU
Nefnir	5MB	300.000	á ekki við
Lemmatizer	72MB	77	10000

Mat á lemmurum

Við sjáum að Nefnir er nákvæmari á óséðum uppflettimyndum og hraðvirkari en taugalemmarinn. Hraði Nefnis er eiginleiki sem taugalemmarar munu aldrei ná á örgjörva (e. CPU). Muninn á nákvæmni má án efa minnka og gera þolnari. Þetta mun hins vegar fela í sér endurhugsun núverandi stefnu og frekari þróunarvinnu. Niðurstaða okkar er að reglubyggður lemmari skuli vera aðallemmarinn fyrir flest verk með taugalemmarann til vara.

I5 – Þáttari

Skýrsla: Miðeind
Aðilar: Miðeind, Háskólinn í Reykjavík

Markmið

Í þessum verkþætti er vinnu frá fyrsta ári haldið áfram við grunnþáttara (e. shallow parser) og fullþáttara (e. full constituency parser) (bæði reglubyggða og tauganetsbyggða) fyrir íslensku.

Grunnþáttarinn nýtist sem hraðvirkari, léttari valkostur fyrir grunnþáttun, þegar ekki er þörf á fullþáttara, til dæmis við útdrátt upplýsinga þar sem borin eru kennsl á tiltekna búta (e. segments) texta og þeir flokkaðir. Þáttarinn samþykkir málfræðilega markað inntak og býr til úttak í samræmi við grunnt setningarfræðilegt merkingarskema (e. shallow syntactic annotation scheme). Þáttarinn samanstendur af liðgerðareiningu (e. phrase structure module) og setningarhlutverkseiningu (e. syntactic functions module). Báðar einingarnar samanstanda af runu stöðuferjalda, sem hver og ein bætir setningafræðilegum upplýsingum við undirstrengi (e. substrings) inntakstextans.

Fullþáttun (e. full constituency parsing) er grunnþörf málrýnieiningar. Þar er reglubyggður þáttari nauðsynlegur þar sem hann leyfir handvirka viðbót sérstaklega merktra „villureglna“ við samhengisfrjálsa málfræði sína, þar sem algengum málfræðivillum er lýst. Reglubyggði þáttarinn er einnig notaður til að útbúa þjálfunargögn fyrir tauganet byggð á fullþáttara. Þess má geta að utan máltækniáætlunarinnar er þegar farið að nota reglubyggða þáttarann í hugbúnaði.

Tauganetsþáttarinn, öfugt við reglubyggða þáttarann, umber málfræðilegar villur og býr til áætlað þáttunartré jafnvel fyrir setningar sem eru ekki réttar. Hann gæti því nýst frekar en reglubyggði þáttarinn þegar kemur að því að þátta raddbeiðnir og annað inntak sem kemur beint frá notanda, óbreytt. Aftur á móti er ekki hægt að aðlaga málfræði hans að sértækum notkunardæmum; slík aðlögun krefst marktæks (e. nontrivial) þjálfunarsetts fyrir hvert notkunardæmi.

Varða 6 – lýsing

M5: Tauganetsþáttarinn hefur verið þjálfaður á uppfærðri og bættri þjálfunarmálheild. Tauganetið verður þjálfað uns það nær samleitni (e. converge) (þ.e. stærð þjálfunarmálheildarinnar er ekki beinlínis skilgreind fyrirfram). Grunnlína fyrir tauganetsþáttara er skilgreind. Grunnþáttarinn hefur verið gefinn út hjá CLARIN. Reglubyggði fullþáttarinn er uppfærður í hirslu CLARIN.

M6: Tauganetsþáttarinn er gefinn út hjá CLARIN í Docker-gám sem auðvelt er að setja upp. Markmiðið er að hafa svipaða eða betri F-einkunn en fyrir reglubyggða þáttarann (þ.e. >= 75,17% mælt með evalb).

Afurðir

Öllum markmiðum vörðunnar var náð.

Reglubyggði grunnþáttarinn er aðgengilegur á CLARIN (http://hdl.handle.net/20.500.12537/122) sem IceParser 1.5.0. Minniháttar breytingar hafa verið gerðar frá útgáfu, og nýjustu útgáfu IceParser má finna á GitHub (https://github.com/hrafnl/icenlp). Reglubyggði fullþáttarinn er aðgengilegur á CLARIN (http://hdl.handle.net/20.500.12537/147), á GitHub (https://github.com/mideind/ReynirPackage), og á Python Package Index (PyPI) sem GreynirPackage 3.4.0. Tilreiðarinn sem er innifalinn (verkþáttur I3) hefur einnig verið uppfærður á CLARIN (http://hdl.handle.net/20.500.12537/136) og GitHub (https://github.com/mideind/Tokenizer) sem Tokenizer 3.3.2.
Tauganetsþáttarinn er aðgengilegur á CLARIN (https://repository.clarin.is/repository/xmlui/handle/20.500.12537/149).
Uppfærða og endurbætta þjálfunarmálheildin er aðgengileg á CLARIN (http://hdl.handle.net/20.500.12537/119) og á GitHub (https://github.com/mideind/GreynirCorpus).
Greint er frá grunnlínu fyrir tauganetsþáttarann og niðurstöðum í skýrslunni.

Verkþáttur I5 heldur áfram á þriðja ári.

Skýrsla

1. Grunnþáttari (M5)

Helsta breytingin fólst í því að aðlaga IceParser að nýlega uppfærðri markaskrá fyrir íslensku. Aðrar breytingar fólust aðallega í því að taka á áður þekktum villum og nýjum villum vegna aðlögunar nýrrar markaskrár.

Mat var gert með notkun tveggja prófunarsetta: Prófunarsett með 500 setningum úr fréttagreinamálheild sem Miðeind útvegaði og 509 setningar valdar af handahófi úr málheild Íslensku orðtíðnibókarinnar (IFD) (aðallega bókmenntir), sem voru einnig notaðar við mat á IceParser árið 2007[5].

Prófunarsett fréttagreina var tilreitt með Python tokenizer pakkanum (I3) og svo leiðrétt handvirkt, á meðan prófunarsettið frá IFD kom þegar tilreitt.

Bæði prófunarsettin voru málfræðilega mörkuð með ELECTRA-Base-markara þjálfuðum á Risamálheildinni (I4). Tveir aðilar fóru handvirkt yfir niðurstöður mörkunar og fundu mjög fáar villur, svo mörkunarvillur teljast smávægilegar.

Niðurstöður á mati IceParser eftir fyrrgreindar breytingar eru eftirfarandi:

Prófunarsett fréttagreina frá Miðeind:

Liðgerðareining:		*Setningarhlutverkseining:*	
Heimt:	95,52%	Heimt:	78,36%
Nákvæmni:	95,07%	Nákvæmni:	80,63%

[5] https://dspace.ut.ee/handle/10062/2563

F-gildi: 95,29%	**F-gildi: 79,48%**

Prófunarsett úr málheild Íslensku orðtíðnibókarinnar:

Liðgerðareining:	*Setningarhlutverkseining:*
Heimt: 96,20%	Heimt: 80,43%
Nákvæmni: 96,48%	Nákvæmni: 85,93%
F-gildi: 96,34%	**F-gildi: 83,09%**

2. Reglubyggður fullþáttari (M5)

2.1 Endurbætur

Breytingar voru gerðar á þáttunarskema Greynis til að gera það gagnsærra og reglulegra. Vörpun fyrir aðalliði var skilgreind betur í því ferli. Fyrir utan þetta voru málfræðireglur uppfærðar og leiðréttar þar sem þurfti.

2.2 Mat og niðurstöður

Fyrir vörðu M4 greindum við frá eftirfarandi niðurstöðum:

Heimt út frá svigum: 74,93%	Meðalskörun: 2,96
Nákvæmni út frá svigum: 76,01%	Nákvæmni marka: 94,57%
F-gildi út frá svigum: 75,42%	

Fyrir vörðu M5 greinum við frá eftirfarandi niðurstöðum:

Heimt út frá svigum: 80,74%	Meðalskörun: 2,16
Nákvæmni út frá svigum: 81,75%	Nákvæmni marka: 94,39%
F-gildi út frá svigum: 81,21%	

Ranglega myndaðar setningar (töflugögn, íþróttaúrslit, tíst o.s.frv.) eru teknar úr niðurstöðum. Eins og sjá má hækka heimt og nákvæmni. Í skýrslu M4 nefndum við að algengustu villurnar voru vegna vörpunarskema, sem hefur verið endurbætt.

Aðrar algengar villur eru viðhenging röklíða (e. argument attachment) og viðhenging forsetningarliða (e. PP attachment). Vinna við nýja sagnliðaramma, sem nefnd er í skýrslu M4[6], er vel á veg komin. Gögnin hafa verið auðguð með nýjum reitum með ítarlegri upplýsingum, og samþætt í þáttarann. Endurbætt meðhöndlun sagnliðaramma, sem notar allar upplýsingar sem veittar eru, er í vinnslu.

Til að taka á villumeðhöndlun var trjáleitarforrit útfært og bætt í prófunarpípuna. Þetta gerði okkur kleift að leita að algengum setningagerðum, eins og ákveðnum gerðum undirskipaðra setninga (e. subclauses), og bera saman við þróunarsett gullstaðals og samsvarandi úttak. Þetta hjálpaði gríðarlega við það að leggja áherslu á eitt vandamál í einu og fá heildstæða yfirsýn yfir hvert og eitt.

[6] Þessi vinna er ekki afurð innan máltækniverkefnisins.

3. Þjálfunarmálheildir (M5)

Til að gefa bestu mögulegu þjálfunargögn fyrir tauganetsþáttara var uppbygging GreynirCorpus-trjábankans endurskilgreind.

Prófunarsett gullstaðals (500 setningar) er hugsað fyrir mat. Þróunarsett gullstaðals (4.500 setningar) er hugsað fyrir fínstillingu.

Sérvalið undirsett, kallað silfurmálheildin (e. silver corpus) (603.659 setningar), er hugsuð fyrir sérhæfðari fínstillingu. Setningar fyrir silfurmálheildina voru valdar vegna möguleika þeirra til að veita mikilvægar upplýsingar við þjálfun. Margar gerðir leitarnáms (e. heuristics) voru notaðar til að velja setningar, eins og útilokun setninga ákveðinna sviða, og óþáttaðra, óhástafaðra, erlendra eða stuttra setninga og fyrirsagna.

Heil málheild, kölluð koparmálheildin (e. copper corpus) (10.385.587 setningar), er hugsuð fyrir frumþjálfun, ef þörf er á. Aðeins erlendar, óhástafaðar, og óþáttaðar setningar voru útilokaðar. Stuttum setningum (1.652.938 setningar) og fyrirsögnum (531.855 setningar) var safnað í sér málheildir.

Til að auðvelda hreinsun málheildanna var trjáleitarforritið sem nefnt er í kafla 3.2 einnig innifalið og gefið út með málheildunum. Með þessu gátum við skoðað ákveðnar setningagerðir og tryggt að þær væru þáttaðar eins í öllum skrám gullstaðals. Forritið ætti einnig að nýtast í setningafræðirannsóknum og vonandi á fleiri sviðum.

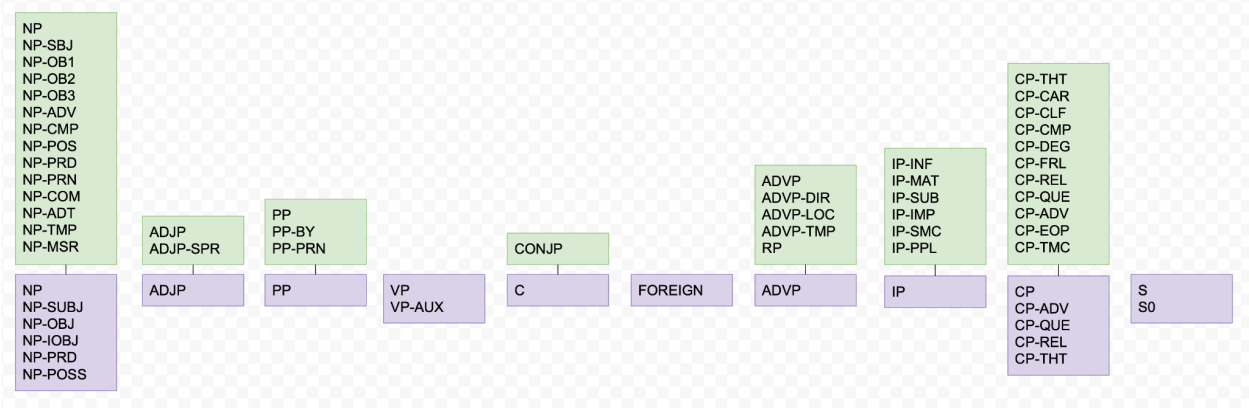
4. Tauganetsþáttari

4.1 Grunnlína (M5)

Til þess að koma á grunnlínu fyrir tauganetsþáttarann var notað tiltækt íslenskt tauganetsþáttunarlíkan[7]. Prófunarsettið úr GreynirCorpus var notað í þessum tilgangi, og hver skrá var þáttuð sjálfvirkt með þáttunarlíkaninu. Úttakinu var svo varpað yfir á almenna skemað, eins og var gert við M3 með reglubyggða fullþáttarann og grunnþáttarann. Mat á úttakinu sýnir F-gildi upp á 87,07 og eftirfarandi niðurstöður grunnlínu:

```
Heimt út frá svigum:    86,31%          Meðalskörun:     0,09
Nákvæmni út frá svigum: 87,85%          Nákvæmni marka: 89,90%
F-gildi út frá svigum:  87,07%
```

[7] http://linguist.is/wp-content/uploads/2020/06/arnardottir2020neural.pdf



Mynd: *Lýsing á vörpun orðasambanda frá tauganetsþáttara til almenna skemans. Grænu kassarnir tákna tauganetsþáttarann og fjólubláu kassarnir almenna skemað.*

Eins og tilfellið var fyrir reglubyggðu þáttarana í skýrslu M3 hefur tauganetsskemað nokkra liði sem eru ekki í almenna skemanu. Hið andstæða á einnig við í sumum tilfellum, þar sem almenna skemað hefur liði sem finnast ekki í tauganetsskemanu.

4.2 Þjálfun og niðurstöður (M6)

Íslenskt BERT-líkan (IceBERT) var valið sem „uppstreymis“-mállíkanið (e. upstream language model). Það var forþjálfað á Risamálheildinni ásamt öðrum málheildum eins og *Fornritin* (gömul rit) frá malfong.is. Almenna nálgunin er svipuð og sú sem notuð er í grunnlínu sem er lýst í kafla 4.1. Tekið skal fram að margmála „uppstreymis“-BERT-líkanið sem grunnlínan notar inniheldur ekki íslensku í forþjálfun sinni.

Lágmarksafbrigði Berkeley-þáttarans var notað sem byrjunarreitur högunar fyrir tauganetslíkanið. Þessi þáttari þarfnast ekki upplýsinga um málfræðilega mörkun í inntaki sínu, og reiðir sig ekki á frekari hönnun eiginleika. Tauganetið var fyrst þjálfað á silfursetti GreynirCorpus trjábankans með upprunalegu merkingarskema. Það var svo fínstillt á 3500 af 4500 setningum úr þróunarsetti gullstaðals.

Mat var gert með sama prófunarsetti úr GreynirCorpus eins og fyrir hina þáttarana, þ.e. með notkun prófunarpípunnar sem þróuð var á fyrsta ári og uppfærðu málheildunum sem lýst er í þriðja hluta. Úttakinu var varpað á, og metið á, almenna skemanu sem skilgreint var á fyrsta ári.

Í þeim tilgangi að bera saman við grunnlínu greinum við frá eftirfarandi niðurstöðum, fyrst eftir einfalda silfurþjálfun (einkunn fyrir ómerkta sviga (e. unlabelled bracketing scores) er inni í sviga):

```
Heimt út frá svigum:    84,91 (86.30)          Meðalskörun:    1,93
Nákvæmni út frá svigum: 82,37 (83.75)          Nákvæmni marka: N/A
F-gildi út frá svigum:  83,59 (84.97)
```

Í kjölfarið, eftir fínstillingu með þróunarsetti gullstaðals, eru niðurstöður eftirfarandi:

```
Heimt út frá svigum:    93,25 (93,99)          Meðalskörun:    0,69
Nákvæmni út frá svigum: 88,28 (88,98)          Nákvæmni marka: N/A
F-gildi út frá svigum:  90,67 (91,39)
```

Þetta er hækkun F-gildis um +3.6 stig frá grunnlínu.

Gildin að ofan eru reiknuð út frá setningum með engum margorða tókum[8] (í úttaki tilreiðara/þáttara Miðeindar), til að geta borið saman við niðurstöður úr grunnlínu Berkeley-þáttarans í 4.1 sem útiloka þær setningar.[9]

Við greinum einnig frá niðurstöðum *evalb* reiknuðum út frá öllum setningum í prófunarsettinu, sem merktar voru hreinar af málfræði- eða stafsetningarvillum, af þeim sem unnu við merkingar. Fyrst, aðeins eftir silfurþjálfun:

```
Heimt út frá svigum:    84,92 (86,47)          Meðalskörun:    1,72
Nákvæmni út frá svigum: 82,06 (83,57)          Nákvæmni marka: N/A
F-gildi út frá svigum:  83,45 (84,98)
```

Eftir fínstillingu með þjálfunarsetti gullstaðals eru niðurstöður eftirfarandi:

```
Heimt út frá svigum:    93,21 (94,11)          Meðalskörun:    0,56
Nákvæmni út frá svigum: 87,65 (88,48)          Nákvæmni marka: N/A
F-gildi út frá svigum:  90,33 (91,19)
```

Fyrir utan betra F-gildi en reglubyggðu þáttararnir á prófunarsetti málfræðilega viðunandi setninga nær tauganetsþáttarinn einnig betri árangri með ranglega myndaðar setningar og býr til þáttunartré fyrir hvaða inntak sem er. Þetta gerir hann hentugan fyrir notkun á ýmsum sviðum, eins og leit að málfræðivillum, sem verður skoðuð frekar á þriðja ári.

[8] Tölvugerð þáttunartré GreynirCorpus innihalda sameinaða tóka eins og „árið 2021“ og „á þessari stundu“ sem margorða atviksliði.
[9] 40% setninga prófunarsetts voru útilokaðar vegna þessarar kröfu.

I7 – Nafnakennsl

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Í upphaflegri verkáætlun máltækniáætlunar var aðskilinn verkþáttur sem sneri að því að stækka Gullstaðal (MIM-GOLD) með nafneiningamerkingum (e. named entity tags). Þessi vinna hefur þegar verið framkvæmd (2019-2020) í verkefni með fjármögnun frá Markáætlun í tungu og tækni. Málheildin sem til varð hefur verið kölluð MIM-GOLD-NER (Nafnakennslamálheildin). Auk merkinganna sjálfra hafa allnokkur nafnakennslalíkön verið þróuð og metin. F1-gildi þeirra líkana sem ná bestum árangri fyrir hefðbundnu flokkana *persóna*, *staðsetning*, *samtök* er 87,9%, 86,7% og 80,2%, í þessari röð. Þessi gildi teljast grunnlína fyrir vinnu sem framundan er.

Á öðru ári er markmiðið að byggja á grunni undanfarandi verkefnis og auka nákvæmni með því að fínstilla grunn-BERT-líkt líkan og greina og flokka einindi sem eru vitlaust skrifuð. Einnig verða forritaskil þróuð sem taka við hráum texta í inntaki og skila þeim texta merktum með nafneiningum.

Varða 6 – lýsing

M5:
NER-kerfi þjálfað með því að fínstilla BERT-líkt líkan sem landfræðiorðabók hefur verið bætt við. Þjónn og forritaskil komin í gagnið. Markmið F1-gilda fyrir hefðbundnu flokkana *persóna*, *staðsetning*, *samtök*, eru grunnlínurnar að ofan + 2%.

M6
NER-kerfi afhent til CLARIN í formi Docker-uppsetningu sem auðvelt er að setja upp, byggt á besta BERT-líka líkaninu, auðgað með landfræðiorðabókum, stækkað með hæfni til að greina og flokka einindi sem eru vitlaust skrifuð.

Afurðir

NER-kerfi afhent til CLARIN í formi Docker-uppsetningu sem auðvelt er að setja upp, byggt á besta BERT-líka líkaninu, stækkað með hæfni til að greina og flokka einindi sem eru vitlaust skrifuð.

Eldri útgáfa kerfisins er aðgengileg á CLARIN með ELECTRA-base líkani (https://repository.clarin.is/repository/xmlui/handle/20.500.12537/118), enn fremur er uppfærð útgáfa komin á CLARIN (http://hdl.handle.net/20.500.12537/159). Báðar útgáfur eru einnig

aðgengilegar á GitHub (https://github.com/cadia-lvl/Icelandic-NER-API) og á Dockerhub (https://hub.docker.com/r/asmundur1/ner_api).
Öllum markmiðum var náð.
Verkþætti I7 lýkur við M6.

Skýrsla

Þjónn bakenda og forritaskil voru þróuð saman og sett upp á Google Cloud. Þessi pakki er til staðar sem Docker-mynd á DockerHub. (https://hub.docker.com/r/asmundur1/ner_api). Kóðinn fyrir forritaskilin er á GitHub, aðgengilegur hér:
https://github.com/cadia-lvl/ELECTRA-base-NER-API

Pakkinn inniheldur hraðvirkari spár þar sem besta líkanið (IceBERT[10]) er notað til að spá fyrir um texta, en einnig hægvirkari blandaðar aðferðir (e. ensemble method), sem notar fjögur mismunandi transformer- (IceBERT, ELECTRA-base, ConvBERT-small og multilingual-BERT) líkön og sameinar þau með CombiTagger. Þetta ferli fól í sér að innleiða kjarnavirkni CombiTagger inn í Docker-mynd sem er einnig hýst á DockerHub: (https://hub.docker.com/r/asmundur1/combitagger).

Aðferðir til að greina og flokka einindi sem eru vitlaust skrifuð voru skoðaðar, en ákveðið var að leggja meiri áherslu á BERT-líku líkönin, þar sem það væri mikilvægara.

Tölfræði prófunarsetts

Prófanarsettið sem er notað er prófunarsettið úr MIM-GOLD-NER málheildinni. Þekktar einingar eru einingar í prófunarsettinu sem koma fram í þjálfunarsettinu. Til einföldunar var nákvæm strengjasamsvörun notuð. MIM-GOLD-NER og hvernig hún skiptist í þjálfunar- staðfestingar-, og prófunarsett má finna á CLARIN (https://repository.clarin.is/repository/xmlui/handle/20.500.12537/140).

	Heildarfjöldi eininga	Þekktar einingar	Óþekktar einingar
Persóna	1329	596	733
Staðsetning	1129	764	365
Samtök	1389	814	575

Afköst líkana

	Nákvæmni	Heimt	F1

[10] IceBERT er RoBERTalíkan þjálfað (af Miðeind ehf.) á bæði RMH og öðrum gögnum frá Icelandic Common Crawl Corpus.

Multilingual BERT	89,28%	88,98%	89,13
Persóna	91,21%	92,10%	91,65
Samtök	86,01%	86,32%	86,17
Staðsetning	91,00%	89,55%	90,27
ConvBERT-small	90,25%	89,88%	90,07
Persóna	93,51%	92,10%	92,80
Samtök	88,00%	87,11%	87,55
Staðsetning	93,42%	93,00%	93,21
ELECTRA-base	91,75%	92,05%	91,90
Persóna	94,05%	95,11%	94,58
Samtök	91,35%	89,70%	90,52
Staðsetning	94,35%	93,27%	93,81
***IceBERT** (Miðeind)*	92,57%	92,90%	**92,73**
Persóna	94,29%	95,71%	**95,00**
Samtök	92,31%	92,44%	**92,37**
Staðsetning	94,33%	92,91%	**93,62**
CombiTagger	93,26%	93,15%	**93,21**
Persóna	94,63%	95,56%	**95,10**
Samtök	92,18%	91,65%	**91,91**
Staðsetning	95,43%	94,24%	**94,83**

F1-gildi grunnlínu (frá fyrri vinnu) afkastabestu líkananna er fyrir hefðbundnu flokkana *persóna*, *staðsetning*, *samtök* og er 87,9%, 86,7% and 80,2%, í þessari röð. Besta einstaka líkanið er IceBERT, sem nær F1-gildum **95,00**, **93,62** og **92,37** fyrir flokkana *persóna*, *staðsetning* og *samtök*, og það nær heildar F1-gildinu **92,73**. Þegar líkönin fjögur eru sett saman í eitt með CombiTagger nær það heildar F1-gildinu **93,21**, þar sem hefðbundnu flokkarnir (persóna, staðsetning, samtök) ná **95,10**, **94,83** og **91,91**, í þessari röð, sem er betri árangur en IceBERT nær á flokkunum persóna og staðsetning, á meðan IceBERT nær betri árangri en CombiTagger á flokknum samtök.

I8a – Merkingargreining – BERT-lík mállíkön

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Þessi verkþáttur var ekki skilgreindur í upphaflegri máltækniáætlun, en áætlunin var aðlöguð í ljósi nýlegrar þróunar í nýjustu aðferðum í tauganetsmállíkönum. Slík líkön eru orðin grunnur að mörgum af bestu útfærslunum fyrir ýmis verkefni málvinnslu (e. NLP).

BERT-lík (Transformer-byggð) tauganetsmállíkön fyrir íslensku verða þjálfuð, í náinni samvinnu við önnur undirverkefni sem bíða slíkra líkana til að bæta við tæknistafla þeirra. Þessi verkefni eru málrýni, vélþýðingar, talgreining, málfræðileg mörkun/lemmun, nafnakennsl og þáttun.

Nokkur BERT-lík líkön af þeim gerðum sem nýlega hafa komið fram verða hönnuð og metin, þeirra á meðal BERT, RoBERTa, ALBERT og ELECTRA. Frammistaða líkananna mun að miklu leyti fara eftir magni, fjölbreytileika og gæðum einmála málheilda sem verða í boði til þjálfunar.

Þess ber að geta að þjálfun BERT-líkra mállíkana á miklu magni gagna krefst umtalsverðs vinnslutíma, hvort sem hann er mældur í FLOPS, heildar CPU/GPU/TPU klukkustundum eða rauntíma. Ítraðar tilraunir og prófanir mismunandi líkana eru því mjög tímafrekar, jafnvel þó að þær séu vel skipulagðar, og þetta þarf að hafa í huga við áætlun mannmánaða.

Varða 6 – lýsing

M5:

- Ýmsar aðferðir orðflísatilreiðingar hafa verið bornar saman.
- Möguleiki þess að auka árangur „neðanstreymis“ (e. downstream performance) með því að bæta textum af vefnum í forþjálfunarmálheild hefur verið skoðaður.

M6:

- Eitt eða fleiri bestu mállíkön hafa verið endurþjálfuð með endurbættum líkanastikum og forþjálfunargögn hafa verið gefin út.
- Grunnlínur fyrir „neðanstreymisverkin“ (e. downstream tasks) NER (F1-gildi: 83,65%) og málfræðileg mörkun/PoS (94,47% nákvæmni) hafa verið endurbættar. Markmiðið er að fækka villum um 25%.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.

- Málheildin The Icelandic Crawled Corpus (ICC), ómerkt textamálheild með u.þ.b. 930M tókum skröpuðum af netinu, er aðgengileg á huggingface.co (https://huggingface.co/datasets/jonfd/ICC).
- Markmiðinu að fækka NER/PoS villum um 25% hefur verið náð.
    - Besta líkanið okkar eins og er nær F1-gildinu 92,84% fyrir NER (sem minnkar villutíðni um 56%), og nákvæmni upp á 96,99% fyrir málfræðilega mörkun (e. PoS tagging) (sem minnkar villutíðni um 46%).
- Besta líkanið úr tilraunum okkar er aðgengilegt á huggingface.co (https://huggingface.co/jonfd/convbert-small-igc-is).
    - Stærri útgáfa besta líkans okkar verður þjálfað í nóvember.

Verkþáttur I8 heldur áfram á þriðja ári.

Skýrsla

Forþjálfunarmálheild

Transformer-byggð mállíkön eru venjulega þjálfuð í tveimur áföngum. Fyrst eru þau *forþjálfuð* á miklu magni ómerktra texta í verki eins og t.d. að spá fyrir um næsta tóka miðað við gefna textarunu á undan. Þegar líkan hefur verið forþjálfað er hægt að *fínstilla* það á hagnýtari verkum, eins og t.d. málfræðilegri mörkun (e. PoS tagging), nafnakennslum (NER), fyrirspurnum eða sjálfvirkum textaútdrætti (gjarnan vísað til sem *„neðanstreymisverk“* í þessu samhengi).

Þó að Risamálheildin (IGC, sjá G1) – stærsta opna einmála málheildin fyrir íslensku – sé smærri en margar forþjálfunarmálheildir fyrir önnur tungumál hefur hún reynst nægilega stór til að forþjálfa líkön sem ná fyrsta flokks niðurstöðum fyrir fjölda íslenskra NLP-verka, eins og málfræðilega mörkun og nafnakennsl.

Það hefur verið sýnt með óyggjandi hætti að stækkun forþjálfunarmálheildarinnar hefur jákvæð áhrif á „neðanstreymisárangur“ (e. downstream performance) fyrir Transformer-byggð líkön. Í þeim tilgangi að búa til stærri forþjálfunarmálheild fyrir íslensku hafa u.þ.b. 930M tókar verið skrapaðir frá völdum vefsíðum, eins og fréttasíðum, vefsíðum stjórnvalda, spjallborðum og bloggum. Samtals höfum við skrapað texta frá 24 íslenskum lénum. Gögnum okkar er svo hægt að bæta við aðrar málheildir, eins og The Multilingual Colossal Clean Crawled Corpus (mC4), margmála gagnasett sem inniheldur u.þ.b. 1,1 milljarð tóka af íslenskum texta. Við sjáum að gagnasett okkar inniheldur töluvert meiri texta á hvert lén en aðrar málheildir vefskriðla (e. web-crawled corpora). Þegar við berum málheildina mC4 við gagnasett okkar sjáum við að það inniheldur aðeins 5,9% tókanna fyrir sömu lén.

ICC hefur verið gefin út með opnu leyfi.

Mat á líkönum og þjálfunarstikum

Við forþjálfum tvær mismunandi gerðir Transformer-líkana, ELECTRA og ConvBERT, og metum áhrifin af því að nota mismunandi forþjálfunarmálheildir, orðflísatilreiðingu og stærð orðaforða.

Líkan

Fyrstu tilraunir okkar voru gerðar með ELECTRA-líkaninu, sem hefur verið sett fram sem einstaklega skilvirkt á gögn, og því vel við hæfi fyrir tungumál með lítið eða meðalmikið magn

gagna (e. low and medium-resource languages). Við gerðum frekari tilraunir með ConvBERT-líkanið, sem er byggt á ELECTRA-líkaninu en innleiðir kviklegan feðmingarvirki byggðan á spönn (e. span-based dynamic convolution operator) inn í sjálfsathyglishausana (e. self-attention heads).

Forþjálfunarmálheild
Til að mæla mögulega kosti þess að stækka forþjálfunarmálheildina búum við til nýtt gagnasett með sameiningu RMH- og mC4-málheildunum og okkar eigin vefskröpuðu gögnum (vísað til neðar sem RMH+). Fyrir mC4-málheildina síum við út öll lén sem eru þegar í RMH, auk flestra erlendra léna, sem við sáum að samanstóðu að mestu af erlendum og vélþýddum texta, auk gagna sem teljast ekki tungumál. Samtals samanstendur nýja forþjálfunarmálheildin af 2,94 milljörðum tóka, á meðan RMH samanstendur af 1,64 milljörðum tóka. Við tökum fram að gagnasett okkar hefur stækkað gríðarlega síðan tilraunir hófust.

Orðflísatilreiðari
Transformer-byggð líkön vinna venjulega inntakstexta með orðflísatilreiðara (e. subword tokenizer) sem skiptir óþekktum orðum í röð þekktra orðflísa (e. subwords). Orðflísatilreiðarar vinna með orðaforða af ákveðinni stærð, yfirleitt lærðum úr forþjálfunarmálheildinni með algrími eins og byte-pair encoding (BPE). Þetta orsakar marktæka fækkun sjaldgæfra og óþekktra orða, sem skýra mikinn fjölda villna í mörgum NLP-líkönum. Því hefur verið haldið fram að BPE-algrímið, sem er almennt valið fyrir orðflísatilreiðingu, sé óákjósanlegt fyrir forþjálfun mállíkana, og að betri kostir séu til, eins og t.d. Unigram-tilreiðingaralgrímið. Í tilraunum okkar berum við saman þessi tvö algrím.

Við þjálfum BPE- og Unigram-tilreiðara með forritasafni tilreiðarans frá HuggingFace. Til samanburðar þjálfum við til viðbótar nokkra Unigram-tilreiðara með forritasafni SentencePiece (SP) frá Google.

Stærð orðaforða
Orðflísatilreiðarar nota almennt orðaforða af stærð milli 30.000 og 50.000 tóka. Takmarkaðar rannsóknir hafa verið gerðar á áhrifum stærðar orðaforða á árangur „neðanstreymis“ (e. downstream performance), sérstaklega fyrir beygingarfræðilega flókin tungumál með lítið eða meðalmikið tiltækt magn gagna. Til að skilja þetta samband betur gerum við tilraunir með forþjálfun líkana með mismunandi stærðir orðaforða: 32.000, 64.000, 96.000 og 128.000 tóka.

Mat
Við þjálfum smáar útgáfur ELECTRA- og ConvBERT-líkananna, sem eru forþjálfuð fyrir eina milljón skrefa með sjálfgefnum stillingum. Forþjálfuðu líkönin eru svo fínstillt á þremur verkum: Málfræðilegri mörkun (PoS), nafnakennslum og venslaþáttun (e. dependency parsing) (DP).

- Fyrir málfræðilega mörkun eru líkönin fínstillt á MIM-GOLD-málheildinni (útgáfu 21.05) fyrir 20 mynstrasöfn með námshraðann 5e-5 og skammtastærð 16, og metið með tíföldum krossprófunum.
- Fyrir nafnakennsl er líkanið fínstillt á MIM-GOLD-NER-málheildinni fyrir 10 mynstrasöfn með námshraðann 5e-5 og skammtastærð 16, og metið með tíföldum krossprófunum.
- Fyrir DP er líkanið fínstillt á Universal Dependencies-útgáfu IcePaHC-málheildarinnar fyrir 200 mynstrasöfn með námshraðann 2e-3 og skammtastærð 5000.

Niðurstöður

	ConvBERT-Small – IGC				ConvBERT-Small – IGC+			
Tilreiðari	**POS**	**NER**	**DP**	**AVG**	**POS**	**NER**	**DP**	**AVG**
BPE (32k)	96,88%	92,03%	84,75%	91,22%	96,84%	92,04%	84,90%	91,26%
BPE (64k)	**97,04%**	92,20%	85,02%	91,42%	96,97%	92,53%	**85,33%**	91,61%
BPE (96k)	96,90%	92,76%	85,10%	91,59%	96,91%	**92,83%**	84,94%	91,56%
BPE (128k)	96,96%	92,62%	**85,21%**	91,60%	96,94%	92,80%	85,27%	**91,67%**
UG (32k)	96,79%	91,71%	84,81%	91,10%	96,89%	91,79%	84,83%	91,17%
UG (64k)	96,88%	92,32%	84,89%	91,36%	96,90%	92,32%	85,09%	91,44%
UG (96k)	97,03%	92,48%	84,93%	91,48%	**96,99%**	92,42%	85,08%	91,50%
SP-UG (32k)	96,83%	91,78%	85,17%	91,26%	-	-	-	-
SP-UG (64k)	96,90%	92,44%	85,00%	91,44%	-	-	-	-
SP-UG (96k)	96,99%	**92,84%**	85,09%	**91,64%**	-	-	-	-
SP-UG (128k)	96,87%	92,75%	85,19%	91,60%	-	-	-	-

	ELECTRA-Small – RMH				ELECTRA-Small – IGC+			
Tilreiðari	**POS**	**NER**	**DP**	**AVG**	**POS**	**NER**	**DP**	**AVG**
BPE (32k)	96,84%	91,23%	84,90%	90,99%	96,91%	91,26%	84,85%	91,01%
BPE (64k)	96,96%	91,88%	84,95%	91,27%	96,98%	91,91%	84,84%	91,24%
BPE (96k)	96,92%	92,23%	**85,11%**	91,42%	96,96%	**92,20%**	**85,01%**	**91,39%**
UG (32k)	96,96%	91,30%	84,77%	91,01%	96,80%	90,97%	84,64%	90,80%
UG (64k)	97,03%	91,78%	84,60%	91,14%	96,98%	91,86%	84,92%	91,25%
UG (96k)	**97,06%**	**92,26%**	85,01%	**91,44%**	**97,02%**	91,84%	84,75%	91,20%

Líkan
ConvBERT-Small-líkön ná alltaf betri árangri en ELECTRA-Smal- líkönin á „neðanstreymisverkum“ (e. downstream tasks), með 0,21% hærri heildareinkunn að meðaltali.

Stærð orðaforða
Stækkun orðaforðans fyrir tilreiðarana hefur gríðarlega jákvæð áhrif á meðalárangur líkananna „neðanstreymis“, með hverfandi áhrifum eftir stærð orðaforða upp á 96þ. Taka skal fram að aukinn árangur „neðanstreymis“ kemur á kostnað aðeins hærri minnisþarfa og hægari ályktunartíma (e. slower inference time).

Orðflísatilreiðari
Líkön þjálfuð með BPE-tilreiðara ná meðalárangri „neðanstreymis“ upp á 91,30%, sem ná örlítið betri árangri en líkön þjálfuð með HuggingFace Unigram-tilreiðaranum, sem ná meðalárangri

upp á 91,22%. Líkön þjálfuð með SentencePiece Unigratilreiðaranum ná bestum árangri, með meðalárangri „neðanstreymis“ upp á 91,45%, sem gefur vísbendingar um að hann sé betri útfærsla á Unigram-algríminu.

Forþjálfunarmálheild
Stækkun þjálfunarmálheildarinnar með vefskröpuðu gögnunum hefur lítil sem engin áhrif á meðalárangur líkananna „neðanstreymis“. Mögulegt er að líkönin sem við mátum séu of smá til að nýju gögnin nýtist nógu vel, eða að gæði íslenska hluta mC4-málheildarinnar séu svo lág að þau hafi neikvæð áhrif á árangur líkananna (t.d. vegna villna við stafakóðun (e. character encoding), tungumálaflokkun, eða tilvist mikils magns gagna sem teljast ekki tungumálagögn o.s.frv.)

I8b – Forþjálfaðar greypingar

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

Markmið

Að gefa út forþjálfaðar orðagreypingar og samhengisháðar greypingar (e. contextualized embeddings) með skýrum ytri stikum (e. hyperparameters) á málheild sem hefur verið vandlega lýst og málvísindalega forunnin (í samræmi við líkön sem gefin eru út á orðagreypingahirslu Nordic Language Processing Laboratory, staðsett á http://vectors.nlpl.eu/repository/). Greypingarnar skal meta með innri matssettum fyrir setninga- og orðhlutafræðilegan skyldleika.

Varða 5 – lýsing

Greypingar gefnar út, byggðar á word2vec, GloVe, fasttext, og/eða ELMo. Skýrsla um niðurstöður mats með mismunandi ytri stikum (e. hyperparameters).

Afurðir

Öllum markmiðum vörðunnar var náð.
Github: https://github.com/stofnun-arna-magnussonar/ordgreypingar_embeddings
CLARIN-IS: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/121
https://repository.clarin.is/repository/xmlui/handle/20.500.12537/120
Hirslurnar innihalda mörg sett af forþjálfuðum greypingum og samsvarandi ytri stikum þeirra, þar sem hvert sett inniheldur bestu greypingar fyrir tiltekinn flokk þess setts.

Við flokkum þessi sett samkvæmt eftirfarandi forsendum:

- Forunnin málheild: Lemmuð eða ólemmuð
- Þjálfunarrammi: GloVe, fastText, eða word2vec
- Sértækt þjálfunarlíkan: Skip-gram eða Continuous Bag of Words (CBOW)
- Matssett: Icelandic Multi-SimLex (MSL) eða IceBATS

Samkvæmt því gefum við út eftirfarandi sett greypinga:

- FastText, sem notar CBOW, með ólemmuðum gögnum, fínstillt fyrir IceBATS
- FastText, sem notar Skip Gram, með lemmuðum gögnum, fínstillt fyrir IceBATS
- FastText, sem notar Skip Gram, með lemmuðum gögnum, fínstillt fyrir MSL
- Word2vec, sem notar Skip Gram, með lemmuðum gögnum, fínstillt fyrir IceBATS

- Word2vec, sem notar CBOW, með ólemmuðum gögnum, fínstillt fyrir IceBATS
- Word2vec, sem notar Skip Gram, með lemmuðum gögnum, fínstillt fyrir MSL
- GloVe, sem notar lemmuð gögn, fínstillt fyrir IceBATS
- GloVe, sem notar ólemmuð gögn, fínstillt fyrir IceBATS
- GloVe, sem notar lemmuð gögn, fínstillt fyrir MSL

Verkþætti I8b lýkur við M6.

Skýrsla

Inngangur

Vinna okkar í M5 og M6 beindist aðallega að greypingunum. Þetta fól í sér uppbyggingu nauðsynlegs ramma til að búa til greypingar með aðferðum sem nefndar eru að ofan, aðlaga þær að þjálfunar- og matsgögnum okkar, og gera sífelldar tilraunir til að finna bestu ytri stika.

Þar að auki fól þjálfun MSL í sér útfærslu ýmissar eftirvinnslu. (Við notum „MSL“ til að vísa til íslenskrar útfærslu Multi-SimLex-matssettsins, og „OMSL“ fyrir upprunalega útgáfu þegar við þurfum að skilja á milli þeirra.) Þjálfun IceBATS krafðist hins vegar ákveðinnar aðlögunar til að mæta mismunandi kröfum ýmissa undirflokka IceBATS-prófa þegar þeim er beitt á íslensku. Áberandi dæmi um þetta er að til að hægt sé að prófa greypingarnar á beygingarfræði mega gögnin sem notuð eru í þjálfun greypinganna ekki vera lemmuð. Við teljum að IceBATS sé fyllilega dæmigert fyrir bæði hið upprunalega BATS og fyrir fjölbreytni íslenskrar málfræði.

Útfærsla

Við þjálfuðum og gáfum út orðagreypingar byggðar á þremur aðferðum: fastText, GloVe, og word2vec. Prófað var að þjálfa samhengisháðar orðagreypingar með ELMo, en því var alveg hætt. Þetta var að hluta vegna þess að væntingar okkar um matseinkunnir gerðu hinar þrjár aðferðirnar mikilvægari (til dæmis eru matseinkunnir fastText MSL oftast tvöfaldar eða þrefaldar miðað við M-BERT), og að hluta til vegna skorts á viðeigandi vélbúnaði.

Hvað varðar ákvæði notar útfærsla okkar af þjálfunarferlinu sjálfu Gensim-forritasafnið ásamt GloVe-útfærslu Stanford NLP Group, á meðan hlutar eftirvinnslu og mats nota hluta úr Scikit-learn og SciPy. Í samræmi við skilgreiningar OMSL höfum við fjarlægt handvirkt vigra utan orðaforða (e. out-of-vocabulary, OOV) fyrir MSL frekar en að nota nálgunaraðferðir fastText úr Gensim, og táknum margra orða segðir (e. MWEs) með meðaltali orðanna sem þær samanstanda af. IceBATS-útfærslan okkar notar hliðrunarviguraðferðina (e. vector offset method) úr Gensim, aðlagaðri til að leyfa mörg rétt svör við hliðstæðuspurningum (e. analogy questions).
Eftirvinnsluaðferðir sem notaðar eru fyrir MSL – sem eiga eingöngu við þjálfunarlotur fastText – eru kallaðar „Mean Center“ (MC), „All But the Top“ (ABTT) og „Uncovec“. Við fylgjum notkun þeirra úr upprunalegu greininni fyrir OMSL[11] að fullu, útfærum virknina eftir skýringum, keyrum allt í sömu röð, og notum sömu gildi í inntaki. Þar sem mat OMSL notaði tvö mismunandi gildi

[11] Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart and Anna Korhonen. 2020. *Multi-SimLex: A Large-Scale Evaluation of Multilingual and Cross-Lingual Lexical Semantic Similarity*

sérstaklega fyrir inntak ABTT (-3 og -10) gerum við það sama, sem skýrir hvers vegna það kemur fyrir tvisvar í niðurstöðum okkar. Okkur fannst eftirvinnsla vera nokkuð óstöðug – viðkvæm fyrir minniháttar breytingum ytri stika, og breytileika í útkomum – en vel vinnunnar virði engu að síður.

Mat: MSL

Niðurstöður okkar á mati MSL eru gefnar sem *p*-fylgniskor Spearmans, alveg eins og í OSML. Erfitt er að meta nákvæman orðafjölda málheildarinnar fyrir hvert tungumálanna 12 sem notuð eru við mat OMSL, en enska málheildin frá 2017 inniheldur 16 milljaðra tóka frá Wikipedia og 600 milljarða tóka frá Common Crawl. Íslenska málheildin sem notuð var fyrir mat MSL er samtals 1,475 milljarðar tóka.

Niðurstöður mats fyrir hverja þjálfunaraðferð og eftirvinnsluaðferðir falla að þeim mynstrum sem gert var ráð fyrir, samkvæmt reynslu sem greint er frá í ýmsum öðrum tilraunum: fastText náði betri árangri en GloVe, word2vec náði betri árangri en fastText, meiri eftirvinnsla skilaði sjaldnar betri árangri, og bestu eftirvinnsluaðferðir í heildina frá OMSL eru meira og minna sömu fyrir mat okkar.

