Samstarf um íslenska máltækni
SÍM

Máltækniáætlun fyrir íslensku

Varða 4 – Lokaskýrsla

1. febrúar 2021

Efnisyfirlit

Formáli

Fyrirvari: Þessi texti er þýðing úr enskri útgáfu M4-skýrslunnar og skal hafa ensku útgáfuna til viðmiðunar ef misræmi er á milli textanna.

Þessi skýrsla dregur saman við Vörðu 4 (M4) afurðir fyrir hvern verkþátt sem skilgreindur var í samningnum milli Almannaróms og SÍM (Samstarf um íslenska máltækni) frá júní 2020 og í *Endurskoðuðum vörðum* (e. Revised Milestones) fyrir annað verkefnisár (Y2) frá nóvember 2020.

Í öllum verkþáttum, utan tveggja, var markmiðum M4 náð að fullu eða með smávægilegum frávikum, og verkþáttunum þremur sem var seinkað fyrir M3 hefur verið lokið og afurðir þeirra afhentar. Eins og tekið er fram í köflum um verkþætti T2 og H3 þurfa þessi undirverkefni af mismunandi ástæðum meiri tíma til að afhenda áætluð talgögn. Hins vegar hafa áætlanir verið gerðar til að ná almennum markmiðum þessara verkefna fyrir M5 og M6 og því lítum við svo á að þau séu ekki á eftir áætlun.

Í byrjun janúar var stofnuð ný vefsíða fyrir SÍM á https://icelandic-lt.gitlab.io/ (enskri útgáfu verður fljótlega bætt við). Tilgangur þessarar síðu er að safna öllum afurðum verkefnisins á einn stað, sem felur í sér speglun allra GitHub-hirslna. Vefsíðan inniheldur einnig upplýsingar um alla þátttakendur í SÍM og tengiliði þeirra, auk ritaskrár ritrýndra vísindagreina sem gefnar eru út af rannsakendum verkefnateyma. Stefnt er að því að ná til fólks í hugbúnaðarþróun og rannsakenda sem eru að leitast eftir nákvæmum tæknilegum upplýsingum um verkefnið, m.a. hirslur fyrir kóða. Að hafa allar hirslur á sama stað auðveldar líka samskipti innan SÍM.

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. Að auki, í samhengi við þá vinnu sem nýlega ráðinn atvinnulífstengill SÍM hefur unnið, munum við skoða verkefnið í heild sinni til að greina mögulegar breytingar á verkþáttum, sérstaklega fyrir þau ár verkefnisins sem eru framundan.

Í janúar stóð Samrómur aftur fyrir lestrarkeppni fyrir grunnskóla til að safna talgögnum. Viðbrögðin voru vonum framar, og leiddu til 790.000 nýrra setningaa í Samróms-gagnasafninu – samanborið við 320.000 sem var safnað frá október 2019 til janúar 2020. Skýrslan um undirverkefni H1 fjallar nánar um þetta.

Eitt vandamál sem hefur ekki verið leyst er geymsla mjög stórra gagnasafna, líkt og talmálheilda og stórra mállíkana, þar sem aðgangur er opinn almenningi. CLARIN getur ekki hýst gögn af slíkri stærðargráðu og verður því að finna aðra lausn, með hýsingu og aðgang að þessum gögnum til frambúðar í huga, lengur en sem svarar til tímaramma núverandi verkefnis.

Líkt og kemur fram í M3 skýrslunni, leggjum við til að afhending Vörðu 5 verði í formi kynningarfundar fyrir meðlimi stjórnar Almannaróms og fulltrúa ráðuneyta, sem svipar til afhendingar M2 í júní 2020.

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 4 - lýsing

Búinn til listi yfir samstarfsfúsa útgefendur. Skýrsla um nýja texta sem safnað er frá fyrri gagnagjöfum. Þrjár óðalsbundnar (e. domain-specific) málheildir skilgreindar.

Afurðir

Öllum markmiðum vörðunnar var náð. Útbúinn hefur verið listi yfir 12 samstarfsfúsa útgefendur, sjá skýrslu að neðan, sem og skýrsla um nýja texta sem safnað hefur verið frá fyrri gagngjöfum. Við höfum skilgreint þrjár undirmálheildir: Alþingismálheild, lagamálheild og fréttamálheild.

Skýrsla

Listi yfir samstarfsfúsa útgefendur

Stjórn Félags íslenskra bókaútgefenda (FIBUT) gátu ekki komist að samkomulagi haustið 2020 sem leiddi til þess að við þurftum að hafa samband við hvern og einn útgefanda og kynna verkefnið fyrir þeim. Síðan í byrjun október 2020 höfum unnið að því hvernig best væri að kynna verkefnið og í framhaldi haft samband við alla viðkomandi bókaútgefendur. Vegna niðurstaðna viðræðna við FIBUT vildum við vera tilbúin fyrir allar mögulegar spurningar frá útgefendum áður en við byrjuðum að hafa samband. Við settum saman kynningu til að nota ef tækifæri gæfist til að sækja útgefendur heim, en vegna faraldursins gátum við ekki farið í neinar heimsóknir. Við settum einnig saman sýnidæmi úr RMH með um 70 bókum frá útgefandanum Skruddu sem við höfðum þegar fengið og unnið. Þetta sýnidæmi átti að vera til þess að veita útgefendum innsýn í það hvernig gögnin yrðu gerð aðgengileg í RMH og hvernig textunum yrði skipt í smærri hluta áður en þeim væri bætt í málheildina. Þegar við byrjuðum að hafa samband við útgefendurna í nóvember virtist kynningin ekki nauðsynleg þar sem fyrstu viðbrögð voru öll jákvæð. Það hefur tekið tíma að fá svör og fjöldi útgefenda hefur ekki svarað með neinum hætti. Við þessu var búist þar sem þessi tími árs er almennt annasamasti tíminn fyrir íslenska bókaútgefendur. Í fyrstu stefndum við á að hafa samband við alla stærstu útgefendurna en þegar bið eftir svörum tók að dragast ákváðum við að hafa einnig samband við þá smærri. Einhverjir útgefendur tjáðu okkur að þeir myndu ekki hafa tíma til að fara yfir málið fyrr en í janúar. Við höfum virt þessar

óskir og beðið þolinmóð og í janúar hófumst við aftur við að hafa samband við útgefendur sem höfðu ekki enn gefið endanlegt svar til að fá vonandi fleiri svör og geta bætt þeim við lista okkar yfir samstarfsfúsa útgefendur. Frá og með 21. febrúar hefur verið haft samband við alla viðkomandi bókaútgefendur og meirihluti þeirra hefur verið jákvæður um að veita aðgang að textum til að bæta í RMH. Staðfestir samstarfsfúsir útgefendur eru eftirfarandi:

Útgefandi	**Staða**	**Útgefnar bækur**
Menntamálastofnun	staðfest	1.150
Bókabeitan ehf.	staðfest	104
Bókaútgáfan Hólar ehf.	staðfest	353
Hið íslenska bókmenntafélag	staðfest	278
Skrudda	staðfest	308
Bókaútgáfan Sæmundur	staðfest	70
Setberg	staðfest	161
Ormstunga	staðfest	132
Bókstafur	staðfest	24
Dimma	staðfest	30
Una útgáfuhús	staðfest	9
Háskólaútgáfan	staðfest	949
Samtals:		**3.568**