Ítarlegar niðurstöður fyrir bæði MSL og IceBATS eru tilgreindar í síðari kafla. Í stuttu máli var **besta fylgniskor 0,539** fyrir MSL, sem náðist með þjálfun þar sem word2vec notar Skip Gram, á lemmuðum gögnum, og með því að beita aðeins ABTT-eftirvinnsluaðferðinni með dd_{A}(-10).

Til samanburðar er miðgildi OMSL-greypingar með hæstu einkunn í hverju tungumáli, þjálfað bæði á Wikipedia- og Common Crawl-málheildunum, 0,510. Miðgildi OMSL-greypinga með hæstu einkunn í hverju tungumáli, aðeins þjálfað á málheild Wikipedia, er 0,439. Besta MSL-einkunnin okkar er marktækt hærri en sú fyrri, og töluvert hærri en sú seinni.

Ef við skoðum hinar tvær aðferðirnar var besta fylgnin með fastText 0,489, sem náðist tvisvar með Skip Gram á lemmuðum gögnum: einu sinni með ABTT(-10)-eftirvinnsluaðferðinni, og aftur með samsetningu eftirvinnsluaðferða, í röð, MC, Uncovec, og ABTT(-10). Besta fylgnin með GloVe var 0,468, með lemmuðum gögnum og ABTT(-10).

Mat: IceBATS

IceBATS er skipt í tvo hópa með tveimur flokkum hvor: afleiðslu- og beygingarfræði annars vegar og alfræðilega og orðfræðilega merkingarfræði hins vegar. Þessi uppbygging er spegilmynd hins upprunalega enska BATS. Einstakir undirflokkar eru ólíkir milli matssettanna tveggja, sem er viðbúið, þar sem að á þessu stigi er ekki lengur hægt að bera tungumálin tvö saman marktækt.

Það varð snemma ljóst við prófanir að ólíkir flokkar þurfa líklega mismunandi stillingar fyrir sem bestu þjálfun, þar sem CBOW virkar sérstaklega vel fyrir beygingarlíkingarnar á meðan Skip Gram virðist almennt vera betra fyrir annað. Auk þess er mikilvægur munur á útfærslu okkar, IceBATS, og þeirri upprunalegu, BATS, að við þjálfuðum mismunandi líkön á lemmuðum og ólemmuðum gögnum. Þannig skiptum við niðurstöðum okkar á matinu fyrst eftir aðferð (fastText, GloVe, eða word2vec), síðan eftir orðaspá (Skip Gram eða CBOW) og magni forvinnslu inntaks (lemmað eða ólemmað).

Áður en við ræðum einstaka einkunnir IceBATS skal nefna sérstaklega nokkur atriði á niðurstöðum mats á IceBATS. Í fyrsta lagi sýnir beygingarfræðiflokkurinn svipaðar en örlítið lægri niðurstöður ef borið er saman við ensku, sem kemur ekki á óvart vegna flækjustigs íslenskra beyginga. Í öðru lagi náðu undirflokkar orða sem innihalda hljóðbreytingar í orðstofnum aðeins lægri einkunnum en aðrir fyrir afleiðslufræði. Þrátt fyrir þetta ná allir undirflokkar afleiðslufræði hærri einkunn en samsvarandi flokkar fyrir ensku. Að lokum höfum við fundið út að lemmun er mjög mikilvæg fyrir þjálfunina og þegar við þjálfum líkönin okkar á lemmuðum gögnum eru einkunnir okkar fyrir merkingarfræði hærri að meðaltali en þær sem eru samsvarandi frá upprunalega BATS-teyminu.

Ítarlegar niðurstöður fyrir IceBATS og samsvarandi ytri stika eru tilgreindar í næsta kafla. Lesið eftir flokkum voru hæstu einkunnirnar **75,6% fyrir beygingarorðapör** með word2vec þjálfuðu á ólemmuðum gögnum, **84,2% fyrir afleiðsluorðapör** með fastText CBOW þjálfuðu á ólemmuðum gögnum, **66,4% fyrir alfræðileg orðapör** með GloVe þjálfuðu á lemmuðum gögnum og **57,5% fyfir orðfræðileg orðapör með** GloVe þjálfuðu á lemmuðum gögnum.

Besta meðaleinkunn sem við fengum fyrir alla fjóra flokka **er 46,6%** með word2vec CBOW þjálfuðu á ólemmuðum gögnum. Þetta er töluvert hærra en upprunalega BATS, sem náði 28,5% með GloVe (hæsta meðaleinkunn okkar með GloVe á ólemmuðum gögnum var 43,4%). Heilt yfir eru meðaleinkunnir hærri ef gögnin eru ólemmuð en þetta er aðeins vegna þess að ef þau eru lemmuð lækka einkunnirnar fyrir beygingarorðapörin svo mikið að þær skekkja meðaltalið. Með því að sleppa beygingarorðapörum var besta meðaleinkunnin fyrir hina flokkana þrjá 56,7% með GloVe á ólemmuðum gögnum. Í stuttu máli virðist GloVe ná bestum árangri heilt yfir með IceBATS, sem kemur á óvart þar sem fjöldi ytri stika sem hægt er að fínstilla er lægri en með hinum aðferðunum.

Það er fræðilega mögulegt að ná hámarksárangri fyrir bæði MSL og IceBATS ef beygingarorðapörin eru hunsuð. Þegar MSL er fínstillt með word2vec og Skip Gram á lemmuðum gögnum er meðaleinkunn IceBATS 41,9% sem er ásættanlega nálægt hámarksárangri fyrir IceBATS. Hins vegar orsakar það að einkunn fyrir beygingarfræði verður 3,9%. Það er því nauðsynlegt að skoða ekki aðeins meðaltalið heldur einstaka flokka til að ná fullnægjandi árangri orðagreypinga fyrir tiltekna notkun.

Endanlegar niðurstöður

Töflurnar tvær fyrir neðan sýna tiltekin gildi tengd þeim settum greypinga sem við höfum gefið út fyrir M6. Tafla 1 sýnir gildi ytri stika sem notaðir voru til að þjálfa greypingarnar, á meðan tafla 2 sýnir allar matseinkunnir fyrir hvert sett.

Til skýringar eru samsetningar fyrir útgefnar greypingar eftirfarandi:

1. FastText, sem notar CBOW, með ólemmuðum gögnum, fínstillt fyrir IceBATS
2. FastText, sem notar Skip Gram, með lemmuðum gögnum, fínstillt fyrir IceBATS
3. FastText, sem notar Skip Gram, með lemmuðum gögnum, fínstillt fyrir MSL
4. Word2vec, sem notar Skip Gram, með lemmuðum gögnum, fínstillt fyrir IceBATS
5. Word2vec, sem notar CBOW, með ólemmuðum gögnum, fínstillt fyrir IceBATS
6. Word2vec, sem notar Skip Gram, með lemmuðum gögnum, fínstillt fyrir MSL
7. GloVe, sem notar lemmuð gögn, fínstillt fyrir IceBATS

8. GloVe, sem notar ólemmuð gögn, fínstillt fyrir IceBATS
9. GloVe, sem notar lemmuð gögn, fínstillt fyrir MSL

Í töflu 1 táknar hver færsla það sett greypinga sem náði hæstu einkunn fyrir þá tilteknu samsetningu aðferða, orðaspá, lemmun inntaks, og fínstillingar matssetts. Sumir ytri stikar eru sameiginlegir fyrir allar þrjár þjálfunaraðferðir á meðan aðrir eru einstakir fyrir hverja aðferð. Þegar ákveðnir ytri stikar áttu ekki við merktum við þá með "-" (bandstrik). Ef einhverjir ytri stikar eru ekki teknir fram í töflunni má gera ráð fyrir sjálfgefnum stillingum fyrir viðeigandi aðferðir og Gensim-útfærslur.

Í röð frá vinstri til hægri eru ytri stikar: stærð vigurs, gluggi, alfa, mynstrasöfn (eða ítranir), mörk neikvæðrar sýnatöku, sýnismark, veldisvísir fyrir neikvæða sýnatöku, lágmarkstákn n-stæðulengdar, hámarkstákn n-stæðulengdar, lágmarksheildartíðni og sléttun í samsvörunarfylki.

Sett #	Stærð	Gluggi	Alfa	Mynstrasafn	Neik. sýnat.	Sýnismark	Veldisv.	Lágm. n	Hám. n	Lágm. heildart.	Sléttun
1	200	2	0.010	13	5	0.00001	0.5	1	2	5	-
2	350	5	0.020	20	13	0.0001	0.5	1	5	5	-
3	300	3	0.025	20	5	0.001	0.1	5	5	5	-
4	350	9	0.050	13	19	0.00001	0.5	-	-	5	-
5	200	1	0.020	20	13	0.0001	0.2	-	-	5	-
6	350	1	0.020	20	13	0.0001	0.4	-	-	5	-
7	300	10	-	10	-	-	-	-	-	-	10
8	350	10	-	20	-	-	-	-	-	-	40
9	350	7	-	15	-	-	-	-	-	-	40

Í töflu 2 fylgjum við sömu lóðréttu röð og í töflu 1. Niðurstöður fyrir hvert sett, allar rúnnaðar niður, eru tilteknar fyrst fyrir matsflokka IceBATS, og eftir það fyrir öll tilbrigði MSL og eftirvinnsluaðferðir MSL. Hæsti árangur fyrir hvern matsflokk er **feitletraður**.

Í röð eru matsflokkar fyrir IceBATS: orðmyndunarleg (e. Derivational) („Der“), beygingarfræðileg (e. Inflectional) („Inf“), alfræðiorðabókarleg (e. Encyclopedic) („Enc“), og orðabókarleg (e. Lexicographic) („Lex“). Fyrir MSL eru þeir: án nokkurrar eftirvinnslu („M1“), með MC („M2“), með ABBT -3 („M3“), með ABBT -10 („M4“), með Uncovec („M5“), og með fleiri en einni þessara aðferða í röð („M1.2.5.3“ og „M1.2.5.4“).

Sett #	Der	Inf	Enc	Lex	M1	M2	M3	M4	M5	M1.2.5.3	M1.2.5.4

1	**.842**	.614	.177	.081	.075	.076	.106	.095	.195	.205	.212
2	.453	.245	.595	.318	.415	.414	.446	.452	.459	.465	.469
3	.450	.195	.488	.224	.480	.480	.482	.489	.487	.488	.489
4	.530	.037	.625	.482	.450	.450	.453	.457	.465	.467	.466
5	.353	**.756**	.429	.321	.390	.390	.399	.399	.420	.421	.419
6	.528	.038	.610	.475	.513	.513	.529	**.539**	.533	.534	.535
7	.462	.008	**.664**	**.575**	.368	.369	.434	.446	.443	.442	.440
8	.235	.716	.461	.320	.323	.323	.385	.414	.391	.399	.398
9	.426	.008	.650	.556	.375	.375	.450	.468	.456	.462	.467

Að lokum

Orðagreypingar má nota á fjölda máta innan fræðigreinarinnar forritun með tungumál, svo sem viðhorfsgreiningu, tungumálaþýðingu, textaleit, upplýsingaheimt, skjalaflokkun o.s.frv.

Að auki eru aðferðirnar sem notaðar voru við þjálfun greypinganna sem fjallað er um í þessari skýrslu vel skjalfestar og vel þekktar, og gætu talist nokkuð aðgengilegar hinum almenna NLP notanda og almennum vélbúnaði sem hann hefur aðgang að. Þetta þýðir að utan hinna ýmissu notagilda greypinganna sem við gefum út er einnig mikið líklegra að þær verði notaðar á þennan máta, og því nytsamlegar fyrir fræðasviðið.

Með því að beita þekktum alþjóðlegum stöðlum hefur verið sýnt fram á að greypingar okkar eru sambærilegar greypingum margra annarra tungumála. Það er von okkar að þær reynist mjög nytsamlegar í frekari rannsóknum og þróun íslenskrar málgreiningar (e. NLP).

V2 – Samhliða málheildir

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

Markmið

Að stækka þær ensk-íslensku samhliða málheildir sem í boði eru og bæta samröðun og síunaraðferðir svo að eins mikill texti og mögulegt er verði raunverulega gagnlegur.

Varða 6 – lýsing

Skýrsla um samröðun og síunaraðferðir. Ný útgáfa ParIce gefin út með nýjum aðferðum. Við stefnum á að fækka slepptum setningum um a.m.k. 20% miðað við þær setningar sem var sleppt í fyrstu útgáfu ParIce.

Afurðir

Öllum markmiðum vörðunnar var náð.

ParIce 2021.10 verður send í hirslu CLARIN fyrir 5. október 2021, ásamt skýrslu um hvernig málheildin var samsett. Hún er aðgengileg á þessari slóð:
https://repository.clarin.is/repository/xmlui/handle/20.500.12537/145

Að auki voru tvö prófunarsett og eitt þjálfunarsett búin til í þetta ferli:

Prófunarsett samröðunar setninga:
https://repository.clarin.is/repository/xmlui/handle/20.500.12537/150.
Prófunarsettinu er lýst í dreifingarpakkanum.

Prófunarsett til að prófa síunaraðferðir (og útdrátt gagna frá sambærilegum málheildum) var búið til fyrir síunarskrefið:
https://repository.clarin.is/repository/xmlui/handle/20.500.12537/151
Prófunarsettinu er lýst í dreifingarpakkanum.

Þjálfunarsett fyrir síunarflokkara:
https://repository.clarin.is/repository/xmlui/handle/20.500.12537/152

Skýrsla

Við bættum gögnum í ParIce-undirmálheildina og hreinsuðum, samröðuðum og síuðum gögnin með skilvirkari aðferðum en áður.

Nýjum EES-skjölum (e. EEA) var safnað og bætt við málheildina. Þetta fjölgar skjölum í þessum hluta frá 8.099 í 17.127. Meðallengd nýju skjalanna er styttri en gömlu og því hefur orðafjöldi ekki tvöfaldast, þó að fjöldi skjala hafi gert það. Í þessari útgáfu bætum við einnig við auðkennum (e. IDs) fyrir málsgreinar til að geta notað samhliða málsgreinar. Hér eru heildartölur fyrir málheildina:

Fjöldi málsgreinapara: 243.217
Fjöldi setningapara: 3.979.092
Orðafjöldi: IS - 21,8M; EN - 23,4M

Við samröðum einnig EES-málheildinni á setningastigi og fjarlægjum öll tvítekin pör. Við síum því næst pörin með þremur skorunaraðferðum, LASER, LaBSE og einkunn byggðri á samröðun orða með CombAlign. Flokkari notar þessar einkunnir til að ákveða hvort samþykkja skal setningapör. Að lokum notuðum við tvær frekari síur, eina sem síar út mjög stutta og mjög langa búta, og aðra sem síar út setningapör sem innihalda óvenjumörg óorð (e. non-word) (40% eða meira). Þetta orsakar 504.783 einstök setningapör í undirmálheildinni eftir síun.

Lítilli undirmálheild, sem samanstendur af fréttum og tilkynningum frá Norrænu ráðherranefndinni, hefur verið bætt við ParIce-málheildina. Hún inniheldur 581 skjalapar sem inniheldur 15.706 setningapör – 10.293 pör eftir síun.

Nýjum gögnum var bætt við frá OpenSubtitles, og þau gögn hreinsuð eins vel og hægt er, þar sem íslensku textarnir þaðan innihalda oft suð (e. noise). Einnig var lítillega bætt við önnur gagnasett. Í heildina innihalda hráu gögnin yfir 8,2 milljónir bútapara, með 3,2 milljónir einstakra bútapara. Eftir síun er málheildin með 1,7 milljón einstakra setninga af góðum gæðum.

	Samtals bútar	Einstakir	Eftir alla síun	skjöl	málsgreinar
ESO	15416	12683	10969	457	[NA - mjög stuttar]
EES	3979092	1145727	504783	17127	243217
OS	3609166	1538238	801531		
Biblía	32752	32752	32752		
Bækur	16764	16212	7135		
Ubuntu	15986	4766	1201		
Ted2020	2399	2392	1662		
Tatoeba	9440	9440	8826		
Hagstofa	2288	2288	1254		
EMA	420297	404666	305943		
KDE4	87575	16911	11799		
Íslendingasögur	28124	25255	10874		
Norðurlandaráð	15706	12802	10293	581	9467
SAMTALS	**8235005**	**3224132**	**1709022**	**18165**	**252684**

Á meðan við endurbyggðum málheildina bjuggum við til prófunarsett af samröðuðum setningum með gögnum okkar, í sama stíl og prófunarsett BleuAlign (http://hdl.handle.net/20.500.12537/150). Við gerðum svo tilraunir með fjölda samröðunaraðferða, m.a. BleuAlign, Vecalign og sérhannaða nálgun byggða á orðasamröðunum. Vecalign náði besta árangrinum á prófunarsettum okkar og við völdum því þá

aðferð. Fyrir síun gerðum við einnig tilraunir með ýmsar aðferðir: bicleaner, BLEU-byggða einkunnagjöf, LASER, LaBSE og einkunn byggða á orðasamröðun. Samblanda af LASER, LaBSE og einkunn byggðri á orðasamröðun náði besta árangrinum á prófunarsettunum og við völdum því þær aðferðir.

V2b – Bakþýðingar

Skýrsla: Miðeind
Aðilar: Miðeind

Markmið

Markmiðið á öðru ári er að bæta aðferðir bakþýðinga sem við lærðum á fyrsta ári, og beita þeim með nýjum líkönum sem til verða, til að auka gæði gervimálheildarinnar sem styður við raunverulegu samhliða málheildina. Forþjálfuðu viðgjafarlausu (BERT-líku) líkönin standa sig betur, en eru mun stærri en bestu líkönin sem notuð voru á fyrsta ári og því ekki við hæfi fyrir bakþýðingar á stórum mælikvarða án breytinga. Þau verða bestuð eftir þjálfun með þekktum aðferðum.

Vörður

M5

- Bestað nýtt líkan tilbúið til að búa til gervimálheild (bakþýdda).

M6

- Annarri kynslóð nýja líkansins, þjálfað á bakþýddum gögnum, lokið.
- Stefnt er að framför um +1.0 BLEU í hvora þýðingarátt yfir líkan grunnlínu án bakþýddra gagna með notkun prófunarsetts utan sviðs.
- Markmið málfimi er að ná tölfræðilega marktækri framför frá óbakþýdda líkaninu við p=0.05 í mannlegu mati.

Afurðir

Öllum markmiðum var náð. Líkön bestuð fyrir ályktunartíma með svokölluðu *layer drop* eru aðgengileg á CLARIN https://repository.clarin.is/repository/xmlui/handle/20.500.12537/128. Uppfærða bakþýdda málheildin er útgefin á CLARIN https://repository.clarin.is/repository/xmlui/handle/20.500.12537/127.

Verkþáttur V2b heldur áfram á þriðja ári.

Skýrsla

Aðferðafræði

Fyrsta tilraun okkar til að þjálfa bestað þýðingarlíkan fyrir hraðvirka þýðingu byggðist á vel þekktum og viðurkenndum aðferðum, þ.e. líkani sem er ekki sjálfsaðhverft. Hins vegar varð það

ljóst að lækkun gæða með þessari aðferð var of mikil, þar sem líkönin náðu verri árangri en líkanið sem greint var frá í M3-skýrslu og bakþýðingar þess myndu því ekki gefa neina framför.

Við skiptum þá um stefnu og notuðum í staðinn líkan sem lýst er í undirverkefni V2a (við M4), fínstillt mBART25 líkan, til að búa til nýjar bakþýðingar. En þar sem bestað líkan er sérstaklega nefnt sem afurð hafa niðurstöður annarrar aðferðar með minnkað líkan verið gefnar út á CLARIN. Þessi aðferð notar *eftirþjálfun* með *layer drop*. Þessu bestaða líkani er lýst betur að neðan.

Til skýringar eru nákvæmar upplýsingar stærðar einmála málheildarinnar sýndar að neðan (tafla úr undirverkefni V2b við M3). Nýja (fínstillta mBART25) líkanið tekur töluvert lengri tíma en það fyrra til að búa til þýðingar úr einmála málheildunum, eða um 3–4 mánuði V100 GPU tíma í báðar áttir. Vegna þessa verulega reiknikostnaðar við að þjálfa og búa til voru frekari bakþýðingar ekki endurteknar.

Að auki, þar sem þýðingarnar eru gerðar með geislaleit (e. beam search), höldum við öllum geislaúttökum (e. beam outputs) (í stað eingöngu efsta geislans (e. top beam)) sem hluta af bakþýddri málheild þessa árs.

Upprunatungumál	**Nafn**	**Stærð** í setningum (notaðar)
Íslenska	Dómstólar	1.8m
Íslenska	Hæstiréttur	2m
Íslenska	Lagatexti	814þ
Íslenska	Vísindavefurinn	268þ
Íslenska	Wikipedia	405þ
Íslenska	Alþingi	6.2m
Íslenska	Annað	350þ
Íslenska	Morgunblaðið	13.6m
Íslenska	Vísir	4.9m
Íslenska	Útvarpsviðtöl og -fréttir (Rás 1 og Rás 2)	1m
Íslenska	**Samtals**	**31.3m**
Enska	Skrapaður fréttatexti	33.4m
Enska	Wikipedia	9.3m
Enska	Europarl	2.0m
Enska	**Samtals**	**44.7m**

	Heildarfjöldi setningapara í bakþýddu málheildinni (eftir síun)	**76m**

Niðurstöður og afurðir

Þar semnú er til handvirkt þýtt prófunarsett innan sviðs sem samanstendur af fréttagögnum kjósum við að nota það sem nýtt viðmið mælinga á gæðum nýju bakþýðinganna. Matssettið sem við notum er það sama og ensk-íslensku prófunargögnin frá fréttaþýðingarverkefni WMT 2021[12]. Þessi matsgögn voru búin til sem hluti af undirverkefni V4d. Niðurstöðurnar í töflunni að neðan sýna töluverða framför frá fyrri BLEU-einkunnum (+1,6 / +0,6 BLEU stig fyrir *en-is* og *is-en* hvert um sig, á prófunarsettinu; fleiri á þróunarsettinu).

Bakþýdd gögn	Matssett	en-is	is-en
Gömlu	newstest2021-dev	25,9	30,4
Nýju	newstest2021-dev	27,8	31,8
Gömlu	newstest2021-test	22,7	32,9
Nýju	newstest2021-test	24,3	33,5

Þar sem umfangsmiklar bakþýðingar teljast nú ómissandi fyrir hágæða vélþýðingarkerfi, eins og augljóst er í greinum um vélþýðingar[13], gerum við ekki samanburð við kerfi sem er þjálfað án bakþýðinga. Gríðarleg lækkun gæða yrði strax augljós. Samanburður mannlegs mats á málfimieinkunnum milli líkans þjálfuðu með bakþýðingum og líkans þjálfuðu án þeirra var, af sömu ástæðum, ekki gerður.

Í blindu mannlegu mati innanhúss á 100 setninga fréttatexta á mBAR- líkönum þjálfuðum með gömlu og nýju bakþýðingunum sáum við að heildareinkunnin (af fimm) fór frá 3,02 upp í 3,25. Eftirfarandi dæmi frá þessum tveimur líkönum sýnir hvernig nýju þýðingarnar, byggðar á endurbættu bakþýðingarmálheildinni (hægri dálkur), eru betri, bæði í merkingu og málfimi:

Upprunalegur texti	Gömlu bakþýðingar	Nýju bakþýðingar
The shortage of qualified HGV drivers, worsened by Brexit and Covid, has left wholesalers unable to get goods to shops, with major dairy producer Arla on Friday admitting it could not get milk to about a quarter of supermarkets last week.	Skortur á hæfum mjólkurbílstjórum HGV, verst settu Brexit og Covid, hefur skilið heildsala eftir ófæra um að koma vörum í verslanir, enda viðurkenndi stóri mjólkurframleiðandinn Arla á föstudaginn að hún gæti ekki sótt	Skortur á hæfum HGV-bílstjórum, sem versnað hefur vegna Brexit og Covid, hefur orðið til þess að heildsalar hafa ekki komist með vörur í verslanir, en stóri mjólkurframleiðandinn Arla viðurkenndi á föstudaginn að hafa ekki komist með mjólk til um

[12] http://www.statmt.org/wmt21/
[13] https://aclanthology.org/D18-1045/ gefur gott yfirlit

	mjólk til um fjórðungs stórmarkaða í síðustu viku.	fjórðungs stórmarkaða í síðustu viku.

Eins og nefnt var að ofan gerðum við tilraunir með stærðarbestað líkan, til að gera málamiðlun milli þýðingargæða og ályktunarhraða (og minnisþarfar). Þjálfun var gerð með mismunandi stigum lagsbrottfalls (e. layer dropout), og eftir þjálfun var annað hvert lag líkansins sem myndast fjarlægt til að búa til bestað líkan fyrir mat. BLEU-árangursmat má sjá í töflunni að neðan. Með því að fjarlægja heilu lögin koma út hlutfallslega stærri net með samsvarandi hraðaaukningu, en einnig minni þýðingargæði. Við prófuðum 20%, 40% og 60% lagsbrottföll við þjálfun, en 40% gaf bestan árangur, í samræmi við það sem hefur áður verið greint frá í fræðunum.[14]

Eftirfarandi tafla sýnir BLEU-árangur þýðingarlíkananna í báðar áttir, *en-is* og *is-en*, fyrir ýmis matssett. Dálkurinn *Öll lög* sýnir upprunalegt líkan með öllum lögum á sínum stað. Dálkurinn *Helmingsstærð* sýnir bestaða líkanið sem var þjálfað með 40% lagsbrottfalli og þar sem annað hvert lag var fjarlægt áður en líkanið var metið.

	en-is		**is-en**	
Matssett	**Öll lög**	**Helmingsstærð**	**Öll lög**	**Helmingsstærð**
IPAC	23,4	20,9	24,2	23,0
ESO	31,0	25,5	33,5	29,9
Bible	25,3	16,3	32,9	25,0
EEA	57,6	43,3	62,5	48,4
Emea2016	61,8	46,1	69,8	54,3
Medical	60,4	41,4	67,4	48,6
OS2018	30,7	27,5	31,3	29,4
Tatoeba	48,5	43,9	57,0	52,3
WMT-2021-dev	27,8	25,6	31,5	29,6
Flores-dev	28,1	26,7	36,2	33,1

Við ályktum að bestuðu líkönin geti verið góður kostur þar sem hraði og minnisþörf eru aðal áhyggjuefnin, frekar en hámarksnákvæmni þýðinga. Við tökum einnig fram að munur árangurs fer eftir því af hvaða gerð textinn er sem skal þýða - hann er meiri fyrir sérhæfða texta (EES/EMEA) en minni fyrir einfaldari texta (Open Subtitles (OS2018), Flores).

[14] https://arxiv.org/abs/1909.11556

V4a – Opin þýðingarvél fyrir íslensku

Skýrsla: Miðeind
Aðilar: Miðeind

Markmið

Markmiðið á öðru ári er að aðlaga, innleiða og fínstilla valdar fyrsta flokks lausnir fyrir taugavélþýðingar milli íslensku og ensku. Þessi vinna byggir á lærdómi frá fyrsta ári, sem sýndi að aðferðir taugavélþýðinga eru mun betri en hefðbundinna tölfræðilíkana (e. statistical models). Sér í lagi áformum við að halda áfram með sambland eftirlitslausrar forþjálfunar þvermála BERT-líkra mállíkana og síaðra runu-til-runu þýðingalíkana.

Varða 6 – lýsing

M5

- Fyrsta ítrun á fínstilltu líkani og frekari ítranir á forþjálfun áður en átt er við fínstillingu.

M6

- Fínstillt líkön aðgengileg til keyrslu með JSON/REST-forritaskilum og vefviðmóti.
- Skýrsla um árangur og mannlegt mat.
- Stefnt er að framför um minnst +1,0 BLEU fyrir hvora þýðingarátt.
- Grunnlína fyrsta árs með notkun bakþýðinga er nú (nóvember 2020) 21,4 BLEU fyrir ensku til íslensku og 24,8 fyrir íslensku til ensku.

Afurðir

Öllum markmiðum var náð. Þjálfuð líkön ásamt leiðbeiningum eru aðgengileg á CLARIN https://repository.clarin.is/repository/xmlui/handle/20.500.12537/125. Skýrsla um árangur er innifalin að neðan.

Verkþáttur V4a heldur áfram á þriðja ári.

Skýrsla

Fínstillt líkön

Við höldum áfram að fínstilla mBART-líkanið sem við greindum frá í M4-skýrslunni. Stærstu framfarirnar koma vegna betri bakþýðingamálheilda, sem lýst er í skýrslunni fyrir verkþátt V2b, og vegna lengri þjálfunartíma. BLEU-einkunnir eru taldar upp í eftirfarandi töflu.

Einkunnirnar eru reiknaðar yfir IPACmatsmálheildina, sem búin var til hjá Miðeind og lýst í grein sem er aðgengileg á arXiv[15].

Líkan	Matssett	en-is	is-en
Transformer base	IPAC	24,8	21,4
mBART	IPAC	26,8	23,5
	WMT-2021 dev	25,9	30,4
mBART-cont	WMT-2021 dev	27,8	**31,8**
mBART-cont-EES	WMT-2021 dev	**29,1**	**31,8**
mBART-rerank	WMT-2021 dev	28,8	31,4

Líkönin fjögur sem við höfum þjálfað eru:

- Transformer-base: Þróað á fyrsta ári.
- mBART: Fínstillt mBART25-byggt þýðingarlíkan. Lýst í M4-skýrslu.
- mBART-cont: Frekari þjálfun mBART með því að bæta endurbættum bakþýðingum við. Lýst í skýrslunni fyrir verkþátt V2b.
- mBART-cont-ees: mBART-cont sérstillt til að þýða EES (e. EEA) reglugerðir. Lýst í skýrslunni fyrir verkþátt V4e.
- mBART-rerank: Tilraunakennt líkan blandaðra aðferða (e. ensemble method) sem notaði myndandi mállíkan til að raða mögulegum þýðingum.

Aðgengilegt fyrir notkun

Besta líkanið fyrir almenna notkun, mBART-cont, er nú aðgengilegt til notkunar fyrir almenning á vefsíðunni https://velthyding.is. Einnig er möguleiki á að nota EES-sérhæfða líkanið (mBART-cont-ees) sem líkan fyrir sérsvið eins og lýst er í kafla V5a. Sama vélþýðingarlíkanið er einnig aðgengilegt í gegnum JSON REST-forritaskil að fyrirmynd Google Translate forritaskilanna, eins og skjalað er hér: https://api.greynir.is/docs (en tilvonandi notendur þurfa að biðja um forritaskilalykil (e. API key) frá Miðeind til að geta notað forritaskilin)[16].

Mannlegt mat

Við gerðum mannlegan samanburð á þýðingarlíkönum okkar við tvö önnur vel þekkt vélþýðingakerfi sem eru opin almenningi. Niðurstöðurnar gefa til kynna að við erum samkeppnishæf við þau stóru kerfi, og gerum betur í áttinni *en-is* en erum meðal þeirra fyrir áttina *is-en*.

[15] https://arxiv.org/abs/2108.05289
[16] Opnun forritaskilanna fyrir almenning myndi kosta gífurlegan tölvukostnað (e. computing resources).

Við framkvæmum matið yfir fjögur mismunandi gagnasett með 200 setningar hvert. Gagnasettin eru á tveimur sviðum, fréttir og EES-reglugerðir, og hvert svið í tvær áttir, *en-is* og *is-en*. Tryggt var að matsgögnin fyndust ekki í þjálfunargögnum, með því að velja fréttir og reglugerðir sem birtar voru eftir að þjálfunargögnin voru undirbúin.

Fjögur kerfi voru metin: mBART-cont, mBART-rerank, Google Translate, og Microsoft Bing Translator. Allar setningar voru þýddar með hverju kerfi sem orsakar fjóra þýðingarmöguleika fyrir hverja upprunalega setningu, samtals 3200 möguleikar.

Þau sem unnu við mat gáfu hverjum þýðingarmöguleika blinda einkunn á bilinu 1–5 með notendaviðmótinu sem lýst er í kafla V5b.

Meðaleinkunn eftir kerfi og gagnasetti (hærri er betri)

	en-is		is-en	
	news	eea	news	eea
mBART-cont	3,30	4,06	3,54	3,88
mBART-rerank	**3,31**	**4,08**	3,46	3,87
Google	3,19	3,49	**3,89**	**4,11**
Microsoft	3,13	3,43	3,23	3,59

Í töflunni sjáum við að líkön okkar ná hæstu einkunnum í áttinni *en-is* en tapa fyrir Google í áttinni *is-en*, en vinna þó Microsoft.

Til að meta tölfræðilegt þol þessara niðurstaðna reiknum við p-gildi fyrir hvert par líkana með núlltilgátunni að líkönin hafi sömu meðaleinkunn. p-gildi nálægt 0 bendir til þess að munurinn á mældum meðaltölum sé líklega raunverulegur á meðan gildi nálægt 1 bendir til niðurstöðu sem er líklega tilviljun.

p-gildi (feitletrað: < 0,05)

		en-is		is-en	
		news	eea	news	eea
Google	mBART-cont	0,10	**0,00**	**0,00**	**0,00**
Google	Microsoft	0,45	0,37	**0,00**	**0,00**
Google	mBART-rerank	0,07	**0,00**	**0,00**	**0,00**
mBART-cont	Microsoft	**0,02**	**0,00**	**0,00**	**0,00**
mBART-cont	mBART-rerank	0,89	0,67	0,24	0,89

Microsoft	mBART-rerank	**0,01**	**0,00**	**0,00**	**0,00**

Samkvæmt töflunni drögum við þá ályktun að það sé ekki marktækur munur á milli mBART-skyldu þýðingarlíkananna okkar tveggja. Í öllum dálkum nema news-en-is er skýr munur milli okkar líkana og hinna tveggja. Í news-en-is-gagnasettinu er mikill munur milli okkar líkana og líkans Microsoft, en munurinn á þeim og líkani Google er ekki jafn skýr.

Þessi gildi voru reiknuð með fallinu `ttest_ind` úr SciPy[17] með `equal_var=False`, sem framkvæmir t-test Welch.

WMT 2021

Teymi Miðeindar tók þátt í íslenskum hluta fréttaþýðingarverkefnisins í WMT 2021[18]. Kerfisgrein okkar er aðgengileg á arxiv.org[19]. Mikið af upplýsingunum í þeirri grein er að finna í þessari skýrslu eða fyrri skýrslum.

[17] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html
[18] http://www.statmt.org/wmt21/translation-task.html
[19] https://arxiv.org/abs/2109.07343

V4b – For- og eftirvinnsla nafneininga

Skýrsla: Miðeind
Aðilar: Miðeind

Markmið

Við munum bæta for- og eftirvinnslu nafneininga og sérstakra tóka byggt á fyrstu tilraunum frá fyrsta ári. Árangursríkar aðferðir verða hugsanlega útvíkkaðar svo þær nái til fleiri flokka, svo sem orðatiltækja sem ættu aldrei að þýðast bókstaflega („hann kom af fjöllum“ - bókstaflega „he came from the mountains“, en réttilega „he was surprised“), fastra orðasambanda og samsettra sagna með því að nota merkingareinræðingar. Á öðru ári mun þetta undirverkefni, til að taka af allan vafa, stefna á að hanna aðferðir og algrím til að taka á slíkum orðasamböndum og sérstökum tilfellum í grundvallaratriðum (með því að nota algengustu tilfellin sem leiðarvísi fyrir þróun), en ekki safna stórtæku gagnasetti slíkra málfræðilegra formgerða í íslensku.

Varða 6 – lýsing

M5
- Viðmið og prófunarsett til að mæla árangur/ávinning tilbúin.

M6
- Markaðar samhliða málheildir tilbúnar, með staðgenglum/mörkum fyrir nafneiningar, sérstaka tóka og margra orða segðir.
- Sérbyggð reglubyggð pípa búin til sem getur leiðrétt rangar þýðingar fyrir sumar nafneiningar (ekki margra orða segðir).

Afurðir

Öllum markmiðum var náð.
- Prófunarsett eru aðgengileg hér: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/130
- Samhliða málheildir eru aðgengilegar hér: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/127
- Viðmið og sérbyggð pípa eru aðgengileg hér: https://github.com/mideind/MT-NE-Pipeline

Verkþætti V4b lýkur við M6.

Viðmið og prófunarsett

Við M5 þróuðum við nokkrar aðferðir til að mæla árangur þýðinga á nafneiningum ásamt prófunarsettum sem má framkvæma þessar mælingar með.

Prófunarsettin voru búin til með því að keyra nafneiningamarkara á báðar hliðar fjölda samhliða undirmálheilda sem við notum til að meta vélþýðingakerfin okkar, og gá hvort þýðingarmarkmiðið innihaldi sömu nafneiningar (*með viðeigandi breytingum*) og upprunaleg þýðingin. Í töflunni að neðan teljum við upp undirmálheildirnar og stærðir þeirra ásamt fjölda nafneininga eftir málheild og flokki[20].

Málheild	#lína	Staðsetning	Annað	Samtök	Persóna	Peningar	Prósent	Dagsetning	Tími
ESO-EN	2000	698	1330	585	264				
ESO-IS	2000	578	872	500	276	3	32	114	1
Bible-EN	1000	329	127	0	950				
Bible-IS	1000	300	127	4	949	1	0	7	0
EES-EN	10000	1163	3230	2778	46				
EES-IS	10000	659	600	1530	62	34	254	1031	0
EMEA-EN	10000	259	3946	322	37				
EMEA-IS	10000	184	1980	174	61	10	954	118	4
OpenSubtitles-EN	10000	246	230	136	1608				
OpenSubtitles-IS	10000	225	171	126	1310	38	9	42	17
Tatoeba-EN	1000	41	33	2	73				
Tatoeba-IS	1000	36	13	3	72	6	0	9	12
WMT-2021-dev-EN	2004	826	443	660	1062				
WMT-2021-dev-IS	2004	789	318	532	1053	23	31	255	19
Flores-dev-EN	997	485	357	174	204				
Flores-dev-IS	997	483	145	145	205	13	19	132	10

Þessi prófunarsett, ásamt nafneiningunum sem finnast í þeim, hafa verið gefin út hjá CLARIN.

[20] Athugið að íslensku og ensku nafneiningamarkararnir búa ekki til nákvæmlega sömu mörk; íslenski nafneiningamarkarinn hefur fleiri flokka.

Mælingarnar, þ.e. viðmiðin, gera ráð fyrir að bæði uppruna- og markmiðshlið hvers pars þýðinga séu mörkuð með nafneiningum. Við getum svo raðað nafneiningunum í grunnsannleikstilvísunina (e. ground truth reference) með nafneiningarnar í úttaki kerfisins með leitarnámi byggðu á stafskiptafjölda (e. edit distance) (t.d. Jaro-Winkler-fjarlægð).

Í töflunni að neðan greinum við frá fjölda samraðana (#Samraðana) sem urðu til og fjölda samraðana deilt með hámarksfjölda mögulegra samraðana (Dekkun). Þessi gildi gefa vísbendingar um hversu vel þýðingarkerfið býr til nafneiningar[21]. Til að mæla réttleika tiltekinnar þýðingar nafneiningar greinum við frá meðal stafskiptaskiptafjöldaskori (meðal fjarlægð, 0 - best, 1 - verst) og það brot nafneininganna sem er nákvæmlega jafnt viðmiðinu (Nákvæmni). Þessi gildi eru gefin upp fyrir allar nafneiningar saman og einnig fyrir hverja gerð nafneininga.

Taflan sýnir að mBART25-cont-líkan annars árs nær betri árangri en hið eldra mBART25. Heilt yfir nær mBART25-cont-líkanið að þýða nafneiningar réttar, eins og sést á minni meðalfjarlægð og hærri nákvæmni. Fyrir áttina enska til íslensku getur líkanið þýtt fleiri nafneiningar, eins og sést í dekkun, en fyrir íslensku til ensku lækkar dekkun lítillega. Sem dæmi er mikilvægt gildi nákvæmni *persónu* frá ensku til íslensku (þar sem líkanið þarf að spá fyrir um rétta beygingu), sem hækkar úr 82,97% í 84,29%.

Mat á þýðingum nafneininga mBART25 og mBART25-cont

	mBART25 en-is	mBART25-cont en-is	mBART25 is-en	mBART25-cont is-en
Samsett				
#Samraðana	15986	16061	20014	19955
Dekkun	90,41%	90,84%	88,39%	88,12%
Meðal fjarlægð	0,043	0,038	0,039	0,036
Nákvæmni	80,01%	81,50%	84,30%	85,33%
Persóna				
#Samraðana	3711	3692	3786	3790
Dekkun	93,12%	92,65%	89,21%	89,30%
Meðal fjarlægð	0,034	0,029	0,029	0,025
Nákvæmni	82,97%	84,29%	87,77%	90,24%
Staðsetning				
#Samraðana	3001	2994	3624	3582
Dekkun	92,22%	92,01%	89,55%	88,51%
Meðal fjarlægð	0,049	0,045	0,053	0,047
Nákvæmni	74,94%	76,59%	81,24%	83,08%
Samtök				
#Samraðana	2744	2773	4206	4178
Dekkun	91,04%	92,00%	90,32%	89,71%
Meðal fjarlægð	0,050	0,043	0,043	0,038
Nákvæmni	77,92%	80,27%	83,12%	84,08%

[21] Nafneiningar sem nafneiningamarkari finnur.

Annað				
#Samraðana	3665	3697	8398	8405
Dekkun	86,73%	87,48%	86,61%	86,69%
Meðal fjarlægð	0,040	0,034	0,035	0,035
Nákvæmni	79,89%	81,44%	84,65%	84,70%
Dagsetning				
#Samraðana	1512	1521		
Dekkun	88,52%	89,05%		
Meðal fjarlægð	0,044	0,044		
Nákvæmni	84,79%	85,47%		
Tími				
#Samraðana	54	56		
Dekkun	80,60%	83,58%		
Meðal fjarlægð	0,240	0,232		
Nákvæmni	29,63%	35,71%		
Peningar				
#Samraðana	102	105		
Dekkun	79,69%	82,03%		
Meðal fjarlægð	0,116	0,113		
Nákvæmni	41,18%	37,14%		
Prósent				
#Samraðana	1197	1223		
Dekkun	92,15%	94,15%		
Meðal fjarlægð	0,027	0,025		
Nákvæmni	88,30%	88,96%		

Markaðar samhliða málheildir

Við bjuggum til stórar samhliða gervimálheildir ásamt upplýsingum um nafneiningarnar í málheildunum með sömu nafneiningamörkurum og notaðir voru fyrir prófunarsettin. Þessar málheildir voru búnar til úr News Crawl[22] og safni íslenskra fréttagreina sem safnað var nýlega. Þessar heimildir voru valdar þar sem þær innihalda mikið af nafneiningum, sérstaklega af gerðinni *persóna*.

News Crawl-málheildin og íslenska fréttamálheildin voru vélþýddar og hvor hlið mörkuð með nafneiningamarkara. Þar sem nafneiningamarkararnir fyrir hvort tungumál skila ólíkum markaskrám var mörkunum varpað á sameiginlega markaskrá fyrir flokkana *staðsetning*, *persóna* og *samtök*. Merking þessara marka milli markara er oftast svipuð.

[22] Árin 2007-2019 sótt frá http://data.statmt.org/news-crawl/. Gögnunum var stokkað og þau látin innihalda gróflega sama fjölda lína og íslenska fréttamálheildin.

Eftir mörkun fjarlægðum við allar línur sem innihéldu engar nafneiningar og allar línur þar sem báðar hliðar innihéldu ekki sama fjölda nafneiningamarka[23]. Þetta tryggir að aðeins nytsamlegar línur standi eftir í málheildunum.

Nytsamleg viðbót við þetta væri að bæta orðatiltækjum/margra orða segðum í þessar málheildir, en það myndi krefjast stórtækrar gagnaöflunar orðatiltækja og samsvarandi þýðinga þeirra sem er utan ramma þessa undirverkefnis á öðru ári.

Gervimálheildirnar sem urðu til geta þjónað sem auka bakþýðingargögn, sérstaklega eftir að þær eru keyrðar í gegnum sérbyggða leiðréttingarpípu (sjá að neðan).

Málheild	**#lína**	**Staðsetning**	**Samtök**	**Persóna**
News Crawl	641.952	236.545	293.027	420.305
Íslenskar fréttagreinar	558.080	299.249	273.121	553.988

Sérbyggð pípa

Við hönnuðum sérbyggða reglubyggða pípu sem getur leiðrétt þýddar nafneiningar. Pípan greinir runu upprunatexta og samsvarandi þýðingar ásamt upplýsingum nafneininga beggja hliða. Fyrir hverja setningu og þýðingu eru allar nafneiningar settar fram til leiðréttingar. Pípan beitir ákveðnum reglum sem geta svo tekið ákvarðanir eftir gerð nafneiningar og uppruna og skilar leiðréttingu þýddrar nafneiningar. Reglurnar eiga að vera sveigjanlegar og auðvelt að bæta við þær.

Sem dæmi útfærðum við reglu sem getur tekið við skrá sem er eins og orðaskrá/orðabók sem er notuð til að skipta út orðum, þ.e. ef lykillinn samsvarar nafneiningu uppruna er þýðingunni skipt út fyrir viðeigandi gildi. Þetta getur verið sérstaklega nytsamlegt til að leiðrétta algeng nöfn samtaka. Við útfærðum einnig reglu sem leiðréttir mörk persóna þegar uppruni er íslenskur. Ef nafneininguna má finna í BÍN (sem inniheldur flest íslensk mannanöfn og föður-/móðurnöfn) beygjum við nafneininguna í nefnifalli fyrir ensku (*„Ég talaði við Heiðu Brjánsdóttur“ -> „I spoke to Heiða Brjánsdóttir“*). Við stefnum á að bæta fleirum reglum fyrir leiðréttingar við í framtíðinni.