Í töflunni fyrir ofan kemur einnig fram fjöldi útgefinna bóka frá hverjum útgefanda, samkvæmt vefsíðum þeirra. Það á eftir að koma í ljós hvort allar þessar bækur verða aðgengilegar og hvort þær verða nytsamlegar fyrir málheildina. Þetta er ástæðan fyrir því að við stefndum á vinnslu 750-1.000 bóka fyrir M5; við vildum ekki setja okkur óraunsæ markmið og töldum þetta viðráðanlegt markmið sem væri hægt að ná. Samstarfsfúsir útgefendurnir hafa, samkvæmt upplýsingum af vefsíðum þeirra, gefið út samtals 2.619 bækur. Við erum enn að bíða eftir

svörum frá eftirfarandi útgefendum sem hafa samtals gefið út a.m.k. 2.338 bækur, sjá töflu fyrir neðan:

Útgefandi	Staða	Útgefnar bækur
IÐNÚ útgáfa	óstaðfest	417
Benedikt bókaútgáfa	óstaðfest	49
Bókafélagið	óstaðfest	265
Sögufélag	óstaðfest	106
Crymogea	óstaðfest	65
Angústúra	óstaðfest	48
Óðinsauga	óstaðfest	20
Edda útgáfa	óstaðfest	110
Forlagið	í bið	óþekkt
Salka / Verðandi	í bið	280
Bjartur - Veröld	í bið	375
Ugla útgáfa	í bið	219
Skálholtsútgáfan	í bið	200
Sögur útgáfa	í bið	77
Útkall ehf.	í bið	27
Drápa / N29	óstaðfest	39

Nýhöfn	í bið	20
Partus forlag	í bið	21
Samtals:		**2.338**

Í töflunni fyrir ofan eru óstaðfestir útgefendur þeir sem haft hefur verið samband við og þeir svarað, en ekki enn staðfest samstarf. Þeir útgefendur sem merktir eru „í bið" hafa ekki svarað með neinum hætti, þrátt fyrir ítrekaðar tilraunir okkar til að hafa samband. Við vonumst til þess að á meðan við byrjum að safna og vinna úr bókum frá útgefendum sem hafa þegar samþykkt samstarf, munum við geta fengið fleiri útgefendur til að taka þátt í verkefninu og í framhaldi af því safna enn fleiri bókum. Við munum einnig leggja áherslu á að fá stærstu útgefendurna til samstarfs. Við erum bjartsýn, sérstaklega vegna þess að nánast engin þeirra svara sem við höfum fengið hafa verið neikvæð.

Hagþenkir, félag höfunda fræðirita og kennslugagna

Eins og tekið er fram í M3-skýrslunni, náðum við ekki samkomulagi við Hagþenki, félag höfunda fræðirita og kennslugagna, fyrir lok október 2020. Í nóvember samþykkti stjórn Hagþenkis að nota sömu aðferð og Rithöfundasamband Íslands notaði síðasta vor við öflun samþykkis frá meðlimum sínum. 1. Desember 2020 var bréf sent til allra meðlima Hagþenkis þar sem verkefnið er kynnt fyrir þeim. Í bréfinu voru rithöfundarnir upplýstir um að ef þeir andmæltu ekki fyrir 1. febrúar 2021, mætti SÁM meta það sem samþykki um að textum þeirra mætti safna og bæta í RMH. Við fengum engin andmæli og einhverjir meðlimir hafa jafnframt haft samband við okkur beinlínis til að lýsa yfir stuðningi sínum og gefa leyfi sitt til söfnunar texta sem þeir hafa skrifað og gefið út. Þetta samkomulag þýðir að við höfum möguleika á því að safna textum frá enn fleiri bókum án þess að þurfa að hafa samband við höfundana til að fá samþykki þeirra. Meðlimir Hagþenkis eru til dæmis höfundar efnis sem Menntamálastofnun hefur gefið út. Við getum einnig talið víst að mikið magn af efni úr fræðiritum er samið af meðlimum Hagþenkis, svo þetta mun flýta fyrir söfnun fræðitexta sem við munum vinna að fyrir næstu vörður, M5 og M6.

Skýrsla um nýja texta sem safnað var frá fyrri upprunum

Nýir textar frá seinasta ári telja samtals um 98 milljón orð.

Undirmálheildir	**Orð**
Fréttaefni	76.143.855

Þingræður	3.955.556
Lög	15.518.986
Dómsúrskurðir	2.820.951
Upplýsingar	219.812
Blogg	5.157
Samtals:	**98.664.317**

Við höfum enn ekki halað niður uppfærðri útgáfu af *Wikipedia*, en á seinasta ári stækkaði hún um rúmlega 2 milljón tóka eftir uppfærsluna. Tvær af þremur vefsíðum fyrir dómsúrskurði hafa breyst síðan 2019 og gefa nú einungis gögn sín á pdf-formi. Við höfum enn ekki sótt þessa texta en við áætlum að ná um 2,9 milljónum orða.

Við áætlum því að viðbót nýrra texta frá fyrri upprunum muni vera um 103 milljón orð.

Við erum nú að safna meiri gögnum frá lagamálheildinni en áður þar sem söfnum nú öllum gögnum sem tengjast einstökum lögum (ólíkar útgáfur, nefndarálit, breytingartillögur o.s.frv.) í stað þess að taka aðeins endanlegar útgáfur laga. Þess vegna koma þau u.þ.b. 15 milljón orð sem safnað var á þessu ári ekki aðeins frá skjölum sem gefin eru út á því ári, heldur einnig skjöl dagsett frá fyrri árum sem hefur ekki verið bætt í málheildina áður.

Þrjár óðalsbundnar málheildir skilgreindar

Við höfum skilgreint þrjár undirmálheildir: alþingismálheild, lagamálheild og fréttamálheild.

1. EFNI OG LÝSIGÖGN

1.1 Alþingismálheild

Alþingismálheildin mun innihalda allar ræður sem hafa fluttar verið á Alþingi síðan 1914 sem eru aðgengilegar á www.althingi.is. Hún mun innihalda eftirfarandi lýsigögn:

Ræður: Titill/mál, dag- og tímasetning, umræðuefni, flokkar og tegund ræðu. Hlekkur á myndbands- og hljóðskrá, þar sem þær eru tiltækar, mælandi.
Mælendur: Fullt nafn, fæðingardagur, og kyn. Fyrir hvert tímabil upplýsingar um stöðu þess einstaklings á Alþingi, hvaða stjórnmálaflokki og umdæmi hann tilheyrði og ef hannn hafði ráðherrasæti í ríkisstjórninni.
Ríkisstjórnir: Listi titla allra ríkisstjórna frá 1917 og hvaða stjórnmálaflokkar mynduðu þær.

Þing og kjörtímabil þeirra: Hvert þing skráð (Heimastjórn, Konungsríkið Ísland og Lýðveldið Ísland) ásamt kjörtímabilum þeirra.

1.2 Lagamálheild
Lagamálheildin mun innihalda skjöl sem tengjast hverju frumvarpi sem lagt er fram á Alþingi og er aðgengilegt á www.althingi.is. Skjölin og frumvarpið sjálft (stundum tvær eða fleiri útgáfur), nefndarálit, og breytingartillögur. Hún mun innihalda eftirfarandi lýsigögn:

Frumvörp: Á hvaða alþingi var frumvarpið lagt fram, skiladagur, ID (auðkenni) umræðu sem það tilheyrir, ítarlegar upplýsingar um persónuna sem lagði fram frumvarpið (id, fullt nafn, fæðingardagur, kyn, flokksaðild…), hlekkur á frumvarpið á www.althingi.is, umræðuefni og flokkar.
Skjöl: Tegund skjals (frumvarp, nefndarálit...), hvenær því var dreift, höfundar (id, fullt nafn, fæðingardagur, kyn, flokksaðild…).

1.3 Fréttamálheild
Fréttamálheildin mun innihalda greinar úr prentmiðlum og vefmiðlum, auk nokkurra umritana frétta úr sjónvarpi og útvarpi. Elstu greinarnar eru frá 1998. Hún mun innihalda eftirfarandi lýsigögn:

Greinar: Titill, dagsetning og tímasetning, nafn höfundar, flokkar, útgefandi, vefslóð (ef um vefmiðil er að ræða).
Útgefendur: Nafn og heimilisfang.
Flokkar: fréttir, íþróttir, matur, börn
2. UPPBYGGING TEI-SKRÁA

2.1 Sameiginlegir eiginleikar
Fyrir hverja undirmálheild er teiCorpus-tag sem inniheldur haus með almennum upplýsingum um málheildina. Hún mun svo annaðhvort innihalda önnur teiCorpus-tög þar sem málheildinni er skipt frekar í smærri hluta, eða TEI-tög sem innihalda texta-tagið. Hver teiCorpus og TEI inniheldur eigin haus með nánari upplýsingum sem gilda um viðeigandi hluta málheildarinnar. Þar sem ein skrá með allri málheildinni væri of stór, notum við í mörgum tilfellum *include-tög* til að vísa í aðrar skrár sem innihalda teiCorpus-tög og TEI-tög.

2.1 TeiCorpus
TeiCorpus hefur eftirfarandi uppbyggingu (sjá Viðauka I fyrir skilgreiningar á öllum merkingum):

```
<teiCorpus>
    <teiHeader>
        <fileDesc>
            <titleStmt></titleStmt>
          <editionStmt></editionStmt>
          <extent></extent>
            <publicationStmt></publicationStmt>
```

```
            <sourceDesc></sourceDesc>
        </fileDesc>
        <encodingDesc>
            <projectDesc></projectDesc>
          <editorialDecl></editorialDecl>
            <tagsDecl></tagsDecl>
            <classDecl></classDecl>
        </encodingDesc>
        <profileDesc>
            <particDesc></particDesc>
        </profileDesc>
    </teiHeader>
    <TEI> … </TEI> [or <teiCorpus> … </teiCorpus>
    ...
</teiCorpus>
```

2.1.2 TEI

teiHeader er breytilegur frá einni undirmálheild í aðra, en skjalstofninn (e. body) mun hafa svipaða uppbyggingu, þar sem textum hefur verið skipt í , málsgreinar (<p>), setningar (<s>) og tóka (<w> fyrir orð og <c> fyrir greinarmerki). Í tilfelli þingræðna notum við <seg> í stað <p> og hver ræða ívafin <u> tagi (segð).

```
<body>
    <text>
        <div>
            <p>
                <s>
                    <w type=[postag] lemma=[lemma]>[wordform]</w>
                    …
                    <c type=[postag]>[punctuation]</w>
                </s>
            </p>
        </div>
    </text>
</body>
```

2.2 Undirmálheildir

Uppbygging skráa er mismunandi frá einni undirmálheild til annarrar.

2.2.1 Alþingismálheild

Málheildin er kóðuð samkvæmt tillögum Parla-CLARIN TEI: https://github.com/clarin-eric/parla-clarin

teiCorpus inniheldur allar almennar upplýsingar en hvert TEI inniheldur allar ræður fyrir einn dag. Í ParticDesc munu allar upplýsingar um mælendur og félög verða geymdar. Í ClassDecl eru geymdar allar upplýsingar um málefni (frumvarp), umræðuefni, sali, tegund mælanda (*guest, chair, regular*), ráðuneyti og kjördæmi.

2.2.2 Lagamálheild

teiCorpus (rótin) mun hafa svipaða uppbyggingu og fyrir Alþingismálheildina, en hvert frumvarp verður ívafin öðrum teiCorpus-tögum sem innihalda haus með lýsigögnum fyrir hvert frumvarp ásamt einni TEI-merkingu fyrir hvert skjal sem tengist því frumvarpi.

2.2.3 Fréttamálheild

teiCorpus (rótin) inniheldur lista útgefenda (í <listOrg> í <particDecl>) og flokka (í <classDecl>). Hún inniheldur önnur teiCorpus-tög fyrir hvern útgefanda sem innihalda eitt TEI-tag fyrir hverja grein.

VIÐAUKI:

teiCorpus	Inniheldur heild TEI-kóðaðrar málheildar, sem samanstendur af einum málheildarhaus og einum eða fleiri TEI-stökum, sem hver inniheldur einn textahaus og texta.
teiHeader	Lýsandi og skilgreinandi lýsigögn tengd stafrænni heimild eða safni heimilda.
fileDesc	(e. file description) inniheldur heildstæða bókfræðilega lýsingu á rafrænni skrá.
titleStmt	(e. title statement) tekur saman upplýsingar um titil verks og þá sem bera ábyrgð á innihaldi þess.
editionStmt	(e. edition statement) tekur saman upplýsingar um eina útgáfu texta.
extent	lýsir áætlaðri stærð texta sem geymdur er á einhverju formi, stafrænu eða ekki, í viðeigandi mælieiningu.
publicationStmt	(e. publication statement) tekur saman upplýsingar um útgáfu eða dreifingu rafræns eða annarskonar texta.
sourceDesc	(e. source description) lýsir heimild(um) rafræns texta, venjulega heimildarlýsing í tilfelli texta sem hefur verið breytt á stafrænt form, eða setning eins og „born digital“ (fæddur stafrænn) fyrir texta án eldri heimild.
encodingDesc	(e. encoding description) skjalar samband stafræns texta og heimilda(r).
projectDesc	(e. project description) ítarleg lýsing tilgang kóðunar rafrænnar skrár, ásamt viðeigandi upplýsingum um ferli samsetningar eða söfnunar.
editorialDecl	(e. editorial practice declaration) veitir upplýsingar um meginreglur ritstjórnar sem beitt var við kóðun texta.
tagsDecl	(e. tagging declaration) veitir upplýsingar um tög í skjali.
classDecl	(e. classification declarations) inniheldur eitt eða fleiri flokkunarkerfi sem skilgreina flokkunarkóða sem notaðir eru annars staðar í textanum.
profileDesc	(e. text-profile description) veitir ítarlega lýsingu á þeim eiginleikum texta sem eru ekki bókfræðilegs eðlis, sérstaklega tungumálum og undirmálum sem notuð eru, aðstæðum sem textinn var búinn til í, þátttakendur og stöðu þeirra.
particDesc	(e. participation description) lýsir greinanlegum mælendum, röddum, eða öðrum þátttakendum í hvers konar texta eða öðrum persónum sem nefndar eru á nafn eða vísað til í texta, útgáfu, eða lýsigögnum.