[23] Nafneiningamarkarar geta gert mistök, og ekki endilega sömu mistök á mismunandi tungumálum.

V4c – Gervimálheild

Skýrsla: Miðeind
Aðilar: Miðeind

Markmið

Lykiláskorun við þjálfun taugavélþýðingakerfa fyrir tungumál með litlum og meðalmiklum gögnum (e. low- and medium-resource) eins og íslensku er safna þjálfunargögnum með nógu víðfeðman orðaforða, sem t.d. inniheldur einnig sjaldgæf orð og orð ákveðinna sviða. Þetta undirverkefni miðar að því að takast á við þetta vandamál með því að búa til samhliða gervimálheildir sem magna upp tilfelli sjaldgæfra orða. Þetta næst með því að þátta íslenska hlið núverandi samhliða málheilda og skipta sjaldgæfum orðum (í flestum tilfellum nafnorðum en hugsanlega líka lýsingarorðum og sögnum) upp í setningatré, rétt beygð, og gera samsvarandi skipti úr orðabók/orðtakabók (e. phrase book) á ensku hliðinni. Skiptin verða aðeins gerð í samhengi þar sem mjög líklegt er að setningarparið sem þau leiða af sér sé enn málfræðilega rétt; en þar sem gervimálheildin verður fyrst og fremst notuð sem bakþýðingarmálfang (ekki sem raunveruleg samhliða málheild) þá þarf nákvæmnin ekki að vera 100%.

Varða 6 – lýsing

M5

- Síunarpípa orðin starfhæf og gögn unnin.

M6

- Líkan þjálfað á tilbúinni málheild með skýrslu um kosti. Markmiðið er að ná framför um 0,3 BLEU-stig fyrir setningar sem innihalda sjaldgæfan orðaforða eða markviss hugtök. Mikilvægi þessara orða mun hafa mikil áhrif á nægð þýðinga, jafnvel þó að væntanleg framför BLEU-einkunnar sé tiltölulega lág.

Afurðir

Öllum markmiðum var náð og verkþættinum er lokið og hann aðgengilegur á CLARIN https://repository.clarin.is/repository/xmlui/handle/20.500.12537/135.

- GreynirTerms skiptitól til að búa til samhliða gervimálheildir: https://github.com/mideind/GreynirTerms
- Dæmamálheildir búnar til með GreynirTerms
- Þessi skýrsla

Verkþætti V4c lýkur við M6.

Skýrsla

Helstu niðurstöður okkar benda til þess að innleiðing gervi- (tilbúinna) þjálfunargagna eins og lýst er hér sé raunhæf nálgun til að hjálpa líkaninu að læra að nota sértæk orð (e.

domain-specific words) þegar þau koma fyrir í óséðum textum, jafnvel þó að það hafi aðeins séð þau í setningafræðilegu samhengi, öfugt við fullt merkingarlegt samhengi. Við prófuðum mismunandi aðferðir við að innleiða tilbúnu þjálfunargögnin, og komumst að því að fínstilling forþjálfaðs tauganetsþýðingarlíkans á blöndu af gervi- og raunverulegum gögnum virkar best. Á þróunarsetti innan sviðs lærir líkanið um tvo þriðju hugtakanna sem það hefur aðeins séð í fínstillingu, en lækkar samt ekki BLEU-einkunn.

Í heildina lofar þessi aðferð góðu við þjálfun tauganetsþýðingarlíkana sérsviða á sviðum þar sem góðir orðalistar eru fyrir hendi. Aðferðin er skalanleg, en krefst leitarnáms og gagna í hæsta gæðaflokki (tvímála málheildir og orðalistar). Fyrir almennara notagildi (betra almennt tauganetsþýðingarkerfi) gætu líkönin sem fínstillt eru á gervigögum einnig nýst til að búa til betri bakþýðingar. Þó að líkanið nái ekki að læra öll hugtökin er augljós ávinningur af 50% aukningu réttra þýðinga sjaldgæfra, áður óséðra hugtaka, sem er erfitt að finna í samhliða þjálfunargögnum.

Þarfir

Eftirfarandi úrræði eru notuð í tilraununum:

- GreynirTerms skiptitól
- Listi algengra hugtaka (íslenskra og enskra) sem koma fyrir í sniðmátamálheildinni (sem kemur með GreynirTerms)
- Orðalisti með hugtökum sérsviða á íslensku og ensku (í okkar tilfelli listi stjörnufræðihugtaka, *Orðaskrá úr stjörnufræði*[24])
- Sniðmátamálheild til að nota fyrir skiptingar (IPAC[25])
- Málheild sérsviðs fyrir mat (textar Stjörnustöðvar Evrópulanda á suðurhveli (ESO))
- Þjálfunargögn fyrir tauganetsþýðingarkerfi á almennu sviði
- Þjálfunarpípa fyrir tauganetsþýðingarkerfi (FairSeq, GreynirSeq)

Gervimálheildin búin til

Eins og lýst var í V4c-kafla í M4-skýrslu bjuggum við til *GreynirTerms*, stoðtól til að skipta hugtökum í gögnum samhliða málheilda. Þetta tól er afhent sem hluti þessa verkþáttar.

Viðeigandi málheild vantaði til að nota sem grunn til að bæta sjaldgæfum orðum inn í. Við völdum samhliða íslensk-enska málheild af ágripum frá *Skemmunni*[26], gagnasafni ritgerða skrifaðar af nemendum fyrir BA- og meistaragráður. Við þurftum einnig orðalista sértækra hugtaka til að nota fyrir skiptingu. Fyrir þetta höfum við valið orðalista stjörnufræðihugtaka, *Orðaskrá úr stjörnufræði*. Þessi orðalisti var valinn þar sem við höfum aðgengi að góðum málheildargögnum þessa fags; samhliða málheild með textum frá Stjörnustöð Evrópulanda á suðurhveli[27] (ESO), u.þ.b. 13.000 setningapör. Þessi gögn innihalda mörg hugtakanna af orðalista stjörnufræðihugtaka, og eru ómissandi til að meta hvernig kerfið lærir ný hugtök þegar það er þjálfað á gervimálheildargögnunum.

[24] https://idordabanki.arnastofnun.is/leit//ordabok/STJARNA
[25] https://arxiv.org/abs/2108.05289
[26] https://skemman.is/?locale=en
[27] https://www.eso.org/public/

Úr ágripunum notuðum við GreynirTerms (lýst í lokaskýrslu M4) til að búa til málheild af 541 sniðmátasetningum, sem nýtist sem grunnur til að bæta við sjaldgæfum hugtökum af lista hugtakapara (íslensk -> ensk) og búa til gervimálheild. Þessar sniðmátasetningar innihalda flestar samsetningar málfræðilegra eiginleika íslenskra nafnorða. Sem dæmi er hér setningapar úr ágripunum, þar sem hugtak úr orðalista stjörnufræðihugtaka hefur verið sett inn (*tifstjarna -> pulsar*):

Í tifstjörnunni er fjallað um jarðminjasvæði á Reykjanesskaga í tengslum við uppbyggingu eldfjallagarðs á svæðinu. -> The topic of this pulsar is the Reykjanesskagi peninsula in southwestern Iceland and the establishment of a future volcanic park in the area.

Hugtökin af orðalista sem notuð voru í þessari tilraun voru takmörkuð við eins orða nafnorð, til að einfalda mat á úttaki, en nálgunin á alveg eins við fyrir margorða nafnorð, og fyrir lýsingarorð og sagnorð ef GreynirTerms er breytt.

Eftir síun orðalista af stjörnufræðihugtökum höfðum við lista 436 hugtakapara, sem hægt var að nota í skiptingar. Þessi pör voru svo sett inn í sniðmátasetningarnar. Hvert hugtak var sett inn 50 sinnum (þ.e. í 50 sniðmátasetningar), sem gerði samtals um 20.000 setningar. Þetta er gervimálheildin sem notuð er í eftirfarandi tilraunum.

Þjálfun tauganetslíkans og síun gagna

Til að skoða og meta hvernig best væri að nota gervigögnin á þjálfunartíma prófuðum við mismunandi aðferðir við að innleiða gervigögnin við forþjálfun og fínstillingu. Til að þjálfa vélþýðingarlíkön okkar notuðum við Greynirseq-breytingarnar á `transformer`-líkanahöguninni, sem er tiltæk í runulíkanaverkfærasettinu (e. sequence modeling toolkit) FairSeq.

Tiltæku samhliða málheildargögnin fyrir þjálfun almenns þýðingarlíkans samanstanda af um 1,8 milljónum setningapara góðra þýðinga, og að auki 45 milljónir setningar sem eru bakþýddar frá ensku til íslensku, með lækkaðri sýnatökutíðni við þjálfun. Upprunalegu samhliða gögnin sem notuð voru fyrir þjálfun tauganetsþýðingarlíkananna innihéldu töluvert magn þeirra hugtaka sem við vildum kenna líkaninu, jafnvel þegar ESO-gögnin voru undanskilin, svo við síuðum burt allar setningar úr þjálfunargögnunum sem innihalda þessi hugtök. Þetta minnkar heildarmagn þjálfunargagna um 5%, en tryggir að hugtökin sem við viljum kenna líkaninu séu örugglega óséð.

Tilraunir

Við þjálfuðum þrjú transformer-líkön á jafn mörgum skrefum og með sömu stikum (virk skammtastærð er 400.000 tókar, líkanið er þjálfað á 24.500 skrefum = 9,8 milljarðar tóka séðir):

1) *no-astronomy*, þar sem setningar með stjörnufræðihugtökum hafa verið síaðar út,
2) *no-astronomy-with-synthetic*, sama og 1), en gervimálheildargögnunum hefur verið bætt við bakþýddu gögnin, og
3) *base*, með öllum upprunalegu þjálfunargögnunum til samanburðar.

Við fínstillum svo þessi líkön með minna magni gagna (gervimálheildinni og öðrum gögnum eins og lýst er), og gefum einkunn með BLEU-mæliaðferð. Í tilraunum okkar er áherslan hins vegar ekki mest á BLEU-einkunnir heldur frekar á að skoða þýðingar þeirra stjörnufræðihugtaka sem líkönin búa til.

Aðalhugmyndin við fínstillingartilraunirnar er að þjálfa *no-astronomy*-líkanið í stuttan tíma á gervimálheildinni. Tilraunir okkar fela í sér fínstillingu líkansins eingöngu á gervimálheildinni, notkun gervimálheildarinnar ásamt upprunalegum þjálfunargögnum, og fínstillingu á blöndu upprunalegra þjálfunargagna og gervimálheildar. Til stýringar fínstillum við einnig *no-astronomy* á málheild upprunalegu ágripanna, þar sem hún er jafnt formleg sem fræðileg, en án hugtakanna sem við viljum að líkanið læri. Sem viðmið fínstillum við einnig *base-model*-líkanið með ágripunum og þjálfunarmálheild ESO.

Líkönin fá einkunn á fjórum þróunarsettum: *eso*, *eso-terms*, *ees*, og *abstracts (ágrip)*. Við höfum þegar nefnt gögn ESO og ágripin; auk þeirra metum við á reglugerðum EES (*ees*), og á undirsetti ESO-gagnanna (*eso-terms*) sem samanstendur af u.þ.b. 3000 setningum þar sem minnst eitt stjörnufræðihugtak úr gervimálheildinni kemur fyrir í réttu samhengi.

		BLEU4 á dev				
	Líkan	**eso**	**eso-terms**	**ees**	**abstracts**	**Rétt hugtök í eso-terms**
Forþjálfuð líkön	base (öll upprunaleg gögn notuð)	38,06	41,2	52,05	27,33	41,9%
	no-astronomy	29,36	27,59	44,71	23,18	5,1%
	no-astronomy-w-synthetic	29,38	27,92	44,74	24,7	7,6%
Fínstillt líkön	no-astronomy+abstracts	28,37	26,84	45,57	24,98	5,2%
	no-astronomy+ESO	40,02	45,44	45,03	23,04	38,7%
	no-astronomy+synthetic, 600 skref	27,68	27,1	43,11	24,19	67,8%
	no-astronomy+synth-and-en-is	29,47	28,26	44,89	26,41	11,1%
	No-astronomy+equal-ratio-synth-and-en-is, 600 skref	28,98	27,92	49,92	25,08	39,1%
	No-astronomy+equal-ratio-synth-and-en-is, 1000 skref	28,79	27,95	50,02	24,99	52,1%

Tafla 1. Niðurstöður þjálfunar íslensk-enskra tauganetsþýðingarlíkana á gervi- og stýringargögnum.

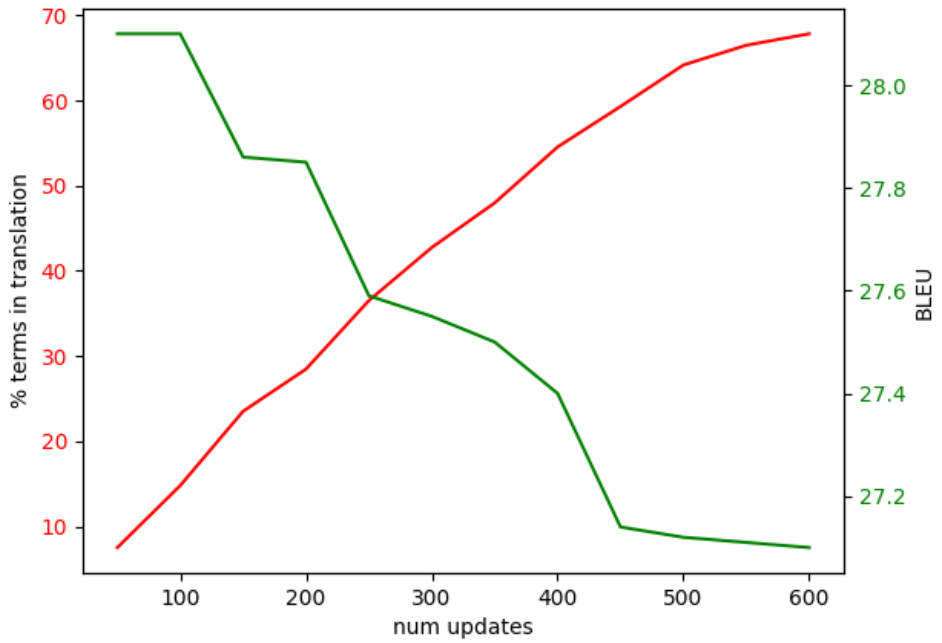
Niðurstöður og umræða

Aðaltilraunirnar og niðurstöður þeirra eru taldar upp í töflu 1. Við mælum BLEU-einkunn á fjórum þróunarsettum. Fyrir *eso-terms*-settið mælum við hversu mörg hugtök eru „rétt“ þýdd, þ.e. með markmiðshugtaki úr orðalistanum[28].

Síðasti dálkurinn fyrir forþjálfuðu líkönin sýnir okkur að *base*-líkanið þýðir þegar um 42% hugtakanna rétt, á meðan *no-astronomy*-líkanið (gulmerkt) nær aðeins um 5% hugtakanna rétt. Líkanið *no-astronomy-w-synthetic*, sem var að auki forþjálfað á gervigögnunum, lærir aðeins örlítið meira en *no-astronomy*, og er reikningslega óviðráðanlegt.

[28] Tekið skal fram að þýðing getur verið rétt þó að tiltekið hugtak sé ekki notað; setninguna má t.d. umorða eða nota samheiti eða vísa til orðs með fornafni.

Fyrir líkönin með gervigögnunum þarf gætilega stillingu stika, tíðar uppfærslur eftirlits og náið eftirlit ferlisins til að ofþjálfa ekki. Þetta eftirlitsferli er sýnt í mynd 1, þar sem græna línan táknar hvernig BLEU-einkunnir lækka eftir því sem við fínstillum frekar á gervigögnunum, á sama tíma og líkanið lærir fleiri hugtök.



Mynd 1. Samband milli lærðra stjörnufræðihugtaka í *eso-terms*-settinu (rautt) og BLEU-einkunna (grænt) við fínstillingu no-astronomy á gervigögnunum (*no-astronomy+synthetic* í töflu 1).

Til að læra meira en þriðjung hugtakanna þurfum við í þessari þjálfunartilraun að þjálfa þangað til BLEU-einkunnin er lægri en hún var fyrir fínstillingu (27,59 á eso-terms fyrir *no-astronomy*). Þó að BLEU sé ekki fullkomin mæliaðferð tökum við hana sem vísbendingu um að gæði þýðinga fari lækkandi þegar einkunnin byrjar að lækka. Á sama tíma sjáum við að hlutfall rétt þýddra hugtaka fer hækkandi nokkuð hratt (sjá *no-astronomy+synthetic* við mismunandi skref), svo það virðist vera samband milli þessara tveggja þátta við fínstillingu svona lítils gagnasetts.

Til að sporna við þessu gerðum við tilraunir með fínstillingu á upprunalegu þjálfunargögnunum ásamt gervigögnunum, í mismunandi hlutföllum. Það hlutfall sem virkaði best var 50% gervigögn og 50% upprunaleg þjálfunargögn, og þjálfun fyrir 1000 uppfærslur með virki skammstastærð 8000 tóka. Niðurstöðurnar eftir 600 uppfærslur og 1000 uppfærslur má sjá í töflu 1 undir *no-astronomy+equal-ratio-synth-and-en-is* (gulmerkt). Notkun þessarar blöndu gervi- og raunverulegra þjálfunargagna virðist gera líkaninu bæði kleift að auka gæði þýðinga almennt og læra hugtök frá merkingarlegum „bull“ gögnum.

V4d – Þróunar- og prófunargögn

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

Markmið

Til að gera mat þýðingarvéla í hönnun þolnara (e. more robust), þá er þörf fyrir handyfirfarin þróunar- og prófunarsett. Slík sett búum við til úr textum bæði innan og utan sviða. Með það að markmiði að koma íslensku inn í eitt af sameiginlegu verkefnunum á WMT 2021 verða prófunarsett utan sviða sett saman, að öllum líkindum með því að þýða fréttatexta. Handvirku mati verður beitt á ákveðnum stöðum á meðan hönnun stendur. Ensk-íslenskur/íslensk-enskur orðalisti verður settur saman sem hjálpargagn við þjálfun vélþýðingarkerfa. Slíkur orðalisti nýtist einnig til samröðunar við gerð samhliða málheilda.

Varða 6 – lýsing

Gefa út orðalista. Framkvæma handvirkt mat vélþýðingarkerfa í þróun.

Afurðir

[Frá M5] Útgáfa þróunar- og prófunargagna sérsviða.
[M6] Útgáfa orðalista. Framkvæmd handvirks mats á vélþýðingarkerfum í þróun.
[M6] Þróunar- og prófunargögn WMT21, aðgengileg á https://github.com/cadia-lvl/WMT21-data/ (og á síðu WMT21, http://statmt.org/wmt21/translation-task.html)
Öllum markmiðum hefur verið náð.

Verkþætti 4d lýkur við M6.

Skýrsla

Þróunar- og prófunargögn

Þróunar- og prófunargagnasett hafa verið búin til með gögnum innan sviðs frá ParIce-málheildinni. Þessi sett eru handmerkt og byggjast upp af textum eftirfarandi undirmálheilda ParIce:

EMA
Prófun: 3.279 setningapör

Þróun: 3.000 setningapör

OS

Prófun: 3.888 setningapör
Þróun: 3.500 setningapör

ESO

Prófun: 47 skjöl; 820 málsgreinapör; 1.252 setningapör
Þróun: 47 skjöl; 751 málsgreinapör; 1.240 setningapör

Norden

Prófun: 50 skjöl; 620 málsgreinapör; 823 setningapör
Þróun: 45 skjöl; 580 málsgreinapör; 790 setningapör

EEA

Prófun: 14 skjöl; 2.240 málsgreinapör; 3.109 setningapör
Þróun: 13 skjöl; 2.108 málsgreinapör; 3.100 setningapör

Settin verða aðgengileg á íslensku CLARIN-hirslunni fyrir 8. október:
https://repository.clarin.is/repository/xmlui/handle/20.500.12537/146

Ensk-íslenskur/íslensk-enskur orðalisti var búinn til. Við notuðum ýmsar aðferðir til að búa til mögulega lista:

- **orðagreypingar þvert á tungumál** byggðar á einmála málheildum;
- **samröðun orða** byggð á samhliða málheildum, gervi- (bakþýddum) samhliða málheildum og setningapörum sem dregin voru úr Wikimatrix og ParaCrawl;
- **velting (e. pivoting)** úr ISLEX-orðabókunum (íslensk-skandinavísk tungumál, auk frönsku og rússnesku): Við notuðum tiltækar orðabækur frá skandinavískum millistigstungumálum yfir á ensku. Þetta voru m.a. Wiktionary, Apertium Freedict og fleiri.
- **vélþýðing** orðalista

Við prófuðum skilvirkni þessara ýmsu aðferða og bjuggum svo til lista af möguleikum samkvæmt skilvirkustu samsetningunum. Þeir möguleikar voru svo metnir handvirkt. Skilyrði fyrir skráningu í endanlegum orðalista voru að orð í ensk-íslensku pari væri möguleg þýðing andstæða orðsins, sama hvort það var sjaldgæft eða ekki. Í þeim tilfellum þar sem voru margar þýðingar sama orðs röðum við þeim með sjálfvirkri einkunnagjöf, sem ætti að setja algengustu þýðinguna efst. Einkunnin er byggð á því hversu margar aðferðir okkar stinga upp á hverju pari.

Útgefni listinn inniheldur 193.604 pör.
Flestar þýðingarnar eru 1 á móti 1, en það eru líka þýðingar með margra orða einingum á annarri eða báðum hliðum:

is-en
1 á móti 1: 143.854 pör
1 á móti n: 45.040 pör

n á móti 1: 1.623 pör
n á móti n: 3.131 pör

69.704 mismunandi íslensk orð eru talin upp í orðalistanum og 65.617 mismunandi ensk orð. Nánari lýsing uppbyggingar orðalistans er að finna í útgáfupakkanum.
Orðalistinn verður aðgengilegur á CLARIN fyrir 8. október:
https://repository.clarin.is/repository/xmlui/handle/20.500.12537/144

Handvirkt mat

Ennfremur var handvirkt mat gert á vélþýðingarkerfum í þróun innan máltækniáætlunarinnar. Þrír aðilar mátu þýðingarúttak fjögurra mismunandi kerfa, tvö sem eru í þróun innan verkefnisins, auk Google Translate og Microsoft Translator. 200 setningar voru metnar fyrir hvert tveggja sérsviða (fréttir; EES-reglugerðir) í hvora átt (*en-is*; *is-en*). Greint er frá niðurstöðum í skýrslunni fyrir verkþátt V4a.

WMT21

Íslenska var tekin með í fyrsta skipti í fréttaþýðingarverkefni WMT21 á þessu ári. Hluti af þessari vinnu var söfnun þróunar- og prófunarsetta og handvirk þýðing þeirra milli íslensku og ensku. Ensku upprunalegu gögnin (2000 setningar) voru veitt af skipuleggjendum WMT21, og þýdd yfir á íslensku. Íslensku upprunalegu gögnunum (2000 setningum) var safnað úr útgefnum greinum frá íslenskum fréttasíðum 26. júlí 2020, og þau þýdd yfir á ensku.

Fréttaþýðingarverkefnið fékk fjölda innsendinga frá mismunandi teymum fyrir þýðingaráttirnar en-is og is-en, og þær má skoða á GitHub síðu WMT. Handvirkri yfirferð innsendinganna er nýlega lokið og hefur ekki verið gefin út, en BLEU-niðurstöður efstu teymanna voru mjög glæsilegar, sem og úttak kerfisins.

WMT-gögnin verða einnig nytsamleg í komandi vinnu við tauganetsþýðingar milli íslensku og ensku þar sem hágæða samhliða þróunar- og prófunarsett frétta eru af skornum skammti.

V4e – Þýðingarvél fyrir sérsvið

Skýrsla: Miðeind
Aðilar: Miðeind

Markmið

Að tryggja að vélþýðingarkerfi máltækniverkefnisins séu nothæf og viðeigandi fyrir ákveðin sérsvið, á þessu stigi aðallega sem tól fyrir atvinnuþýðendur. Þróun og sannprófanir kerfanna verða gerðar í nánu samstarfi við atvinnuþýðendur/skrifstofur þýðenda. Sú reynsla sem safnast mun hafa áhrif á stillingar, hönnun og útfærslu vélþýðingarkerfa innan máltækniverkefnisins frá þriðja ári og áfram, til að hámarka nytsemi þeirra í raunverulegum tilfellum.

Varða 6 – lýsing

M5

- Vélþýðingarkerfi fyrir sérsvið þjálfað, prófunarumhverfi tilbúið.

M6

- Prófanir hafa verið gerðar, skýrsla um niðurstöður tilbúin.
- Markmiðið er framför BLEU-einkunnar um 0,3 miðað við fyrir fínstillingu, fyrir valið sérsvið.

Afurðir

Ein afurð stendur eftir: mat á Trados Studio. Samkvæmt tímalínu okkar (sjá skýrslu) verður þessari vinnu lokið fyrir 15. nóvember 2021. Þetta er vegna tafa við uppsetningu Trados Studio 2021 hjá utanríkisráðuneytinu. Aðrar afurðir eru tilbúnar:

- Vélþýðingarkerfi fyrir sérsvið er aðgengilegt á velthyding.is
- Greint er frá niðurstöðum gæðamats í þessari skýrslu

Verkþáttur V4e heldur áfram á þriðja ári.

Skýrsla

Þessi vinna var unnin í samvinnu við Þýðingamiðstöð utanríkisráðuneytisins (ÞMST). ÞMST sér um vinnu við þýðingu reglugerðartexta sem þarf að þýða yfir á íslensku sem hluta af EES-samningnum, og hefur með þessari vinnu safnað stórum þýðingarminnum með ensk-íslenskum reglugerðartextum.

Samkvæmt samkomulagi okkar við ÞMST veita þau okkur aðgang að þýðingarminni sínu til að þróa tauganetsþýðingarkerfi fyrir sérsvið sem þau geta svo prófað og metið, til að sjá hvort sjálfvirkar þýðingar geta hjálpað þýðendum við vinnu sína. Vinna við mat er tvíþætt, eins og lýst er að neðan, og samanstendur af gæðamati mismunandi tauganetsþýðingarkerfa í vefviðmóti auk mats í vinnuumhverfi þýðenda, til að meta hvort breyting eftir vélþýðingu (e. post-editing) er möguleiki fyrir ÞMST.

Af þeim markmiðum sem skilgreind voru fyrir vörðu M6 hefur ekki náðst að klára eina afurð: Mat á Trados Studio. Seinkunin kemur til vegna samhæfivandamála útgáfa og seinkunar útfærslu nýrrar útgáfu Trados Studio hjá ÞMST, eins og lýst er í hlutanum *Mat á viðbót*. ÞMST uppfærir í nýju útgáfuna í september 2021, og þá mun mat fara fram við fyrsta tækifæri.

Öðrum afurðum fyrir bæði M5 og M6 hefur verið lokið; tauganetsþýðingarkerfi fínstilltu á EES-reglugerðum hefur verið lokið, og nær ~2 BLEU-stigum hærra en ósérhæft líkan, prófunarumhverfi eru tilbúin, og gæðamat hefur verið framkvæmt og greint frá því.

Þýðingarlíkan fyrir sérsvið

Samhliða gögnin frá ÞMST voru notuð til að þjálfa tauganetsþýðingarkerfi fyrir sérsvið EES-reglugerða. Þetta var gert með því að fínstilla *mbart-cont*-líkanið sem lýst er í V4a og þessum samhliða gögnum, um ein milljón.

Eftirfarandi eru niðurstöður BLEU-einkunna (sacreBLEU) fyrir fínstillta *mbart-cont-EES*-líkanið og *mbart-cont*-líkanið fyrir fínstillingu, yfir EES- (e. EEA) prófunarsett okkar:

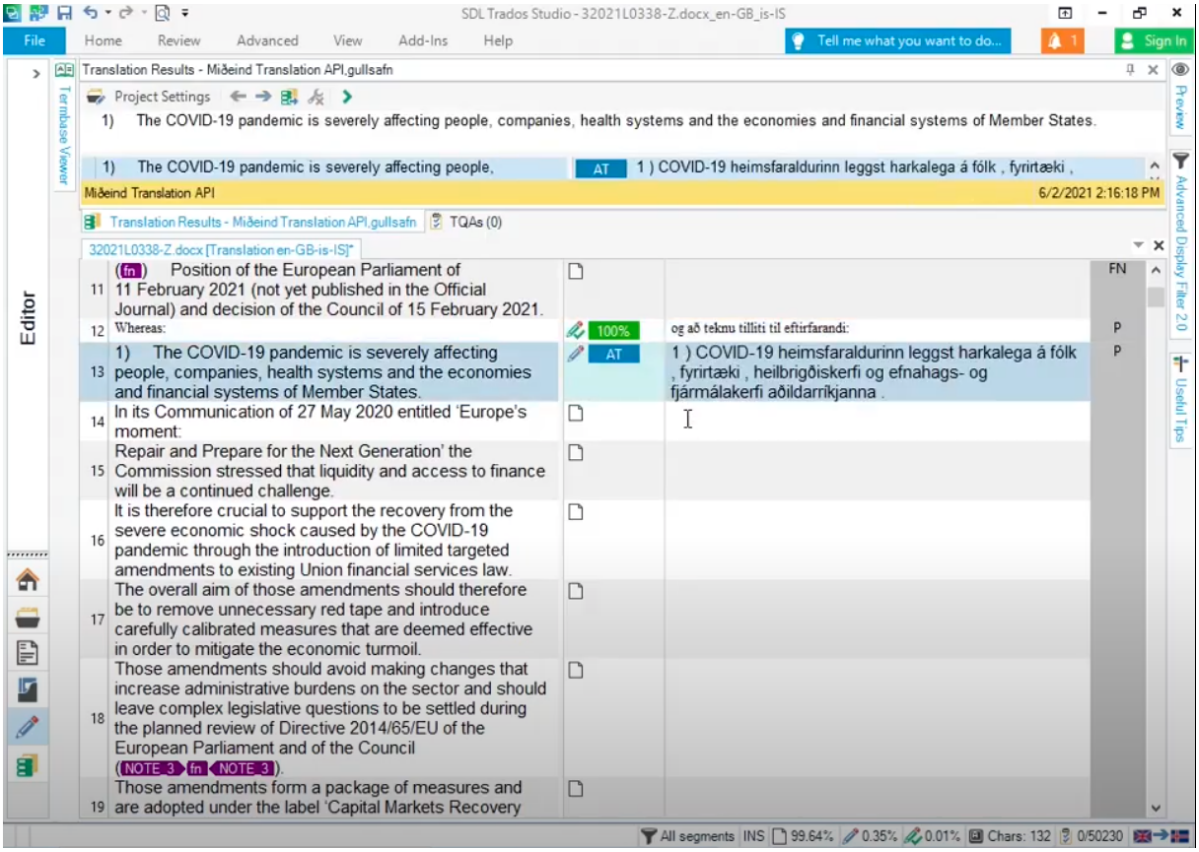
Líkan	**Matssett**	**en-is**	**is-en**
mBART25-ENIS-cont	EEA test	57,6	63,2
mBART25-ENIS-cont-EES	EEA test	59,5	65,6

Við sjáum að hækkun BLEU er +1,9 fyrir en-is áttina og heil +2,4 fyrir is-en. Þessi líkön hafa verið gefin út fyrir þýðingar á https://velthyding.is, eins og lýst er í V5a.

Trados-viðbót

Við hönnuðum viðbót fyrir SDL Trados Studio 2021-þýðingarumhverfið, sem er vinsælt meðal atvinnuþýðenda. Viðbótin útfærir þýðingarvél Miðeindar á velthyding.is sem upprunaþýðingatillögu inn í venjulegt vinnuferli þýðenda. Viðbótin er byggð á dæmaviðbót frá SDL[29].

[29] https://github.com/RWS/Sdl-Community/tree/master/MTConnectorsGooglePlugin



Sýn á Trados Studio umhverfið með þýðingarviðbótina virka.
Á meðan „100%“ þýðir að þýðingin sé 100% samsvarandi við þýðingarminnið,
þýðir „AT“ að þýðingin sé úr viðbótinni (sjálfvirk þýðing).

Mat

Það er tvennt mismunandi sem þarf að prófa. Í fyrsta lagi metum við gæði þýðinganna sjálfra, mat sem svipar til þess sem notað er í kafla V4a. Í öðru lagi metum við áhrif viðbótarinnar í Trados Studio á vinnuferli þýðenda. Ástæðan fyrir skiptingu prófana er til að reyna að aðskilja áhrif þýðingargæða og sérkenni notendaviðmótsins.

Gæðamat þýðinga

Þetta mat svipar til þess sem notað er í kafla V4a. Hér eru það atvinnuþýðendur sem meta og sérfræðingar viðfangsefnisins í stað akademísks málvísindafólks og íslenskusérfræðinga. Við metum aðeins valið sérsvið, EES-reglugerðir, og aðeins í eina átt, EN->IS, þar sem þýðendur ÞMST þýða nánast eingöngu frá ensku yfir á íslensku.

Þetta mat notar eftirvinnslumatsskala með 1 til 4 stjörnur; sem er skali með skynjaða eftirvinnslutilraun sem notaður er í verkefninu European Quality Translation 21 machine translation project[30]. Þýðendur voru beðnir um að gefa þýðingum einkunn eftir því hversu miklar breytingar þeir teldu nauðsynlegar til að gera textana nothæfa í vinnu sinni. Fjórar stjörnur þýða að þýðinguna megi nota óbreytta eða með minniháttar breytingum, og ein stjarna þýðir að

[30] https://www.qt21.eu/

þýðinguna þurfi að endurskrifa frá grunni. Vísað er til skýrslunnar fyrir verkþátt V5b fyrir skjáskot af matsviðmóti fyrir EES-reglugerðirnar.

Fyrir matið voru 250 upprunasetningar þýddar úr ensku á íslensku með fjórum mismunandi tauganetsþýðingakerfum: mBART-cont-líkanið, fínstillta mBART-cont-EES-líkanið, tilraunakennda mBART-rerank-líkanið, og Google Translate. Þrír faglegir matsaðilar fóru yfir allar þýðingar fyrir hvern uppruna í blindprófun.

Einkunnir (skalinn er 1-4, hærri er betri)

	Meðaltal	**1/4**	**2/4**	**3/4**	**4/4**
Google Translate	2,19	216	400	258	68
mBART-cont	2,89	70	212	409	251
mBART-rerank	2,90	71	203	420	248
mBART-cont-EES	**2,98**	65	164	442	271

Welch t-test p-gildi, núlltilgátan er að enginn munur sé milli einkunna kerfa. p-gildi lægri en 0,05 eru feitletruð

Kerfi A	**Kerfi B**	**p-gildi**
Google	mBART-cont	**0,00**
Google	mBART-cont-EES	**0,00**
Google	mBART-rerank	**0,00**
mBART-cont	mBART-cont-EES	**0,04**
mBART-cont	mBART-rerank	0,92
mBART-cont-EES	mBART-rerank	0,05

Líkanið sem fínstillt var fyrir EES-gögn (mBART-cont-EES) nær að meðaltali 2,98 af 4 stigum, og er hæsta einkunnin í þessu mati. Næst eru mBART-líkönin tvö fyrir almennt svið (cont og rerank), með um 2,9, og Google er í seinasta sæti með meðaleinkunnina 2,19. Almenna mBART-cont-líkanið er þegar þjálfað á EES-reglugerðum, og er því nokkuð hæft til að þýða þessa textagerð. Fínstilltu líkönin ná hærri einkunnum hér, og munurinn er tölfræðilega marktækur ($p<0,05$). Við höllumst að því að þessi munur sé ekki tilviljun ef við skoðum úttakið sjálft og tökum mið af BLEU-einkunnum sem greint er frá hér að ofan.

Í heildina fær fínstillta EES-líkanið *4* af 4 (þýðingu má nota með engum eða litlum breytingum) í meira en fjórðungi tilfella, og fær *1* aðeins í 7% tilfella, með *3* sem algengustu einkunn. Þetta er öfugt fyrir Google-líkanið, sem fær *4* í aðeins 7% tilfella, og algengasta einkunnin er *2*.

Fyrstu viðbrögð þeirra sem mátu voru að gæði þýðinganna komu skemmtilega á óvart hvað varðar setningafræði, málfræði og notkun hugtaka, jafnvel við þýðingu langra flókinna setninga.

Mat á viðbót

Vegna tafa við uppsetningu SDL Trados Studio 2021 hjá ÞMST höfum við ekki getað klárað þetta mat. ÞMST notar eldri útgáfu SDL Trados Studio sem samræmist ekki viðbótum fyrir 2021 útgáfuna, svo það er ekki hægt eins og er fyrir þýðendur að framkvæma matið. Það reyndist erfitt að fá þróunarleyfi fyrir eldri útgáfuna þar sem SDL selur þau ekki lengur. Ákvörðun var tekin um að nota 2021 útgáfuna þar sem búist var við að uppfærslan yrði komin í notkun í tæka tíð.

Forprófanir tæknimanns hjá ÞMST gefa til kynna að viðbótin virki eins og búist er við og við munum geta lokið matinu þegar þýðingarumhverfin hafa verið uppfærð. Þegar ný útgáfa Trados Studio verður komin mun valinn hópur þýðenda prófa viðbótina á raunverulegum skjölum. Umhverfið mun vera sett upp til að vera hlynnt þýðingum úr þýðingarminninu, með tauganetsþýðingu sem viðbúnað þegar þýðingarminnið gefur ekki góða samsvörun.

ÞMST bíður nú eftir tækniaðstoð frá Trados til að uppfæra hugbúnaðinn, og ef gert er ráð fyrir að nýja útgáfan verði sett upp hjá ÞMST í síðasta lagi í fyrri hluta október (tímalína staðfest af ÞMST) getum við tekið október og byrjun nóvember frá fyrir notendaprófanir. Við teljum að einn mánuður ætti að vera nægur tími fyrir þýðendurna til að fá tilfinningu fyrir ferlinu. Þegar prófunum er lokið munu þátttakendur svara spurningalista þar sem þeir gefa álit sitt á vélþýðingunum og ferlinu. Við stefnum því á að skila þessum hluta tilbúnum við miðjan nóvember. Þetta er þó háð utanaðkomandi þáttum, eins og hvort uppfærslunni verður örugglega lokið innan þessa tímaramma, og vinnuálag ÞMST á þeim tíma.

V5a – Viðmót fyrir þýðingarvél

Skýrsla: Miðeind
Aðilar: Miðeind

Markmið

Bæta veflæga vélþýðingaviðmótið fyrir daglega notkun, sem upphaflega var hugsað fyrir fagnotendur en einnig með almenning í huga. Sérstaklega að leyfa val (úr fyrirfram ákveðnu setti) sviðs (e. domain) í samstarfi við atvinnuþýðendur á hverjum stað. Þegar ákveðið svið er valið gefur það þýðingarkerfinu merki um að aðlaga framleiddan markmiðstexta (e. target text) að stíl þess sviðs.

Varða 6 – lýsing

M5

- Upphleðsla skjala og vefsíðuþýðingar útfærðar og aðgengilegar á vefnum.

M6

- Þýðingar sérsviðs útfærðar og aðgengilegar á vefnum.

Afurðir

Öllum markmiðum hefur verið náð.

Verkþáttur V5a heldur áfram á þriðja ári.

Skýrsla

Upphleðsla skjala

Upphleðsla skjala er studd á https://velthyding.is. Studd skráarsnið eru *.txt og *.docx. Hægt er að velja skjöl með „Hlaða upp“-hnappnum eða með því að draga og sleppa. Virkni Word-skjala er útfærð með Mammoth-pakkanum[31].

Vefsíðuþýðingar

Vefsíðuþýðingaviðbótin er aðgengileg á https://github.com/mideind/GreynirTranslateChromePlugin og má setja upp samkvæmt meðfylgjandi leiðbeiningum.

[31] https://github.com/mwilliamson/mammoth.js/

Þó að viðbótin virki viljum við leggja áherslu á að hún sé aðeins sönnun á gildi hugmyndar (e. proof of concept) sem er ekki ætluð fyrir víðtæka almenna notkun og er því ekki auglýst á https://velthyding.is. Vefsíður innihalda almennt mikið magn texta, og þýðing þeirra krefst mikils reikniafls sem við getum því miður ekki veitt í langan tíma án endurgjalds.

Myndband sem sýnir viðbótina er aðgengilegt hér: https://drive.google.com/file/d/11d0As2QtydbYu3t3qm7RJa0pHyWJ-PNG/view.

Þýðingar sérsviða

Líkönin sem voru fínstillt fyrir V4e (textar EES-reglugerða) eru aðgengileg á https://velthyding.is. Þýðingarlíkan má velja með því að smella á táknmynd stillinga (tannhjól) og velja valmöguleikann „Þýðingarlíkan“. Þessi listi kemur frá bakendanum. Bakendinn sér svo um að veita tiltekið líkan í gegnum JSON REST-forritaskil sín.

V5b – Lýðvirkjun

Skýrsla: Miðeind
Aðilar: Miðeind

Markmið

Endurbætur á lýðvirkjun. Notendur munu geta gefið mat sitt á gæðum þýðinga, sérstaklega með einkunnagjöf valmöguleika og nægð og málfimi til að geta A/B prófað mismunandi líkön. Notendur munu geta, annaðhvort nafnlaust eða innskráðir, séð og valið möguleika þýðinga, auk þess að geta bætt inn eigin þýðingum. Slík innslegin og/eða leiðrétt gögn frá notendum verða geymd til að auka samhliða málheildir og til viðbótar í þjálfunarumferðir í framtíð. Söfnuðum gögnum verður meðal annars beitt til að hjálpa við síun bakþýddra gagna.

Varða 6 – lýsing

M5

- Veflægt viðmót fyrir mat og yfirferð tilbúið

Afurðir

Öllum markmiðum hefur verið náð og verkþættinum er lokið.

- Matstól lýðvirkjunar komið í gang á velthyding.is

Verkþætti V5b lýkur við M6.

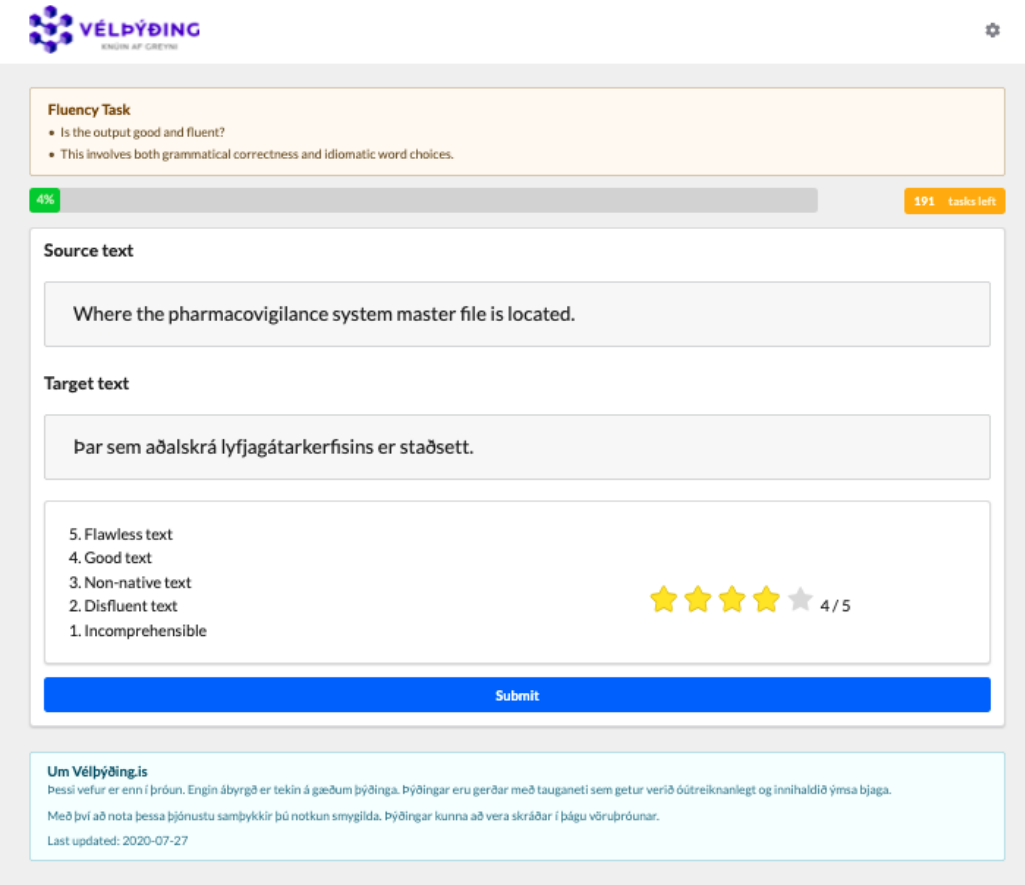
Skýrsla

Matstól lýðvirkjunar er komið í gang á velthyding.is og hefur þegar verið notað fyrir mannlegt mat í V4a og V4e. Tólið gefur kost á ýmsum matsaðferðum, byggðum á *málfimi* og *nægð*, *samanburði*, og *beinu mati*. Eftirfarandi er lýsing á kerfishögun, og á notendaviðmótinu eins og það birtist endanlegum notanda.

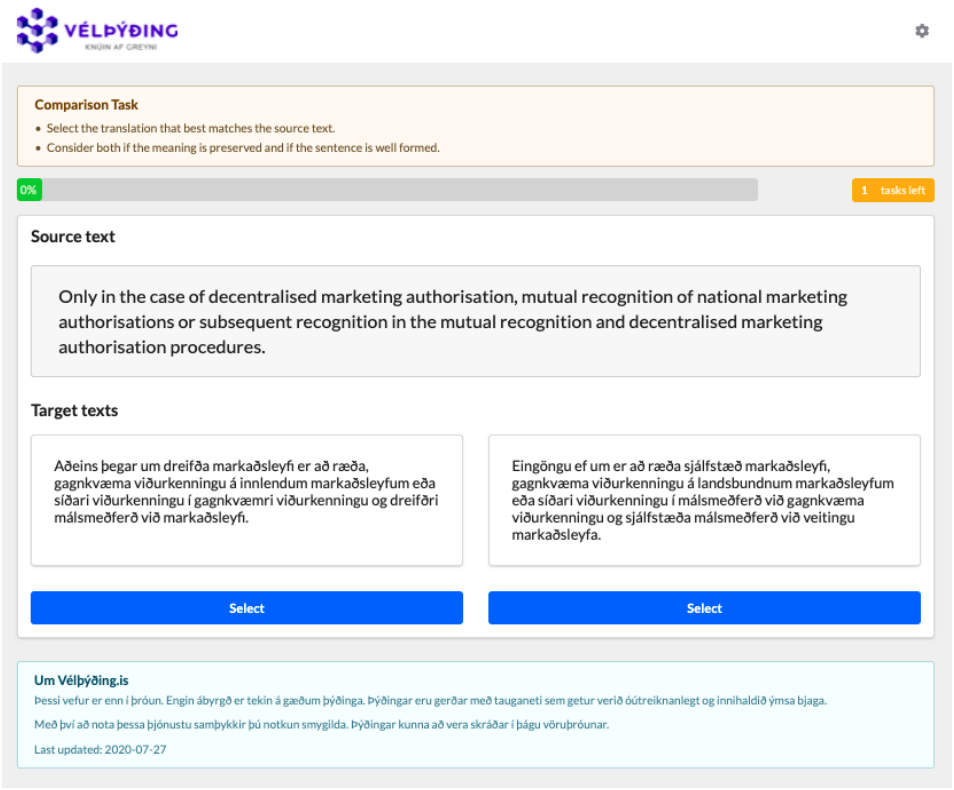
Högun

Tólið er sett fram á undirsíðu fyrir vefsíðu vélþýðinga Miðeindar, velthyding.is, þar sem notendur geta skráð sig og fengið aðgang að umsögnum sínum. Framendinn er forritaður í ReactJS, og bakendinn notar Django-kerfið. Skráðir notendur á velthyding.is með réttindi stjórnenda geta sent inn gagnasett, búið til matsframkvæmdir (e. evaluation campaigns) og úthlutað framkvæmdum á notendur. Framkvæmd er skilgreind sem mat á ákveðnu gagnasetti, og má úthluta á marga notendur. Framkvæmd getur verið opin eða lokuð (e. public or private), og

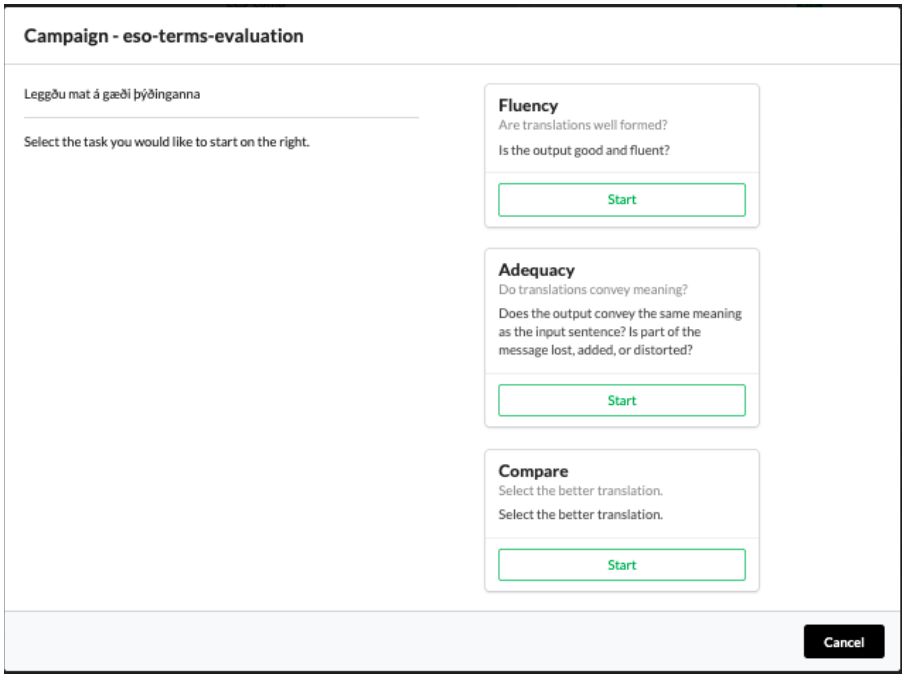
henni getur verið úthlutaður tímarammi (byrjunar- og endadagsetning). Gagnasett inniheldur upprunatexta og einn eða fleiri markmiðstexta. Þetta veitir möguleika á samanburði mismunandi gagnasetta, annaðhvort með einkunnagjöf hvers pars uppruna/markmiðs (mynd 1) eða með því að bera saman tvö markmið og velja betri þýðinguna (mynd 2). Mismunandi matsaðferðir (málfimi og nægð, samanburð, og beina úttekt) má einnig setja saman í eina framkvæmd (mynd 3).



Mynd 1. Mat á málfimi. Notandinn raðar þýðingum eftir málfimi þeirra.



Mynd 2. Samanburður tveggja þýðinga. Notandinn velur þá þýðingu sem honum þykir betri.

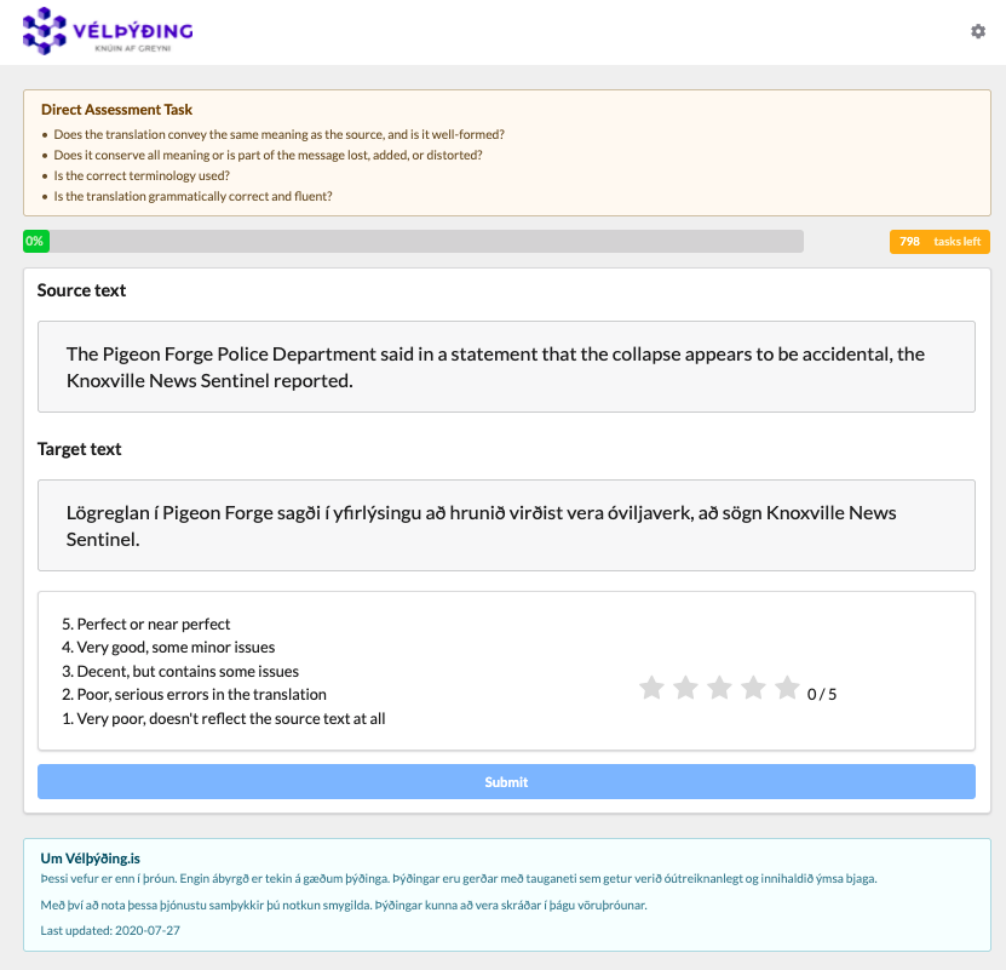


Mynd 3. Framkvæmd samsett af málfimi, nægð og samanburðarúttektum.

Matsaðferðir

Þarfir matstólsins fólu í sér útfærslu fyrrnefndra umsagna um málfimi, nægð og samanburð. Mæliaðferðir fyrir málfimi og nægð, skilgreindar fyrir ARPA MT-áætlunina[32], eru orðnar staðlaðar matsaðferðir fyrir vélþýðingar. Þær fjalla um tvo aðskilda þætti þýðingar, og þá þætti er hægt að vega og sameina í eitt mæligildi, á meðan samanburður nýtist til að skoða tvær þýðingar hlið við hlið.

Að auki höfum við útfært möguleika á beinni úttekt í tólinu, til að meta almenn gæði þýðingar setningar (mynd 4). Þetta er matsaðferðin sem notuð er í mannlegu mati tauganetsþýðingarkerfa eins og lýst er í V4a.

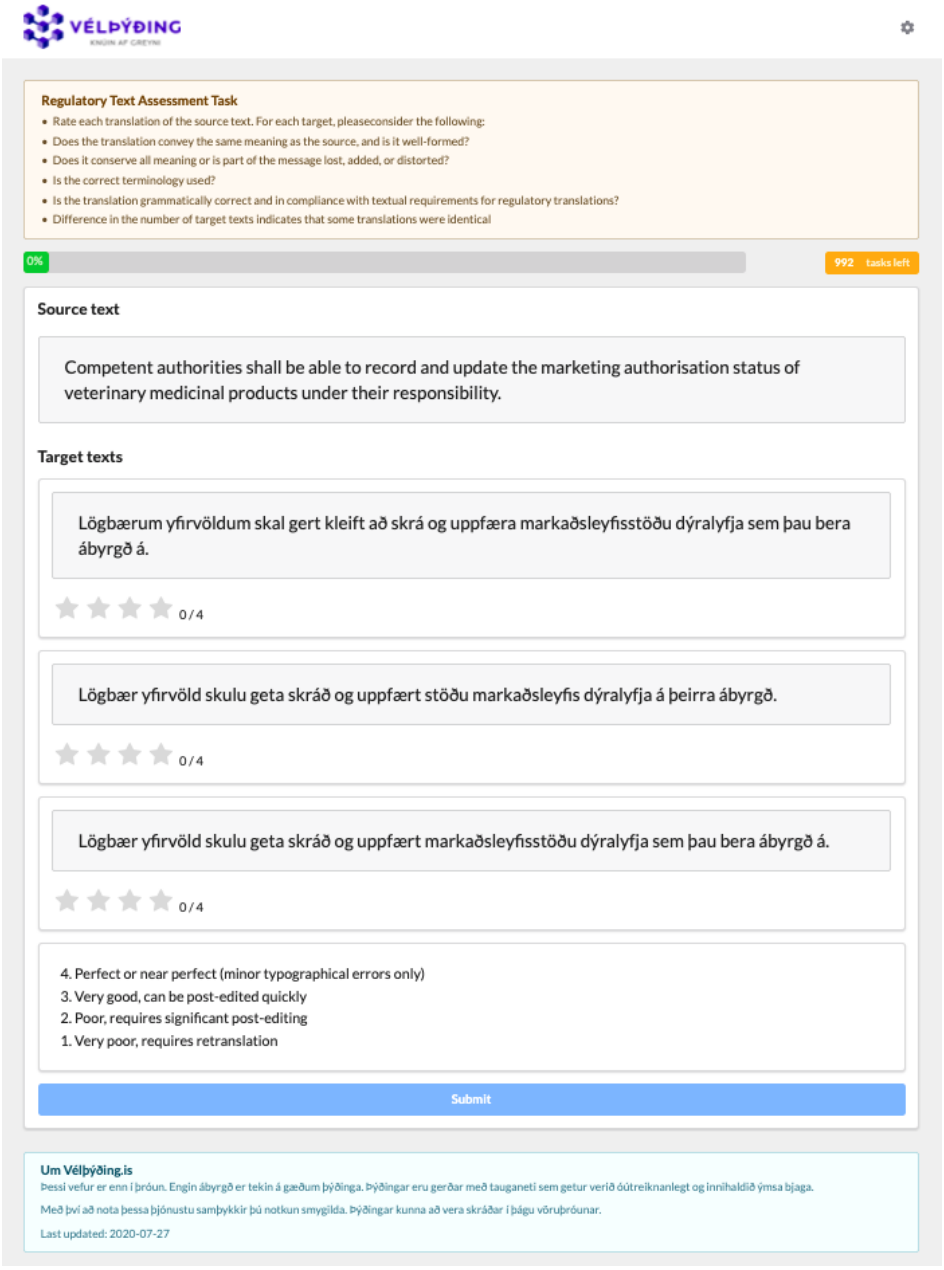


Mynd 4. Bein úttekt er almennt mat á gæðum þýðingar.

[32] https://dl.acm.org/doi/pdf/10.3115/1075812.1075840

Viðbót 2: Mat á EES-þýðingum

Fyrir þýðingar sérsviða (V4e) gerðum við beinnar úttektar framkvæmd á þýðingum á EES-reglugerðum frá fjórum vélþýðingarkerfum með tólinu. Að beiðni notenda, þýðenda við Þýðingamiðstöð utanríkisráðuneytisins, útfærðum við eina viðbót við tólið: eina síðu þar sem allar þýðingar eins upprunatexta eru birtar saman til að auðvelda samanburð markmiða og spara tíma þegar viðmiðunarefni er notað (mynd 5).



Mynd 5. EES-úttektarframkvæmd, hverju markmiði gefin einkunn á einni síðu.

Fyrstu viðbrögð við tólinu

Notendur í matsferlunum tveimur voru beðnir um að gefa álit sitt á tólinu, og í heildina voru svörin jákvæð; notendur sögðu notendaviðmótið vera auðvelt í notkun. Einn notandi lagði til að nytsamlegt væri að geta gefið einkunn með flýtilyklum (e. keyboard shortcuts). Þessi tillaga hefur verið skráð sem eiginleiki sem verður útfærðir í framtíðinni.

Í upprunalegu áætluninni er tólinu lýst sem matstóli fyrir lýðvirkjun. Þetta væru sannarlega góð not á tólinu, en ekkert slíkt átak hefur enn verið skilgreint eða framkvæmt. Áður en lagt verður í svo stórt verkefni er gott að hafa reynsluna af því að ljúka tveimur litlum matsverkum fyrir þessa vörðu.

L2 – Sérhæfðar villumálheildir

Skýrsla: Háskóli Íslands
Aðilar: Háskóli Íslands, Miðeind

Markmið

Að safna sérhæfðum villumáheildum. Ákveðnir samfélagshópar gera öðruvísi villur en aðrir; börn og unglingar, lesblindir, fólk sem hefur íslensku sem annað mál, o.s.frv. Þar sem slíkar villur geta verið óalgengar í almennri villumálheild þarf sérhæfðar málheildir til að geta tekið á þeim.

Varða 6 – lýsing

M5
Söfnun texta frá börnum vel á veg komin. A.m.k. 1.000 villur úr þessum textum hafa verið greindar og merktar. Stærðarmarkmið fyrir sérhæfðar villumálheildir endanlega skilgreind.

M6
Sérhæfðum málheildum hefur verið lokið og standast kröfurnar sem skilgreindar voru í M5, og sendar til CLARIN.

Afurðir

Öllum markmiðum vörðunnar var náð. Stærðarmarkmið fyrir hverja undirmálheild, skilgreind við M5, eru 20.000 villur í Villumálheild íslensku sem annað mál, 5.000 villur í Íslensku lesblinduvillumálheildinni og 5.000 villur í Villumálheild íslensks barnamáls.
Sérhæfðu villumálheildirnar þrjár hafa verið gefnar út á CLARIN hver um sig og á GitHub sem ein heild (https://github.com/antonkarl/iceErrorCorpusSpecialized).

- Villumálheild íslensku sem annað mál er aðgengileg á CLARIN: http://hdl.handle.net/20.500.12537/131
- Íslenska lesblinduvillumálheildin er aðgengileg á CLARIN: http://hdl.handle.net/20.500.12537/132
- Villumálheild íslensks barnamáls er aðgengileg á CLARIN: http://hdl.handle.net/20.500.12537/133

Verkþáttur L2 heldur áfram á þriðja ári, með áherslu á annars máls texta og lesblindutexta.

Skýrsla

Þetta undirverkefni felur í sér að búa til villumálheildir fyrir þrjá aðgreinda hópa málnotenda: annarsmálshöfum íslensku, einstaklingum með dyslexíu og börn. Þessir hópar endurspeglast í málheildunum þremur sem gefnar eru út, Villumálheild íslensku sem annað mál, Íslensku

lesblinduvillumálheildarinnar og Villumálheild íslensks barnamáls, sem rætt er nánar um í næstu undirköflum.

1. Merkingarferli
Aðferðin við að búa þessar málheildir til byggist að miklu leyti á aðferðinni sem notuð var til að búa til Íslensku villumálheildina, sem var ítarlega lýst í M3-skýrslu. Eina breytingin á merkingarferlinu er í söfnun texta, þar sem þeim þurfti að safna beint frá höfundum. Merkingarferlið sem kemur í kjölfar textasöfnunar notar lagskipta aðferð sem endar í safni XML-skjala á TEI-formi með endanlegum villumerkingum.

Fyrir textasöfnun var búið til rafrænt eyðublað sem tryggir upplýst samþykki höfundar texta. Þessi rafrænu eyðublöð voru notuð við söfnun texta frá þeim sem læra íslensku sem annað mál og frá lesblindum, með mismunandi eyðublöðum fyrir hverja málheild til að safna viðeigandi upplýsingum um höfund. Höfundar geta valið hvort textarnir eru birtir nafnlaust eða ekki. Annað eyðublað var búið til fyrir söfnun texta skrifaða af börnum, þar sem foreldrar/forráðamenn þurftu að gefa samþykki sitt. Þetta eyðublað er annaðhvort prentað út og undirskrift foreldris/forráðamanns innifalin, eða sent rafrænt, þar sem tölvupósturinn sjálfur jafngildir undirskrift. Þetta var gert eftir ráðfæringu við lögfræðinga.

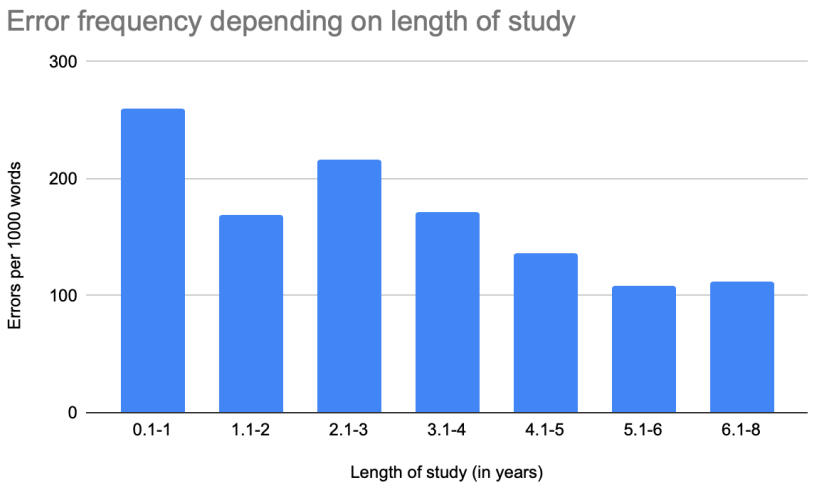
2. Merkingarskema
Merkingarskemað eru það sama og notað var fyrir Íslensku villumálheildina, þróað frekar til að gera grein fyrir villum sem birtast helst í sérhæfðu villumálheildunum. Það samanstendur af þremur þrepaskiptum stigum: Aðalflokkar, undirflokkar, og villukóðar notaðir við merkingu. Aðalflokkarnir eru alls sex, undirflokkarnir eru 31 og villukóðarnir eru 253. Merkingarskemað er aðgengilegt á https://github.com/antonkarl/iceErrorCorpusSpecialized/blob/master/errorCodes.tsv.

Sniðið á XML-skránum er það sama og í Íslensku villumálheildinni og því er lýst í M3-skýrslu.

3. Sérhæfðar villumálheildir

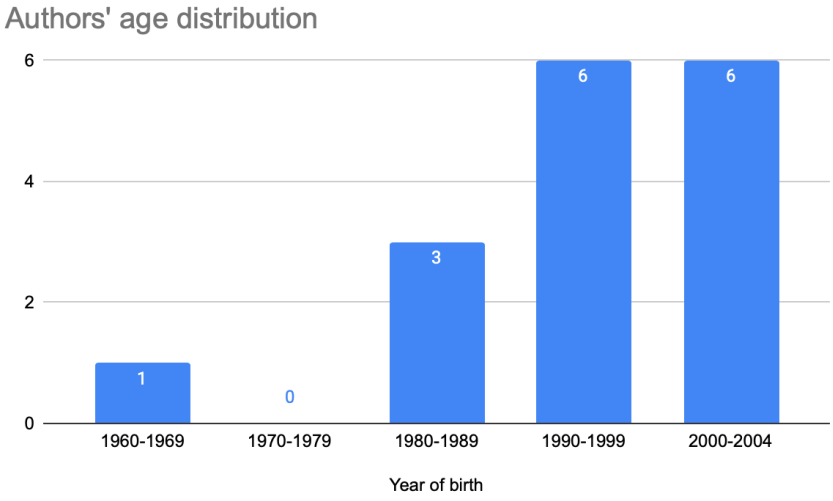
3.1 Villumálheild annars máls
Villumálheild íslensku sem annað mál samanstendur af 76 ritgerðum með 14.926 breytingarspönnum og 21.842 flokkuðum villum og er gefin út á CLARIN.

Rafræna eyðublaðið biður um ýmsar upplýsingar um höfundinn og textana sem gefnir eru, sem gerir ítarlega greiningu textanna mögulega. Súluritið að neðan sýnir hversu margar villur á hver 1000 orð voru í textum á mismunandi stigum tungumálanámsins. Lárétti ásinn sýnir í hversu mörg ár viðkomandi höfundur hefur lært tungumálið, tekið saman í eins árs tímabil. Súluritið sýnir, með einni undantekningu, að villutíðni lækkar eftir því hversu lengi viðkomandi hefur lært íslensku og að við fimm ár jafnist villutíðnin út. Móðurmál höfundar er ekki tekið með í þetta súlurit, en það getur haft áhrif á villutíðnina þar sem sum tungumál eru líkari íslensku en önnur.



3.2 Lesblinduvillumálheild
Íslenska lesblinduvillumálheildin samanstendur af 26 textum með 3.352 breytingarspönnum og 5.730 flokkuðum villum og er gefin út á CLARIN.

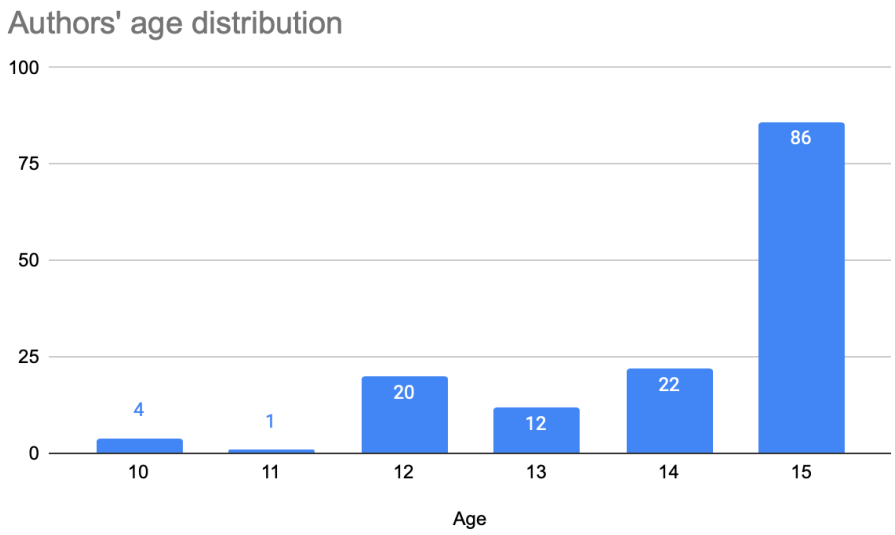
Textarnir voru skrifaðir af 16 mismunandi höfundum með dyslexíu. Á eyðublaðinu kom skýrt fram að höfundar þurfi að hafa formlega dyslexíugreiningu. Engum aldurstakmörkunum var beitt við söfnun texta, og eyðublað með samþykki söfnunar fyrir barnatexta var notað fyrir söfnun texta frá höfundum sem ekki höfðu náð lögaldri. Súluritið að neðan sýnir aldursdreifingu höfunda. Meirihluti höfunda fæddist milli 1990 og 2004, aðeins fjórir höfundar voru eldri en það.



3.3 Villumálheild barnamáls
Villumálheild íslensks barnamáls samanstendur af 119 textum með 5.079 breytingarspönnum og 7.817 flokkuðum villum og er gefin út á CLARIN.

Aldursbil textahöfunda var 10–15 ár, fæddir 2005–2011. Þar sem söfnun texta hófst 2020 eru fyrrnefnd ár skilgreind. Súluritið að neðan sýnir aldursdreifingu höfunda, og sýnir að flestir textar

koma frá höfundum nær efri mörkum aldursbilsins. Þetta gæti orsakað færri villur í hverjum texta.



4. Tölfræðilegur munur

4.1 Villur á hver 1000 orð
Þegar villur á hver 1000 orð í sérhæfðu villumálheildunum eru bornar saman við almennu villumálheildina er stórtækur munur. Almenna villumálheildin er með 45,76 villur á hver 1000 orð á meðan sérhæfðu villumálheildirnar í heild eru með 171,73, rétt tæplega fjórum sinnum fleiri villur. Innan sérhæfðu villumálheildanna er líka munur á villutíðni, lesblindumálheildin er með hæstu villutíðnina en villumálheild annars máls þá lægstu. Villumálheild annars máls er með 153,52 villur á hver 1000 orð á meðan barnamálsmálheildin er með 208,77 og lesblindumálheildin 214,27, sem er stórt stökk frá villumálheild annars máls.

4.2 Tíðni undirflokka
Til að fá yfirlit yfir villurnar sem þessir mismunandi hópar gera, skoðum við fimm algengustu undirflokkana fyrir sérhæfðu villumálheildirnar þrjár og berum þá saman innbyrðis og við almennu villumálheildina.[33] Villumálheild íslensku sem annað mál er í sérflokki hvað þetta varðar vegna þess að hún er eina málheildin sem hefur tvo málfræðilega aðalflokka í fimm algengustu aðalflokkunum. Þessar villur tengjast tungumálainnsæi og því er rökrétt að þær séu algengari í textum skrifuðum af þeim sem nota íslensku sem annað mál. Í hinum málheildunum eru flokkar sem tengjast réttritun algengari. Allar málheildirnar deila algengasta undirflokknum, *greinarmerkjasetning* (e. punctuation), en sá flokkur er algengari í villumálheild barnamáls og almennu villumálheildinni.
Almenna villumálheildin og villumálheild barnamáls deila fleiri svipuðum villum, varðandi kommur og gæsalappir, á meðan lesblindumálheildin er með fleiri villur tengdar stöfum sem vantar og klofnum samsetningum.

[33] Töflurnar eru aðgengilegar á https://github.com/antonkarl/iceErrorCorpusSpecialized/tree/master/subcatFrequency

L7 – Kerfisbundin þróun málrýnis

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

Markmið

Að þróa málrýnikerfi fyrir íslensku sem nær yfir skilgreint safn villna, sem borin eru kennsl á og forgangsraðað á meðan á verkefninu stendur.

Markmið verða skilgreind á ítraðan hátt eftir forgangsröðun villutegunda sem eru skilgreindar í undirverkefni L3, með gögnum frá undirverkefnum L1 og L2. Almenna markmiðið er að finna þessar villur og veita nothæfar útskýringar, tillögur og/eða leiðréttingar, eftir því sem við á.

Varða 6 – lýsing

M5: Málrýnir getur borið kennsl á og lagt til leiðréttingar fyrir (og valkvætt leiðrétt sjálfvirkt, þar sem hægt er að gera ráð fyrir) stóran hluta setts af algengustu málfræðivillunum, eins og rangt fall frumlags ópersónulegra sagna. Við stefnum á að heildar F-gildi fyrir 15 algengustu málfræðivilluflokkana verði 50%. Kröfur samþættingar í ritstýringarumhverfi Kjarnans skilgreindar og innleiddar staðbundið. Umfang mögulegrar eigindlegrar endurgjafar gert skýrt.

M6: Fyrsta útgáfa skipanalínutóls málrýnis fyrir almenning gefin út á CLARIN, sem getur borið kennsl á og lagt til leiðréttingar fyrir sett algengustu stafsetningar- og málfræðivillna sem finnast í íslenskum textum. Fyrir 20 algengustu undirflokka í hverjum yfirflokki stefnum við á að F-gildi verði 70% fyrir stafsetningarvilluflokka og 60% fyrir málfræðivilluflokka. Útgáfa málrýnisins fyrir almenning er aðgengileg á vefnum, sem gerir notendum kleift að skrifa eða senda inn texta til yfirferðar og merkingar. Málrýni hefur verið samþætt inn í ritstjórnarumhverfið. Prófanir hafa verið gerðar. Skýrsla um niðurstöður og þarfir notenda tilbúin.

Afurðir

Öllum markmiðum vörðunnar var náð.

- Við náum F-gildinu 61,43% fyrir stafsetningarvillur í heild. Við náum F-gildinu 70% fyrir 21 villukóða, og yfir 60% fyrir 25.
- Við náum F-gildinu 21,21% fyrir málfræðivillur í heild, en aðeins 17 villukóðar eru í prófunarsettinu. Eins og nefnt var í M4 er villa í prófunarforritinu sem olli því að þegar endurskoðaðar vörður voru skilgreindar var spá í hærra lagi, sérstaklega fyrir málfræðivillur.
- Skipanalínutólið í GreynirCorrect var uppfært til að greina málfræðivillur. Ný útgáfa pakkans GreynirCorrect er aðgengileg á CLARIN http://hdl.handle.net/20.500.12537/148, sem og á GitHub (https://github.com/mideind/GreynirCorrect).
- Málrýnirinn er aðgengilegur á vefnum á https://yfirlestur.is/.

- Samþættingu í ritstjórnarumhverfi Kjarnans er lokið og hún rædd í þessari skýrslu.

Verkþáttur L7 heldur áfram á þriðja ári.

Skýrsla

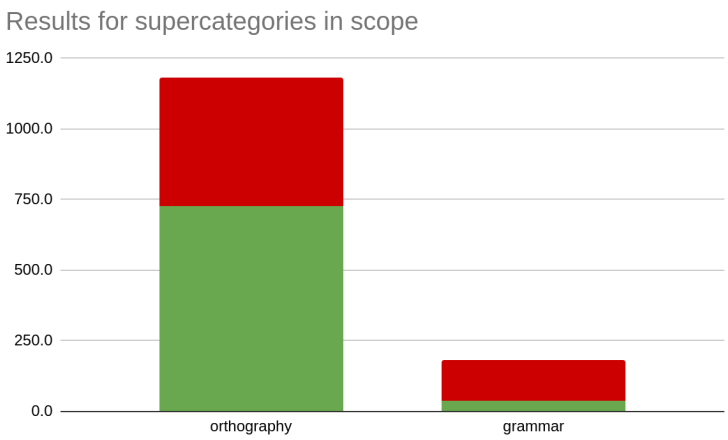
1. Yfirlit
Á öðru ári var áhersla lögð á að finna og merkja málfræðireglur, öfugt við fyrsta ár þegar áherslan var á að finna stafsetningarvillur og aðrar samhengisháðar villur. Greinarmunurinn á slæmu orðalagi og raunverulegum villum er mun óljósari fyrir málfræðivillur en fyrir þær villur sem tekið hefur verið á hingað til. Til viðmiðunar notuðum við *íslenskan málstaðal*, eins og hann birtist í *Ritreglum* og *Stafsetningarorðabókinni* (aðgengilegar hjá SÁM). Þar sem við rákumst á óljós eða sjaldgæf tilfelli leituðum við hjálpar frá Stofnun Árna Magnússonar (SÁM). Í sumum tilfellum leiddi það til endurflokkunar og/eða sköpunar nýrra villuflokka.

Við skilgreinum stranga og óstranga málrýni. Ströng málrýni stýrist eingöngu af íslenskum málstaðli, en fyrir óstranga málrýni merkjum við líka nokkrar setningagerðir þar sem við höfum ekki áþreifanlegar reglur frá málstaðli, en erum fullviss um (a) að þær séu rangar og/eða (b) að þær séu á gráu svæði en notendur kynnu að meta merkinguna.

Milli M4 og M5 var nýju lagi flokkunar bætt í Íslensku villumálheildina, og þau eru því þrjú. Undir málfræðivillum (efsta lagi), höfum við 14 undirflokka, sem hver hefur einn eða fleiri villukóða. Af þeim 14 voru 7 metnir sem a.m.k. að hluta til innan ramma. Í öðrum kafla skoðum við nánar niðurstöður og meðhöndlun fyrir hvern þeirra.

2. Mat og niðurstöður
2.1 Heildarniðurstöður
Súluritið að neðan sýnir niðurstöður fyrir tvo yfirflokka sem metnir voru innan ramma, *málfræði* og *stafsetning*. Aðrir yfirflokkar eru að mestu metnir sem utan ramma, þar sem þeir taka á óljósari vandamálum. Grænu/rauðu stæðurnar tákna tilvik hvers yfirflokks í villumálheildinni, þar sem græni hlutinn táknar F-gildið fyrir yfirflokkinn. Taflan sýnir tíðni og F0.5-gildi viðeigandi yfirflokka.

Flokkur	Tíðni	F-gildi
Stafsetning	1182	61,5
Málfræði	215	18,5
Samtals	1397	54,9

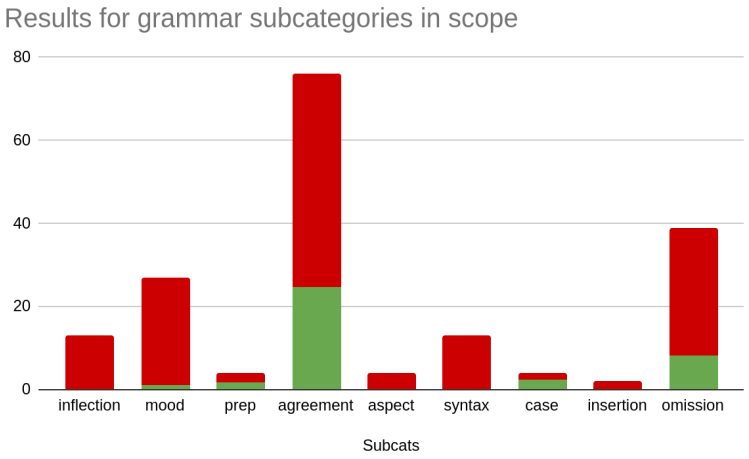


Yfirflokkurinn *stafsetning (e. orthography)* er langstærstur og samsvarar nokkurn veginn vinnu fyrsta árs. Kerfið nær góðum árangri fyrir flesta undirflokka stafsetningar, en sumir villukóðar innan

undirflokkanna eru erfiðari en aðrir. Undirflokkurinn *greinarmerkjasetning (e. punctuation*) inniheldur algengustu villukóðana, sem margir hverjir ná mjög háum F-gildum. Af þeim 60 villuflokkum innan ramma sem koma fyrir í prófunarsettinu náum við F-gildi yfir 70% fyrir 21, og yfir 60% fyrir 25. Sumum villukóðum er enn ekki tekið á, sem draga niðurstöðurnar niður fyrir allan undirflokkinn. Hið sama gildir fyrir marga aðra undirflokka.

Yfirflokkurinn *málfræði* samsvarar nokkurn veginn vinnu fyrsta árs. Súluritið að neðan sýnir niðurstöður fyrir undirflokka innan ramma undir *málfræði*, með sömu grænu/rauðu skiptingu og í súluritinu að ofan. Taflan að neðan sýnir F-gildi þessara undirflokka. Af þeim 17 villuflokkum innan ramma sem koma fyrir í prófunarsettinu tökum við á 8. Taka skal fram að við tökum einnig á mörgum kóðum sem koma ekki fyrir í prófunarsettinu. Tveir flokkanna ná F-gildi yfir 57%, og fimm þeirra yfir 25%. Bæði fjöldi villukóða og heildartíðni þeirra í villumálheildinni eru mikið lægri en fyrir *stafsetningu*, svo erfitt er að draga einhverjar ályktanir frá þessum tölum. Villukóðar innan þessa undirflokks eru einnig grófgerðari en þeir sem eru innan *stafsetningar*.

Undirflokkar	Tíðni	F-gildi
inflection	13	0,0
mood	27	3,1
prep	37	9,7
agreement	76	32,2
aspect	4	5,0
syntax	13	0,0
case	4	62,5
insertion	2	0,0
omission	39	21,3
Samtals	215	19,7



2.2 Beygingarvillur (inflection)
Villur þar sem annaðhvort er röng orðmynd eða röngum beygingarreglum beitt. Tekið hefur verið á algengum beygingarvillum með „villureglum“ í málfræðinni, eins og þegar -r í nafnorðum sem enda á -ir er haldið í öllum orðmyndum (þolfalli eintölu *læknirinn → lækninn)*, og eignarfalli fleirtölu *karlana → karlanna*.

2.3 Háttarvillur sagnorða (mood)
Villukóðunum var skipt í fíngerðari kóða eftir því hvort háttur stýrist af aukatengingu, sögn í aðalsetningu, eða óljósari atriðum eins og merkingarlegu samhengi, sem er frekar orðalagsvilla en málfræðivilla. Líkt og með margar aðrar málfræðivillur eru mörg tilfelli þar sem engar viðeigandi leiðbeiningar voru í íslenskum málstaðli, svo ekki var hægt að benda á hátt sem greinilega villu.

Þar sem hægt var skilgreindum við óströnga tilfelli. Tilfelli þar sem þarf samhengi utan setningarinnar voru metin utan ramma eins og er. Fjöldi *rangra neikvæða* (e. false negative) bendir til þess að það séu mörg tilfelli sem við höfum ekki tekið á.

2.4 Forsetningarvillur (prep)

Forsetningarvillur eru tilfelli þar sem röng forsetning er valin. Þeim er skipt í þrjá villukóða, sem fara eftir forsetningunni sem um ræðir. Tveir fyrri eru fíngerðari en sá þriðji. Meðhöndlun fyrstu tveggja hefur verið bætt í málrýnitólið, með því að nota tilvik þeirra í Íslensku villumálheildinni.

Þriðji villukóðinn er víðtækari en hinir tveir og mikið algengari, svo honum var skipt í þrjá fíngerðari kóða eftir andlagi forsetningarinnar. Einn þessara kóða er skilgreindur utan ramma eins og er þar sem hann er samhengisháður, og meðhöndlun fyrir annan þeirra hefur verið bætt í málrýninn. Við erum eins og er að vinna í þeim þriðja, sem byggir mjög á rökliðagerð sagn- og nafnorða.

2.5 Samræmisvillur (agreement)
Við náum að taka vel á samræmi frumlags og sagnorðs. Við erum að vinna í samræmi andlags og sagnfyllingar og samræmi innan nafnliða.

2.6 Horfsvillur í sagnorðum (aspect)
Einn villukóði er skilgreindur sem innan ramma, aðrir velta of mikið á merkingu. Sá kóði lýsir villum þar sem sagnorð í framvinduhorfi ætti að vera í einfaldri nútíð, til dæmis *ég er að sitja > ég sit.* Tekið var á þessum villum með því að skoða setningagerðirnar eftir þáttun.

Í mörgum tilfellum hefur eðli sagnar áhrif á hvort hægt er að nota hana í framvinduhorfi. Þar sem við höfum ekki gögnin til að greina á milli þeirra eins og er, voru allar slíkar setningagerðir merktar sem hugsanlegar villur. Þetta hefur auðvitað áhrif á niðurstöðurnar, þar sem við fáum fjölda *falskra jákvæðra* (e. false positive).

2.7 Setningafræðivillur (syntax)
Stór flokkur sem inniheldur ýmsar setningafræðilegar villur. Greining á ruglingi atviksorða og lýsingarorða veltur aðallega á merkingarlegu samhengi og er því strembið utan takmarkaðs ramma eins og nafnliða. Villur í 'hvor/hver annar' eru umdeilanlegar, þar sem flestir málnotendur nota það í raun á rangan hátt. Ekki hefur verið tekið á þessum villum.

Villum í nafnháttarmerkjum er tekið á fyrir þekktar hjálparsagnir. Önnur tilfelli innan villukóðanna velta of mikið á merkingu og eru taldar utan ramma. Flokkunum hefur ekki verið skipt upp, svo þetta hefur áhrif á niðurstöður okkar.

Öðrum villum hefur ekki verið tekið á sem slíkum, en þær valda því að setningar verða ótækar og líklega óþáttanlegar, svo kerfið mun hugsanlega finna þær sem villur og láta notandann vita.

2.8 Fallvillur (case)
Þessar villur skarast á við beygingarvillur, en það þótti réttlætanlegt að hafa sér flokk fyrir þær þar sem þær eru nógu tíðar og mikilvægar. Kerfið tekur nokkuð vel á þeim, þökk sé góðri meðhöndlun á falli frumlaga.

2.9 Aðrar endurbætur (M6)
Þáttarinn sem er undirliggjandi var endurbættur sem hluti af verkþætti I5.

Til að bregðast við prófunum notenda, sem ræddar eru í kafla 3, útfærðum við breytingar sem bættu stórlega upplifun notenda en endurspegluðust ekki endilega í niðurstöðum sjálfvirks mats, en höfðu ekki neikvæð áhrif á þær. Sem dæmi leiðréttum við ekki lengur óþekkt orð með

hástöfum þar sem þau eru líklegast nafneiningar, erlend eða íslensk, sem við höfum ekki í gögnum okkar.

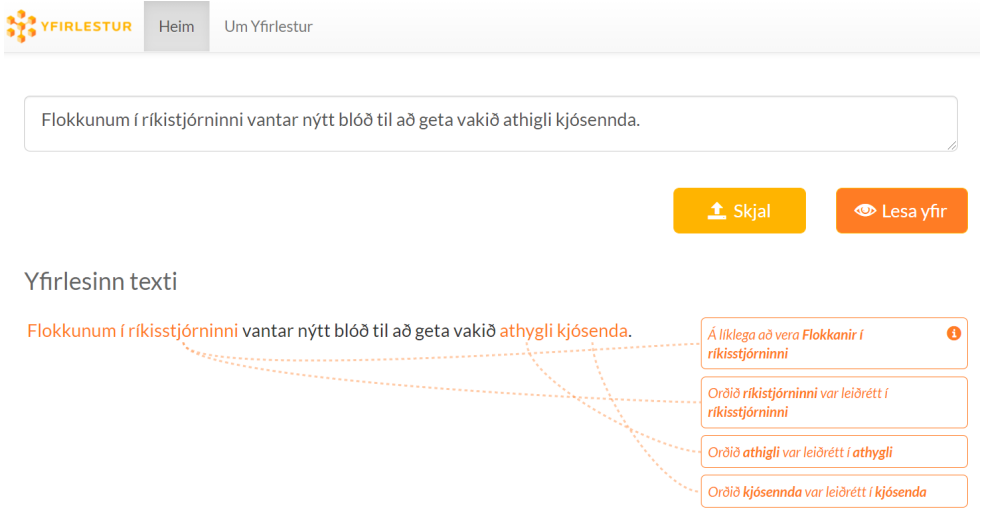
Verkalistinn fyrir M6 nefnir skoðun á BERT-líkum tauganetum fyrir stafsetningartillögur. Í tilraunum okkar var BERT-líkan fyrir íslensku þjálfað á verkefni sem felst í því að fylla inn orð sem vantar úr setningu (e. masking task). Þetta líkan getur lagt til rétta stafsetningu en var ekki innleitt í vefsíðuna yfirlestur.is þar sem það er (1) nokkuð þungt í keyrslu hvað varðar örgjörva- og minnisþarfir; og (2) líkanið notaði orðflísatóka en við teljum að betri árangri megi ná með notkun kóðunar á stafa- eða bætastigi sem verður skoðað á þriðja ári. Þessar tilraunir voru því settar í hlé til þriðja árs, þegar við munum beita gagnabreytingaraðferðum til að búa til gervivillumálheildir fyrir þjálfun tauganetslíkana fyrir villugreiningar á stafa- eða bætastigi.

2.10 Nálægasta gull (e. closest gold)
Til að meta árangur tólsins án viðmiðs var úttak 100 setninga úr þróunarsettinu handyfirfarnar samkvæmt nálægustu gullmæliaðferð (e. closest gold metric)[34].

Þó að úrtakið sé smátt bendir það til þess að tólið nái betri árangri en sjálfvirku mælingarnar gefa til kynna, þar sem stafsetningarvillur ná F0.5-gildi upp á 86,48% og málfræðivillur 51,47%. 14% setninganna voru óþáttaðar og þessar tölur fást þegar engar villur eru gefnar fyrir óþáttaðar setningar.

3. Vefviðmót

Vefviðmótið er komið upp og er í notkun á https://yfirlestur.is og er reglulega uppfært með nýjustu hugbúnaðaruppfærslum. Vefsíðan tekur við algengustu gerðum textaskráa. Eins og er gefur hún úttakið upp á vefsíðunni í staðinn fyrir í skrá. Þegar verkþáttur L9 verður tilbúinn munum við uppfæra þessa virkni fyrir algengustu gerðir skráa.