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.

Varða 4 - lýsing

Ljúka högun gagnasafnsins og prófa með gögnum sem þegar hefur verið safnað. Afurðunum fyrir alla fjóra hluta BÍN-gagnasafnsins eru lýst hér á eftir, sbr. ítarlega lýsingu á hlutunum í Endurskoðuðum vörðum, dags. 27.11.2020.

Afurðir og staða

Högun gagna fyrir næstu kynslóð BÍN er tilbúin. Unnið er að greiningunni. Staða hinna fjögurra hluta BÍN er eftirfarandi, fyrst nýjasta viðbótin, Ritmyndir, sem hefur verið megináhersla verksins síðan í nóvember. Vinna við aðra hluta BÍN heldur áfram frá M3. Einum nýjum hluta BÍN hefur verið bætt við, lista íslenskra skammstafana. Þessi viðbót er nauðsynleg þar sem skammstafanir birtast sem ákvæði í samsettum íslenskum orðum.

- **Ritmyndir**: Vörðu fjögur hefur verið náð. Högun gagnasafnsins er lokið og hefur verið prófað með gögnum sem þegar hefur verið safnað, sbr. ítarlega skýrslu hér að neðan.

- **Orðföng:** Greiningu 26.000 ósamsettra orða (þ.e. grunnorð og afleidd orð) er lokið. Unnið er að gerð gagnasafnsins og verið að vinna gögnin fyrir innflutning. Vinna við tengingar við Rökliði sagna stendur yfir. Vinna við tengingar við styttingarlista (**IceAbbr.**) stendur einnig yfir, eins og lýst er að neðan.

- **Rökliðir sagna (VS):** Fyrsta bunka lokið og hann tilbúinn til innflutnings. Högun gagnasafnsins er lokið.

- **BÍN** (Beygingarlýsingarhluti BÍN): Greining villna í BÍN hefur verið endurunnin í tengslum við vinnu í **Ritmyndum** til að gera kerfin fullkomlega samhæfð. Samvinna við teymið sem vinnur við Villumálheildina hefur reynst ómissandi, og stefnt er á að gera greiningu þeirra samhæfanlega við greininguna í BÍN. Skipting rangra mynda milli Ritmynda og meginútgáfu BÍN er slík að einstakar beygingarmyndir eru settar í Ritmyndir, á meðan heil beygingardæmi eru sett í BÍN, réttilega dæmd röng. Því hefur nýr flokkur beygingardæma verið búinn til , þ.e., heil beygingardæmi óstaðlaðra orðmynda sem ekki voru í gögnunum fyrir. Þetta á t.d. við um ranglega myndaðar samsetningar, eins og *jarðaber* (ákvæði *jarða*, ef.ft.) fyrir *jarðarber* (ákvæði *jarðar*, ef.et.). Einkunnagjöf óstaðlaðra orða og orðmynda heldur áfram, með vísunum í viðeigandi málsgreinar í Ritreglum.

- **Python pakki gagna:** Þetta er afurð í M5. Pakkinn mun innihalda BÍN-gögn á *Kristínarsniði*, sjá. https://bin.arnastofnun.is/DMII/LTdata/k-format/

- **Listi íslenskra skammstafana** (**IceAbbr.,** Skammstafanaskrá): Nýrri útgáfu áður útgefins lista íslenskra skammstafana hefur verið lokið og verður útgefinn á vefsíðu Clarin. Listinn inniheldur 4.881 skammstafanir, skipt í gildar skammstafanir, styttar orðmyndir, o.s.frv., með undirliggjandi texta eða útskýringu.

Skýrsla um ritmyndir

Ritmyndir: Ritmyndir innihalda stafsetningarvillur, beygingarvillur, innsláttarvillur, eldri afbrigði og aðrar sjaldgæfar og óstaðlaðar orðmyndir sem eru undanskildar í BÍN. Algengar villur sem finnast í gegnum öll beygingardæmin má bæta við meginútgáfu BÍN til notkunar í máltækni með leiðréttingum og villugreiningu sem hönnuð er fyrir Ritmyndir, án þess að vera útgefin á BÍN-vefsíðunni. Villur sem finnast eingöngu í einstökum beygingarmyndum er bætt í Ritmyndir. Villur eru leiðréttar og tengdar við uppflettimyndir BÍN og rétt málfræðilegt mark er haft með þegar það er mögulegt. Hugbúnaðarhögun til að innihalda þessar myndir hefur verið lokið og nýja gagnasafnið er nú keyrt í prófunarmynd.

Gögnin fyrir hverja orðmynd er í 14 reitum: 1 orðmynd/villa, 2 stöðluð/leiðrétt mynd, 3 uppruni, 4 aldur (ár), 5 einkunn, 6 tegund/málsnið, 7, stafsetning, 8 innsláttarvilla, 9 beyging, 10 málfræðilegt mark, 11 BÍN-uppflettimynd, 12 BÍN-auðkenni uppflettimyndar, 13 sýnileiki, 14 skýringar. Flokkunin og upplýsingarnar eru fyrir orðmyndina/villuna og/eða tenginguna milli villunnar og leiðréttrar myndar. Upplýsingarnar um uppruna, aldur, einkunn og tegund/málsnið hafa að gera með orðmyndina/villuna. Stafsetning, innsláttarvilla og beyging hafa að gera með hvað er leiðrétt, þ.e. Sambandið milli villunnar og leiðréttu myndarinnar, uppflettimyndar hennar og málfræðilegs marks. Villan og leiðrétt mynd tilheyra alltaf sama málfræðilega marki.

Til að finna réttar uppflettimyndir og mörk til að tengja við gögnin, hefur viðmót stjórnanda glugga fyrir forvinnslu sem finnur allar beygingarmyndir í BÍN sem passa við leiðrétta mynd og setur inn upplýsingar í reiti 10, 11 og 12 í gögnunum.

Þegar tvö eða fleiri mörk í sama uppflettiorði passa við beygingarmyndina, er reitur fyrir mark hafður auður, sem táknar að myndin er óljós innan þeirrar uppflettimyndar.

Ef leiðrétt mynd passar við beygingarmyndir í tveimur eða fleiri uppflettimyndum, er ný færsla búin til fyrir hverja uppflettimynd og ritstjórinn verður að velja hvorri skal halda. Til að aðstoða við þetta sýnir viðmótið orðflokkinn (kyn í tilfelli nafnorða) og réttmætiseinkunn uppflettiorðsins.

Ritstjórinn getur einnig valið að taka BÍN-auðkenni (DMII-ID) uppflettimyndarinnar í inntaki svo að orðmyndin sé sjálfkrafa tengd réttu orði. Leiðrétt mynd verður að samræmast beygingarmynd þeirrar uppflettimyndar. Reit leiðréttrar myndar má einnig skilja eftir auðan ef ekki er hægt með neinu móti að leiðrétta villuna, þ.e. Ef leiðréttingin þurfa annað málfræðilegt mark. Ef orðmyndin er ekki í raun villa, heldur sjaldgæf mynd sem ætti ekki að leiðrétta, er hún ekki leiðrétt og málfræðilegu marki er bætt við handvirkt.

Eftir að gögnin hafa verið keyrð í gegnum forvinnslu, löguð og leiðrétt eftir þörfum, má bæta þeim í gagnasafnið. Þegar því er lokið má skoða og breyta hverri færslu í ritstjórnarviðmóti, sem inniheldur einnig leitartól. Þar er það tengt meginútgáfu BÍN og leit í BÍN-ritstýringartóli, sem gerir vinnu gagnanna auðveldari.

Vinna við greininguna er vel á veg komin, þ.e., á skilgreiningum mismunandi flokka stafsetningarvillna, innsláttarvillna og beygingarvillna.

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 4 - lýsing

5.000 færslum hefur verið bætt í orðabókina (færsla=einstök umritun). Prófunarsett fyrir atkvæðaskiptingu og innsetningu áherslna tilbúin. Ritstýringartól orðabókarinnar getur nú spilað hljóðritanir í gegnum talgervil.

Afurðir

1. 5.133 nýjar færslur í framburðarorðabókinni.
2. Prófunarsett fyrir sjálfvirka atkvæðaskiptingu og innsetningu áherslna fyrir öll framburðartilbrigði.
3. Spilunarmöguleiki (e. play-back feature) fyrir ritstýringartól orðabókarinnar.

Öllum afurðum vörðunnar var skilað.

Íslensk framburðarorðabók á github: https://github.com/grammatek/iceprondict
CLARIN: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/99

Skýrsla

Nýju g2p-líkönin sem þróuð voru á fyrsta ári hafa nýst til að flýta handvirka leiðréttingarferlinu, þar sem þau búa til umritanir fyrir hvert framburðartilbrigði og hljóðritanirnar eru almennt nákvæmari en með gamla líkaninu. Þetta þýðir að minni vinna fer í leiðréttingu, þar sem áður þurfti að hljóðrita hvert framburðartilbrigði handvirkt, sem var ólíkt hefðbundnum framburði. Eins og er fylgir val nýrra færslna í orðabókina tíðnilistum frá Risamálheildinni (RMH) (G1). Markmiðið er að 30.000 algengustu orðin í RMH verði í orðabókinni fyrir lok verkefnisins.

Á fyrsta ári var megináhersla á hljóðritun orðalista með góðri dreifingu mögulegra samsetninga rittákna í íslensku, til þess að búa til þjálfunarsett fyrir sjálfvirkar einingar rittákns-til-fónems (g2p). Þann tíma sem eftir er af verkefninu verður tveimur nýjum flokkum orða bætt við orðabókina: a) orðum sem fylgja ekki almennum g2p reglum í íslensku, og b) algengum erlendum orðum og skammstöfunum. Fyrir orð með óreglulegum framburði verða upplýsingar úr BÍN (G4) notaðar. Gagnasafnið merkir þessi óreglulegu orð með tilteknu fráviki frá almennri reglu sem á við, og því má með auðveldum hætti draga þessi orð úr gagnasafninu. Verkþáttur G1 hefur afhent tíðnilista sérnafna og skammstafana auk algengra orða. Þessir listar verða notaðir til að velja erlend orð fyrir orðabókina.

Vinna er hafin við að finna út hvernig skal hljóðrita ensk orð með íslenska hljóðritunarstafrófinu. Enska CMU framburðarorðabókin[1] hefur verið notuð til að bera kennsl á ósamræmi milli enskrar hljóðritunar og líklegrar íslenskrar hljóðritunar enskra orða. Bein vörpun milli hljóðritunarstafrófa er að öllum líkindum ekki fullnægjandi, þar sem íslendingar bera gjarnan ensk orð fram, sérstaklega sérhljóð, eftir réttritun auk þess sem þeir vita um enskan framburð. Við munum halda þessari vinnu áfram á yfirstandandi verkári og hljóðrita algengustu ensku orðin úr RMH handvirkt, en eftir stendur hvort einhvers konar sjálfvirkar hljóðritunarreglur eða líkön væri hægt að hanna.

Fyrir atkvæðaskiptingu og áherslumerkingu prófunarsetta, voru sömu 1.000 orðin valin úr RMH eins og fyrir prófanir orðskiptingatól (sjá M3 skýrslu: undirverkefni G5, Orðskiptingar), þ.e. Orð númer 10.001-11.000 í tíðnilistanum. Atkvæðaskipting og áherslumerking í íslensku er nokkuð regluleg og því bjuggum við aðeins til prófunarsett en ekki stór þjálfunarsett fyrir þessi verk. Þróun sjálfvirks algríms til atkvæðaskiptingar og áherslumerkingar er verk innan verkþáttar T10 sem verður farið í á núverandi verkári. Atkvæðaskiptingin fylgir stuðla- og rímaðferðinni (e. onset-rhyme approach), þar sem hvert atkvæði byrjar á samhljóði ef það er mögulegt. Þessi regla er frábrugðin orðskiptingareglunni, þar sem orðskipting er oft framkvæmd á þann hátt að byrjun nýrrar línu eigi að vera sérhljóði. Í íslensku er áherslan alltaf á fyrsta atkvæði, og svo til skiptis án áherslu og með aukaáherslu, t.d. 1-0-2-0 (1: aðaláhersla, 2: aukaáhersla, 0: engin áhersla). Hins vegar getur aukaáherslan færst til í samsetningum, sem orsakar mynstur eins og 1-0-0-2 eins og í ‘kennaraborð’, þar sem ‘kennaranum’ hefur reglubundið mynstur 1-0-2-0. Að auki má deila um hvort þriggja atkvæða orð sem samanstendur aðeins af rót og beygingarendingu ætti að hafa aukaáherslu merkta á síðasta atkvæðinu, eins og í ‘kennara’. Við ákváðum að sleppa áherslumerkingu á þriðja atkvæði í slíkum tilfellum. Þrátt fyrir regluleg áherslumynstur í heildina, er ákveðið ósamræmi varðandi samsetningar sem hafa eins atkvæðis ákvæði. Almenna reglan er að fyrsta atkvæði hvers kjarnaorðs í samsettu orði ætti að hafa áherslu. Hins vegar vilja samsett orð með eins atkvæðis ákvæði falla í víxlmynstrið: *bús-áhöld* mynstur: 1-0-2.