Skjáskot af `yfirlestur.is` sem sýnir innsleginn texta notanda merktan með stafsetningar- og málfræðivillum og tillögum.

[34] https://aclanthology.org/2021.eacl-main.231

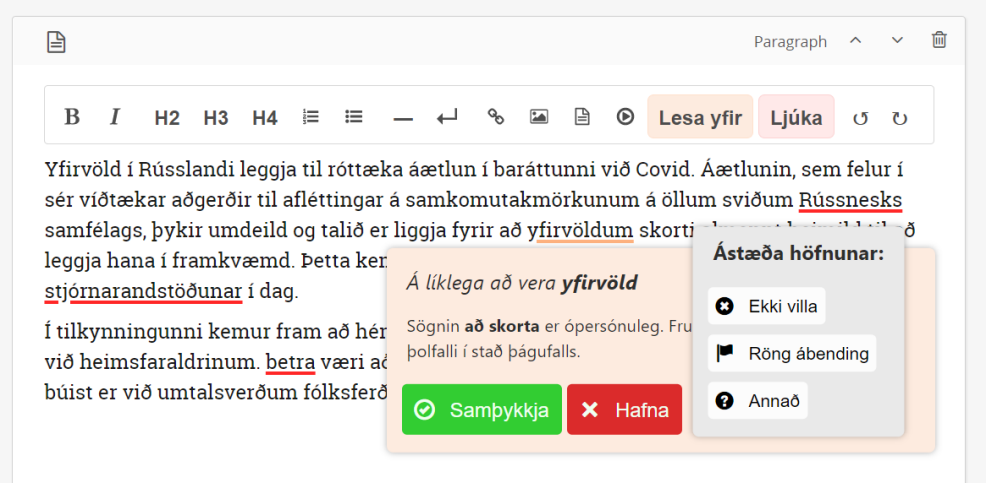
4. Mat notenda Kjarnans

Undirpakki innan L7 er notkunardæmi í samstarfi við veffréttamiðilinn Kjarnann (https://kjarninn.is/). Markmið verkþáttarins er að innleiða málrýnitólið í ritstjórnar- og vinnsluferli fréttasíðu. Við öðlumst reynslu með notkunardæmi í raunheimi, bæði hvað varðar tæknileg atriði og hvað varðar raunverulegar þarfir og forgangsröðun í krefjandi umhverfi fyrir málrýni.

4.1 Þróun
Merkingarviðbótin fyrir ritstjórnarumhverfi Kjarnans virkar með samskiptum við fyrirliggjandi forritaskil Yfirlestur.is, sem keyra málrýni GreynirCorrect, sem er viðhaldið af Miðeind. Grunnvirkni viðbótarinnar er að senda innihald fréttagreinar sem verið er að skrifa í forritaskil Yfirlestur.is fyrir greiningu/leiðréttingu villna, merking innihaldsins með tillögum samkvæmt svari frá forritaskilum og að veita notendaviðmót til að samþykkja eða hafna tillögunum. Auk grunnvirkni fyrir samþykki/höfnun tillaga tekur viðmótið á upplýsingum endurgjafar, sem notaðar eru til að safna gögnum fyrir sjálfar notendaprófanirnar.

Ritstjórnarumhverfi Kjarnans notar Wagtail-efnisstjórnunarkerfið (e. CMS), vinsælt umhverfi sem notar Django-rammann fyrir heildarvirkni. Til að geta beitt mikilli textameðhöndlun notar Wagtail Draftail-textaritilinn, sem er afbrigði DraftJS-textaritilsins sem þróaður er af Facebook. Öll þróun málrýniviðbótarinnar fyrir ritstjórnarumhverfi Kjarnans þurfti því að samræmast þessum ólíku römmum og forritaskilum.

Það er engin almennt aðgengileg útfærsla á stafsetningar- og málfræðimerkingaviðmóti í boði fyrir Draftail-textaritilinn, svo við byggðum á tiltækum textaramma ritilsins til að birta merkingarviðmótið. Undirliggjandi textaeiginleikar DraftJS í Wagtail-ritlinum gefa möguleika á nauðsynlegri efnismeðferð til að skipta út villum og leggja áherslu á þær. Viðmótið sjálft birtist í gegnum React-rammann fyrir Javascript.



Skjáskot úr ritstjórnarumhverfi Kjarnans, sem sýnir Wagtail-ritilinn, merkingar í röð og viðmót merkinga og endurgjafar, með gögnum úr forritaskilum Yfirlestur.is

Fyrir betri samþættingu í Draftail-textaritilinn var GreynirCorrect (og forritaskil Yfirlestur.is) endurraðað til að skila bókstafaspönnum greindra villna, í stað þess að skila aðeins spönnum viðeigandi tóka. Þessi breyting á eiginleikum forritaskila Yfirlestur.is mun gera framtíðarútfærslur forritaskilanna auðveldari.

Það er raunverulegur möguleiki á að þessi nýja viðbót muni nýtast í öðrum íslenskum tilfellum fyrir Wagtail-efnisstjórnunarkerfið. Wagtail er þegar notað af mörgum vel kunnum fjölmiðlasamtökum á Íslandi, auk fjölda ríkisstofnana, og eftirspurn eftir íslenskri stafsetningar- og málfræðileiðréttingu í fremstu röð er mikil á þessum sviðum. Wagtail-efnisstjórnunarkerfið var valið af þessari ástæðu og hönnuðir Draftail-ritilsins hafa síðan sýnt verkefninu áhuga.

4.3 Endurgjöf (M5/M6)
Söfnun á endurgjöf notendaprófana með Kjarnanum var gerð með sérstöku kalli í forritaskil Yfirlestur.is. Þegar notandi skoðar merkingu er gögnum um viðeigandi villu safnað og þau send í forritaskil Yfirlestur.is og skráð í gagnagrunni innanhúss til skoðunar hjá Miðeind.

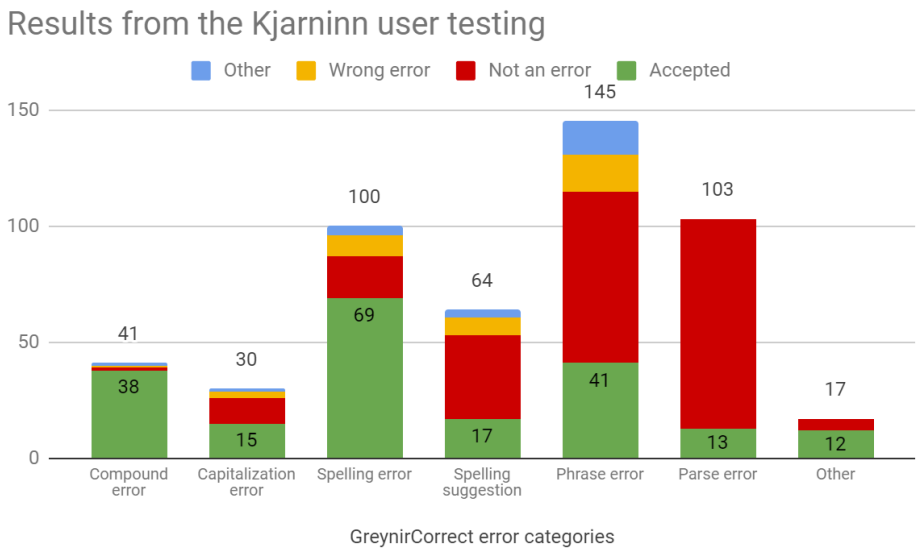
Ramminn fyrir gögn sem er safnað var takmarkaður við eftirfarandi 6 atriði fyrir hverja villu:

1. Villukóða GreynirCorrect
2. Orðaspönn villunnar
3. Upprunalegan inntakstexta
4. Texta leiðréttingar/tillögu
5. Heil setning sem inniheldur villuna
6. Endurgjöf notenda
    1. Leiðrétting var samþykkt
    2. Leiðréttingu var hafnað. Ástæða:
        1. Upprunalegur texti er réttur
        2. Upprunalegur texti er rangur en leiðrétting/tillaga er röng
        3. Annað

Atriði 1–5 samanstanda af samhengisupplýsingum um villuna sem um ræðir. Atriði 6 er raunveruleg endurgjöf frá notanda. Ef leiðréttingu/tillögu er hafnað er notandinn beðinn um frekara val, þar sem hann greinir frá ástæðu höfnunar.

4.4 Mat og niðurstöður (M6)
Gagnaöflunaráfangi notendaprófana fór fram milli 19. júlí og 27. september 2021. Vegna samhliða uppfærslu GreynirCorrect er samsetning endurgjafargagnanna örlítið mismunandi milli fyrri og seinni hluta notendaprófana. Til hægðarauka vísum við til fyrri hluta tímabilsins sem *alfa-prófunarfasa* og síðari helmings sem *beta-prófunarfasa* þegar þörf krefur.

Samtals var 500 tillögum um leiðréttingar safnað á meðan á notendaprófunum stóð. Hlutfall samþykktra og hafnaðra leiðréttinga kemur fram í súluritinu að neðan, skipt eftir villuflokkum GreynirCorrect. Grænu súlurnar tákna rétt jákvæðar (e. true positive, samþykktar leiðréttingar), rauðu súlurnar er rangt jákvæðar (e. false positive, leiðréttingum sem var hafnið), gulu eru rangt greindar villur og bláu eru hafnanir af öðrum ótilgreindum ástæðum.

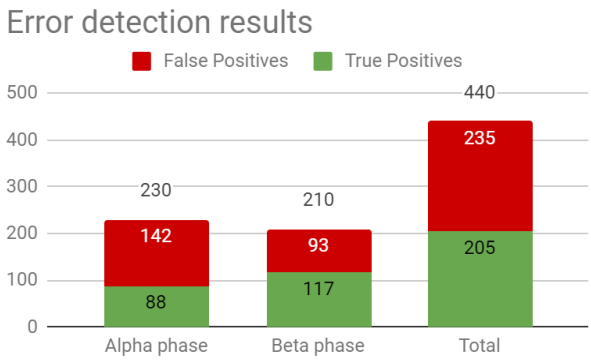


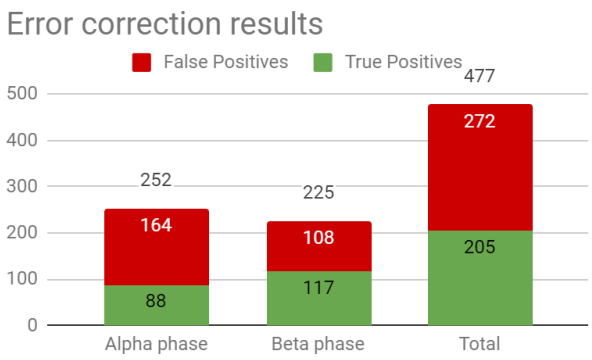
Flestir villuflokkar GreynirCorrect komu fram í söfnuðum endurgjöfum, en nokkrir komu aðeins fyrir örfáum sinnum og þeim er hópað saman sem *other* í töflunni að ofan. Þessir flokkar eru greinarmerkjavillur (e. punctuation errors), skammstafanavillur (e. abbreviation errors) og villur tengdar óþekktum orðum (e. unknown word errors). Flokkurinn með bannorðaviðvörunum kom aldrei fyrir í söfnuðum endurgjöfum.

Eins og rætt var í kafla 2.1 tekur GreynirCorrect eins og er betur á stafsetningarvillum en málfræðivillum. Þetta endurspeglast í endurgjöfum notenda, þar sem samsett orð (e. compound), hástafaritun (e. capitalization), og stafsetningarvillur (e. spelling errors) eru í flestum tilfellum samþykkt. Stafsetningartillögur (e. spelling suggestions) eru tilvik þar sem við erum ekki alveg örugg með leiðréttingargildi okkar, svo við merkjum þær aðeins sem villur en reynum ekki að leiðrétta þær, svo það kemur ekki á óvart að þar náum við verri árangri. Liðavillur (e. *phrase errors),* sem ná yfir ýmsar málfræðivillur, er oftast hafnað. Hins vegar var það í minnst 25 þessara tilfella vegna of gráðugrar reglu um liðavillu í hugbúnaði í alfa-prófunum, sem var bætt í miðjum notendaprófunum.

Flokkurinn þáttunarvilla (e. parse error) táknar tilfelli þar sem inntakssetningu var ekki hægt að þátta samkvæmt samhengisfrjálsu málfræðinni, jafnvel með villureglum. Í þessum tilfellum var notandinn beðinn um að meta hvort þetta gæti verið vegna stafsetningar-/málfræðivillna eða ekki. Samkvæmt endurgjöfinni eru aðeins nokkur tilfelli sem orsakast af raunverulegum villum í setningunni; í öðrum var þáttunarvilla vegna einhvers sem vantaði í samhengisfrjálsu málfræðina. Taka skal fram að fréttatextar innihalda stundum næstum orðréttar tilvitnanir úr tali sem fylgja ekki málfræðireglum.

Ekki er mögulegt að reikna út F-gildi fyrir notendaprófunarfasa. Þess í stað greinum við aðeins rétt og röng jákvætt (e. true positives, false positives) úr söfnuðu gögnunum. Rangt jákvætt er skilgreint sem aðeins *hafnaðar leiðréttingar* við mælingu á villuábendingum og bæði *hafnaðar leiðréttingar* og *rangar villuábendingar* við mælingu á villuleiðréttingum. Þetta er sýnt í súluritunum að neðan.





Í heildina skilaði GreynirCorrect röngum niðurstöðum fyrir meira en helming endurgjafargagnanna. En GreynirCorrect náði þó betri árangri í heildina í beta-prófunum en í alfa-prófunum. Við þessu mátti búast vegna fyrrnefndrar liðareglu sem var of gráðug. Þar sem þáttunarvillum virðist alltaf hafa verið hafnað í endurgjöf úr alfa-prófunum hefur það líka áhrif á þetta. Sem slíkar túlkum við endurgjöf úr beta-prófunum sem raunverulega virkni hugbúnaðarins í ritstjórnarumhverfi Kjarnans.

Söfnun og greining endurgjafar notenda mun halda áfram á þriðja ári og aðferðir til að fækka röngum jákvæðum (e. false positives) verða rannsakaðar.

L9 – Málrýni í ritvinnslukerfum

Skýrsla: Háskóli Íslands
Aðilar: Háskóli Íslands

Markmið

Samþætting á íslenskum málrýni í algengum forritum eins og Microsoft Office, Google Docs, Adobe InDesign o.s.frv., auk Apple MacOS. Verkþátturinn sem skilgreindur er fyrir annað ár er sambærilegur L8, undirbúningsvinna fyrir samþættingarvinnu sem gerð verður á þriðja ári.

Varða 6 – lýsing

M5
Þarfir fyrir MS Office, MacOS og Google Docs skilgreindar

M6
Vegvísir fyrir samþættingu málrýnis í MS Office, MacOS og Google Docs tilbúinn.

Afurðir

Öllum markmiðum vörðunnar var náð. Þörfum og vegvísi er lýst að neðan.

Verkþáttur L9 heldur áfram á þriðja ári.

Skýrsla

1. Yfirlit

Samþætting GreynirCorrect, málrýnisins sem þróaður er í L7, inn í ritvinnsluhugbúnað er almennt gerð í hverju tilviki hugbúnaðar fyrir sig. Samþætting getur farið eftir kerfinu, með ólíkum þörfum t.d. forrits sem keyrir á MacOS og sama forrits sem keyrir á Windows. Vegna þessa, til að ná jafnvægi milli dekkunar og aðgengi notenda, höfum við takmarkað samþættingu GreynirCorrect við tvö algengustu ritvinnsluumhverfin þvert á kerfi; Google Docs og MS Office-svítuna.

Helsti kostur þess að einblína á Google Docs og MS Office er útbreiðsla þessara tveggja umhverfa í daglegri notkun. Samþætting hér mun tryggja tafarlaust aðgengi mikils fjölda fólks að GreynirCorrrect. Auk þess gefa bæði umhverfin kost á aðgengilegum römmum til að þróa viðbætur eiginleika, í gegnum Google Apps Script og með samspili við Office Javascript-forritaskilin, í þessari röð. Loks eru þessir valmöguleikar ekki háðir stýrikerfum. MS Office-svítan veitir möguleika á þróun viðbóta sem virka þvert á kerfi, óháð því hvort forritið er keyrt á Windows eða MacOS. Hið sama gildir fyrir Google Docs, sem virkar óháð stýrikerfi þar sem það er veflægur hugbúnaður.

Verkþátturinn mun byrja með áherslu á þróun fyrir Google Docs, og MS Office eftir það. Innan MS Office-svítunnar mun megináherslan vera á MS Word, og umfang samþættingar við annan Office-hugbúnað (Powerpoint, Outlook, o.s.frv.) mun skýrast við þróun.

1.1 Þróunarvettvangur: Google
Google veitir opinn þróunarvettvang fyrir Google Docs-viðbætur í formi vefsamþætts þróunarumhverfis (e. IDE) fyrir Google Apps Script. Google veitir einnig ítarlega skjölun á þróunaraðferðum og leiðbeiningum fyrir útgáfu viðbóta fyrir almenning.

1.2 Þróunarvettvangur: MS Office
Sá vettvangur sem Microsoft mælir með til þróunar viðbóta fyrir MS Office er Node.js, sem býr til viðbót í formi vefapps sem er keyrt af MS Office-forriti. Microsoft veitir Yeoman-gjafa (e. generator) til að meðhöndla Node.js-verkefni sem ætluð eru fyrir slíka innleiðingu (sjá: https://github.com/OfficeDev/generator-office), með samþættingu við flest þróunarumhverfi (e. IDE).

1.3 Tímalína
Í samræmi við áætlaðar vörður þriðja árs er bráðabirgðatímalínan fyrir verkþátt L9 eftirfarandi.

Þróun samþættingar í Google Docs verður fyrsta áherslan, með bráðabirgðaútgáfu tilbúna við M7. Samhliða þróun eru smáatriði notendaprófunarfasa mótuð, og allar verklagsreglur tilbúnar við M7.

Í kjölfar M7 hefst þróun á samþættingu í MS Office, þar sem MS Word verður fyrsta áherslan. Virkni þvert á stýrikerfi milli Windows og MacOS verður tryggð. Samhliða þessu verða notendaprófanir á Google Docs-viðbótinni framkvæmdar. Eftir því sem vinnu við MS Office-samþættinguna miðar áfram verða kerfisbundnar prófanir á viðbótinni gerðar innanhúss. Við M8 mun notendaprófunum Google vera lokið, og bráðabirgðaútgáfa MS Office-viðbótarinnar tilbúin.

Í kjölfar M8 verður Google Docs-viðbótin endurbætt með niðurstöðum úr notendaprófunarfasa og sömu niðurstöður notaðar þar sem á við fyrir MS Office. Við M9 munu bæði MS Office- (þvert á stýrikerfi) og Google Docs-viðbæturnar vera tilbúnar og útgefnar á CLARIN.

2. Virkni
Samþætting GreynirCorrect verður aðallega veflæg, og mun virka í gegnum stöðluð samskipti við forritaskil Yfirlestur.is, sem keyrir GreynirCorrect (https://yfirlestur.is). Við búumst við einhverri ónettengdri virkni frá GreynirCorrect. Upplýsingar um mögulega ónettengda virkni verða skýrari eftir því sem þróun miðar áfram.

Það er hugsanlegur flöskuháls í virkni samþætta málrýnitólsins þegar kemur að notendaumferð og vinnslugetu. Forritaskil Yfirlestur.is eru eins og er sniðin að notkun frá vefsíðunni Yfirlestur.is og lokuðum verkefnum, t.d. undirverkefni notendaprófana Kjarnans í L7, ásamt notkun innanhúss. Ef notkun forritaskilanna verður útbreidd og tíð af almenningi í gegnum t.d. viðbót fyrir hugbúnað, verður vinnslugeta að vandamáli, sem þarfnast þá nýstárlegra lausna.

3. Undirbúningsvinna fyrir Kjarnann

Fyrrnefnt undirverkefni L7, notendaprófanir Kjarnans, hafði þegar leitt í ljós margar þarfir fyrir innleiðingu forritaskila Yfirlestur.is í hugbúnað þriðju aðila, og hafa forritaskilin nú verið bætt fyrir almenna notkun. Undirverkefnið nýtist einnig sem prófunarumhverfi fyrir almenna þróun viðmóts sem virkar sérstaklega með forritaskilum Yfirlestur.is. Endurgjöf notenda úr undirverkefninu mun bæði bæta forritaskilin enn frekar og vera nytsamlegt í þróunarferli fyrir samþættingu í ritvinnsluforriti.

4. Prófanir notenda

Þróun viðbótar sem ætluð er fyrir almenning þarfnast notendprófunarfasa. Þar sem fyrsta áhersla L9-verkþáttarins verður Google Docs verða notendaprófanir gerðar sérstaklega á þeirri samþættingu. Endurgjöf úr notendaprófunarfasa mun svo leggja sitt af mörkum fyrir bæði sjálfa viðbót Google Docs, og MS Office-samþættinguna, sem á þeim tímapunkti verður í þróun og prófunum innanhúss.

Umfang og verklag notendaprófana mun fylgja þeim reglum sem beitt er í notendaprófunum Kjarnans á forritaskilum Yfirlestur.is fyrir stafsetningar- og málfræðileiðréttingu. Þessi hópur notenda verður líklega fenginn frá Háskóla Íslands í gegnum staðlað ferli. Smáatriði notendaprófunarfasa verða skilgreind síðar innan verkefnisins. Ólíkt notendaprófunum Kjarnans í L7 verður áhersla notendaprófana á virkni viðbótarinnar sjálfrar, en ekki forritaskila Yfirlestur.is.

L10 – Aðlögun að máltæknihugbúnaði

Skýrsla: Grammatek
Aðilar: Grammatek

Markmið

Að hafa útgáfu málrýnisins sem má nota sem einingu (e. module) innan stærri máltæknihugbúnaðar. Einingin verður að bjóða upp á sveigjanlegar stillingar til að koma til móts við þarfir mismunandi hugbúnaðar. Við munum innleiða útgáfu málrýnis til að samþætta í talgervil, en á sama tíma kanna þarfir annars hugbúnaðar sem gæti notið góðs af sjálfvirkum leiðréttingum stafsetningarvillna, t.d. vélþýðingar og leitar.

Varða 6 – lýsing

Nauðsynlegar breytingar á grunnvirkni útfærðar, einingu pakkað fyrir talgervingu, prófunartilvik tilbúin. Niðurstöður MOS-prófana[35] á talgervilskerfi með og án málrýnieiningar tilbúnar. Málrýnieining fyrir talgervil, viðmið fyrir aðlögun einingarinnar fyrir aðra notkun tilbúin og gefin út á CLARIN.

Afurðir

- Talgervilsframendaþjónusta sem inniheldur málrýni
  https://github.com/grammatek/tts-frontend-service

Eins og lýst var í endurskoðuðum vörðum í sambandi við talgervil Android (T5) verður þessum verkþætti lokið með seinkun. Ber helst að nefna að MOS-prófanir verða gerðar í samvinnu við T8, almennt ferli mats fyrir talgervla. Öllum afurðum vörðunnar verður skilað fyrir 31. desember.

Verkþætti L10 lýkur við M6.

Skýrsla

Forvinnsla texta fyrir talgervla þarf að undirbúa hráan inntakstexta fyrir breytingu yfir í hljóðritun og svo skila hljóðrænni framsetningu inntaks fyrir talgervilinn. Í undirbúningi fyrir hljóðritun er textanormun mest áberandi einingin en málrýni getur einnig verið mikilvægt skref fyrir betri upplifun notenda.

Forvinnslupípunni er lýst nánar í T5, hér lýsum við hlutverki málrýnis í þeirri pípu.

[35] Mean-Opinion-Score: mælikvarði á heildargæðum samkvæmt mannlegu mati, með einkunnum frá 1 (slæm) upp í 5 (frábær)

Almennt séð gegnir málrýnir því hlutverki að aðstoða fólk við að skrifa betri texta: leiðrétta innsláttarvillur, greina stafsetningar- og málfræðivillur og benda á það sem mætti orða betur, til dæmis. Máltæknihugbúnaður hefur að hluta til aðrar þarfir; til dæmis hafa leitarvélar ávinning af því að inntak og lyklar innihaldi sem fæstar innsláttar- og stafsetningarvillur. Talgervilskerfi hafa í raun sérstakt samband við stafsetningarvillur: margar stafsetningarvillur eru vegna þess að fólk skrifar það sem það heyrir, þ.e. skrifar eftir framburði. Þessar villur hafa ekki áhrif á talgervilskerfi svo lengi sem hægt er að búa til réttan framburð með g2p-einingunni. Í mannlegu mati geta „slæmar“ setningar því hljómað fullkomlega skiljanlegar þegar þær lesast af talgervli, en aðrar minniháttar stafsetningarvillur geta haft mikið meiri áhrif:

hvar hefur það verið sannað með rannsóknum að það sé í lagi að neita fíkkniefna, ef þú villt neita þeirra þá er það þitt val en ef þú hefur lesið alla þá pisla og greynar um notkun á eiturlyfjum þá myndir þú skilja það sem ég er að meina
Dæmi 1: Engar þessara villna hafa áhrif á úttak talgervilskerfis

Þar verður látið reyna á samstarf vinstra megin við mijðu
Dæmi 2: Einföld stafavíxlun velur óskiljanlegum framburði

Eins og nefnt var að ofan er textanormun mikilvægasta eining forvinnslu texta fyrir talgervla. Þetta felur í sér að breyta tölum og styttingum í réttar myndir og þar sem það mistekst gæti málrýnir hjálpað við leiðréttingu þeirra mynda:

Áfangastaðir WOW air eru nú 31 talsins, 23 innan Evrópu en 8 talsins í Norður Ameríku.
Áfangastaðir wow air eru nú þrjátíu og eins talsins, tuttugu og þrjú innan Evrópu en átta talsins í Norður Ameríku. [eins->einn; þrjú->þrír]
Dæmi 3: Fyrsta setningin er hrátt inntak í talgervilskerfi, önnur er úttak úr normun. Textanormarinn getur ekki greint að tölurnar séu háðar frumlaginu „áfangastaðir“ og notar hvorugkyn í stað karlkyns.

Við höfum útfært tengingu málrýnisins í framendapípu talgervils og það virðist taka vel á stafsetningarvillum eins og þeim í dæmi 2. Við þurfum að skoða hvort bæta má sjálfvirka leiðréttingu villna falls/kyns/talna þegar greining hæðis mistekst. Fyrir MOS-prófanir munum við leiðrétta handvirkt þær villur sem málrýnir tekur ekki enn á til að fá samanburð á úttaki talgervils á leiðréttum textum gegnt textum með stafsetningarvillum.

H1 – Upptökur með Samrómi

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að eiga stórt safn upptaka af stuttum segðum til að þjálfa talgreina. Segðirnar verða 3-7 sekúndur hver, eða um 2-12 orð (eins mörg og er þægilegt að koma fyrir á skjá snjallsíma) og koma frá ýmsum hópum mælenda: mismunandi aldur, mállýskur o.s.frv.

Varða 6 – lýsing

M5:
• 200.000 segðir unglinga teknar upp
• 30.000 segðir annars máls mælenda (e. L2 speakers) teknar upp
• 20.000 segðir teknar upp af fullorðnum með því að endurtaka talaðar segðir
• 20.000 segðir teknar upp af börnum með því að endurtaka talaðar segðir
M6:
• 50.000 segðir annars máls mælenda (e. L2 speakers) teknar upp
• 30.000 segðir teknar upp af fullorðnum með því að endurtaka talaðar segðir
• 30.000 segðir teknar upp af börnum með því að endurtaka talaðar segðir

Afurðir

Fimm af sjö markmiðum vörðunnar náðust.

- 200.000 segðir unglinga teknar upp
- 30.000 segðir annars máls mælenda (e. L2 speakers) teknar upp
- 50.000 segðir annars máls mælenda (e. L2 speakers) teknar upp
- 20.000 segðir teknar upp af fullorðnum með því að endurtaka talaðar segðir
- 20.000 segðir teknar upp af börnum með því að endurtaka talaðar segðir

Lýðvirkjunarkeppni sem haldin var í grunnskólum, *Grunnskólakeppni,* heppnaðist gríðarlega vel og náði þessum vörðum næstum ein og sér.

Fyrstu 100.000 fullgildu segðirnar, Samromur 21.05, hafa verið gefnar út á openSLR, https://openslr.org/112/.

Hins vegar náðust ekki vörður M6 fyrir endurteknar talaðar segðir.

- 30.000 segðir teknar upp af fullorðnum með því að endurtaka talaðar segðir
- 30.000 segðir teknar upp af börnum með því að endurtaka talaðar segðir

Hugbúnaðarútfærsla söfnunar tók lengri tíma en búist var við og var ekki tilbúin fyrr en rétt fyrir M6. Við stefnum á að ljúka M6 markmiðum 3. janúar 2022. Útgáfa verður svo á þriðja ári þar sem þessi verkþáttur heldur áfram.

Aðgerðaáætlun fyrir eftirstandandi söfn Herma
Þar sem markmið þessarar söfnunar var að ná til yngri barna yfir sumarið höfðum við samband við Íþrótta- og tómstundasvið Reykjavíkur (ÍTR) varðandi hugsanlegt samstarf þar sem það heldur námskeið fyrir börn yfir sumarmánuðina og námskeið eftir skóla, og Skema sem heldur námskeið fyrir börn. Langan tíma tók að fá svör frá ÍTR og áður en við gátum gert nokkuð með þeim skall á önnur bylgja COVID og aðgengi að skólum og námskeiðum var takmarkaður. Hins vegar náðum við að vinna með Skemu síðasta sumar með góðum árangri og við ætlum að vinna með þeim aftur í nóvember í vetrarfríinu.

Við ætlum einnig að auka vitund á þessari gerð framlags og beina sjónum að ólæsum börnum, blindum, sjónskertum eða fólki með lestrarvandamál. Jafnvel að slökkva á öllum öðrum gerðum framlags þar til þessum vörðum er náð.

Til að ná markmiðinu á næstu þremur mánuðum munum við hafa samband við skóla þar sem þeir eru byrjaðir, biðja Skema um hjálp yfir vetrarfríið, og auglýsa á samfélagsmiðlum Samróms.

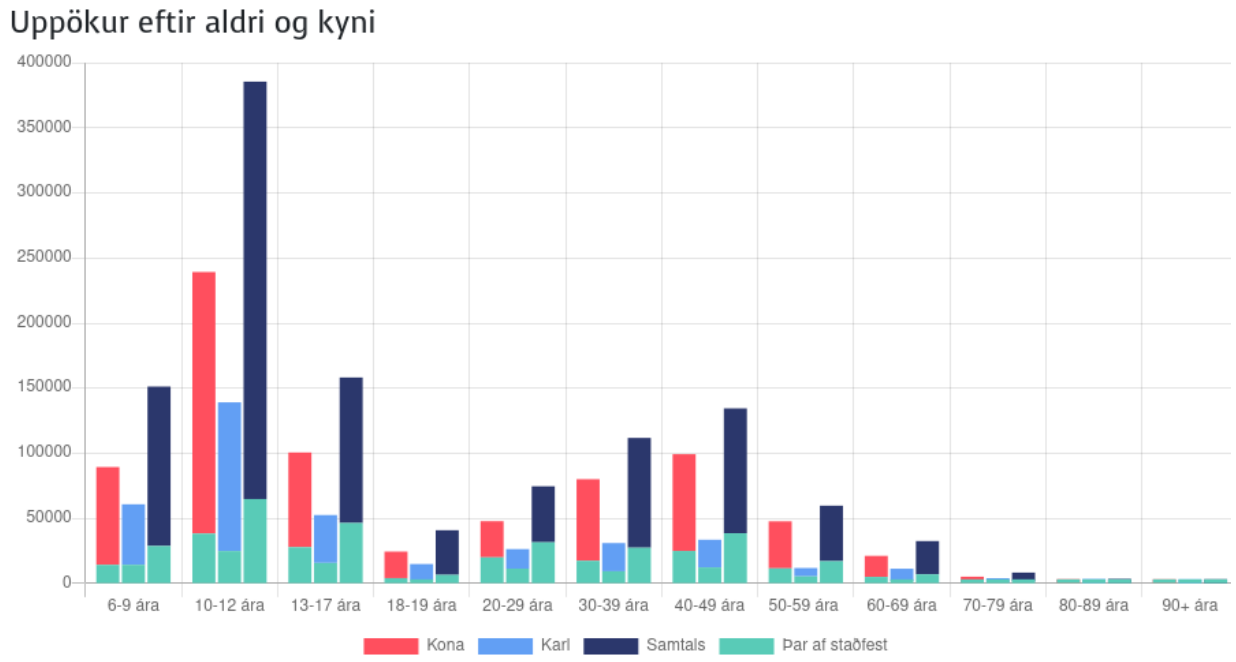
Skýrsla

Í þessum vörðuáfanga höfum við bætt um 500.000 staðfestingaratkvæðum í söfnunarumhverfið, um 193.000 eru handunnin atkvæði frá ráðnu starfsfólki sem veldur 132.000 nýjum staðfestum segðum. Einn gagnasettspakki hefur verið gefinn út og annar er á leiðinni. Vinna hefur verið lögð í að safna segðum endurtekins tals.

Megináherslan hefur verið á yfirferð tiltækra gagna og söfnun segða með endurteknu hljóði. Fyrir yfirferð hafa þrjár aðferðir verið notaðar: lýðvirkjun, ráðið yfirferðarfólk og sjálfvirk yfirferð með tóli sem kallast Marosijo. Fyrir ráðið yfirferðarfólk var umhverfið uppfært til að leyfa auðveldari notkun með flýtivísum (e. keyboard shortcuts). Sjálfvirka yfirferðin mun styðja við yfirferð lýðvirkjunar. Fyrir endurtekið hljóð unnum við með sumarnámskeiðum til að auglýsa og safna segðum frá börnum. Fyrir fullorðna réðum við sumarnemendur bæði fyrir upptökur og til að auglýsa.

Gögn (tölfræði, söfnunarferli, gerðir)

Almenn tölfræði gagna
Til að sjá gögn í rauntíma má alltaf skoða vefsíðu okkar. Hér kemur stutt yfirlit núverandi gagna. Í heildargögnum eru 1.166.895 segðir sem safnað hefur verið í Samróm, sem gera um 1.700 klst. og af þeim er búið að yfirfara og samþykkja næstum 400 klst. Enn þarf að yfirfara mikið magn gagna, um 750.000 segðir. Af yfirförnum gögnum eru um 64% samþykktar segðir. Í súluritinu að neðan má sjá dreifingu aldurs og kyns.



Dreifing aldurs og kyns. Rautt: konur, blátt: karlar, dökkblátt: samtals og grænt: yfirfarið

Eins og sjá má á súluritinu að ofan eru áhrifin af vel heppnuðu Grunnskólakeppninni í janúar greinileg, og dregur aldursdreifinguna nær börnum. Tíðni kvenna í kynjadreifingu er einnig áberandi þar sem erfiðara hefur verið að fá karla til þátttöku. Fyrir annars máls mælendur var 20.000 segðum safnað frá fullorðnum og 50.000 frá börnum.

Segðir teknar upp með endurtekningu talaðs hljóðs
Fyrir söfnun segða með endurtekningu talaðs hljóðs sömdum við við Skema. Þau aðstoðuðu okkur við að fá börn til að taka upp endurtekið hljóð. Nemendur í sumarstarfi tóku einnig upp endurtekið hljóð og auglýstu söfnunina. Það eru til um 20.000 segðir fyrir börn og 21.000 fyrir fullorðna í gagnagrunninum.

Ráðið yfirferðarfólk
Til að auka magn yfirfarinna gagna var ráðið yfirferðarfólk og þau þjálfuð í því hvað gerir segð gilda eða ógilda. Vanalega þarf segð tvö jákvæð eða tvö neikvæð atkvæði til að teljast gild eða ógild, en þar sem þau voru þjálfuð fengu þau ofuratkvæði, þ.e. aðeins eitt atkvæði frá þeim þarf til að dæma segð gilda eða ógilda. Einkunnir Marosijo voru notaðar til að velja pakka sem yfirferðarfólk lagði áherslu á. Til að auka afköst yfirferðarfólks enn frekar notuðu þau flýtivísa (e. keyboard shortcuts). Alls yfirfóru þau 193.000 klippur.

Uppsetning

Við byrjun M5 höfðum við u.þ.b. 1,1 milljón segða og ljóst var að einhvers konar sjálfvirk leið til yfirferðar var nauðsynleg. Tól sem kallast Marosijo var notað til að yfirfara segðirnar og reikna út einkunn frá 0,0 upp í 1,0 fyrir hverja og eina. Því lægri sem einkunnin er, því líklegra er að upptakan verði slæm. Því hærri, því líklegri að hún verði góð.

Segðir með háa einkunn voru margar rangt jákvæðar (e. false positives, háar einkunnir þrátt fyrir að vera ekki góðar segðir), eða um 20%. Þessar segðir voru að mestu góðar en í þær vantaði oft byrjun eða endi hljóðsins. Oft voru þær líka bornar fram vitlaust. Segðir með lága einkunn voru einnig með svipaða prósentu af rangt neikvæðum (e. false negatives, lágar einkunnir þrátt fyrir að vera góðar). Þessar segðir voru góðar en náðu stundum lágri einkunn vegna hávaða í bakgrunni (sem er í lagi fyrir talgreinisþjálfun). Af þessari ástæðu var reynt að fækka fjölda rangt jákvæðra og neikvæðra með því að bæta Marosijo-líkanið sjálft og með notkun Montreal Forced Aligner-tólsins.[36] Þetta bætti árangur ekki að neinu leyti. Lokaniðurstaðan var að halda upprunalegum niðurstöðum úr Marosijo og vera meðvituð um rangt jákvæða og neikvæða. Eftirfarandi reglur voru skilgreindar og atkvæðin búin til út frá þeim:

1. Segðir með einkunn 0,9 og hærri fá **jákvætt atkvæði**.
2. Segðir með einkunn á bilinu 0,01 til 0,3 fá **neikvætt atkvæði**.
3. Segðir með einkunn á bilinu 0,01 til 0,3 og er ekki hægt að samraða með MFA fá **neikvætt ofuratkvæði**.
4. Segðir með einkunn 0,0 fá **neikvætt ofuratkvæði**.
5. Segðir sem eru tómar fá **neikvætt ofuratkvæði**.
6. Segðir sem eru ekki innan ofantalinna einkunnasviða, en innan einkunnasviðsins 0,301 til 0,899, fá **ekkert atkvæði**.
7. Segðir sem hafa ekki verið yfirfarnar af Marosijo og eru ekki tómar fá **ekkert atkvæði**.

Fyrir venjuleg atkvæði þarf tvö til að samþykkja (eða hafna) upptöku að fullu. Fyrir ofuratkvæði þarf aðeins eitt. Vegna tilvistar rangt jákvæðra og neikvæðra var ákveðið að nota helst ekki ofuratkvæði á þau heldur aðallega venjuleg atkvæði. Ofuratkvæði voru aðeins gefin segðum sem uppfylltu reglur 3 - 5.

Uppfærsla söfnunarvettvangs
Bæta þurfti vinnuferli til að meðhöndla segðir teknar upp með endurteknu hljóði. Við köllum þennan verkferil *Herma*. Fyrir þetta þurfti meiriháttar endurbyggingu vettvangsins til að meðhöndla þessa þriðju gerð framlags. Enn fremur þurfti að búa til viðeigandi setningar og hljóðklippur til að þetta gæti virkað. Notað var tól búið til í HR til að finna setningar í Risamálheildinni og hljóðklippur búnar til með Amazon Polly.

Niðurstöður/Umræða

Sjálfvirk yfirferð
Atkvæðin sem fengust úr sjálfvirkri yfirferð sem lýst er að ofan voru:

	Fjöldi	Niðurstaða
Jákvæð atkvæði	435.550	**4.070 samþykkt**
Neikvæð atkvæði	15.386	**109 hafnað**
Neikvæð ofuratkvæði	60.363	**60.352 hafnað**

[36] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

Niðurstöður frá jákvæðum/neikvæðum venjulegum atkvæðum bornar saman við hversu mörg atkvæði voru gefin. Þetta er mikið lægra þar sem aðeins 4.070 segðir höfðu þegar eitt jákvætt atkvæði frá fólki þegar atkvæðin voru skráð. Vegna sjálfvirkrar yfirferðar þýðir þetta að 435.550 - 4.070 = 431.480 segðir hafa nú eitt jákvætt atkvæði og þurfa aðeins eitt atkvæði frá fólki til að teljast samþykktar. Handvirk yfirferð minnkar því um nákvæmlega helming. Hið sama gildir fyrir neikvæð venjuleg atkvæði.

Ítarlegri samantekt og umræðu um niðurstöður sjálfvirkrar yfirferðar má lesa í hirslunni samromur-tools.[37] Einnig voru Venn-myndir gerðar til að veita betri yfirsýn yfir tölfræðina.[38]

Að lokum

Þó að við fáum einhverjar yfirferðir í gegnum lýðvirkjun er uppistaða yfirfarinna gagna frá vinnu eins og frá þessu sumri. Atkvæðin 193.000 orsaka 128.000 samþykktar segðir tilbúnar til notkunar. Sjálfvirku atkvæðin 500.000 frá Marosijo-tólinu orsökuðu 4.000 samþykktar klippur tilbúnar til notkunar. Að því sögðu flýttu sjálfvirku atkvæðin frá Marosijo-tólinu vinnunni um helming fyrir áframhaldandi yfirferð úr lýðvirkjun og fjarlægðu margar ógildar segðir í gagnasafninu svo sú vinna sem fór í tólið borgaði sig vel. Einnig voru Marosijo-einkunnir notaðar til að velja stóran hluta af því sem yfirferðarfólki var gefið til að fara yfir. Tólin sem búin voru til fyrir það má nota aftur með reglulegu millibili til að bæta sjálfvirkum yfirferðaratkvæðum við nýjar segðir. Þó að markmiði vörðunnar fyrir M6 hvað varðar endurteknar segðir hafi ekki verið náð er kerfið nú komið upp og má nota til áframhaldandi söfnunar.

Fyrsta útgáfa Samrómsgagnanna er komin út. Önnur útgáfa sem inniheldur 131 klst. barnagagna (á aldrinum 4–17 ára) er í útgáfuferli á LDC.

Það eru nokkur möguleg næstu skref. Átak þarf í því að bæta gagnasöfnunarumhverfið fyrir notendur með skjálesara (helsta leiðin til að vafra á vefnum fyrir blinda eða sjónskerta). Til að safna fleiri endurteknum segðum þyrfti einnig að ráðast í markaðsherferð fyrir ólæs börn. Það eru miklir möguleikar í gagnasafni sem hægt er að pakka í komandi útgáfur eftir því sem við fáum fleiri yfirferðir. Að lokum ætti því að ráðast í frekara yfirferðarátak eins og gert var síðasta sumar.

[37] https://github.com/cadia-lvl/samromur-tools/blob/master/QualityCheckPostProcess/readme.md
[38] https://drive.google.com/drive/folders/11qU3vMLtYhNgd6WgOCgcl7HgHYgXfu97

H2 – Umritun efnis úr sjónvarpi og útvarpi

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

Markmið

Að eiga umritanir sjónvarps- og útvarpsefnis til að þjálfa talgreina fyrir fjölmiðla. Á fyrsta ári er markmiðið að umrita 100 klukkustundir af umræðuþáttum úr útvarpi og sjónvarpi, umrita eða framkvæma samröðun hljóðs-til-texta með forumrituðum gögnum frá textavarpssíðu 888 (fréttir), og umrita 25 klukkustundir af afþreyingarefni.

Varða 6 – lýsing

- 250 klukkustundir af útvarpsþáttum umritaðar
- 250 klukkustundir af samröðuðu útvarps- og sjónvarpsefni tilbúnar til notkunar í talgreini

Endanleg útkoma eru þá alls 500 klukkustundir af umrituðu og samröðuðu sjónvarps- og útvarpsefni.

Afurðir

Báðum markmiðum vörðunnar var náð, 250 klukkustundir af samröðuðu útvarps- og sjónvarpsefni tilbúnar til notkunar í talgreini, og 250 klukkustundir af útvarpsþáttum umritaðar. Gagnasettin verða gefin út á þriðja ári þar sem þessi verkþáttur heldur áfram. Samröðunarhirslan fyrir útvarps- og sjónvarpsefni er https://github.com/cadia-lvl/alignment-and-segmentation

Verkþáttur H2 heldur áfram á þriðja ári. Helstu verk: yfirferð, pökkun og útgáfa gagnasetta.

Skýrsla

Þegar sjálfvirkir talgreinar (e. ASRs) eru gerðir fyrir útsendingargögn notar fólk oft fundin gögn eins og gögn innan sviðs: umritanir útvarpsefnis, skjátexta, myndatexta, og texta textavélar. Þó að þessi gögn séu mikils virði þar sem þau eru þegar til eru þau takmarkandi. Texti úr textavél hefur engar tímamerkingar þar sem hann er færður handvirkt línu fyrir línu. Svo er vandamálið við skjátexta og myndatexta er að þeir eiga að passa á lítið svæði neðst á skjánum svo þeir eru oft styttir og umorðaðir, og endurtekningar fjarlægðar, sem gerir þá ekki samkvæma hljóðinu. Í öllum þessum fundnu gögnum er einnig það vandamál að erlend tungumál eru með. Þetta getur verið hljóð í erlendu máli og skjátexti á viðkomandi máli, í okkar tilfelli á íslensku.