Prófunarsettið var búið til fyrir hverja framburðarmállýsku og hefur eftirfarandi upplýsingar fyrir hverja færslu:

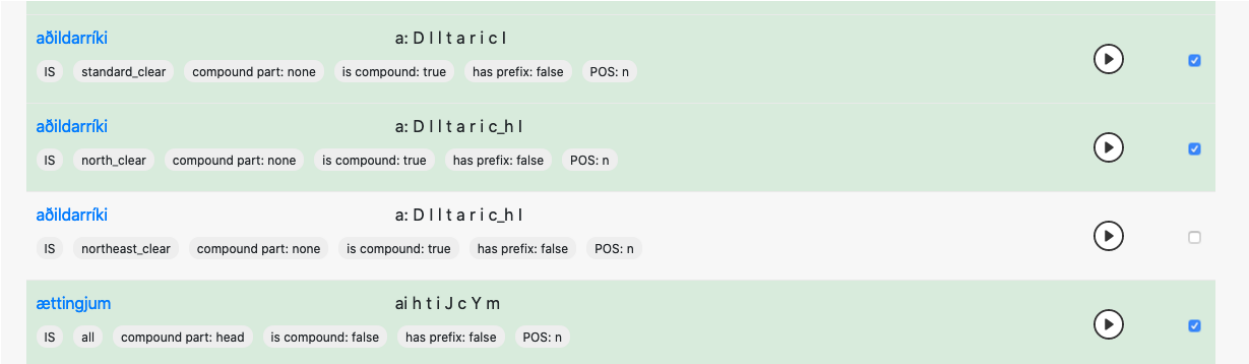
```
Orð: einkahlutafélög
Hljóðritun: ei N k_h a l_0 Y t_h a f j E l 9 G
Framburðartilbrigði: northeast_clear
Atkvæðaskipting: ei N.k_h a.l_0 Y.t_h a.f j E.l 9 G
Atkvæðaskipting með áherslumerkingum: ei1 N.k_h a0.l_0 Y1.t_h a0.f j E1.l 90
G
```

Áherslumerkingin er gerð í samræmi við form CMU orðabókarinnar, en við ákváðum hins vegar að merkja aðeins áherslu og enga áherslu þar sem aðaláhersla kemur aðeins fram á fyrsta

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

atkvæði. Ef það þarft frekari sundurgreiningu má auðveldlega breyta áherslumerkingum innan orða yfir í aukaáherslumerkingu sjálfkrafa.

Ritstýringartólið er nú með tengingu í talgervil til að spila aftur valda hljóðritun. Tólið tengist við Amazon Polly eins og er, en um leið og nýjar raddir verða gefnar út í undirverkefni T7, mun bakendinn byrja að nota raddir innan verkefnisins.



Skjáskot úr PEDI-ritstýringartóli*, sem inniheldur nú spilunarhnapp fyrir hverja færslu.*

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 4 - lýsing

- Sameiginlegt líkan fyrir mörkun og lemmun með BERT-líku mállíkani. Markmiðið er að nákvæmni mörkunar verði >= 96%.
- Bráðabirgðavillugreiningarskýrsla fyrir mörkun og lemmun.
- Fyrri helmingur Gullstaðals (MIM-GOLD) auðgaður með lemmum.

Afurðir

Öllum afurðum hefur verið skilað.

- Líkan sem er fært um að marka og lemma texta samtímis hefur verið útbúið. Líkanið er byggt á forþjálfuðu mállíkani (transformer, ELECTRA-small). Meðalnákvæmni mörkunar líkansins á Gullstaðlinum er 96,95%, sem er umfram þá 96% nákvæmni sem krafist er.
- Sjá skýrsluhluta fyrir villugreiningu.
- Meira en 500.000 tókar hafa verið yfirfarnir handvirkt.

Líkaninu má hala niður á https://repository.clarin.is/repository/xmlui/handle/20.500.12537/98

Skýrsla

Fyrri helmingur MIM-Gold hefur verið endurskoðaður og leiðréttur handvirkt, u.þ.b. 520.000 uppflettimyndir (e. lemmas). Unnið er að annarri endurskoðun þar sem meiri áhersla verður lögð á stöðlun.
Nokkur atriði varðandi uppflettimyndir í MIM-Gold þurfti að leysa og samræma. Til að nefna fáein:

> Það getur verið óljóst hvernig skal lemma sérnöfn, sérstaklega ef þau innihalda ákveðinn greini eða tvö eða fleiri orð (t.d. *Prikið*, *Rauða ljónið*).
>
> Oft er erfitt að greina á milli lýsingarorða og lýsingarháttar þátíðar (t.d. *sannfærður* í samhenginu *Ég var ekki sannfærður*).
>
> Stafsetningarvillur geta haft áhrif á lemmarann á ýmsan hátt, hversu langt skal fara í að taka á leiðréttingum þeirra?

Við höfum komið upp regluverki fyrir sum þessara óljósu tilfella, en önnur munu líklega vera óljós áfram (t.d. lýsingarorð og lýsingarháttur þátíðar).

Líkanið (ABLTagger, Steingrímsson o.fl.[2]) úr M3 hefur verið endurbætt frekar til að innleiða forþjálfað mállíkan til að veita samhengisháðar orðagreypingar og styðja samhliða lemmun. Lemmunareiningin er ekki gefin út á þessu stigi en uppfærð útgáfa markarans, sem vísað er til sem v2.0, er aðgengileg á CLARIN[3].

Við innleiðingu forþjálfaða mállíkansins voru nokkrar breytingar gerðar á högun ABLTagger v1.0. Veigamestu breytingarnar voru meðal annarra viðbót afgangstenginga/sleppitenginga (e. residual connections) við meginútgáfu BiLSTM og notkun GRU (e. Gated Recurrent Unit) fyrir endurkvæmnisnet (RNN, e. Recurrent Neural Network) bókstafa. Þetta gerði okkur kleift að kvarða líkanið í stærð, sem jók nákvæmni mörkunar í 95,78%, algild aukning 0,19 prósentustiga frá 95,59%.

Forþjálfaða mállíkanið er ELECTRA-small[4] sem inniheldur um 14M stika og styður setningar sem innihalda að hámarki 126 orðflísar[5]. Líkanið er forþjálfað á RMH (Risamálheildinni) og við vísum í undirverkefni I8 fyrir nánari lýsingu. Að styðjast aðeins við samhengisháðu orðagreypingarnar fyrir mörkun eykur þegar árangur mörkunar, sem nær 96,69% nákvæmni, en við komumst að því að með því að bæta öllum áður gerðum orðabreypingum við (character RNN, DMII, forþjálfað FastText) til að búa til eitt orðagreypingarlíkan bætti árangur frekar, í 96,95%[6] sem greint var frá. Enn stærra forþjálfað mállíkan mun bæta árangur enn frekar og draga úr þörfinni á auka orðagreypingum fyrir mörkun.[7] Við tókum einnig eftir því að RNN bókstafagreyping (e. character RNN embedding) var nauðsynleg fyrir lemmun.

Lemmunareiningin er RNN með sjálfvirku aðhvarfi (e. autoregressive RNN) sem tekur sem viðbótarinntak sameinaðrar orðmyndar auk athyglisvigurs (e. attention vector) bókstafa RNN tókans sem skal lemma. Högun ABLTagger v2.0 er svipuð og LemmaTag[8] og við vegum tap lemmarans þannig að það verði í svipuðu magni og tap mörkunar. Þó að við höfum vegið tapið, tókum við eftir því að árangur lemmarans hafði slæm áhrif á mörkun, sem stangast (að einhverju leyti) á við fræðigreinar. Þetta þarfnast nánari skoðunar.