Umritun útvarpsefnis

Ný útgáfa Tiro-ritilsins[39] til notkunar í gagnasöfnun Creditinfo var kynntur fyrir Creditinfo í febrúar 2021 til að fækka tímamerkingarvillum. Hins vegar eru enn einhver vandamál með tímamerkingar. Á endanum notuðu allir umritarar ritilinn, og sendu tilkynningar um villur og endurgjöf til hönnunarteymis Tiro. Frá því var öll umritun Creditinfo gerð með ritlinum. Næst var afhendingarferli búið til bæði fyrir efni til umritunar og fyrir gagnasafnið til HR. Umritanirnar sem fengust frá Creditinfo frá apríl 2021 og eftir það voru gerðar með talgreini Tiro og teljast hágæða. Þessar umritanir eru tilbúnar til útgáfu eins og þær eru.

Vinnsluferlinu má gróflega skipta í fjóra hluta:

1. Umritanirnar frá Tiro virðast vera nákvæmar og bútun þarf ekki að bæta. Aðeins þurfti að breyta sniði umritana á staðlað snið lýsigagna.
2. Umritanir Creditinfo af útvarpsþáttum áður en ritill Tiro var tekinn í notkun þarf að bæta. Fyrst og fremst þarf að breyta sniði þeirra yfir á sama staðlaða snið lýsigagna.
3. Hljóðskrárnar þarf svo að búta í segðir fyrir hvern einstakan mælanda. Þessar segðir þurfa að vera nægilega langar til notkunar í talgreini á Kaldi.
4. Að lokum höfum við einnig 900 klukkustundir frá RÚV sem þarf einnig að búa til sniðin lýsigögn fyrir og búta niður í segðir. Hins vegar eru umritanirnar ekki samkvæmar hljóðinu og hafa ekki mælendamerkingar svo endanlegt magn segða er minna en 900 klukkustundir.

Gögn

Við höfum tvær tegundir útsendingargagna, texta samkvæma hljóði (CreditInfo) og skjátexta (Kistan). Vegna ólíkra vandamála sem þessar tvær gerðir gagna hafa valdið notuðum við mismunandi aðferðir til að búa gögnin til fyrir talgreiningu.

CreditInfo

Creditinfo hefur skilað 470 klukkustundum hljóðs eins og er á Terra sem ættu samkvæmt tímamerkingum þeirra að vera 162 klukkustundir segða sem þarf að samraða. Þessi fjöldi klukkutíma er besta tilfelli þar sem þagnir eru í þessum segðum sem þarf að sía út. Að auki eru 74 klukkustundir umritaðs efnis á Tiro-ritlinum. Við höfum greint umritanirnar og komist að þeirri niðurstöðu að upplýsingarnar þarf að vinna aftur til að teljast hæfar í talgreinisþróun. Við höfum yfirfarið hluta af umrituðum gögnum á Tiro-ritlinum, sem virðast passa betur í hiklausa notkun en gögn Creditinfo sem voru afhent fyrr. Við fengum hjálp frá nemendum í sumarstarfi sem fóru í gegnum 29 hljóðskrár og löguðu þær. Nemendurnir gerðu nýjar tímamerkingar þar sem hver segð inniheldur aðeins einn mælanda. Samanlögð lengd þessara 29 skráa er 14,14 klukkustundir.

Eins og er höfum við 88,14 klukkustundir samraðaðra gagna í háum gæðum frá Creditinfo í gegnum forritaskil Tiro og úr vinnu sumarstarfsmanna, sem er nokkuð langt frá markmiðinu um 250 klukkustundir.

[39] tal.tiro.is/

Samröðun og bútun útvarpsefnis

162 klukkustundir efnis Creditinfo voru afhentar milli júlí 2020 og febrúar 2021. Þessar umritanir hafa í núverandi mynd alvarlega galla sem gera þær ónothæfar í þjálfun talgreina. Helstu gallarnir eru að tímamerkingarnar eru næstum alltaf rangar. Jafnvel ef þær væru lagaðar eru oft margir mælendur í hverri segð milli tímamerkinga. Jafnframt er ósamræmi í því hvernig þeir sem umrita greina milli mælenda í þessum segðum og eru þær upplýsingar hafðar með í umrituninni sjálfri. Upplýsingar mælanda þarf því að sía út sem við þurfum að gera sérstaklega fyrir hvert snið umritana.

Við höfum verið að vinna að hagræðingu samröðunarferlisins og að setja þessar umritanir á snið sem hægt er að nota í þjálfun talgreina. Þessu ferli má gróflega skipta í tvö undirferli. Það fyrsta er að setja umritanirnar á snið þar sem við höldum aðeins upplesnum texta og ekki upplýsingum mælanda eða öðrum upplýsingum sem eru ekki sagðar. Upplýsingar mælanda þarf að halda aðskildum frá segðunum sjálfum. Annað er að finna tímamerkingar fyrir þessar segðir í hljóðskránni.

Mikil hindrun var að endursníða umritanirnar, þar sem við þurfum að keyra mismunandi skriftur fyrir mismunandi gerðir sniða og við getum jafnvel ekki verið viss um að hafa séð allar gerðir. Þegar umritanirnar voru komnar á snið sem hæfir þjálfun talgreina, þ.e.a.s. aðeins umritunin sjálf, hófumst við handa við að finna út hvernig við gætum staðsett umritanirnar í hljóðskránni þar sem tímamerkingarnar í umritunum Creditinfo eru ónothæfar. Vænlegasta lausnin var að fá nýja sjálfvirka umritun úr talgreini Tiro sem getur gefið tímamerkingu fyrir hvert orð í umritun. Síðan fundum við hvar var líklegast að gamla umritunin myndi birtast í umritun. Fyrsta skrefið var að finna fyrsta orðið í gömlu umrituninni sem passaði fullkomlega við umritun Tiro og tímamerkin fyrir byrjun þess orðs haldið sem byrjun segðarinnar. Sama aðferð var notuð fyrir seinasta orðið og þannig var hægt að nota upplýsingar um fyrstu og seinustu tímamerkingu til að búta hljóðskrána. Eftir nokkrar tilraunir höfum við þróað aðferð sem finnur hvar í umritun Tiro líklegast er að gefinn bútur komi fyrir. Eftir að hafa fundið þetta notuðum við Levenshtein-fjarlægðina milli orða í bútnum og umritun Tiro til að hnitmiða fyrsta og seinasta orðið sem passa fullkomlega á milli. Þessi orð eru svo notuð til að snyrta bútinn svo við höldum sem mestu af honum en geymum jafnframt upplýsingarnar um tímamerkingu.

Dæmi um snyrtingu búta má sjá hér að neðan:

Upprunalegur bútur:

Já sko sumir sagnfræðingar fást alla ævi við það liðna en þú fylgist líka með samtímanum og hefur alltaf gert e byrjum á aðgerðum stjórnvalda og u viðbrögðum okkar fólksins. Sýnist þér almennt að að ráðamönnum hafi tekist hafi tekist vel upp og að almenningur hafi hagað sér eðlilega miðað við aðstæður?

Snyrtur bútur:

alla ævi við það liðna en þú fylgist líka með samtímanum og hefur alltaf gert e byrjum á aðgerðum stjórnvalda og u viðbrögðum okkar fólksins. Sýnist þér almennt að að ráðamönnum hafi tekist hafi tekist vel upp og að almenningur hafi hagað sér eðlilega miðað við aðstæður?

Upprunalegur bútur:

Já það er líka svolítið e erfiðara að að u ætlast til þess að almenningur fari eftir lögum þegar ráðherrar e gera það ekki.

Snyrtur bútur:

erfiðara að að u ætlast til þess að almenningur fari eftir lögum þegar ráðherrar

Við höfum prófað þessa aðferð á nokkrum hljóðskrám sem innihalda þátt af Morgunvaktinni á Rás 2. Fyrir hverja skrá skoðuðum við hversu mörg orð glötuðust við snyrtingu búta, hlutfall glataðra orða og orðvillutíðni (e. WER) milli bútanna og snyrtu bútanna. Eftirfarandi tafla sýnir þessar niðurstöður:

Dags. útsendingar	Fjöldi búta	Samtals orð	Glötuð orð	% glataðra orða	Meðal orðvillutíðni	Lengd
11.08.2020	175	10091	1868	18,5%		44 mín
22.10.2020	133	6196	1213	19,5%	36,4%	26 mín
28.12.2020	51	5148	479	9,3%	17,5%	25 mín
04.01.2021	119	8881	951	10,7%	21,9%	52 mín
18.01.2021	70	6757	488	7,2%	15,9%	38 mín

Þessar niðurstöður sýna að það eru einhver gögn sem glatast við snyrtingu bútanna miðað við orð í umritun Tiro. Magn þeirra gagna sem glatast er erfitt að meta þar sem það er mismunandi eftir hljóðskrám. Umritanir með fleiri búta glata fleiri orðum þar sem oftar er snyrt.

Stærsti gallinn við þessa aðferð eins og stendur er að oft eru mörg orð snyrt af byrjun og enda bútanna þar sem við höldum aðeins tímamerkingum ef orðin í bútnum og umritun Tiro eru nákvæmlega þau sömu. Tölurnar í töflunni má því líta á sem versta tilfelli.

Kistan

Innan gagna Kistunnar höfum við .vtt-umritanir, textavarpsumritanir og handrit. Þetta eru alls 928 klukkustundir hljóðgagna. Af þeim gögnum eru 295 klukkustundir .vtt, 436 klukkustundir textavarp og 197 klukkustundir handrit. Hins vegar inniheldur þetta gagnasett hljóðskrár sem hafa engan ritaðan texta af neinu tagi. Um helmingur textavarpsgagnanna eru með texta svo þannig fáum við 213 klukkustundir. .vtt-skrárnar eru með 252 klukkustundir með texta. Þannig höfum við minnst 465 klukkustundir gagna með hljóði og texta frá .vtt-gögnum og textavarpsgögnum. Í heildina er þetta 27% minnkun gagna sem við getum notað fyrir samröðun.

Samröðun og bútun útvarps- og sjónvarpsefnis

Með notkun samröðunarkerfisins sem þróað var í M4 og líkönum þróuðum fyrir Alþingistalgreinirinn söfnum við öllum textum frá hverjum þætti sem einni stórri segð. Síðan normum við, hreinsum og fjarlægjum greinarmerki úr textanum. Næst drögum við út atriði með háa upplausn, MFCC (Mel Frequency Cepstral-stuðlar) fyrir afkóðunarlíkönin. Við búum til búta

og endurbútum þá svo með greini utan sviðs. Þetta er mögulegt vegna þess að orðvillutíðni okkar utan málfræði er svo lág. Síðan eru skrárnar endursniðnar fyrir gagnasettið.

Kerfið hefur verið bætt til að virka með texta textavélar, ekki bara .vtt-skrám. Hingað til hefur þessi aðferð skilað um 60 mín. samraðaðra gagna fyrir hverjar 100 mín. umritaðra gagna.

Við útfærðum einnig sjálfvirka samræðugreind og samröðunarforskriftir til að úthluta auðkennum mælenda fyrir hvern þátt. En það var ekki fullnægjandi án sannreyndra kennsla mælenda því getum við ekki hópað saman segðir frá sömu mælendum þvert á gagnasettið. Vegna takmarkaðs tíma höfum við ákveðið að sleppa auðkennum mælenda alfarið á þessu stigi.

Nú höfum við 250 klst. samraðaðar og bútaðar af tiltækum gögnum úr Kistunni á Kaldi-sniði sem hæfir talgreiningu. Við munum þurfa að búta hljóðgögnin og búa til gagnasettið.

Niðurstaða og framtíðarvinna

Niðurstaðan er sú að við höfum útfært aðferð sem getur dregið tímamerkingar fyrir búta úr umritunum Creditinfo. Það glatast einhver gögn en við vinnum að því að minnka þetta tap gagna og auka nákvæmni. Þegar þessi viðbót er gerð við aðferðina munum við geta keyrt hana á öllum hljóðskrám og umritunum Creditinfo til að fá nothæf gögn til þjálfunar talgreina. Þriðja ár mun vera notað til að yfirfara, pakka og gefa út gagnasettin.

Með notkun þessara aðferða, skrifta og líkana sem við höfum þróað má breyta öðrum svipuðum gögnum í gagnasett talgreiningar. Gögn CreditInfo og 250 klukkustundir samraðaðs sjónvarpsgagnasetts RÚV geta búið til samræðugreindir og talgreina sem búa til nýja sjálfvirka skjátexta eða skráðar umritanir útvarpsþátta. Þau má einnig nota á öðrum sviðum og í aðra taltækni.

H3 – Umritun samræðna og fyrirspurna

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

Markmið

Að umrita samræður og fyrirspurnir til að eiga talmálsgögn með frjálsara flæði. Magn safnaðra gagna verður skilgreint þegar upptökubúnaður hefur verið settur upp og fyrstu upptökum lokið. Samtöl verða tekin upp á fyrsta og öðru ári, en upptökur fyrirspurna hefjast og þeim lýkur á öðru ári.

Varða 6 – lýsing

M5: 20 klst. af samræðugögnum teknar upp og umritaðar. Fyrirspurnamálheild búin til.
M6: 20 klst. af lesnum fyrirspurnum teknar upp.

Afurðir

Markmiðum vörðunnar hefur ekki verið náð. Vegna villna í söfnunarferlinu þurftum við að safna 33 klst. af upptökum þar til á síðustu stundu til að fá 20 klst. útgefanlegt gagnasett. Þetta er yfir 50% meira en búist var við. Þessi viðbótargagnaöflun ýtir hluta af prófarkalestri og þar með pökkun til október. Hefðbundna umhverfið er aðgengilegt á https://spjall.samromur.is

Yfir 20 klst. lesinna fyrirspurna hafa verið teknar upp. Hins vegar þarf enn að yfirfara þær. Fyrirspurnagagnasettinu er búið að pakka og verður gefið út fyrir 31. desember 2021.

Verkþætti H3 lýkur við M6.

Skýrsla

Skýrslan að neðan er yfirlit söfnunarvinnu okkar fyrir samræður og fyrirspurnir.

Yfirlit samræðna

Formáli

Oft eru sjálfvirkir talgreinar (e. ASRs) gerðir úr formlegum ritum eins og fréttagreinum, bókum, og þingræðum. Þetta eru almennt niðurskrifaðir og formlegir textar. Þetta gerir talgreinum almennt erfitt að greina tal í samræðum þar sem það hefur aðra eiginleika en formleg ræða. Samræður innihalda oft meiri hnökra og óklaraðar setningagerðir. Því virka mállíkön óvart verr sem búin eru til með formlegum og hárréttum textum. Til að vinna á þessu vandamáli í ensku hafa rannsakendur talgreina fyrir ensku búið til sértæk samræðugagnasett, eins og Switchboard, AMI, og CALLHOME, sem innihalda samræður frá gagnasettum funda og símtala. Þetta eru þrjú samræðugagnasett sem mikið eru notuð og vísað í innan taltækni. Smærri alþjóðleg gagnasett eru einnig aðgengileg í OpenASR20. Við höfum mótað gagnaöflunaraðferðir

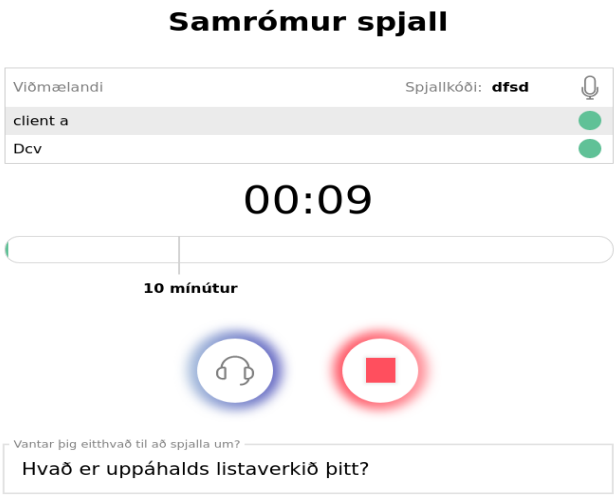
okkar og uppbyggingu gagnasetts fyrir Spjallróm eftir CALLHOME- og OpenASR 2020-gagnasettunum. Þannig er Spjallrómur fyrsta íslenska samræðugagnasettið sem hæfir talgreiningu.

Uppsetning

Sem fyrsta skref gagnaöflunar í nútímanum bjuggum við til Spjall, spjallsetur á vefnum. Það notar Vo-IP með eftirfarandi forritasöfnum: Web Audio API, MediaRecorder, og WebRTC. Upprunalegi hugbúnaðurinn var stöðugt uppfærður með áherslu á tæknilega eiginleika og upplifun notenda sem myndi skila bestu gögnunum.

Þessir þrír tæknilegu eiginleikar eru stuðningur hljóðnema, hljóðböggun (e. audio chunking), og þolin spjallrýmisauðkenni (e. chatroom IDs). Í fyrsta lagi eru hljóðnemar studdir fyrir Chrome, Firefox, og Safari. Í öðru lagi skiptir hljóðböggun hljóðupptökum í sundur og skilar þeim til vefþjóns þar sem baggarnir eru settir saman í eina heila hljóðskrá. Hljóðböggun hefur bjargað fimm annars glötuðum samræðum. Í þriðja lagi færa spjallrýmisauðkenni notanda alltaf í sama spjallrými jafnvel ef notandi aftengist eða vefsíður eins og Facebook bæta rakningu við hlekkina.

Þrjár mikilvægar notendaupplifanir eru tillögur að umræðuefni, takmörkuð stærð spjallrýma og notendatilkynningar. Í fyrsta lagi birtast tillögur að umræðuefni í slembiröð til að stinga upp á umræðuefni til að auðvelda notendum að halda samræðum gangandi. Í öðru lagi tryggir takmörkun stærðar spjallrýma við tvo að aðrir þátttakendur komi inn og trufli upptökuna. Einnig kemur það í veg fyrir upptökur einstakra notenda. Að lokum láta tilkynningar notanda vita af hverju hnappur er óvirkur, hvað olli villu, og þegar þeir reyna að fara úr flipanum áður en samræða hefur verið send inn. Við sendum einnig báðar hljóðskrár með einum hnapp og gefum möguleika á flýtivísum.



Mynd 1. Þetta er spjallforritið Spjall með notendalista, skeiðklukku, upptökustjórnborð, og hugsanlegt umræðuefni.

Fyrir söfnuð gögn bjuggum við til síðu með stigatöflu og tölfræði sem getur síað upptökur. Þær má sía að hluta (ein hljóðskrá) eða alls (tvær hljóðskrár). Allir þessir eiginleikar valda því að kláraðar upptökur eru um 87%, sem er nógu stöðugt til að safna heilum 40 mínútna samræðum.

Gögn

Söfnunarferli
Gögnunum var safnað frá september 2020 til september 2021. Um miðjan júní höfðum við þegar safnað 21 klst. samræðugagna. Þar sem stór hluti var slæm gögn héldum við áfram söfnun til að fá 20 klst. útgefanlegra gagna. Samræðurnar voru teknar upp á vefnum. Í umhverfinu geta þátttakendur boðið öðrum þátttakendum í spjall. Þegar þeir koma í spjallrými hafa þeir samþykkt skilmála söfnunar og gefið upp lýðfræðilegar upplýsingar: aldur og kyn. Innan spjallrýmisins ýta þeir á hnapp til að deila hlekk spjallrýmisins með vini. Þegar tveir þátttakendur eru í rýminu taka þeir upp og spjalla í 10–40 mín. Að lokum senda þeir upptökurnar til vefþjónsins þar sem þær fara í röð fyrir sjálfvirka talgreiningu í talgreiningarvefþjónustu Tiro.

Söfnun samræðna hefur verið stór áskorun. Við reyndum lýðvirkjun, héldum keppni, réðum þátttakendur, og bókuðum samræður. Árangur lýðvirkjunar var takmarkaður og úr urðu örfáar samræður. Næst héldum við keppni. Þetta þýddi að sjö ráðningaraðilar söfnuðu eins mörgum samræðum og mögulegt var á þriggja vikna skeiði. Þeir fengu þátttakendur til liðs við sig með því að láta orðið berast. Þó að þetta hafi gefið okkur næstum fimm klukkustundir gagna var mikið af þeim fullt af villum. Næst réðum við fólk. Hver starfsmaður var takmarkaður við aðeins tvær samræður til að sporna gegn því að gagnasettið greini aðeins talmynstur örfárra. Þetta gaf okkur meiri gögn. Að lokum gerðum við lista yfir viljuga þátttakendur. Af listanum úthlutuðum við þeim félaga og bókuðum upptökur. Þetta gaf okkur meiri gögn. Þessi síðasta aðferð hefur skilað langmestu magni gagna á sem stystum tíma. Í heildina var vinna við ráðningarnar mjög tímafrek og ekki skalanlegar.

Það erfiðasta var að fólki fannst almennt óþægilegt að tala á upptökum og jafnvel ef því fannst það í lagi þótti samfélagsneti þeirra það ekki. Því gátu einstakir þátttakendur ekki tekið þátt þó að þeir vildu.

Villur
Á miðri leið við söfnun fyrstu 20 klukkustundanna skoðuðum við gögnin. Í flestar skrár vantaði hljóð að hluta til, allt að tíu mínútur af þögn. Lausn okkar við þessum óviðbúnu þögnum var að bæta hljóðböggun við sem bjargaði fimm samræðum og vandamálið er nú horfið. Einnig voru notenda- og tæknilegar villur. Dæmi um notendavillur eru ósamræmdar lýðfræðilegar upplýsingar og einstaka upptökur með of mörgum röddum. Tæknilegar villur komu einnig upp. Sumar þeirra var ekki hægt að leysa og skiluðu engum gögnum. Sjálfvaldir vafrar fyrir Android-tæki sofnuðu oft sem olli því að upptakan stoppaði og hélt ekki áfram. Stundum sendust upptökur ekki til vefþjónsins þó að umhverfið segði svo hafa gerst.

Umritanir
Umritanirnar fylgja sömu viðmiðunarreglum og H4-fyrirlestrarnir. Hver samræða er sjálfvirkt umrituð og svo yfirfarin af tveimur manneskjum. Hljóðið er flutt í gegnum talgreini Tiro sem er búinn til með blöndu útgefinna íslenskra gagnasetta, Málróms og Þingræðna Alþingis, og einkagögnum. Því næst eru umritanirnar yfirfarnar af tveimur manneskjum. Yfir fimmtán klukkustundir hafa verið yfirfarnar einu sinni og þurfa aðra umferð. Restina þarf enn að yfirfara tvisvar. Eftir það eru umritanirnar sjálfkrafa dregnar út. Vandamál með útdregnu umritanirnar úr viðmóti Tiro gera það erfitt að greina frá heildarlengd samræðna í gagnasettinu. Því greinum við aðeins frá heildarlengd hljóðs.

Árangur

Gagnasettið Spjallrómur þarfnast breytinga við að taka út eftirfarandi persónugreinanleg gögn: nöfn þeirra sem yfirfara og ráða, hlekki í upprunalegar hljóðskrár, þegar minnst er á nána ættingja, kennitölur, persónuleg heimilisföng, og nafneiningar sem tengjast þátttakendum. Að lokum, eftir að búið er að taka út allar þessar upplýsingar, er gagnasettinu pakkað.

Til að samræða verði notuð í gagnasettið, þarf eftirfarandi að standast: Hver þátttakandi hefur sinn eigin hljóðnema og rödd hvers þátttakanda er tekin upp sér. Hver samræða skal innihalda tvær hljóðskrár, eina fyrir hvorn mælanda. Þátttakendur skulu hafa íslensku að móðurmáli eða með færni nálægt því. Lýðfræðilegar upplýsingar fyrir hvern mælanda innihalda aldur og kyn. Samræðulengd er 10–40 mínútur.

Allt í allt höfum við 32 klukkustundir og 56 mínútur af samræðugögnum, góðum og slæmum. Þetta er yfir 50% meira en varðan kveður á um. Hins vegar hafa yfir 25% gagnanna aðeins eins hljóðskrá, sem þýðir að aðeins eru 23 klukkustundir og 42 mínútur af gögnum sem nýtast í sjálfvirka talgreiningu. Innan þess undirsetts er svo enn að finna skrár með slæmu hljóði. Því er útgefið gagnasett 20 klukkustunda langt. Það inniheldur fleiri en 50 samræður og mælendur, með aðskilda hljóðskrá fyrir hvern mælanda. Mælendur eru karl- og kvenkyns og aldur þeirra er á bilinu 18 til 99 ára.

Niðurstaða

Spjallkerfið er nú komið í gang og má nota það fyrir áframhaldandi söfnun gagna af okkur á vefsíðunni og af öðrum í gegnum kóðahirsluna. Þetta kerfi hefur þann kost að ekki þarf lengur að fá fólk inn í hljóðver og safna þar gögnum í persónu fyrir samræður. Frekari endurbætur hugbúnaðarins myndu takmarka aðgengi umhverfisins að stærri tækjum öðrum en símum. Annar möguleiki væri að bæta við biðstofu, sem gerði notendum kleift að tengjast óþekktum notendum í kerfinu og byrja bara að spjalla. Þetta krefst mikils fjölda notenda í kerfinu á sama tíma. Lokið verður við málheildina Spjallróm og hún gefin út á CLARIN.

Yfirlit fyrirspurna

Formáli

Google assistant, Siri, Cortana, þetta er allt tækni sem notar sérhæfða talgreiningu á fyrirspurnum. Það eru engar tiltækar fyrirspurnamálheildir fyrir þjálfun íslenskra talgreina. Tilgangur þessa verkþáttar er að búa til málheild með fjölbreyttum fyrirspurnum, bæði sérhæfðum fyrir raddstjórnun og úr tali og texta.

Gögn (tölfræði, söfnunarferli, tegundir)

Ákveðið var að nota gagnasöfnunarumhverfi Samróms, sem búið var til í undirverkefni H1, til að safna fyrirspurnunum. Fyrirspurnir voru búnar til bæði úr algengustu spurningum á Google (með viðbættum tilbrigðum, skiptingu nafna/staða) og skrapaðar úr Risamálheildinni. Þeim var svo bætt við lista segða til söfnunar með forgangsröðun umfram aðrar segðir. Þetta gefur okkur sjálfkrafa umritanirnar en þarfnast í staðinn staðfestingar. Í töflunni að neðan er tölfræði aldurshópa úr söfnuðu gögnunum.

Aldurshópur	0–19	20–39	40–59	60–79	80+
%	32,2	27,5	33,0	6,4	0,5

Gögnin samanstanda af 37,8% körlum og 62,2% konum. Alls er 31 klukkustund gagna aðgengileg úr gagnasettinu.

Uppsetning (fyrir handyfirferðarferlið)

31 klukkustund gagna hefur verið valin úr fyrirspurnum söfnuðum í Samrómi. Þetta er 11 klukkustundum meira en markmið vörðunnar segja til um en búist er við að einhverjar klukkustundir glatist vegna óhæfra segða. Segðirnar hafa verið settar í yfirferðarpakka sem verður farið yfir.

Niðurstaða

Notkun núverandi notendahóps og Samróms sem söfnunarvettvangs reyndist mjög vel. Það gerði okkur kleift að safna miklu magni fjölbreyttra gagna. Af 31 klukkustund sem valin var fyrir fyrirspurnagagnasettið hafa um 8 klukkustundir verið samþykktar. Eina sem stendur eftir er yfirferð a.m.k. 12 klukkustunda frekari gagna, pökkun þeirra og útgáfa.

H4 – Umritun fyrirlestra

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

Markmið

Að þróa Kaldi-forskrift fyrir íslensku. Forskriftin ætti að vera sjálfstæð og gögnin aðgengileg, annaðhvort sem hluti af forskriftarskránum sjálfum eða sjálfvirkt sóttar innan þeirra. Forskriftin ætti að fela í sér hin hefðbundnu tæknilegu skref með HMM (Hidden Markov-líkön) byggð á þrístæðum, talgreiningu byggða á LDA/MLLT/SAT til tauganeta með seinkun og tvíátta LSTM.

Varða 6 – lýsing

M5: 50 klukkustundir fyrirlestra umritaðar frá 5 fræðasviðum

Afurðir

Markmiðum fyrir vörðu M5 hefur verið náð. Samtals hafa safnast 51 klukkustund og 26 mínútur af umrituðum og prófarkalesnum fyrirlestrum sjö fræðasviða frá tveimur háskólum. Gagnasafninu hefur verið safnað og það umritað en ekki gefið út. Hins vegar er útgáfa áætluð fyrir lok ársins í samstarfi við talgreiningarteymi HR.

Verkþætti H4 lýkur við M6.

Skýrsla

Formáli

Þessi hluti gagnasöfnunar fyrir talgreina fól í sér söfnun og umritun 50 klukkustunda gagnasafns sem inniheldur háskólafyrirlestra frá fimm fræðasviðum. Söfnun og umritun hófst á fyrsta ári og lýkur á öðru ári samkvæmt áætlun verkefnisins.

Gögn (tölfræði, söfnunarferli, tegundir)

Gögnunum var safnað úr fyrirlestrum og þau umrituð handvirkt með hjálp talgreiningarkerfis Tiro. Síðasta skref umritunarferlisins var prófarkalestur sem var unninn af faglegum prófarkalesara. Ritill Tiro var notaður í öllum tilfellum fyrir umritun sem orsakar samræmt gagnasafnsúttak sem gerir eftirvinnslu auðveldari. Eftirfarandi tafla gefur yfirsýn yfir fræðasviðin og lengd fyrir hvern fyrirlesara. Gögnin voru gefin úr fyrirlestrum tveggja stærstu háskóla á Íslandi, Háskóla Reykjavíkur og Háskóla Íslands.

Fræðasvið	Háskóli	Fyrirlesari	Klst. (K:MM)
Viðskiptafræði	HR	Hinrik Jósafat Atlason	0:19

Viðskiptafræði	HR	Katrín Ólafsdóttir	5:35
Tölvunarfræði	HR	Hrafn Loftsson	9:41
Tölvunarfræði	HR	María Ólafsdóttir	5:35
Tölvunarfræði	HR	Steinunn Gróa Sigurðardóttir	0:28
Verkfræði	HR	Ingunn Gunnarsdóttir	6:43
Verkfræði	HR	Jón Guðnasson	4:32
Lögfræði	HÍ	Ása Ólafsdóttir	7:09
Málfræði (íslenska)	HÍ	Eiríkur Rögnvaldsson	7:07
Sálfræði	HR	Brynja Björk Magnúsdóttir	3:01
Íþróttafræði	HR	Sveinn Þorgeirsson	4:59
Samtals			**51:26**

Tafla: *Yfirlit yfir lengd í klst. og mínútum af hljóði fyrirlestra sem hafa verið umritaðir. Taflan inniheldur fræðasvið, nafn fyrirlesara og nafn háskóla, Háskólinn í Reykjavík (HR) og Háskóli Íslands (HÍ).*

Uppsetning

Gagnasafninu var safnað frá ólíkum fræðasviðum sem uppfylla kröfur settar í áætluninni. Fyrstu umritanir voru gerðar aðallega af nemendum, helst nemendum með kunnáttu á viðeigandi sviðum. Ítarlegar leiðbeiningar voru skrifaðar af einum af prófarkalesurunum í byrjun umritunarferlisins til að tryggja samræmi milli vinnu mismunandi einstaklinga.

H7 – Þróun á almennum talgreini

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að þróa almenna Kaldi-forskrift fyrir íslensku. Forskriftin skal vera sjálfbirgin (e. self contained) og málföng til reiðu, annaðhvort sem hluti af forskriftarskránni sjálfri eða aðgengileg til sjálfvirks niðurhals innan frá. Forskriftin skal innihalda hefðbundin skref aðferða með þrístæðum byggðum á HMM, LDA/MLLT/SAT-byggðum talgreini að tímaseinkuðum tauganetum og tvístefnu-LSTM.

Varða 6 – lýsing

Fullgerð forskrift fáanleg sem Docker-mynd þjálfuð með Samrómi. Markmiðið er að ná orðvillutíðni undir 10%.

Afurðir

Fyrsti hluti gagnasetts Samróms hefur verið gefinn út undir ákveðnu nafni og hann sendur til LDC og OpenSLR til að gera niðurstöður okkar endurtakanlegar fyrir þriðju aðila. Nafnið á gagnasettinu er samromur_21.05. Aðeins gögn úr gagnasettinu voru notuð í forskriftina og prófunarsettið var valið til þess að það innihéldi eins margar málfræðilega einstæðar segðir og hægt er. Markmiðinu varðandi fullgerða forskrift var náð. Tvístefnu-LSTM og TDNN-hljóðlíkan voru útfærð í Kaldi sem styðst við HMM sem er þegar til. Hvað varðar árangur dróst TDNN á eftir tvístefnu-LSTM þar sem TDNN náði bestu orðvillutíðninni, 15,36%. Besta orðvillutíðnin **12,98%** náðist með notkun LSTM. Þessar niðurstöður fengust á samromur_21.05. Þessi árangur í orðvillutíðni er bæting frá M4 (21,82%), en er samt hærri en vonast var eftir (10%), sem við teljum vera vegna flækjustigs prófunarundirsettsins samromur_21.05. Docker-mynd var búin til sem inniheldur forskrift til þjálfunar alls kerfisins og Kaldi-útgáfu.

LSTM-forskriftin er aðgengileg á:
https://github.com/cadia-lvl/samromur-asr/tree/s5_tdnn_lstm/s5_tdnn_lstm

Málheildin er aðgengileg á:
https://openslr.org/112/

Docker-myndin er aðgengileg á:
https://hub.docker.com/repository/docker/michalborsky/samromur-asr

Verkþætti H7 lýkur við M6.

Skýrsla

Skref við útfærslu LSTM- og TDNN-aðferðanna eru eftirfarandi: 1) Uppfærsla kóða fyrir fyrri HMM-aðferð. 2) Hljóðlíkan þjálfað með notkun tvístefnu-LSTM. 3) Þjálfun mállíkans. Eftirfarandi texti er lýsing fyrir LSTM, þar sem það var betur uppbyggt, en einnig þar sem uppbyggingin er sú sama fyrir TDNN, fyrir utan endanlegar tölur yfir orðvillutíðni.

Þegar núverandi LSTM-aðferðin var búin til höfðu hljóðgögnin sem notuð voru fyrir fyrri HMM-aðferð breyst, sérstaklega í þeim hluta sem notaður var fyrir matssett sem hafði upprunalega lengdina 3klst. 48 mín. en hefur nú lengdina 15 klst. 51 mín. Tafla 1 greinir frá tölfræði um mælendur og segðir í núverandi skiptingu. Skriftur voru búnar til fyrir sjálfvirkan undirbúning málfanga fyrir þessa aðferð til að hafa gagnasettið á stöðluðu KALDI[40] s5-sniði.

Eftirfarandi tafla lýsir niðurstöðum gagnasettsins Samrómur 21.05:

	Lengd	Fjöldi mælenda	Fjöldi segða	Orðvillutíðni
Mat	15 klst. 51 mín.	1.641	10.000	12,98%
Þróun	15 klst. 16 mín.	2.108	9.995	11,48%

Þegar hljóðlíkönin voru búin til var skrefum fyrir LSTM og TDNN bætt við fyrri HMM-aðferðina sem lýst var í M4-skýrslu. Þessi nýju skref fela aðallega í sér að auðga þjálfunarmálheildina með hraðatruflun (e. speed perturbation) í tveimur mismunandi hlutföllum með tilliti til upprunalegs hraða (0,9 og 1,1) og reikna svo út iVector-a fyrir alla málheildina (þ.m.t. auðguðu gögnin) og að lokum þjálfun LSTM-netsins. Árangur þessa ferlis gaf orðvillutíðnina 11,48% í þróunarsettinu og 12,98% í matssettinu. Til samanburðar náðist orðvillutíðnin 13,43% og 17,28% á prófunarsettum Alþingis og Málróms, sem var líklega vegna þess að ekkert þeirra gagna var notað í þjálfun eða stillingu, aðeins við mat.

Fyrir mállíkanið notuðum við í grunninn sama mállíkan og í fyrri HMM-aðferð sem lýst er í M4-skýrslu, sem er 5,3 GB að stærð og hefur meira en 44 milljónir lína af texta. Með því að setja saman þennan texta með segðunum úr þjálfunarsettinu var mögulegt að ná öllum orðum fyrir orðasafnið sem búið var til við M4. Þetta orðasafn er samtals með 953.545 færslur, en þrátt fyrir það inniheldur það ekki fjölda orða sem Sequitur gat ekki breytt í fónem. Til að taka á þessu vandamáli var annað G2P-tól sem kallast Epitran[41] notað. Þannig var orðunum sem Epitran breytti bætt í tiltæka orðasafnið og nýja orðasafnið er nú greypt inn í Kaldi-forskriftina sem hefur samtals 968.331 færslur. Viðbæturnar í orðasafnið höfðu jákvæð áhrif á þann fjölda orða sem finnast ekki í orðasafninu, en hann fór úr 2,4 í fyrri aðferð í 0,3 í núverandi aðferð.

Docker-myndin til að þjálfa þetta kerfi var byggð á grunni Kaldi docker-myndarinnar. Forskriftinni var bætt við ásamt nauðsynlegum tólum og gögnum til að þjálfa mállíkanið til að gera sjálfstæða

[40] Kaldi-tólasettið – https://github.com/kaldi-asr/kaldi
[41] Epitran G2P-tólið – https://github.com/dmort27/epitran

útgáfu. Aðeins gagnasettið samromur þarf að vera tengt þegar mynd er búin til. Stærð myndarinnar er 9,2GB.

H8 – Vefviðmót fyrir talgreiningu

Skýrsla: Tiro
Aðili: Tiro

Markmið

Að þróa vefviðmót fyrir talgreiningu sem nota má í hugbúnaðarvörum. Fyrsta skrefið er þróun vefmiðmóts sem leyfir notendum að velja úr N-bestu tillögum (e. N-best hypothesis), en gefur annars hráar niðurstöður talgreinis. Annað skrefið mun útvega umfangsmeiri eftirvinnslu á úttakinu, þar á meðal normun texta og samræðugreind (sjá úttak úr H14). Þriðja skrefið er vefviðmót fyrir talgreiningu í rauntíma

Varða 6 – lýsing

M5: Vefviðmót fyrir talgögn með auðguðu úttaki
M6: Vefviðmót fyrir talgreiningu í rauntíma

Afurðir

Öllum markmiðum vörðunnar var náð. Tiro Speech Core-þjónustan hefur verið endurbætt með viðbót „auðgaðs úttaks“ sem inniheldur öfuga textanormun (e. inverse text normalization , ITN), sjálfvirka greinarmerkingu og samræðugreind. Þjónustan styður nú einnig greiningu í rauntíma með tvístefnustreymi.

Kóðinn og skjölun forritaskilanna eru agðgengileg í Git-hirslunni sem hýst er á Github á https://github.com/tiro-is/tiro-speech-core og mörkuð útgáfa er aðgengileg á https://github.com/tiro-is/tiro-speech-core/releases/tag/M6. gRPC-þjónustan er aðgengileg á grpcs://speech.tiro.is:443 og REST/JSON-þjónustan er aðgengileg á https://speech.tiro.is.

Verkþætti H8 lýkur við M6.

Skýrsla

Þremur stigum „auðgaðs úttaks“ var bætt við Tiro Speech Core: sjálfvirk greinarmerking, öfug textanormun (ITN) talna og algengra styttinga, og samræðugreind.

Sjálfvirka greinarmerkingin í Tiro Speech Core er byggð á notkun ELECTRA-líkansins úr undirverkefni H12 (eða sambærilegu líkani), sem er HuggingFace Transformer (PyTorch). Til að geta keyrt ályktun (e. inference) úr C++ var nauðsynlegt að breyta því yfir í TorchScript[42] með því að rekja mat þess. ELECTRA-líkanið notar WordPiece-tóka, svo að tilreiðara sem framkvæmir nauðsynlegar breytingar var bætt við.

[42] TorchScript: https://pytorch.org/docs/1.9.0/jit.html

Endurskrifunarmálfræði þýdd inn í OpenFST WFSTs (e. weighted finite state transducers) er notuð fyrir öfuga textanormun talna og algengra styttinga. Vinnan er byggð á handskrifaðri Thrax [43]-málfræði fyrir talgreiningarkerfi Alþingis[44].

Stuðningur við samræðugreint úttak er byggður á vinnu úr undirverkefni H14. Aðferðirnar sem notaðar voru þar henta ekki þeim stutta biðtíma sem talgreining í streymi krefst svo að ekki er stuðningur við samræðugreind í þeim hluta forritaskilanna.

Talgreiningin í rauntíma er útfærð með svipaðri aðferð og greining fyrir talgögn, nema að því leyti að forritaskilin taka aðeins við hráu 16 bita PCM-hljóði. Hún notar sömu gerð af Kaldi-líkönum („chain“-líkön) og notar sömu aðferðir fyrir afkóðun. Í stað þess að mata afkóðarann með heilli hljóðskrá tekur afkóðarinn við böggum af streymdum hljóðgögnum frá biðlara og leitar reglubundið að uppfærðum tilgátum. Hann sendir þær ef biðlarinn bað um hluta niðurstaðna, og ef hann bað um að vita hvort hugsanlegur endapunktur hefur verið greindur. Þegar endapunktur er greindur er endanleg tilgáta send til biðlara og hljóðbaggar sem á eftir koma eru unnir sem nýir bútar/nýjar segðir.

[43] Thrax: http://www.opengrm.org/twiki/bin/view/GRM/Thrax
[44] https://github.com/cadia-lvl/kaldi/tree/8301eb0/egs/althingi

H9 – Talgreining í snjallsímum

Skýrsla: Tiro
Aðili: Tiro

Markmið

Að hanna talgreiniviðmót fyrir Android- og iOS-lyklaborð. Á öðru ári verður áherslan á þróun frumgerðar Android-lyklaborðs með einföldu talviðmóti. Unnið verður í samstarfi við L8 – Málrýni í snjalltækjum (3 mánuðir úr H9 og 3 mánuðir úr L8). Á öðru ári munum við kanna aðferðir til að bæta talgreini og málrýni í lyklaborð snjalltækis, bæði sem vefþjónustu og sem einingu í tæki (e. on-device module). Við munum búa til frumgerð lyklaborðs með möguleika á rödd til texta fyrir Android með notkun vefþjónustu.

Varða 6 – lýsing

M5: Hönnun íslenskrar talgreiniþjónustu fyrir Android. Þarfir fyrir talgreini í tæki.
M6: Frumgerð íslenskrar talgreiniþjónustu fyrir Android í gegnum vefþjónustu.

Báðar vörðurnar voru endurskoðaðar þegar undirverkefni T5 (Talgervlar í snjalltækjum) var fært yfir á annað ár og L8 á þriðja ár. Hönnun og útfærsla sýndarlyklaborðs var færð til þriðja árs með L8 og hönnun og frumgerð þjónustunnar hafðar áfram á öðru ári.

Afurðir

Markmiðum vörðunnar var náð. Afurðir vörðunnar voru uppfærðar þegar undirverkefni T5 (Talgervlar í snjalltækjum) var fært yfir á annað ár og L8 (Málrýni í snjalltækjum) á þriðja ár.

Verkþáttur H9 heldur áfram á þriðja ári.

Skýrsla

Þarfir

Greining streymis og niðurstöður

Notandi sem slær inn orð á lyklaborði snjallsíma býst við tafarlausum viðbrögðum, þ.e. innslegnir stafir birtast samstundis á skjánum og orðin myndast með því að bæta við fleiri stöfum. Notandinn mun búast við sams konar upplifun þegar inntak er tal. Því er hörð krafa að styðja tvístefnustreymi hljóðs til talgreinisins og niðurstöðurnar frá honum í notendaviðmótið.

Stuttur biðtími

Þegar notandi byrjar að slá inn texta með lyklaborði snjallsíma býst hann við því að engin áberandi töf sé frá því að inntaksreitur er valinn og texti sleginn inn þar til textinn byrjar að birtast í þeim reit. Þetta þýðir að fyrir talað inntak er nauðsynlegt að tengingartími, fyrir greiningu

gegnum net, og hleðslutími, fyrir greiningu í tækinu, sé haldið í algjöru lágmarki (biðtími <500 ms frá því að byrjað er að tala í viðmótinu).

Mjög stórt orðasafn

Notandi sem notar tal sem inntak býst við að hvaða orð sem er skiljist af talgreini, eins og ef um væri að ræða innslátt staf fyrir staf. Þetta er þó ekki raunhæft jafnvel með þróuðustu talgreinum, sérstaklega fyrir tungumál með eins flókna beygingar- og orðhlutafræði og íslenska. Það er þó hægt að nálgast það með mjög stóru orðasafni með orðmyndum eða aðeins minna orðasafni með orðflísaeiningum (e. subword units) (t.d. með notkun BPE[45]).

Mikil nákvæmni

Dæmigerður notandi sem notar tal sem inntak býst alltaf við því að það sem hann segir sé það sem er tekið inn. Þetta er þó ekki alltaf satt fyrir aðrar gerðir inntaks og notendur eru vanir því að laga inntak sitt eftir á, annaðhvort handvirkt eða með hjálp sjálfvirkra leiðréttingartækja. Talgreinirinn þarf að hafa tök á *ásættanlegri* nákvæmni, sem er í venjulegu tilfelli orðvillutíðni upp á 10% að hámarki.

Hönnun á háu stigi

Android-kerfi veita möguleika á útfærslum þriðju aðila á mörgum þjónustum sínum, t.d. talgervlinum fyrir lesinn texta (e. text-to-speech service)[46] og, í okkar tilfelli, talgreiningarþjónustuna[47]. Með því að nota sérbyggðan talgreini til að útfæra `android.speech.RecognitionService` og skrá hana við Android-kerfið höfum við veitt hvaða neytanda á tækinu sem er, app eða þjónustu sem notar venjulega `android.speech.SpeechRecognizer`, aðgengi að þessum sérbyggða talgreini.
Fyrsta útfærsla þjónustunnar mun nota ský/veflæga þjónustu til að gefa niðurstöður greiningar. Hún mun nota forritaskilin Tiro Speech Core gRPC sem þróuð voru í undirverkefni H8[48]. Viðmót `RecognitionService` er einfalt:

```
// Tilkynnir þjónustunni að hún eigi að byrja að hlusta eftir tali.
void   onStartListening(Intent recognizerIntent, RecognitionService.Callback listener)
// Tilkynnir þjónustunni að hún eigi að hætta að hlusta eftir tali.
void   onStopListening(RecognitionService.Callback listener)
// Tilkynnir þjónustunni að hún eigi að hætta við talgreininguna.
void   onCancel(RecognitionService.Callback listener)
```

Þessar þrjár aðferðir eru notaðar af Android-kerfinu á viðeigandi tímum og útfærsla okkar beinir eða skilar niðurstöðum aftur til þess sem hringdi með því að keyra aðferðir `RecognitionService.Callback`[49]. Það hefur eftirfarandi viðmót:

```
void   beginningOfSpeech()
void   bufferReceived(byte[] buffer)
```

[45] Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909 (2015).
[46] TextToSpeechService: https://developer.android.com/reference/android/speech/tts/TextToSpeechService
[47] RecognitionService: https://developer.android.com/reference/android/speech/RecognitionService
[48] Tiro Speech Core: https://github.com/tiro-is/tiro-speech-core
[49] RecognitionService.Callback: https://developer.android.com/reference/android/speech/RecognitionService.Callback

```
void    endOfSpeech()
void    error(int error)
AttributionSource     getCallingAttributionSource()
int     getCallingUid()
void    partialResults(Bundle partialResults)
void    readyForSpeech(Bundle params)
void    results(Bundle results)
void    rmsChanged(float rmsdB)
```

Með því að nota talgreini á netinu losum við um meirihluta flækjustigs við útfærslu þjónustunnar. Þó þarf hún a.m.k. að gera eftirfarandi:

- Hefja og meðhöndla hljóðupptöku
- Tengjast við talgreiningarþjónustuna (SRS) með réttum stillingum: tungumál, hljóðkóðun, gerð mállíkans, hljóðtökutíðni o.s.frv.
- Meðhöndla og beina villum til þess sem hringdi með `error(int error)`
- Framkvæma einhverja lágmarks greiningu á raddvirkni á tækinu
- Senda stöðugt hljóðbagga yfir netið í talgreiningarþjónustuna
- Taka stöðugt við hluta af og/eða endanlegum niðurstöðum frá talgreiningarþjónustunetinu og áframsenda þær til þess sem hringdi í gegnum `partialResults(Bundle partialResults)` eða `results(Bundle results)` (hugsanlega sniðið).
- Gefa möguleika á forritsvirkni til að breyta ýmsum stillingum

Hljóðupptaka

Android-þjónusta eða -forrit notar `android.media.AudioRecord`[50] til að taka upp bagga af hljóði með hljóðnemanum. Þegar greiningarlota er hafin verður að gera hljóðupptökuaðferð tilbúna með hljóðkóðun og hljóðtökutíðni sem bæði Android og talgreiningarþjónustan styðja. Hljóbaggarnir sem hafa verið teknir upp þurfa að vera stuttir, <300ms, og sendir í netmeðhöndlarann eins fljótt og mögulegt er.

Stjórnun tenginga

Talgreiningarþjónustan ætti að styðja bæði upplesturshams, eða ótakmarkaða greiningu, og talleitarham, sem endar eftir fyrstu greindu segð. Tengingarstjórinn þarf að meðhöndla og halda á lífi tengingum við talgreiningarþjónustuna. Þetta felur í sér að varpa tungumálinu, kóðun og tökutíðniupplýsingum yfir á rétt snið fyrir talgreiningarþjónustuna og taka á mögulegum rofum tengsla og endurtengingar í kjölfarið í upplesturshams.

Greining á raddvirkni

Til að spara bandvídd þegar veflægur talgreinir er notaður í upplesturshams er mikilvægt að nýta greiningu á raddvirkni til að ákveða hvenær skal senda upptekið hljóð til talgreiningarþjónustunnar.

Senda bagga og taka við (hluta af) niðurstöðum á ósamhangandi hátt

Hörð krafa er um að niðurstöðum, hvort sem þær eru endanlegar eða aðeins að hluta, sé skilað með minnsta mögulega biðtíma aftur til þess sem hringdi. Því þurfum við aðskildar ósamstilltar einingar til að senda tiltæka hljóðbagga (sem greinast með raddvirkni) í talgreiningarþjónustuna

[50] AudioRecord: https://developer.android.com/reference/kotlin/android/media/AudioRecord

og taka við og meðhöndla niðurstöður. Móttökueiningin meðhöndlar öll svör, eða niðurstöður, úr talgreiningarþjónustunni. Þetta felur í sér þýðingu og áframsendingu net- og stillingarvillna til þess sem hringdi, og áframsendingu hlutaniðurstaðna og endanlegra niðurstaðna til þess sem hringdi. Hún mun einnig þurfa að framkvæma einhverja samhengismeðvitaða sniðmótun á niðurstöðum þegar hún er í upplestursham áður en niðurstöður (eða hluti þeirra) eru sendar.

Notendaviðmót stillinga

Notendaviðmót stillinga þar sem notandinn getur a.m.k. breytt slóðinni á talgreiningarþjónustu sem notuð er. Þetta gerir notandanum kleift að nota sjálfhýsta útgáfu af Tiro Speech Core.

Útfærsla frumgerðar

Frumgerðin er aðallega skrifuð í Kotlin með Bazel sem byggingarkerfi. Hún útfærir Android-viðmótið `android.speech.RecognitionService`. Önnur forrit á tækinu geta notað frumgerðarþjónustuna annaðhvort með því að biðja um venjulegt `SpeechRecognizer` eftir að hafa stillt þjónustuna sem sjálfvalda í tækinu eða með því að biðja sérstaklega um hana með nafni.

Talgreiningarþjónusta Tiro Speech Core gRPC (undirverkefni H8) sem er uppi á speech.tiro.is er notuð til að útfæra greininn í tækinu. Þegar forrit biður um talgreiningu er tvístefnu-tengingu komið á milli tækisins og þjónsins og bagga af 16 bita PCM-hljóði er streymt til þjónsins á meðan tækið hlustar eftir niðurstöðum frá þjóninum samhliða.

Auk þjónustunnar var útfært dæmi um forrit sem meðhöndlar tvær af fyrirætlunum Android-talgreiningar fyrir notendur; veffyrirspurn og tal sem inntak. Forritið opnast þegar forrit biður um aðra þessara fyrirætlana, kveikir á þjónustunni og skilar svo niðurstöðum til þess sem hringdi. Í tilfelli veffyrirspurnar áframsendir þetta notandann á skráð forrit fyrir veffyrirspurnir, t.d. vafra eða leitarforrit Google.

H13 – Líkan fyrir orðmyndir og samsett orð

Skýrsla: Háskólinn í Reykjavík
Aðili: Háskólinn í Reykjavík

Markmið

Markmiðið er að rannsaka og þróa orðflísamállíkön (e. sub-word unit language model) fyrir talgreini.

Varða 6 – lýsing

Mat og skýrsla (fræðigrein) um notkun orðflísamállíkans fyrir talgreini.

Afurðir

Við höfum þróað margar aðferðir fyrir orðflísamállíkön (SWUM), og ber þar helst að nefna byte-pair encoding (BPE) og Unigram aðferðirnar. Ítarleg lýsing og greining á SWUM-aðferðum, og samanburði þeirra við hefðbundna talgreiningu á orðastigi (kölluð orðastigsgrunnlíkan) fyrir íslensku og ensku, hefur verið gerð sem meistararitgerð sem var varin með góðum árangri í júní 2021.

Við höfum sýnt fram á að orðflísaaðferðir séu betri en líkön á orðastigi með nægjanlegu magni gagna fyrir gagnasett með mörg orð utan orðaforða, eða á LVCSR-sviðinu. Við vinnum enn að því að þýða það efni sem náðist og gefa út sem grein í fræðiriti. Ástæðan er sú að við erum að meta aðferðirnar fyrir tékknesku og spænsku líka, og leggjum áherslu á þá hæfni orðflísamállíkana að greina rétt orð utan orðaforða.

Hlekkur á meistararitgerð D. E. Mollberg:
https://skemman.is/handle/1946/39412

Hlekkur á hirslu orðflísamállíkans:
https://github.com/cadia-lvl/samromur-asr/tree/master/s5_subwords

Verkþætti H13 lýkur við M6.

Skýrsla

Útfærsla orðflísamállíkana hefur kvíslast frá hirslu almenna talgreinikerfisins (H7). Þetta þýðir að við höfum notað gagnasettin Samróm og Alþingi fyrir íslensku, og LibriSpeech fyrir ensku. Við höfum þjálfað talgreiningarkerfi með stigvaxandi magni hljóðgagna, byrjuðum með 10 klst., og tvöfölduðum magn í hverju skrefi, þar til við náðum 640 klst. Segðirnar voru valdar af handahófi,

en hvert skref innihélt allar segðir úr fyrra skrefi. Prófunarsettið var skorðað, sem og textamálheildin fyrir þjálfun mállíkans. Íslenskt mállíkan var fengið úr hluta Risamálheildarinnar sem gefin var út árið 2018. Fyrir ensku notuðum við LibriSpeech-textamálheildina. Fyrir orðflísamállíkönin notuðum við snyrt líkan 6-stæðu fyrir afkóðun og 8-stæðu fyrir endurskorun (e. re-scoring). Fyrir grunnlíkön (e. base models) notuðum við 3- og 4-stæður fyrir afkóðun og endurskoðun.

Samtals þjálfuðum við og mátum 30 hljóðlíkön og 88 mállíkön. Eftirfarandi texti varpar ljósi á áhugaverðustu og efnilegustu niðurstöðurnar, sem er lýst ítarlega í ritgerðinni. Fyrir íslensku nær grunnorðastigslíkanið betri árangri en orðflísastigslíkönin þegar hljóðgögn eru af skornum skammti, en víkur um sæti fyrir Unigram-líkaninu við 80 klst. hljóðgagna, og BPE-líkaninu við 320 klst. Fyrir 640 klst. undirsettið hefur grunnlíkanið orðvillutíðnina 9,09% ± 0,08 og BPE er örlítið undir staðalfrávikinu, 8,89% ± 0,07. Unigram-líkanið nær töluvert betri árangri með orðvillutíðnina 8,22% ± 0,07. Grunnlínuaðferðin hefur lægri orðvillutíðni fyrir 10 til 80 klst. gagna en eftir það eru líkönin innan staðalfrávika hvert við annað. 640 klst. grunnlínulíkanið hefur orðvillutíðnina 4,06% ± 0,09, BPE-líkanið hefur orðvillutíðnina 4,14% ± 0,09 og Unigram hefur orðvillutíðnina 4,12% ± 0,09. Myndirnar að neðan teikna ferla orðvillutíðni sem fall af gögnum.

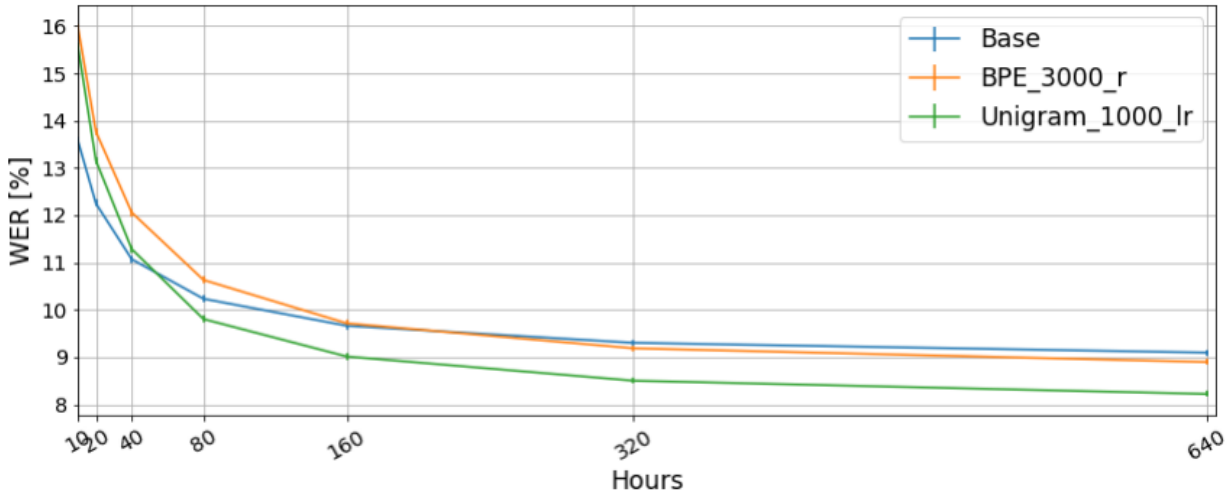


Figure 4.1.1: The WER for the Icelandic models. The graph shows the results for the word-level base phoneme model, the model with BPE subword units and the model with Unigram subword units. The standard deviation is depicted as vertical lines.

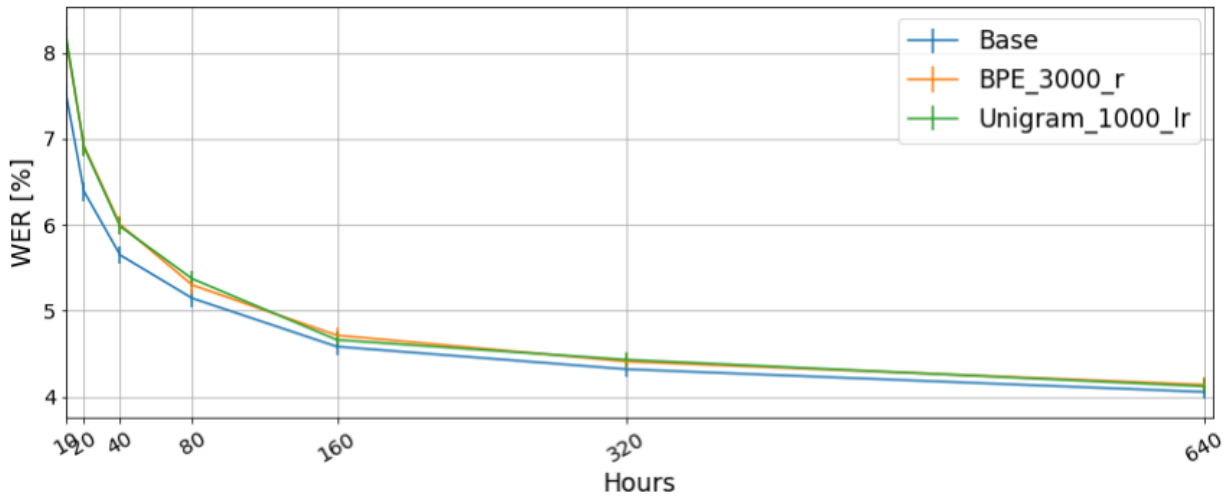


Figure 4.1.2: The WER for the LibriSpeech models. The graph shows the results for the word-level base phoneme model, the model with BPE subword units and the model with Unigram subword units. The standard deviation is depicted as vertical lines.

Við höfum einnig metið fjölda orða utan orðaforða sem myndast rétt af 640 klst. orðflísamállíkönunum, sem er tekið saman í töflunni að neðan. Taka skal fram að fjöldi orða utan orðaforða í LibriSpeech-málheildinni er mjög lágur, með aðeins 47 orð og aðeins 5 þeirra rétt umrituð.

	Orð utan orðaforða alls	Fjöldi orða utan orðaforða sem myndast rétt	Hlutfall árangurs
LibriSpeech	47	5	10,64 %
Alþingi	98	36	36,73 %
Samrómur	3014	1363	45,22 %

H15 – Hljóðlíkön fyrir barna- og unglingaraddir

Skýrsla: Háskólinn í Reykjavík
Aðili: Háskólinn í Reykjavík

Markmið

Markmiðið er að þróa hljóðlíkön sérstaklega fyrir börn og unglinga. Þessar raddir hafa talist erfiðari þar sem tíðnieinkenni (e. frequency characteristics) þeirra eru önnur en radda fullorðinna.

Varða 6 – lýsing

Hljóðlíkön fyrir barnaraddir þjálfuð og metin.

Afurðir

Einfalt GMM-HMM-hljóðlíkan byggt á þrístæðum (e. triphones) var þjálfað, sem nær orðvillutíðninni 43,71%. Auk þess var TDNN-LSMT-hljóðlíkan þjálfað, sem nær orðvillutíðninni 24,47%. Málheildin með tali barna var sniðin sem sjálfstæð málheild með nafninu „Samrómur Children“ og er hún tilbúin til útgáfu hjá Linguistic Data Consortium (LDC).

Grein var skrifuð um gerð hljóðlíkananna og verður send inn á LREC 2022. Hún er aðgengileg hér: https://drive.google.com/file/d/1kgOmOID3ADzb9fL08-9bW6fOM91sZaNt/view?usp=sharing.

Verkþætti H15 lýkur við M6.

Skýrsla

Á meðan á þessari vörðu stóð voru sumarnemendur ráðnir til að fara yfir segðir í málheildinni Samrómi. Á einhverjum tímapunkti voru þeir beðnir um að leggja áherslu á hljóðskrár sem innihalda tal barna til þess að fá stærra safn af þeirri gerð gagna. Við lok ágúst fóru sumarnemendurnir og gögnunum sem höfðu skapast og farið var yfir var safnað og þau sniðin sem sjálfstæð málheild. Svona varð málheildin Samrómur Children til.

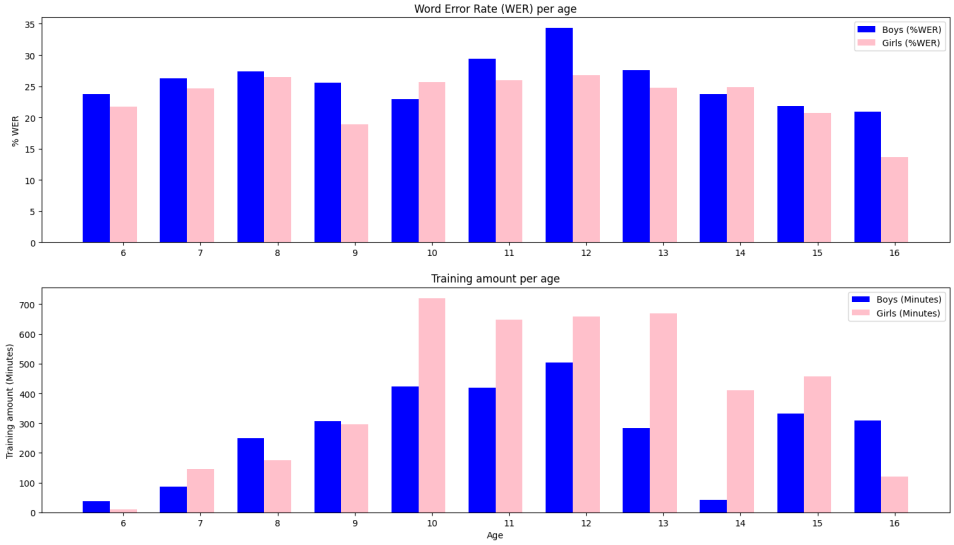
Eins og með fyrri útgáfur Samrómsgagna þurfti að skipta málheildinni í þjálfunar-, prófunar, og þróunarhluta en samkvæmt fræðigreinum er aldursvið eitt mikilvægasta sérkennið þegar unnið er með tal barna vegna þróunar raddlíffæra með tímanum. Af þeirri ástæðu er ekki hægt að búa til þróunar- og prófunarhluta Samrómur Children af handahófi eins og með útgáfu Samróms fullorðinna; þetta er vegna þess að handahófskennd skipting tryggir ekki að gögnin skiptist jafnt milli mismunandi aldurshópa. Því miður er aldursdreifing tals barna nokkuð ójöfn og sumir

aldurshópar eru með töluvert meiri gögn en aðrir. Til að taka á þessu voru báðir hlutarnir (prófun og þróun) valdir handvirkt. Prófunarsettið, sem er mikilvægara en þróunarsettið, inniheldur nákvæmlega fimm mínútur af hljóði úr hverju aldursviði (frá 6 til 16 ára) og frá hvoru kyni (strákum og stelpum). Þróunarsettið var búið til úr mælendum með óþekkt kyn þar sem ekki var nægilegt magn venjulegra gagna til að tryggja sama magn hljóðs eins og fyrir prófunarsettið.

Eftirfarandi tafla sýnir tölfræði fyrir málheildina Samrómur Children.

	Lengd	**Hljóðskrár**	**Fjöldi mælenda**
Prófun	1 klst. 50 mín.	1.714	625
Þróun	1 klst. 50 mín.	1.489	34
Þjálfun	127 klst. 25 mín	134.394	2.517
Samtals	**131klst. 06 mín.**	**137.597**	**3.176**

Þar sem málheildin Samrómur Children var tilbúin var byrjað að búa til hljóðlíkön í Kaldi. Sem fyrsta skref var einfalt GMM-HMM-hljóðlíkan byggt á þrístæðum (e. triphones) þjálfað, sem náði orðvillutíðninni 43,71%. Eftir það náðist 24,47% orðvillutíðni með hluta af TDNN-LSMT Kaldi-forskriftinni sem þróuð var fyrir H7 við M5. Þessar niðurstöður eru slæmar miðað við þær sem náðust á Samrómi fullorðinna, en taka verður til greina að samkvæmt fræðigreinum er tal barna töluvert flóknara viðureignar en tal fullorðinna. Auk þessa verður að muna að gögnin í Samrómur Children hafa ekki jafna dreifingu, sem veldur enn meiri flækju. Eftirfarandi súlurit sýna tölur um hvernig orðvillutíðni og magn gagna dreifist á aldurshópana.



Neðra súluritið sýnir magn þjálfunargagna (í mínútum) milli aldurshópanna frá 6 til 16 ára. Efra súluritið er svipað en sýnir orðvillutíðni. Eins og sjá má er orðvillutíðnin í jafnvægi þrátt fyrir að þjálfunargögn séu það ekki. Þessi hegðun er gild uppötvun, a.m.k. innan ramma rannsóknarhóps okkar, þar sem þetta kom á óvart og kemur ekki svo skýrt fram í fræðigreinum. Áframhaldandi rannsókn á þessum skrýtnu niðurstöðum mun líklega leiða til rannsóknargreinar

utan ráðstefnugreinarinnar sem er í vinnslu, og stefnir á að birta málheildina Samrómur Children eins og hún er.

T1 – Upptaka á einingavalsgögnum

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

Markmið

Að eiga upptökur af mismunandi röddum fyrir einingavalstalgervla. Markmiðið er að ná fjölbreytni í aldri og framburðarmállýskum, með jafnt hlutfall karl- og kvenradda. Átta raddir verða teknar upp, fjórar karlraddir og fjórar kvenraddir, hver þátttakandi skal lesa í a.m.k. 20 klukkustundir. Í þessum verkþætti verður unnið náið með verkþætti T7, við undirbúning handrita til upplesturs.

Varða 6 – lýsing

M5: Gefa út á CLARIN
M6: Á ekki við

Afurðir

Öllum afurðum er lokið. Gagnasettið var gefið út á CLARIN við byrjun M5 og er nú aðgengilegt endurgjaldslaust í hirslu CLARIN á https://repository.clarin.is/repository/xmlui/handle/20.500.12537/104.
Þessum verkþætti er því lokið.

T2 – Gögn fyrir raddblöndun

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

Markmið

Að eiga safn af upptökum til þjálfunar fyrir raddblöndun. Þetta felur í sér upptökur margra mælenda með svipaðar raddir sem blandast vel til þess að búa til stikaða talgervla. Markmiðið er að taka upp raddir 40 þátttakenda (20 kvenna og 20 karla), þar sem hver þátttakandi les í allt að 2 klukkustundir.

Varða 6 – lýsing

M5: Allir 40 þátttakendur hafa lokið upptökum, eftirvinnslu gagna lokið og þau eru tilbúin til útgáfu
M6: Útgáfa gagnanna

Afurðir

Markmiðum vörðunnar hefur ekki verið alveg náð. Allir 40 þátttakendur hafa lokið upptökum, eftirvinnslu gagna er lokið og þau eru í ferli þess að vera pakkað fyrir útgáfu. Auk þess hafa allar upptökur verið yfirfarnar minnst einu sinni af yfirlesara.

Gögnin verða gefin út fyrir 31. október þessa árs, sem lýkur þá þessum verkþætti.

Verkþætti T2 lýkur við M6.

Skýrsla

Allir 40 þátttakendur hafa lokið upptökum. 20 karl- og 20 kvenraddir.

Alls voru 57.399 upptökur gerðar. Af þeim var farið yfir 56.225 og þær merktar sem góðar eða slæmar af staðfestingaraðilum. Sumar upptökurnar var farið yfir oftar en einu sinni. Alls voru 1936 upptökur merktar sem slæmar minnst einu sinni, um 3,5%. Staðfestingaraðilar voru sammála um 97% upptaka sem farið var yfir oftar en einu sinni. Óyfirfarnar upptökur eru hluti af hliðarpakka sem inniheldur lengri setningar sem voru strangt til tekið ekki ætlaðar fyrir talgervingu.

Fjöldi setninga á þátttakanda er á bilinu 979 til 1929. Þetta kemur til vegna ýmissa þátta svo sem skipulagsvandamála, talhraða, skilvirkni upptaka og nákvæmni o.s.frv. Yfir helmingur þátttakenda er með minnst 1400 upptökur hver sem eru um 2 klst. tals eða meira.

Þátttakendur lesa af fyrirfram skrifuðum handritum svipuðum þeim sem notuð voru í T1.

Þátttakendur lásu frá einu af tíu mismunandi handritum. Fyrri hluti allra tíu handritanna innihélt sömu setningar – fyrstu 2000 setningarnar notaðar í T1. Seinni hlutinn innihélt einstakar setningar fyrir hvert handritanna tíu. Hver þátttakandi las jafnt í 1. hluta og 2. hluta skipaðs handrits. Auk þess las hver þátttakandi sett lengri setninga sem safnað var sérstaklega til að ná yfir mismunandi framburðarmállýskur og hljóðeinkenni.

T3 – Nýting annarra ganga

Skýrsla: Háskólinn í Reykjavík
Aðili: Háskólinn í Reykjavík

Markmið

Að safna hágæða upptökum með texta sem þegar eru til og hægt væri að nota fyrir talgervingu, og þannig auka magn aðgengilegra gagna fyrir þróun talgerviskerfa. Þegar slíkar upptökur hafa fundist, t.d. frá Hljóðbókasafni Íslands, útvarpsupptökum með handritum o.s.frv., þarf að leysa leyfismálin. Það sama gildir um eldri upptökur talgervla. Eftir söfnun gagnanna sem aðgengileg eru samkvæmt viðeigandi leyfum þarf að samraða þeim og sníða til notkunar við þróun talgervils.

Varða 6 – lýsing

M6: Hágæða upptökur sem nota má í þróun talgervilskerfa hafa verið fundnar. Samkomulag í höfn um söfnun allra gagna frá þeim eigendum sem eru viljugir til samstarfs.

Afurðir

Borin hafa verið kennsl á hágæða upptökur sem nota má við þróun talgervilskerfa og þær greindar. Hins vegar var ákveðið að það kæmi sér betur að leggja áherslu á þróun heldur en frekari gagnasöfnun, svo þessi verkþáttur mun ekki halda áfram á þriðja ári.

Skýrsla

Vinna við þennan verkþátt hófst við M6 og var ætluð sem undirbúningur fyrir tilvonandi gagnasöfnun frá utanaðkomandi áttum til notkunar í talgervingum á þriðja ári. Listi mögulegra gagnagjafa var undirbúinn með líklegustu aðilum fyrir þessa gerð efnis. Þeir voru:

- Ríkisútvarpið, eða RÚV (e. The Icelandic National Broadcasting Service)
- Hljóðbókasafn Íslands (e. The Icelandic Audiobook Library)
- Storytel – Stærsta hljóðbókaveitan á Íslandi í dag

Auk þeirra var haft samband við samstarfsfélaga SÍM hjá Stofnun Árna Magnússonar (SÁM) til mögulegs samstarfs. Álits þeirra var leitað til að spyrjast fyrir um framvindu þeirra við söfnun textagagna frá útgefnum bókum á Íslandi.

Greint verður frá framvindu og niðurstöðum þessara viðræðna hér.

Ríkisútvarpið (RÚV)
Tengiliður okkar innan SÍM varðandi aukna efnissöfnun RÚV er Helga Lára Þorsteinsdóttir. Borin voru kennsl á mögulega gjafa efnis fyrir talgervingu eins og:

- Hljóðupptökur þegar útgefinna bóka, almennt lesnar af höfundum en þó ekki alltaf. Dæmi um þetta er stórt gagnasett sem inniheldur upptökur af Halldóri Laxness að lesa sum verka sinna.
- Þættir með handriti sem er aðgengilegt og sem fylgja því orðrétt.
- Tal fréttaþula í útsendum fréttum með skjátextum.

Það kom í ljós að mikið magn slíkra gagna er geymt í safni RÚV og þau má frá sjónarhorni RÚV nota eins og áætlað er hér.

Hljóðbókasafn Íslands
Þetta safn er ríkisstofnun sem veitir hljóðupptökur af textum fyrir fólk með hamlanir sem gera þeim ómögulegt að lesa. Það vinnur samkvæmt strangri undanþágu frá höfundarréttarlögum og hefur ekki nein réttindi yfir textunum. Safnið er í mjög viðkvæmri stöðu og þarf að fylgja ströngum vinnureglum þegar kemur að gögnunum. Niðurstaða okkar var að það hentaði ekki nógu vel fyrir gagnasöfnun af þessu tagi.

Storytel
Storytel er vinsæl hljóðbókaþjónusta á Íslandi sem framleiðir mikið af hljóðbókum á ensku. Við höfðum samband varðandi mögulega veitingu gagna en þau virtust ekki mjög áhugasöm um verkefnið. Aðaláhyggja þeirra var mögulegur kostnaður við sendingu gagnanna og vinna við að afla leyfum frá hagsmunaaðilum. Við teljum að við getum nálgast þau aftur með ítarlegri áætlun og fjármagn þegar söfnun hefst á ný.

Höfundarréttur
Leyfi þarf frá öllum hagsmunaaðilum sem taka þátt í öllu ferlinu. Þeir hagsmunaaðilar sem um ræðir fara eftir gerð efnis og þarf að skoða í hverju tilfelli. Þetta eru hagsmunaaðilar eins og höfundar textanna, útgefendur textanna, lesendur textanna og framleiðendur upptakanna. Tilfelli þar sem höfundarréttur hefur fyrnst væri auðleyst vandamál, en það eru ekki mörg tilfelli þar sem 95 ár þurfa að líða frá dauða höfundar.

Við höfðum samband við Steinþór Steingrímsson til að fá álit hans á vinnu við söfnun texta sem fer fram innan Stofnunar Árna Magnússonar (SÁM). Vinna þeirra við að ná samstöðu frá sambandi höfunda og félagi útgefenda var stoppaður af einum útgefanda. Því er ætlun SÁM að sneiða hjá þeim tiltekna útgefanda og safna gögnum frá þeim sem ekki hafa hafnað þátttöku.

Eitt áframhaldandi markmið var að sækja lista texta sem SÁM safnar samþykki fyrir og finna hvort hljóðupptökur eru til af þeim frá einhverjum hljógagnagjafa okkar. Í þessum tilgangi fengum við lista frá SÁM.

RÚV gaf okkur einnig lista tiltækra upptaka hljóðbóka til samanburðar.

Árangur / Umræður

Ekki haldið áfram þar sem þetta var tekið af þriðja ári.

Niðurstaða og næstu skref

Þessi söfnun mun vera lagalega flókin þar sem margir hagsmunaaðilar tengjast hverri hljóðbók/skrá. Söfnun allra löglegra leyfa mun taka tíma og vinnu.

Næsta skrefið gæti verið að nota tiltæk gögn frá RÚV og hafa samband við Storytel með ítarlega áætlun og fjármagn til að dekka vinnu þeirra.

T4 – Vefgáttir fyrir talgervla

Skýrsla: Tiro
Aðili: Tiro

Markmið

Að búa til vefþjónustu fyrir talgervla. Frumgerðin sem gefin var út á öðru ári tekur við stöðluðu inntaki fyrir stutta texta og fyrsta (opinbera/útgefna) útgáfan inniheldur forvinnsluskref, m.a. tilreiðslu og textanormun.

Varða 6 – lýsing

M5: Frumgerðarútgáfa tekur við styttri stöðluðum textum (setningum)
M6: Útgefin útgáfa sem inniheldur tilreiðslu og normun texta

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Talgervilsþjónustan er komin í gang á https://tts.tiro.is, sem inniheldur einnig skjölun forritaskila. Kóðinn hefur verið gefinn út í Git-hirslu á https://github.com/tiro-is/tiro-tts og merkt útgáfa fyrir M6 hefur verið gerð aðgengileg á https://github.com/tiro-is/tiro-tts/releases/tag/M6.

Verkþáttur T4 heldur áfram á þriðja ári.

Skýrsla

Við völdum að útfæra undirsett forritaskilanna fyrir Amazon Polly. Þjónustan kallast Tiro TTS og er útfærð með Python með Flask-rammanum (e. framework). Eins og stendur er stuðningur við tvær gerðir bakenda: Fastspeech2/MelGAN (FS) og staðgengil Amazon Polly. Alls eru fjórar raddir aðgengilegar á https://tts.tiro.is. Tvær FS-raddir úr undirverkefni T7 þróaðar af Háskólanum í Reykjavík, *Diljá* og *Álfur*, og íslensku Polly-raddirnar tvær, *Dóra* og *Karl*.

Talgervill Tiro útfærir undirsett Polly-forritaskilanna sem eru skjöluð í kóðanum og á https://tts.tiro.is. Undirsettið inniheldur eins og er þarfir helstu notenda þessara forritaskila, WebRICE (T6) og Símaróms (T5). Þjónustan tekur við bæði skjali með ónormuðum texta eða SSML[51] og skilar hljóði (MP3, Ogg Vorbis eða hráu 16-bita PCM) eða talmerki, sem gefa til kynna hliðrun bætis og tíma fyrir hvert talgervt orð sem beðið er um. Eins og stendur er stuðningur við SSML takmarkaður við `<phoneme>`-tagið þar sem sá sem hringir getur notað undirsett X-SAMPA til að stjórna framburði einstakra orða eða segða sem beðið er um.

[51] Speech Synthesis Markup Language: https://www.w3.org/TR/speech-synthesis11/

FS-raddirnar frá talgervli Tiro geta notað normun studda af gRPC-þjónustu með niðurstöðum úr undirverkefni T9[52] eða einfalda normun, sem tekur bara á greinarmerkjum og nokkrum óprentanlegum Unicode-stöfum. Þjónustan er í gangi á https://tts.tiro.is með það fyrrnefnda.

[52] https://github.com/grammatek/tts-frontend-service

T5 – Talgervlar í snjallsímum

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

Markmið

Að hafa nothæfa frumgerð af íslensku talgervilsappi fyrir Android til að brúa það bil sem myndaðist þegar eina íslenska talgervilsappið var fjarlægt af Google Play Store.

Varða 6 – lýsing

Tvær frumgerðir af íslensku talgervilskerfi eru í gangi til notkunar á Android-tækjum: önnur þeirra keyrir á vefþjóni, og veitir bestu gæði sem eru í boði eins og er, og hin er hönnuð til að keyra á tæki, hugsanlega með fórnum gæða. Hvort kerfið fyrir sig býður upp á eina kvenkyns og eina karlkyns rödd. App er komið upp á Android sem inniheldur framendapípu, tilbúið í Beta-prófanir.

Afurðir

- Forrit fyrir Android sem nýtist sem kerfislæg talgervilsþjónusta til að lesa alla texta á Android-tækinu á íslensku með talgervilsröddum af netinu (undirverkefni T7). https://github.com/grammatek/simaromur.
- Google Play Store: https://play.google.com/apps/testing/com.grammatek.simaromur

Raddir á tæki er enn verið að undirbúa til að útfærra í forritinu, og því hafa markmið ekki náðst að fullu.

Verkþáttur T5 heldur áfram á þriðja ári.

Skýrsla

Appið Símarómur var prófað í lokuðum Beta-prófunum í sumar, en opin Beta-útgáfa var gefin út á Google Play Store í byrjun september. Appið er sífellt prófað af meðlimum Blindrafélagsins. Endurgjöf er almennt mjög jákvæð, en „ofurnotendur“ kvarta þó undan töfum radda af netinu.

Android veitir forritaskil fyrir talgervingu (e. TTS) þannig að sérhannaða rödd megi útfæra í kerfinu og virkja kerfislægt. Við þróuðum forritið „Símaróm“ sem veitir slíka talgervilsþjónustu fyrir Android.

Í byrjun lögðum við áherslu á að búa til raddir sem keyra á Android-tækinu sjálfu. Þær þurfa að standast strangari kröfur en raddir sem keyra á vefþjóni. Tauganetsröddin „Álfur“ þarf t.d. > 900 MB af vinnsluminni sem er of stór þörf fyrir Android-tæki. Því var tekin ákvörðun um að nota hefðbundna nálgun, ber þar helst að nefna „clustergen“ og „unit-selection“ til að búa til raddir,

sem krefjast töluvert minna af vinnsluminni. Opni (e. open-source) hugbúnaðurinn „Festival Speech synthesis“ (https://www.cstr.ed.ac.uk/projects/festival/) er mikið notaður til að búa til clustergen og unit-selection-raddir og hefur verið notaður í þessu verkefni til að búa til raddirnar ’Álf’ og ‘Dilja’. Til að geta keyrt á Android verður að nota annan opna rammann „Flite“ (http://www.festvox.org/flite/). Þetta varpar röddunum sem Festival býr til yfir í C-safn sem má nota á Android í gegnum JNI-viðmótið. Því miður eru Festival og Flite nokkuð gömul hugbúnaðartól og mikil vinna hefur verið lögð í að aðlaga þau að nútímarömmum Android, leysa villur og samþætta þau í nútímalega og endurtakanlega samsetningarpípu.

Einnig þurfa raddir á tæki íslenskt textanormunarkerfi auk g2p á tæki.

Í M3 skilaði undirverkefni G6 g2p C++-tólinu g2p-thrax, sem notar regglubyggða nálgun og hefur lágar þarfir. Við umbreyttum g2p-thrax yfir á Android. Aðlaga þurfti samsetningarkerfið auk allra forritasafna sem eru forkröfur (thrax: https://github.com/grammatek/thrax, OpenFST: https://github.com/grammatek/openfst) fyrir Android. Við bjuggum til Android-samsetningarpípur og útgáfur á tvíundarsniði fyrir þær allar og samþættum þær inn í Símaróm. Regina normalizer (undirverkefni T9) var umbreytt yfir á Java til að keyra á Android. Við bættum að auki við unicode-hreinsiferlum.

Staðan nú varðandi raddir á tæki er sú að meirihluta vinnu við forkröfur er lokið. Flite-raddirnar þarf enn að vistþýða á Android og bjóða upp á að sækja þær. Í Símarómi er búið að útfæra val til að sækja rödd og virkja með JNI, en aðlaga þarf skema gagnasafnsins til að geta verið samhliða netröddinni.

Við styðjum tvær karlkyns og tvær kvenkyns netraddir. Netraddirnar eru útfærðar sem netþjónusta á https://tts.tiro.is/v0/. Símarómur sækir lista tiltækra radda og gefur þennan lista á virkan hátt í raddval Android-talgervilsins. Notandinn getur valið milli tveggja tauganetsradda og tveggja hefðbundinna radda fyrir íslensku sem eru: „Álfur“, „Diljá“, „Dóra“ og „Karl“. Nýju taugaraddirnar tvær, „Álfur“ og „Diljá“, framkvæma g2p á vefþjóninum, en textanormun fer fram innan Símaróms sjálfs.

T6 – Veflesari

Skýrsla: Háskólinn í Reykjavík
Aðili: Háskólinn í Reykjavík

Markmið

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

Varða 6 – lýsing

M5: Opinn (e. open-source) veflesarahugbúnaður gefinn út.
M6: Hönnun á hugbúnaði vafraviðbótar. Opinn íslenskur talgervill kynntur utanaðkomandi söluaðilum.

Afurðir

Öllum markmiðum hefur verið náð. Við enda M5 var WebRICE útgefinn ásamt góðri skjölun sem lýsir samþættingarferli fyrir vefhönnuði.
WebRICE-hirslan: https://github.com/cadia-lvl/WebRICE

Öllum markmiðum vörðunnar var náð.
Nauðsynleg skjölun hugbúnaðarhönnunar WebRICE-vafraviðbótar búin til. Hún inniheldur lista krafa, víravirki notendaviðmóts og UML-skjöl. Þessi skjölun hjálpar við þróun og hönnun. Kynning fyrir utanaðkomandi söluaðila er hafin. Ýmsir aðilar hafa byrjað virkar prófanir hugbúnaðarins.

Verkþáttur T6 heldur áfram á þriðja ári.

Skjölun WebRICE-vafraviðbótar:
https://drive.google.com/drive/folders/1OU47doVuUlFA_CtyilS6H9nZeXwy-KEf

Skýrsla

Vinnan í M5 er mjög frábrugðin þeirri sem fer fram í M6. Í M5 var upprunalegu útfærslunni á WebRICE lokið og hún gefin út, á meðan í M6 fólst vinnan eingöngu í hugbúnaðarhönnun fyrir útgáfu vafraviðbótar. Á sama tíma var minniháttar vinna lögð í nýútgefna upprunalega WebRICE. Því er mjög skýr lína milli varðanna tveggja og því skiptist þessi skýrsla í hluta fyrir M5 og annan hluta fyrir M6.

M5

Í byrjun M5 gerðum við notendaprófanir og fengum góða endurgjöf. Samkvæmt henni lærðum við að flestir notendur myndu nota WebRICE á snjallsímum frekar en á tölvum. Tölvur voru þó í öðru sæti. Flestir notendur höfðu áhuga á að nota WebRICE.

Við unnum að endurbótum vefsíðu WebRICE[53]. Við fengum fregnir af því að dæmin virkuðu illa á símum og að takkar spilarans væru of litlir. Við leystum þau vandamál og nú virkar síðan vel á öllum tækjum. Við bættum einnig við skráningu á póstlista á forsíðunni.

Við bættum virkni textaútdráttar svo WebRICE virkar nú betur með fjölbreyttari vefsíðum. Við bættum einnig við möguleika þess að hönnuðir breyti litavali WebRICE-spilarans, sem gerir notendum kleift að aðlaga útlitið að stíl vefsíðna sinna.

Áður takmarkaðist WebRICE við greinar styttri en 600 orð. Var það vegna þess að forritaskil talgervilsins taka aðeins hámark 600 orð. Þökk sé bögguninni sem við útfærðum getur WebRICE nú spilað lengri greinar.

Að lokum samþættum við forritaskil talgervilsins úr undirverkefni T4 í WebRICE. Nú fer allur texti í gegnum þau forritaskil til að búa til hljóð fyrir talgervil.

M6

Formáli

Í þessari vörðu höfum við unnið að almennri hugbúnaðarhönnun WebRICE-vafraviðbótar. Markmiðið var að búa til hönnunarskjölun sem stuðning við útfærsluna sem hefst í M7. Það markmið náðist ekki.
Annað markmið var að auglýsa þegar útgefinn WebRICE og kynna fyrir utanaðkomandi söluaðilum. Það markmið náðist og nokkrir aðilar vita nú af WebRICE og sumir hafa byrjað að prófa hann.

Kostir vafraviðbótar

Allir, óháð tæknikunnáttu, geta sótt og sett upp vafraviðbót. Hönnuðir geta auðveldlega dreift uppfærslum fyrir viðbæturnar sem virkjast sjálfkrafa fyrir notendur.
Annar góður kostur er að notendur geta valið hvaða texta sem er til að lesa, hvar sem er á netinu, svo lengi hann er á íslensku.

Vafrarnir

Vafrarnir sem við lögðum áherslu á eru Google Chrome, Mozilla Firefox, Microsoft Edge og Apple Safari. Þetta eru vinsælustu vafrarnir og þeir hafa góðan stuðning fyrir hönnun viðbóta. Allir vafrar nema Safari útfæra Web Extensions[54]-forritaskilin sem þýðir að viðbætur þeirra geta deilt næstum sama kóðagrunni. Það gerir samtímis þróun mögulega fyrir þá vafra sem sparar mikinn tíma. Safari verður að útfæra sem alveg aðskilið verkefni, en fyrir notandanum ætti endanleg afurð að vera eins og viðbót hinna þriggja vafranna að öllu leyti.

[53] http://webrice.is/

[54] https://extensionworkshop.com/documentation/develop/about-the-webextensions-api/

Hönnunarskjölun

Hönnunarskjölun mun hjálpa okkur þegar hönnun hefst í M7. Hlekkir eru gefnir í neðanmálsgreinum fyrir hvert atriði sem við tölum um í þessum kafla.

Listi yfir kröfur notenda[55]

Búinn var til listi sem skilgreinir kröfur sem vafraviðbót verður að uppfylla. Honum er raðað eftir mikilvægi frá A til C. A-kröfur eru nauðsynlegar fyrir minnstu raunhæfu afurð. B-kröfur eru nauðsynlegar að einhverju leyti en *geta* beðið. C-kröfur eru eiginleikar sem gott væri að hafa. Þessi listi verður leiðarvísir til að sjá hvar þarf að byrja og áminning um forgangsatriði.