Niðurstöður markara og villugreining

[2] Steingrímsson, Steinþór, Örvar Kárason, og Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. Í *Proceedings of Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria.
[3] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/98
[4] Clark, K., Luong, M., Le, Q.V., & Manning, C.D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *ArXiv, abs/2003.10555.*
[5] Lengri setningum er skipt niður og þær unnar sér.
[6] En eykur stærð líkansins umtalsvert.
[7] Við gerðum einnig tilraunir með RoBERTa-byggt mállíkan sem gaf enn betri árangur.
[8] Kondratyuk, D., Gavenciak, T., Straka, M., & Hajic, J. (2018). LemmaTag: Jointly Tagging and Lemmatizing for Morphologically-Rich Languages with BRNNs. *EMNLP*.

Hér fyrir neðan eru niðurstöður nákvæmni mörkunar ABLTagger v2.0 og v1.0 metnar á ýmsum gagnasettum. Fyrsta gagnasettið, sem vísað er í sem Gullstaðall, er **MIM-Gold**[9] og inniheldur um 1 milljón handvirkt markaða tóka. Íslenska orðtíðnibókin (**IFD**)[10] er eldri staðall, sem samanstendur af eldri textum, sem inniheldur um 590.000 handvirkt markaða tóka. Báðum málheildunum hefur verið skipt í 10 hluta, þar sem tíundi hlutinn er notaður fyrir stillingar á forstilltum breytum (e. hyperparameter tuning), og fyrstu 9 hlutarnir fyrir þjálfun og prófanir (níföld krossprófun). Uppgefin nákvæmni mörkunar er því meðaltal yfir 9 hluta. Fyrir **MIM+IFD** notum við alla IFD sem aukaleg þjálfunargögn og prófum á MIM-Gold hlutunum. Þau níu gagnasett sem standa eftir eru undirmálheildir RMH settar saman fyrir M3 í verkefni G1 og eru metnar með útgefna líkaninu sem er þjálfað á MIM+IFD, tíunda hluta.

		ABLTagger v2.0 (%)	ABLTagger 1.0 (%)
MIM-Gold	Alls	**96,95**	95,59
	Þekkt	**97,83**	96,59
	Óþekkt	**87,47**	84,90
IFD	Alls	**97,55**	95,78
	Þekkt	**98,07**	96,72
	Óþekkt	**90,46**	84,13
MIM+IFD	Alls	**97,02**	96,40
	Þekkt	**97,81**	96,85
	Óþekkt	87,23	**90,33**
Sjónvarp og útvarp	Alls	**97,61**	96,69
	Þekkt	**98,10**	97,06
	Óþekkt	91,10	**91,77**
Fréttaefni	Alls	**96,81**	96,07
	Þekkt	**97,66**	97,25
	Óþekkt	**89,05**	85,22
Skoðanir	Alls	**96,78**	95,76
	Þekkt	**97,53**	96,77
	Óþekkt	**88,85**	85,19
Lög	Alls	**95,91**	95,55
	Þekkt	**96,64**	96,54
	Óþekkt	**88,06**	84,85

[9] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/40
[10] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/38

Dómsúrskurðir	Alls	**96,79**	95,09
	Þekkt	**97,45**	96,21
	Óþekkt	**89,41**	82,38
Þingræður	Alls	94,34	**94,88**
	Þekkt	95,79	**96,09**
	Óþekkt	74,43	**78,61**
Íþróttir	Alls	**95,47**	94,87
	Þekkt	**96,69**	96,55
	Óþekkt	**84,40**	79,63
Fræðsluvefir	Alls	**95,46**	94,86
	Þekkt	**96,69**	96,52
	Óþekkt	**84,72**	80,29
Bækur	Alls	**94,86**	94,23
	Þekkt	**95,78**	95,28
	Óþekkt	**81,22**	78,78

Villuflokkun fyrir markara

Fyrir neðan eru fjörutíu algengustu villurnar fyrir ABLTagger v2.0 við mörkun **MIM-IFD**. MIM-IFD er ekki sama gagnasett og MIM+IFD og er notað hér fyrir afturhæfi. Þetta gagnasett samanstendur af þjálfunargögnum frá MIM og IFD, í hverjum hluta, auk prófunargagnanna. Þetta orsakar smærra þjálfunarsett og stærra prófunarsett samanborið við MIM+IFD.

Sæti	Spáð > Rétt	Tíðni	%
1	af > aa	1120	2,37
2	aa > af	991	2,10
3	e > n----s	855	1,81
4	n----s > e	854	1,81
5	nveþ > nveo	577	1,22
6	nhfo > nheo	531	1,12
7	nveo > nveþ	519	1,10
8	nheo > nhfo	476	1,01
9	ct > c	356	0,75
10	nkeo > nkeþ	351	0,74

11	nken-s > n----s	338	0,71
12	c > ct	325	0,69
13	nkeþ > nkeo	321	0,68
14	nheo > nhen	260	0,55
15	n----s > nken-s	256	0,54
16	sng > sfg3fn	252	0,53
17	sfg3eþ > svg3eþ	252	0,53
18	nhfn > nhen	250	0,53
19	sþghen > lhensf	240	0,51
20	nhen > nheo	238	0,50
21	nhen > nhfn	233	0,49
22	sfg1eþ > sfg3eþ	232	0,49
23	foheo > fohen	231	0,49
24	sfg3fn > sng	219	0,46
25	fohen > foheo	209	0,44
26	lhensf > aa	205	0,43
27	nkfo > nkeo	204	0,43
28	sþgken > lkensf	189	0,40
29	x > ta	187	0,40
30	nvfþ > nkfþ	183	0,39
31	svg3fn > svg3en	182	0,38
32	lhensf > sþghen	182	0,38
33	aa > c	180	0,38
34	nkeo > nkfo	180	0,38
35	nkfþ > nvfþ	178	0,38
36	sþgven > lvensf	176	0,37
37	aa > lhensf	174	0,37
38	nhfo > nhfn	158	0,33
39	sfg3eþ > sfg1eþ	156	0,33
40	nvfo > nvfn	154	0,33

	ABLTagger 2.0		ABLTagger 1.0	
markari>rétt	**nr.**	**fjöldi**	**nr.**	**fjöldi**
af > aa	1	**1120**	1	1653
aa > af	2	**991**	2	1421
nveo > nveþ	7	**519**	3	988
nveþ > nveo	5	**577**	4	979
n----s > e	4	**854**	5	886
e > n----s	3	855	6	**770**
nheo > nhfo	8	**476**	7	561
nkeþ > nkeo	13	**321**	8	510
nkeo > nkeþ	10	**351**	9	497
ct > c	9	**356**	10	491
nhfo > nheo	6	531	11	**467**
sng > sfg3fn	16	**252**	12	439
nhen > nheo	20	**238**	13	435
nheo > nhen	14	**260**	14	429
sfg3fn > sng	24	**219**	15	416
sfg3eþ > svg3eþ	17	**252**	16	377
nken-s > n----s	11	**338**	17	366
foheo > fohen	23	**231**	18	337
n----s > nken-s	15	**256**	19	331
nhen > nhfn	21	**233**	20	326
fohen > foheo	25	**209**	21	316

fpkfþ > fphfþ	44	**144**	22	311
nvfn > nvfo	41	**151**	23	302
fpkfe > fphfe	68	**112**	24	294
c > ct	12	325	25	**293**

Bráðabirgðavillugreining fyrir lemmara (e. lemmatizer)

Lemmarinn (sem vitnað er til sem Lemmarans) var þjálfaður á óleiðréttum lemmuðum gögnum, sömu gögnum og verið er að leiðrétta handvirkt, með handvirkri skiptingu í þjálfun/prófun. Samkvæmt þeim sem fara yfir, innihalda gögnin ekki margar villur (<1%) svo þau ættu að nægja fyrir tilraunaútgáfu lemmarans (e. proof-of-concept lemmatizer). Lemmarinn er þjálfaður samhliða málfræðilega markaranum. Lemmarinn styðst við lærða orðmynd sem inniheldur upplýsingar um málfræðileg mörk. Þessi framsetning er ekki nógu upplýsandi ef lemmarinn er þjálfaður einn.