Hönnun notendaviðmóts[56, 57]

Til samræmis var ákveðið að halda hönnun notendaviðmóts eins nálægt þeirri sem þegar var gerð fyrir upprunalega WebRICE eins og hægt er. Þessi skjöl sýna hönnunina, eiginleika og útskýra vinnuferli notanda fyrir endanlega afurð.
Þegar hönnun hefst munu þessi skjöl nýtast sem leiðarvísir og halda hönnun okkar við eitthvað sem *líkist* og *gerir* það sem upprunalega var ákveðið. Þó verður að hafa í huga að þetta er *leiðarvísir* en ekki óhaggandi mynd af endanlegri afurð. Skjölin voru búin til með hugbúnaði frá diagrams.net.[58]

Hliðarspjald

Víravirkið sýnir, auk annara hluta, hliðarspjald með textareit. Það má opna til vinstri eða hægri. Þegar notandi velur texta og ýtir á ‘Spila’-takkann afritast línurnar sjálfkrafa yfir í textareitinn þar sem viðbótin mun yfirstrika orðin þegar þau eru lesin.

UML-skjöl[59]

Nokkur UML-skjöl voru búin til. Þau lýsa áætlaðri högun kerfisins, tæknilegri uppsetningu og flæði stjórnar.
Við bjuggum til tvær gerðir UML-skjala, flæðirit[60] og skýringarmyndir.[61][62][63]
Yfirlit[64] var einnig skrifað. Það gefur ítarlegri skýringar á lögunum sem talað er um í skjölunum.
Þessi skjöl munu hjálpa þegar kemur að því að setja upp fyrsta kóðagrunn þegar útfærsla hefst.

[55] https://docs.google.com/document/d/1koGE3Bq9zEf8e4Xqajj771CVV1B-j_r9LYn-ooH5tW4/edit
[56] https://drive.google.com/file/d/1ARABE6FlzmZUR07Fkw54J9t-I-HFBZCS/view
[57] https://drive.google.com/file/d/12opJddqVPpgHuOL-beSwdrJsL9avaVfy/view
[58] https://www.diagrams.net/
[59] https://drive.google.com/drive/folders/1N4gPEk_QVj9GNMY2-nAca5kMZhVSoEJZ
[60] https://drive.google.com/file/d/1VZVzDlayG7xSDitd_M05FaW-7_nR0Hjn/view
[61] https://drive.google.com/file/d/1Qro00jGGfSTla4uC6fhRkHqE0Fg3mTMb/view
[62] https://drive.google.com/file/d/1q6J5zI4-Y6blzFeQfGLSTDlIRbLqcWL4/view
[63] https://drive.google.com/file/d/1pvC2OATy1LR0f2qP_MPANI2lN1hRYzR6/view
[64] https://docs.google.com/document/d/16WAGYrkGnEONRlyOqe1pIurg9ViQ3s3nyJxTrkz6Q5A/edit

SSML – rannsókn og dæmi

SSML er víða notað fyrir talgervla og er orðið staðall sem flest talgervilskerfi nota. Þar sem við viljum að WebRICE verði útfært samkvæmt núverandi stöðlum gerðum við smá rannsókn[65] til að sjá hvernig hann getur notið góðs af því að vera með stuðning við talgervil.

WebRICE myndi njóta góðs frá SSML á margan hátt. Það gæti hjálpað við að fá tölfræði með innbyggðum mörkum og veita betri raddþjónustu almennt. Það má nota til að stjórna hlutum eins og tónhæð, tónhæðarlínu, tónhæðarsviði, hraða, lengd og hljóðstyrk. Allt mjög nytsamlegt fyrir veflesara eins og WebRICE.

Markaðssetning og auglýsing

Auglýsing WebRICE er hafin og heldur áfram í M7. Við höfum leitað til ríkisstofnana eins og Þjóðskrár og Skattsins auk Íslandsbanka, einum þriggja stærstu viðskiptabanka á Íslandi og kynnt WebRICE fyrir þeim.
Tæknileg fræðigrein um WebRICE, eiginleika hans, útfærslur og þróunaráætlun framtíðar, var skrifuð og send til Skýs.[66] [67] Hún hefur enn ekki verið gefin út.

Áætlunin fyrir M7 í október er að hitta og tala við aðila fimm til sjö vinsælustu íslensku vefsíðnanna (eins og visir.is, mbl.is og frettabladid.is), auk vefhönnunarstofa og fyrirtækja leiðandi á sínu sviði. Þessi áætlun leggur áherslu á að hitta um 15 áhrifamikla hugsanlega notendur, sölu-/hagsmunaaðila og kynna WebRICE fyrir þeim með ítarlegri nálgun sem mun vonandi leiða til hraðari fjölgunar notenda WebRICE. Þessu til stuðnings og til að ná til annarra áhugasamra aðila munum við fara í auglýsingaherferð sem felur í sér útvarpsviðtal (á Bylgjunni eða Rás 2) og fleiri fréttagreinar (líklegast á mbl.is og/eða visir.is). Þegar hefur komið ein grein á frettabladid.is.[68]

[65] https://docs.google.com/document/d/1vwKLngLEygUhDl3CrWqTkUEiXoVFf1wwZWFZ_c7E-wo/edit
[66] WebRICE Ský grein
[67] https://www.sky.is/
[68] https://www.frettabladid.is/frettir/bylting-i-maltaekni-nyr-islenskur-veflesari-kynntur-til-leiks/

T7 – Forskriftir að röddum

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að gefa út opnar (e. open-source) forskriftir til hönnunar einingavals- og stikaðra talgervla. Í byrjun munu sérfræðingar vinna með raddupptökur sem þegar eru aðgengilegar (en ekki endilega opnar). Eftir því sem vinnu við upptökur nýrra radda vindur fram (T1 og T2) verða þær notaðar við þróun og í lokaútgáfu forskrifta verða gögnin sem notuð voru til þróunar einnig gefin út, til að aðrir hönnuðir geti endurskapað vinnuna.

Varða 6 – lýsing

M5: Ákveða hvað önnur útgáfa verður (Ossian og Merlin fyrir Talróm)
M6: Útgáfa annarrar útgáfu SPSS (Statistical Parametric Speech Synthesis)-forskriftar

Afurðir

Öllum markmiðum hefur verið náð.

1. Fastspeech2-forskriftin er uppfærð með uppfærslum í upprunalegu hirslunni[69].
2. Auk þess hefur MozillaTTS verið breytt til að hlaða Talrómsgögnunum auðveldlega inn til að nota öll talgervilskerfin sem eru í boði þar.

Þessi verkþáttur heldur ekki áfram en afurðir úr T13 verða gerðar aðgengilegar á svipaðan hátt.

Skýrsla

Þó að Máltækniskýrslan hafi lagt til Ossian eða Merlin fyrir þennan hluta ákváðum við að hvorugur kostir væri lengur við hæfi. Báðir nota eldri tækni og eru lítið sem ekkert studdir lengur og frekari þróun virðist hafa stöðvast. Áherslan í dag virðist nokkuð dreifð og ekkert eitt kerfi er augljós kostur til notkunar (eins og Kaldi fyrir talgreiningu). Valið á MozillaTTS var marksækin þar sem þróun í hirslunni er stöðug og inniheldur mörg bestu líkönin sem eru í boði.

Miklum tíma var endurúthlutað fyrir T5 í M6, þar sem vörpun einingavalsforskrifta yfir í Android-samhæft snið reyndist erfiðari en talið var. Vinnan fól í sér aðlögun Flite, C-byggðu talgervingarforriti, til að virka með íslenskum röddum hönnuðum í Festvox til notkunar með talgervilskerfinu Festival Speech Synthesis System. Kóðagrunnur Flite er talsvert illa skjalaður og nokkrum villum og óvæntri hegðun verður að taka á. Áframhaldandi vinna í T13 verður gerð

[69] https://github.com/ming024/FastSpeech2

innan hirslnanna sem taldar eru að ofan, svo að allar endurbætur verða gerðar aðgengilegar þegar þær eru að fullu útfærðar.

T8 – Mat á gæðum talgervla

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík, Hljóðbókasafn Íslands

Markmið

Að veita nauðsynlegan stuðning við þróun talgervla með því að meta frammistöðu fyrir ýmsar útfærslur talgervla. Eftir bakgrunnsrannsókn verða þrjár til fimm aðferðir valdar og innleiddar í hugbúnaði. Prófanir eru svo skipulagðar með því að undirbúa raddirnar, tímasetja prófanir og fá hlustendur. Niðurstöðurnar eru einnig metnar og greint frá þeim.

Varða 6 – lýsing

M5: Frumgerð vettvangs huglægs mats. Matsumferðir huglægs mags skipulagðar
M6: Vettvangur huglægs mats tilbúinn. Fyrstu umferð huglægs mats lokið. Kerfi til að meta niðurstöður tilbúið og í notkun.

Afurðir

Öllum markmiðum hefur verið náð. Hljóðbókasafn Íslands átti í fyrstu að útvega þátttakendur í fyrstu umferðir huglægs mats, en það hefur verið fært á þriðja ár fyrir mat á T7, T13 og L10 á stærri mælikvarða.

- Umhverfi huglægs mats LOBE (https://github.com/cadia-lvl/LOBE)
- Tvær umferðir huglægs mats (sjá eftirfarandi skýrslu og T13)
- Huglægt mat gert á gögnum T1. Grein: Sigurgeirsson, A., et al. (2021): Talrómur: A large Icelandic TTS corpus. *NoDaLiDa*. (https://aclanthology.org/2021.nodalida-main.50/)

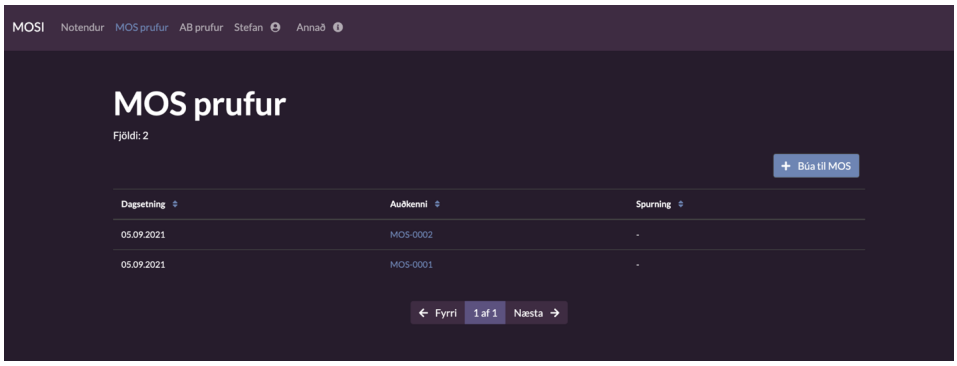
Verkþáttur T8 heldur áfram á þriðja ári.

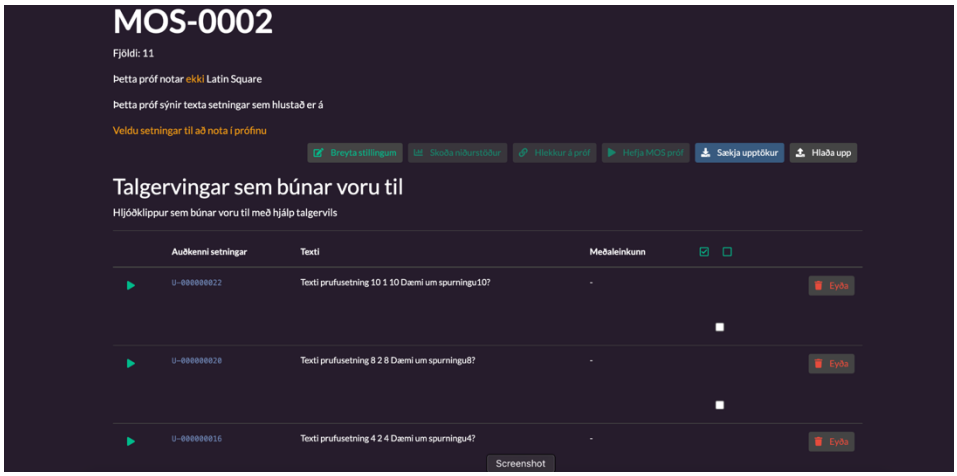
Skýrsla

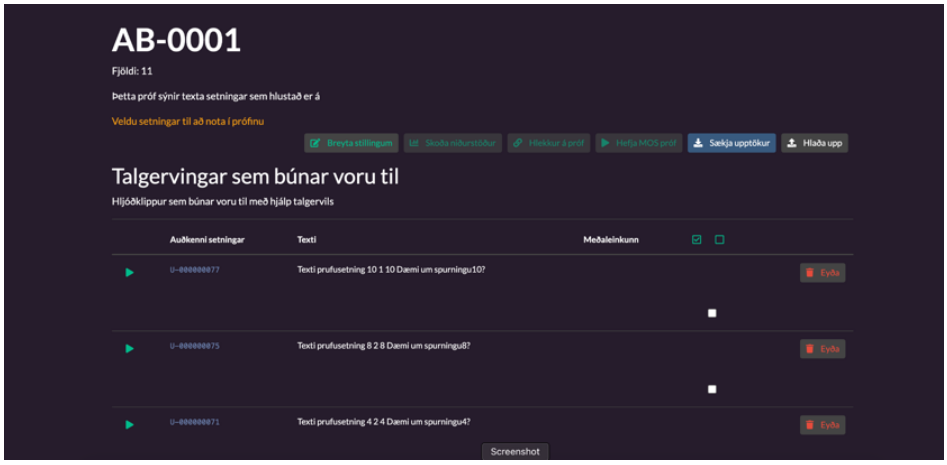
LOBE hefur verið í stöðugri þróun ásamt því að villur hafa verið lagaðar og annað tengt viðhaldi á forritinu. Virkni mats hefur verið endurbætt með viðbót MOS-prófana, ítarlegri tölfræði og síðu fyrir niðurstöður.
Samskiptaumhverfi var útfært fyrir sumarnemendur til að eiga samskipti og deila framvindu, með það að markmiði að gera vinnuna skemmtilegri og auka skilvirkni. Samanburður á skilvirkni sumarnemenda yfir tvö ár virðist benda til þess að hún aukist þegar vinnan er gerð að leik.

Vinna við nýtt sjálfstætt matsumhverfi sem kallast MOSI er komin í gang. Þetta snýst aðallega um að taka matshlutann úr LOBE og gera matsumhverfið að algjörlega aðskildri einingu. Umhverfið verður tilbúið í lok október þetta ár og gefið út á GitHub, með skjölum til að setja umhverfið upp, leiðbeiningum fyrir þróun og keyrslu mats í umhverfinu. Líkt og LOBE gerir

umhverfið einfalda tölfræðigreiningu á niðurstöðum, sem nægja í flestum tilfellum. Notendur geta séð MOS (Mean Opinion Scores, meðalmatseinkunn) eftir þátttakendum, segðum, og mikilvægast, eftir líkani/afbrigði. A/B-prófanir eru einnig útfærðar til að prófa smávægilegan mun gæða milli líkana.
MOSI virkar eins og LOBE þar sem notendur senda inn hljóðbúta sem þeir vilja meta með MOS-, AB- og ABX-prófunum. Það er möguleiki á viðbót annarra matsaðferða í framtíðinni ef þurfa þykir. MOS-prófanir eru tilbúnar en enn er unnið við AB- og ABX-prófanir.







T9 – Forvinnsla texta, textanormun og áherslugreining

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

Markmið

Að eiga textanormunarkerfi fyrir íslensku sem undirbýr texta fyrir hljóðritun. Þetta felur í sér að skrifa út tölur og dagsetningar, skrifa út skammstafanir o.s.frv., 4. verður t.d. fjórði. Íslenska er áskorun þegar kemur að normun texta, sérstaklega vegna fallbeygðra talna: höfuðtölur frá 1-4 og allar raðtölur beygjast eftir falli og kyni. Þetta þýðir að tölur geta haft allt að 15 mismunandi framburðarmyndir eftir samhengi. Textanormun hefur verið meðhöndluð með stöðuferjöldum (e. finite state transducers) og tauganetum. Við munum kanna báðar aðferðir fyrir íslensku. Mjög nákvæm normun er talgervlinum ákaflega mikilvæg, þar sem rangar framburðarmyndir talna og annarra óstaðlaðra orða (e. non-standard words) virka mjög pirrandi fyrir þann sem hlustar.

Varða 6 – lýsing

M5: Þjálfunargögn tilbúin, gerð normunarreglna lokið.
M6: Textanormunarkerfi aðlagað til að keyra á Android-tækjum og sem hluti af framendapípu á vefþjóni.

Afurðir

- Þjálfunargögn fyrir textanormun (http://hdl.handle.net/20.500.12537/153)
- Almennt textanormunarkerfi fyrir íslensku, byggt á reglulegum segðum og málfræðilegri mörkun (https://github.com/cadia-lvl/regina_normalizer)
- Textanormun sem hluti af þjónustu framenda talgervils (https://github.com/grammatek/tts-frontend-service)
- Textanormunarkerfi sem hluti af TTS-appi á Android (sjá verkþátt T5)
- Útgefin grein: Sigurðardóttir, H.S., Nikulásdóttir, A.B., og Guðnason, J. (2021): Creating Data in Icelandic for Text Normalization. Í *NoDaLiDa* (pp. 404-412).

Öllum markmiðum var náð.
Verkþáttur T9 heldur áfram á þriðja ári.

Skýrsla

Við kynnum Regínu, reglubyggt kerfi sem getur sjálfvirkt normað gögn fyrir talgervilskerfi. Við mörkuðum handvirkt fyrstu normuðu málheildina á íslensku, 40.000 setningar, og þróuðum

Regínu, TN-kerfi byggt á reglulegum segðum. Nýja kerfið nær 89,82% nákvæmni miðað við handvirkt mörkuðu málheildina á óstöðluðum orðum og sýndi marktæka framför í nákvæmni borið saman við eldri normunarkerfi fyrir íslensku.

Textanormun er nú óaðskiljanlegur hluti talgerviskerfis. Ótakmarkaðir inntakstextar geta innihaldið svokölluð óstöðluð orð (e. non-standard words (NSWs)), sem er ómögulegt fyrir tölvu að lesa án þess að breyta í venjulega strengi bókstafa og greinarmerkja. Þessum óstöðluðu orðum er skipt í táknfræðilega flokka sem eru m.a. styttingar, tölur og sérstafir.

Mikilvægi textanormunar fyrir talgervla er ekki augljós þó að notagildi hennar sé þekkt. Flest orð þarf ekki að norma, og því eru normuð gagnasöfn og ónormaðar hliðstæður þeirra næstum eins. En ef óstöðluð orð eru ekki þanin (e. expanded) stekkur talgervilskerfi yfir þau orð sem gerir textann ónákvæman og ófullkominn.

Til skýringar skulum við líta á dæmi um setningu fyrir og eftir normun.

```
Hæsti tindur Esjunnar er 914 m. (Esjan's highest peak is 914m.)

Hæsti tindur Esjunnar er níu hundruð og fjórtán metrar. (Esjan's
highest peak is nine hundred and fourteen meters.)
```

Aðferðafræði

Íslenska er tungumál með beygingarkerfi, þar sem hvert orð getur haft margar myndir eftir samhengi. Til dæmis væri hægt að þenja töluna 2 sem *tveir, tvo, tveimur, tveggja, tvær,* eða *tvö*, eftir falli næsta orðs. Raðtalan *2.* getur svo verið *annar, annan, öðrum, annars, önnur, aðra, annarri, annarrar, annað, öðru, annars, aðrir, annarra*, eða *aðrar*. Aðeins fyrstu fjórar tölurnar (einn, tveir, þrír, og fjórir) hafa þessa fjölbreytni beyginga.

Kerfið byggt fyrir þessa rannsókn notar reglulegar segðir og málfræðireglur til að meta hvernig skal þenja orð. Því hefur verið gefið nafnið Regína. Fyrsta skref Regínu er að keyra reglur fyrir þenslu styttinga, mælieininga, peninga, vefhlekkja, og rómverskra talna í gegnum ónormaða textann. Reglurnar fyrir mælieiningar taka á forsetningum. Til dæmis gæti þetta hjálpað þegar grunnútgáfan af *km* er *kílómetrar*. Ef við segjum *til 2 km* notar Regína forsetninguna *til* til að þenja orðið í eignarfalli, *kílómetra*. Næsta skref er að keyra þennan þanda texta í gegnum málfræðilegan markara. Í stað þess að lesa *km* sem styttingu (og gefa mark sem slíka) greinir markarinn nú orðið *kílómetra* sem og að það sé í eignarfalli. Nú geymir Regína málfræðileg mörk fyrir hvert orð. Næst er táknfræðilegur flokkur eftirstandandi óstaðlaðra orða greindur. Reglur fyrir tölur gilda fyrir höfuð- og raðtölur, tugabrot og almenn brot. Í þessu skrefi skoða orðin sem merkt eru sem tölur mark næsta orðs. Tölum sem koma ekki á undan lýsingarorði eða nafnorði eru gefin sjálfgefið fall. Lokaskrefið í kerfinu er að keyra textann í gegnum reglur fyrir aðra táknfræðilega flokka: tíma, niðurstöður íþrótta, tölustafi, bókstafi, dagsetningar, og tákn. Til samanburðar var normaða textanum endurraðað með handvirkt merkta textanum, þar sem hverri setningu og hverju orði raðað saman til að halda skipulaginu greinilegu.

SEMIOTIC CLASS	ACCURACY [%]	# examples
ALL	99.51	729,673
PLAIN	99.94	626,541
CARDINAL	86.87	8,456
ORDINAL	87.24	1,653
DIGIT	51.45	241
DECIMAL	74.36	197
FRACTION	33.33	39
LETTERS	96.05	3,576
ABBREVIATIONS	80.72	1,675
ROMAN NUMERALS	33.66	104
MONEY	46.89	352
WLINK	99.14	348
MEASURE	61.96	1,559
TIME	80.36	713
DATE	97.75	7,937
SYMB	88.36	1,735
PUNCT	99.93	74,547

Table 1: Results for general news

SEMIOTIC CLASS	ACCURACY [%]	# examples
ALL	98.45	12,106
PLAIN	99.98	8,923
CARDINAL	96.84	538
ORDINAL	91.89	74
DIGIT	0.0	1
DECIMAL	0.67	3
FRACTION	0.0	1
LETTERS	99.06	106
ABBREVIATIONS	75.0	20
WLINK	100.0	1
MEASURE	60.0	5
TIME	100.0	2
DATE	88.4	43
SYMB	90.91	88
SPORT	84.55	893
PUNCT	100.0	1,408

Table 2: Results for sports news

Niðurstöður

Gagnasettið með almennum fréttum hafði 729.763 orð, og 701.088 þeirra þurftu ekki normun. Grunnlína kerfisins án nokkurrar vinnu var því 96,08%. Þeim 28.675 orðum sem eftir voru var skipt í höfuð-, raðtölur, tugabrot, tölustafi, almenn brot, stafarunur, styttingar, vefhlekki, mælieiningar, tíma-, dagsetningar, og tákn. Nákvæmni og stærð hvers flokks koma fram fyrir almennar fréttir og íþróttir í töflunum til vinstri.

Regína virkar vel og skilar ekki villandi niðurstöðum. Handvirkt merktu gögnin urðu óhjákvæmilega að þróunargagnasetti, þar sem það er alltaf sýnilegt hönnuði Regínu. Hins vegar er þetta eingöngu vandamál við að bera kerfið saman við málheildina. Regína mun vera notuð til að norma texta fyrir talgervla. Þó að tiltekin þensla orðs breytist eftir aðilum merkir það ekki ranga normun.

Kerfinu er lýst ítarlegar í Sigurðardóttir (2021)[70].

Normaranum, upprunalega forritaður í Python, var breytt yfir í Java til að keyra á Android-tækjum (sjá verkþátt T5). Útfærsla normarans í talgervilskerfi í notkun hefur gengið vel, raunveruleg gögn hafa skilað nýjum prófunardæmum og við höfum samhliða lagað minniháttar villur. Einnig er normarinn aðlagaður og samþættur í pakka framendaþjónustu talgervils sem má keyra á þjóni til að undirbúa inntak fyrir talgervilsvél.

[70] Sigurðardóttir, H. S., Nikulásdóttir, A. B., og Guðnason, J. (2021). Creating Data in Icelandic for Text Normalization. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 404-412).

T10 – Sjálfvirk hljóðritun

Skýrsla: Grammatek
Aðilar: Grammatek, Háskólinn í Reykjavík

Markmið

Að eiga einingu sem umritar normaðan texta sjálfvirkt yfir á hljóðritunartákn. Annars vegar verður einingin tengd talgervilseiningu til að hljóðrita sjálfkrafa óþekkt orð sem koma inn. Hins vegar má nota eininguna til að útbúa nýjar orðabækur, t.d. ef of mörg orð vantar frá sérhæfðu sviði í kjarnaorðabók. Þessi verkþáttur er mjög nátengdur G6, þróun framburðarorðabókar.

Varða 6 – lýsing

M5: Sjálfvirkar aðferðir atkvæðaskiptingar og áherslumerkingar útfærðar og prófaðar
M6: g2p-kerfi, með atkvæðaskiptingu, er útfært til að keyra á Android-tækjum og sem hluti af framendapípu á vefþjóni.

Afurðir

- Endurbætt LSTM-líkön fyrir g2p (https://github.com/grammatek/g2p-lstm)
- g2p-kerfi sem inniheldur atkvæðaskiptingu og áherslumerkingu, Python-einingu sem hluta af tts-frontend-service: https://github.com/grammatek/tts-frontend-service
- g2p-kerfi, með atkvæðaskiptingu og áherslumerkingu sem hluti af framendapípu talgervils virkar í Android, sjá verkþátt T5

Öllum markmiðum var náð.
Verkþáttur T10 heldur áfram á þriðja ári.

Variant	M3	15.000	20.000	25.000
NORTH	29.20	26.00	26.30	**25.50**
NORTH	4.57	3.83	4.07	**3.88**
NORTH (corr.)	23.75	15.83	14.53	**13.33**
NORTH (corr.)	3.65	2.40	2.31	**2.06**
NORTHEAST	N/A	27.86	27.15	**26.75**
NORTHEAST	N/A	4.36	4.19	**4.10**
NORTHEAST (corr.)	26.75	15.93	15.83	**13.83**
NORTHEAST (corr.)	4.05	2.63	2.47	**2.14**
SOUTH	N/A	22.30	21.80	**20.10**
SOUTH	N/A	3.57	3.35	**3.06**
SOUTH (corr.)	21.14	15.83	14.33	**13.33**
SOUTH (corr.)	3.36	2.32	2.20	**2.04**
STANDARD	26.10	21.90	22.00	**20.70**
STANDARD	4.07	3.40	3.41	**3.14**
STANDARD (corr.)	22.04	14.93	14.43	**12.53**
STANDARD (corr.)	3.30	2.30	2.22	**1.93**
SIGMORPHON	**7.33**	8.67	8.89	8.22
SIGMORPHON	**1.54**	1.96	2.00	1.84
SIGMORPHON (corr.)	8.00	2.22	2.89	**2.00**
SIGMORPHON (corr.)	1.43	0.45	0.57	**0.41**

Table 1: Model scores, in terms of WER (upper row) and PER (lower row).

Nokkur g2p-líkön voru þjálfuð með gögnum úr framburðarorðabókinni (verkþætti G6) og með sömu stillingum LSTM kóðara-afkóðara eins og g2p-taugalíkönin sem lýst er í M3. Niðurstöðurnar lofa góðu, líkön þjálfuð á 15.000, 20.000 og 25.000 handahófskenndum færslum ná stöðugt betri árangri yfir öll framburðartilbrigði, borið saman við grunnlíkan okkar úr M3 (þjálfað á ~5.700 orðum.) Stækkun þjálfunarsettsins í 30.000 færslur skilaði ekki betri árangri.

Öll líkön voru prófuð á sömu prófunarsettum og voru notuð í M3, þ.m.t. prófunarsettið sem kom frá ráðstefnu Sigmorphon 2020. Sigmorphon-settið skilar betri niðurstöðum heilt yfir þar sem prófunarsett okkar voru gerð til að dekka víða dreifingu samsetninga rittákna og er því erfiðara viðureignar fyrir g2p-líkan.

Á meðan á prófunum stóð leiddi handvirk skoðun villuúttaks í ljós að nokkrar eftirstandandi villur voru vegna mismunandi umritunaraðferða, þar sem sumir leiðarvísar okkar hafa breyst síðastliðið ár. Prófunarsettin hafa því verið handvirkt „leiðrétt," sem leiddi til frekari endurbóta og sanngjarnari samanburðar.

Einhver vinna hefur einnig verið lögð í g2p-umritun enskra orða. Eins og lýst var í verkþætti G6 var reglubyggt tól sérhæft til að umrita ensk orð samkvæmt íslenskum hljóðritunarreglum útfært með Thrax-málfræði. Með því að nota þetta reglubyggða tól og ferli handvirkra leiðréttinga voru búin til gullgögn af enskum orðum bornum fram á íslenskan máta, og þau svo notuð til að reyna að þjálfa LSTM-líkön eins og þau sem lýst var að ofan.

LSTM g2p models

	1	2	3	4	5	6	7	8	Mean
WER(1)	34.80	35.24	**31.5**	32.16	34.14	35.46	35.68	40.75	34.97
PER(1)	9.79	9.20	**8.51**	9.27	8.79	10.22	9.46	11.62	9.61
WER(2)	88.55	88.99	87.67	**86.56**	90.31	89.43	88.99	90.53	88.88
PER(2)	40.87	41.78	40.05	40.31	42.70	**38.74**	41.50	42.45	41.05

Rule-based g2p systems

	1	2	3	4	5	6	7	8	Mean
WER(1)	45.15	46.48	51.10	**42.07**	46.48	43.61	46.48	51.76	46.64
PER(1)	12.49	12.67	14.26	**11.72**	12.13	12.00	12.31	14.56	12.77
WER(2)	90.90	91.63	90.31	**89.21**	90.97	91.41	92.07	92.29	91.11
PER(2)	41.15	41.58	41.26	40.73	42.38	**39.86**	41.31	41.97	41.28

Table 2: WER and PER scores of systems applied to the same eight test sets of English words. Top table shows WER and PER scores for our LSTM g2p models, (1) compared with simply using an available LSTM g2p model trained on Icelandic data (2). Bottom table shows scores of our rule-based system (1), compared with the existing rule-based system for Icelandic g2p (2).

Þó að þessar reglur séu enn ekki nálægt fyrsta flokks reglum fyrir íslenskt (eða enskt) g2p eru bæði reglubyggða tólið og LSTM töluverð endurbót með því einfaldlega að leyfa íslenskum g2p-líkönum að meðhöndla ensk orð.

Atkvæðaskiptingu og áherslumerkingu var bætt í framendavinnslu talgervilsins samkvæmt aðferðafræði sem lýst var í M3. Niðurstöðurnar sýna ~25% villutíðni, að mestu vegna þess að greiningu samsettra orða vantar, sem verður bætt við á þriðja ári.

T13 – Stikaðir talgervlar

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að eiga stikaðan talgervil fyrir íslensku sem byggir á nýjustu aðferðum. Vinna á öðru ári mun fela í sér ítarlegar rannsóknir, tilraunir og undirbúningsvinnu fyrir þróun hágæðakerfis.

Ákvæði: (Upptaka íslenskra radda í T1 og T2, frammistaða talgervils í T8, forvinnsla texta í T9)

Varða 6 – lýsing

M5:
Uppsetning málfanga fyrir íslensku fyrir SPSS. Útgefinn hugbúnaður nýjustu gerðar SPSS-aðferðar fyrir íslensku. Skýrsla og hönnun um hvernig skuli stjórna hljómfalli og áherslu.

M6:
FastSpeech2-afturvirkni útfærð fyrir íslensku og ensku með mismunandi eiginleikum úttaks. Samstæðir talgervlar útfærðir fyrir íslensku og ensku. Skýrsla um gæðamat fyrir þá bestu og nýjar nálganir við SPSS.

Afurðir

Í M5 náðum við markmiðum okkar með prófunum á sex fyrsta flokks talgervlum fyrir bæði íslenska mælendur (Talrómur[71]: Álfur (Mælandi F) og Diljá (Mælandi C)) og enska mælendur (LJSpeech[72]: Linda (F)). Fjórir Mixed-Excitation-byggðir talgervlar (STRAIGHT[73], WORLD[74], Griffin-Lim[75] og MagPhase[76]) og einn taugabyggður talgervill (WaveNet[77]). Til að talgerva hvern talbút notuðum við aðferðarfræði afritstalgervingar (e. copy synthesis) fyrir hvern talgervil og mátum frammistöðu þeirra, bæði með hlutlægu (PESQ[78] og Visqol[79]) og huglægu mati, meðalmatseinkunn (MOS).

Eftirfarandi hirsla inniheldur hlekki á rannsóknarritgerðir og upprunalegar Git-hirslur allra talgervla sem við prófuðum fyrir ensku og íslensku:
https://github.com/cadia-lvl/Vocoders-state-of-the-art/blob/master/README.md

---

[71] Talrómur – https://repository.clarin.is/repository/xmlui/handle/20.500.12537/104
[72] LJSpeech – https://keithito.com/LJ-Speech-Dataset/
[73] STRAIGHT – https://github.com/HidekiKawahara/legacy_STRAIGHT
[74] WORLD – https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder
[75] Griffin-Lim – https://github.com/bkvogel/griffin_lim
[76] MagPhase – https://github.com/CSTR-Edinburgh/magphase
[77] WaveNet – https://github.com/r9y9/wavenet_vocoder
[78] PESQ – https://www.itu.int/rec/T-REC-P.862
[79] Visqol – https://github.com/google/visqol

Hlekkur á meistararitgerð G. Fang þar sem rætt er um þessa mismunandi talgervla: https://skemman.is/handle/1946/39414

Í M6 náðum við ekki markmiði okkar, sem var að útfæra talgervla sem við höfðum prófað í M5 í Fastspeech2-líkanið. í fyrstu töldum við nokkuð einfalt að breyta Fastspeech2, en sáum seinna hversu erfitt var að aðgreina talgervilhlutann frá Fastspeech2-líkaninu. Fyrir M7 viljum við prófa önnur talgervilslíkön eins og Tacotron 2[80], þar sem rannsakendur hafa þegar skilið talgervil frá kjarnalíkaninu. Við teljum mjög líklegt að þetta náist fyrir 31. desember.

Verkþáttur T13 heldur áfram á þriðja ári.

Skýrsla

M5

Fyrir afritstalgervingu þarf ekki að þjálfa Mixed-Excitation-byggða talgervla fyrir notkun, en það þarf fyrir tauganetsbyggða talgervla. Við þjálfuðum tvö líkön með Wavenet, eitt fyrir LJSpeech (12.980 talbútar) og Mælanda F (19.809 talbútar). Við þjálfuðum LJSpeech-líkan okkar (Wavenet EN) með 340.000 skrefum með 2.032 þjálfunartapi og samsvarandi þróunartapi, 2.204. Í öðru WaveNet-líkani okkar (Wavenet IS) notuðum við fyrsta líkanið okkar sem grunn eða viðmið og þjálfuðum að auki 220.000 skref með íslenskri málheild Mælanda F, sem skilar þjálfunartapi 1.279 og þróunartapi 1.624.

Úttak úr afritstalgervingu fyrir hvern talgervil er metið með Visqol, PESQ og MOS. Matsaðferðir Visqol og PESQ eru kóðabyggðar, á meðan mat MOS þarfnast mannlegra hlustenda.

Hlutlægt mat (PESQ og Visqol)

Fyrir talgervla byggða á Mixed-Excitation mátum við 1000 talgervða talbúta fyrir hvern mælanda, og fyrir taugabyggðan talgervil mátum við 10 talgervða talbúta fyrir hvern mælanda. Við völdum bara 10 talgervða talbúta frá taugabyggðum talgervli af því að það tók hvert líkan um 45 mínútur að búa til 70 sekúndur af talgervðu tali.

Viscol er hlutlæg matsaðferð, sem ber saman Mel-rófsmynd milli grunnsannleiks (e. ground truth) tals og tilbúins tals. Skali matsins er á bilinu 0 – slæmt til 5 – náttúrulegt (grunnsannleikur).

		STRAIGHT	WORLD	Griffin-Lim	MagPhase	Wavenet EN	Wavenet IS
LJspeech	Mean	4.295	4.263	4.377	**4.7320**	**4.451**	3.500
	Std	0.083	0.078	0.088	0.0002	0.052	0.070
Speaker F	Mean	4.127	4.132	4.269	**4.7320**	3.268	**4.239**
	Std	0.120	0.115	0.119	0.0004	0.232	0.110
Speaker C	Mean	**4.330**	**4.314**	**4.421**	**4.7320**	3.337	3.973
	Std	0.106	0.105	0.084	0.0003	0.461	0.115

Taflan að ofan sýnir að MagPhase-talgervillinn nær hæstu einkunn fyrir alla mælendur með einkunnina 4,7320 ± 2×10^{-4}.

[80] Tacotron2 – https://github.com/NVIDIA/tacotron2

PESQ er önnur hlutlæg matsaðferð og helsti tilgangur hennar er að meta hlutfall óhljóða og tals, en rannsakendur hafa einnig notað það fyrir mat á talgervlum[81]. Skali matsins er á bilinu -0,5 – slæmt til 4,5 – náttúrulegt.

		STRAIGHT	WORLD	Griffin-Lim	MagPhase	Wavenet EN	Wavenet IS
LJspeech	Mean	3.388	3.398	3.542	4.499	**3.818**	2.648
	Std	0.135	0.124	0.158	0.006	0.047	0.050
Speaker F	Mean	3.238	3.336	3.458	**4.500**	1.936	**3.453**
	Std	0.157	0.144	0.120	0.005	0.534	0.185
Speaker C	Mean	**3.513**	**3.409**	**3.633**	**4.500**	2.286	3.282
	Std	0.129	0.148	0.124	0.001	0.410	0.106

Taflan að ofan sýnir að MagPhase-talgervillinn nær hæstu einkunn fyrir alla mælendur með einkunnina 4,500 $\pm 5 \times 10^{-3}$.

Huglægt mat (MOS)

Fyrir bæði Mixed-Excitation- og tauganetsbyggða talgervla mátum við fimm talgervðar ræður fyrir hvern mælanda. Í MOS-mati okkar völdum við 12 þátttakendur, fjóra kvenkyns og átta karlkyns, þar sem helmingur hefur unnið á sviði vinnslu náttúrulegra tungumála og þekkja til MOS-prófana.

MOS er huglægt mat, þar sem talgæði eru metin af manneskjum með náttúrulegum tölum frá 1 til 5, þar sem 1=slæmt, 2=lélegt, 3=ásættanlegt, 4=gott og 5=frábært.

		STR	WOR	GL	MP	WN EN	WN IS
LJspeech	Mean	4.456	4.080	3.955	**4.674**	4.366	2.640
	Std	0.316	0.211	0.307	0.190	0.392	0.262
Speaker F	Mean	4.020	3.762	3.492	4.510	1.728	**4.712**
	Std	0.450	0.206	0.264	0.357	0.308	0.117
Speaker C	Mean	3.674	3.395	3.225	**4.298**	1.546	3.528
	Std	0.430	0.145	0.156	0.206	0.251	0.400

Samkvæmt mennskum hlustendum náði MagPhase hæstu einkunn fyrir LJspeech (4.6740.190) og Mælandi C (4.2980.206), fyrir Mælanda F, Wavenet IS-líkanið náði besta árangri með einkunnina (4.7120.117).

M6

Við ákváðum að innleiða talgervilinn WORLD inn í FastSpeech2[82], þar sem við höfðum þegar útfært bæði FastSpeech2- og WORLD-talgervlana fyrir íslenska og enska mælendur. Í þessu ferli lentum við í ýmsum tæknilegum erfiðleikum. Má þar nefna uppsetningu almennilegs umhverfis fyrir Fastspeech2 og högun fyrir Fastspeech2+WORLD. Betri nálgun gæti verið að skoða hvað aðrir rannsakendur hafa unnið að með önnur talgervilskerfi áður en við leggjum alla áherslu á Fastspeech2-líkanið.

[81] Wavenet – https://www.csd.uoc.gr/~nagaraj/papers/WaveNet.pdf
[82] FastSpeech2 – https://github.com/ming024/FastSpeech2

Samstarf við atvinnulífið

Skýrsla: Grammatek
Aðilar: Grammatek, allir aðilar SÍM

Markmið

Að undirbúa samstarf við atvinnulífið/viðskiptaaðila á seinni stigum máltækniverkefnisins.

Skýrsla

Verkefni samstarfs við atvinnulífið hefur verið nálgað bæði með ytri og innri leiðum. Helsta markmiðið var að auka vitund á máltækniverkefninu meðal fyrirtækja og stofnana á Íslandi, leitast eftir beinu sambandi og ræða möguleika máltækni í fjölbreyttu samhengi. Öll fyrirtæki eða opinberar stofnanir á Íslandi sem gætu hugsað sér að hafa íslensku í tæknilegum lausnum ættu að vita af þeirri vinnu sem fer fram innan máltækniverkefnisins.

Tengslamyndun

Myndun tengsla við fyrirtæki á Íslandi og erlendis eða undirbúningur tengsla á hærra stigi við stór alþjóðleg fyrirtæki hefur verið mikilvægur hluti vinnunnar á öðru ári. Fjölmargir fjarfundir fóru fram, þar sem faraldurinn útilokar enn fundi augliti til auglitis að mestu leyti þar til nýlega. Tengslin voru annaðhvort á víðum grunni, þar sem almenn vinna SÍM var kynnt, eða hnitmiðaðri á verk ákveðinna hópa. Í þeim tilfellum voru meðlimir hópa hafðir með í viðræðum til að auðvelda bein samskipti milli hópanna og hugsanlegra samstarfsaðila í atvinnulífinu. Við héldum einnig áfram viðræðum varðandi sérþarfir: sjón- og heyrnarskerta, börn með sérþarfir í sambandi við samskipti o.s.frv.

Alþjóðleg fyrirtæki

Frá hlið SÍM felst undirbúningur fyrir tengsl við alþjóðlega aðila nú í skilgreiningu og söfnun efnis sem gæti vakið áhuga stórra alþjóðlegra máltæknifyrirtækja. Stóru málgagnasöfnin, merkingarfræðigögnin, einingar ákveðinna tungumála og málheildir eru þær afurðir sem nýtast hvað mest fyrirtækjum sem þegar hafa eigin kerfi, og leitast við að bæta íslensku í safnið sitt. Ráðuneytið og Almannarómur munu undirbúa og sjá um bein samskipti við stóru fyrirtækin, SÍM mun veita upplýsingagögn og taka þátt í viðræðum á seinna stigi.

Við höfum hins vegar þegar verið í sambandi við tvö alþjóðleg fyrirtæki sem sérhæfa sig í talgervlum: Acapela og SpeakUnique. Þessir tengiliðir leggja áherslu á raddbanka fyrir íslensku og við höfum leitt þessar viðræður fyrir hönd þeirra sérstöku notendahópa sem þurfa þessa þjónustu. Niðurstaða þeirrar vinnu þarf þó að vera í höndum samtaka eða stofnana sem munu nota þjónustuna fyrir hönd skjólstæðinga sinna.

Almennt séð teljum við að stór og fjölbreytt talgagnasett og meðfylgjandi gögn og einingar sem máltækniverkefnið hefur skilað verði sérstaklega mikils virði þegar kemur að samningum eða auglýsingu verkefnisins á alþjóðlegum grunni.

Máltækniráðstefna

Almannarómur og SÍM héldu ráðstefnu í maí sem sérstaklega beindist að íslensku atvinnulífi. Ráðstefnan var hálfs dags fjarviðburður, með beinu streymi á mikilvægum útsendingarvefsíðum. Meðlimir SÍM undirbjuggu þó nokkra fyrirlestra í beinu streymi, en meirihluti undirbúningsins var á formi 13 myndbanda. Þessi myndbönd voru öll fagmannlega framleidd með gaumgæfilega skrifuðum handritum. Hóparnir innan SÍM fengu það verk að kynna kjarnaverkefnin og hinn ýmsa hugbúnað til hugsanlegra notenda í atvinnulífinu og opinbera geiranum. Þó nokkrir hópar voru með fólk frá fyrirtækjum/stofnunum í myndböndum sínum, sem sýnir fram á hið nána samstarf sem þegar á sér stað og ýtir undir mikilvægi máltækni á ýmsum sviðum. Nýlega voru öll myndböndin gefin út á YouTube-rás: https://www.youtube.com/watch?v=n6Nj1vD2arg&list=PLkcNpSDhybnuRf3gOkcUEFBkHoANiz3rA .

Það varð eftirtektarverð aukning á áhuga frá atvinnulífinu eftir ráðstefnuna og heilt yfir er gríðarlegur munur á því hvernig fyrirtæki horfa á vægi máltækni fyrir starfsemi sína og forrit sín samanborið við byrjun verkefnisins árið 2019.

Leiðbeiningasíður

Safn leiðbeininga hefur verið útbúið og verður gefið út á upplýsingasíðu SÍM á næstu vikum: https://icelandic-lt.gitlab.io/. Þetta efni er viljandi útbúið af aðila sem vinnur ekki á sviði máltækniþróunar. Þetta ýtir undir skjölun afurða og tólin eru keyrð og prófuð af höfundi leiðbeininganna. Leiðbeiningarnar eru ætlaðar forriturum almennt og við vonum að þær lækki þröskuldinn fyrir fólk sem langar að „fikta“ með máltækni.

Næstu skref

Á þriðja ári verður enn meiri áhersla lögð á samstarf við atvinnulífið og tækniyfirfærslu. Við teljum að grunnurinn sé fyrir hendi: vitund og áhugi atvinnulífsins og opinbera geirans færist í aukana á Íslandi, tengsl og tengslanet eru að myndast, við erum þegar að vinna með fyrirtækjum og opinberum aðilum að hugbúnaðarverkefnum (sjá t.d. V4 og L7 í skýrslunni). Einnig eru verkefnið og afurðir kjarnaverkefna stöðugt að verða þekktari og aðgengilegri með auglýsingum og leiðbeiningum, og það sem mestu skiptir, að gæði gagna og tóla eru í mörgum tilfellum tilbúin í útfærslu í tilraunaverkefni iðnaðarins. Vinna við tækniyfirfærslu mun því halda áfram samfellt á þriðja ári.