Nákvæmni Lemmarans borin saman við Nefnir[11] má sjá í töflunni að neðan.

	Lemmarinn (%)	Nefnir (%)	Tókafjöldi
Alls	**97,54**	97,31	93125
Óþekktar (orðmyndir)	85,05	**87,54***	7951
Þekktar (orðmyndir)	**98,71**	98,22*	85174
Óþekktar (uppflettimyndir)	83,90	**84,90***	4802
Þekktar (uppflettimyndir)	**98,28**	97,98*	88323

Takið eftir að tókafjöldi er fjöldi tilfella, þ.e. ekki einkvæmni. Flestar óþekktu uppflettimyndanna eru einnig óþekktar orðmyndir. Skilgreiningin á óþekktum og þekktum orðmyndum og uppflettimyndum á bara við um Lemmarann þar sem þær eru byggðar á þjálfunargögnunum, og því merkjum við samsvarandi gildi með stjörnu (*). Þessar tölur eru samt sem áður gefnar fyrir Nefni þar sem þær gætu gefið vísbendingar um áætlaða aukningu nákvæmni ef við bætum uppflettimyndum úr BÍN við þjálfunargögnin.[12] Við sjáum að þessi tilraunaútgáfa lemmarans (e. proof-of-concept lemmatizer) nær örlítið meiri nákvæmni en Nefnir en Nefnir nær oftast betri árangri á óþekktum orðum.

[11] Ingólfsdóttir, S., Loftsson, H., Daðason, J.F., & Bjarnadóttir, K. (2019). Nefnir: A high accuracy lemmatizer for Icelandic. *ArXiv, abs/1907.11907*.
[12] Nefnir er þjálfaður á BÍN (DMII).

Við söfnuðum einnig algengustu villunum sem tilraunaútgáfa lemmarans gerir til að fara handvirkt yfir úrtak þeirra villna.

Sæti	Spáð > Rétt	Tíðni	% af villum
1	það > hann	18	0,79
2	sá > það	16	0,70
3	sá > hann	12	0,52
4	ok (conjunction) > og	11	0,48
5	vera > er	9	0,39
6	það > sá	9	0,39
7	Norður-kórea > Norður-Kórea	9	0,39
8	það > hún	9	0,39
9	hann > það	9	0,39
10	það > því	8	0,35
11	því > það	7	0,31
12	elvis > Elvis	7	0,31
13	kvartun > kvörtun	7	0,31
14	þunga > þungun	6	0,26
15	mikið > mikill	5	0,22
16	þá > hann	5	0,22
17	hún > hann	5	0,22
18	getnaðarvarn > getnaðarvörn	5	0,22
19	Delamater > DeLamater	5	0,22
20	Mcbain > McBain	5	0,22
21	- > --	5	0,22
22	Evrópuráð > Evrópuráðið	5	0,22
23	1 > 1.	4	0,17
24	listi > list	4	0,17
25	sá > þá	4	0,17
Alls		189	8,25%

Flestar villur eru óalgengar. Margar þessara (kerfisbundnu) villna tengjast spáðu marki, þ.e. markarinn gerir villu og velur rangt mark, og lemmarinn lemmar því orðið ranglega. Þetta er tilfellið í 1, 2, 3, 4, 5, 6, og fleiri. Í töflunni hér fyrir neðan sjáum við sundurliðun mörkunar- og

lemmunarvillna. Í mörgum áðurnefndum villum fær markarinn einfaldlega út rangt kyn[13]. Augljóslega eru sumar villur einfaldlega rangar og ætti að leiðrétta eins og 7, 13, 14, 18 og 24 auk rangra falla í villum 12, 19, 20.

	Uppflettimynd rétt	Uppflettimynd röng
Mark rétt	88633 / 95,18%	1291 / 1,39%
Mark rangt	2201 / 2,36%	1000 / 1,07%

Önnur kerfisbundin villa sem við tókum eftir var að lemmarinn gat ekki klárað orð sem enduðu á bandstrikum („-“). Til dæmis, „stjórnunar- og skipulagsmál“, þar sem „mál“ er sleppt. Þetta gæti verið eitthvað sem endurbættur lemmari gæti meðhöndlað.

Við sjáum að hægt er að bæta atriði í lemmaranum. Sérstaklega í tilfellum þar sem markið er rétt en uppflettimyndin röng. Einföld lausn á þessu gæti verið að styrkja líkanið með uppflettimyndum skilgreindum í BÍN. Auðvitað gæti meiri nákvæmni mörkunar einnig leitt til betri árangurs.

[13] Í sumum þessara tilfella þarfnast rétt metið fall víðara samhengis. Til dæmis þarfnast „þeim brá“ víðara samhengis en gefið er í setningunni. Markari þjálfaður á mörgum röðuðum setningum í einu myndi geta náð þessu.

I5 - Þáttari

Skýrsla: Miðeind
Aðilar: Miðeind, Háskólinn í Reykjavík

Markmið

(Frá fyrsta ári:) Markmið þessa verkþáttar er fjórþætt. Í fyrsta lagi að meta tvo núverandi þáttara fyrir íslensku, með áframhaldandi þróun þeirra að leiðarljósi. Þáttararnir tveir eru IceParser, grunnþáttari byggður á stöðuferjöldum (e. finite-state transducers), og Greynir, djúpþáttari byggður á samhengisfrjálsri málfræði. Í öðru lagi, að vinna með öðrum teymum að þörfum fyrir sjálfvirka þáttun. Í þriðja lagi að aðlaga UD-þáttarann að íslenskri UD-málfræði og í fjórða lagi að þróa tauganetsþáttara þjálfaðan á trjábanka byggðum úr úttaki Greynis.

Varða 4 - lýsing

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.

M4

- Reglubyggði fullþáttarinn hefur verið aukinn og bættur í samræmi við niðurstöður mats á fyrsta ári og sýnir mælanlegar framfarir í nákvæmni þáttunar í samhengi við grunnlínu fyrsta árs (t.d. F-gildi út frá svigum > 75,17% eins og hefðbundið evalb-tól sýnir).

Afurðir

Öllum markmiðum vörðunnar var náð.
Reglubyggði fullþáttarinn er aðgengilegur á GitHub.
Þáttaða málheildin sem notum var við þróun og prófun er aðgengileg á GitHub.
Prófunarsvíta er aðgengileg á GitHub.
F-gildið fyrir reglubyggða fullþáttarann er 75,42% á núverandi (stærra) prófunarsetti.

Athugið að hluti af vinnu þessa undirverkefnis endurspeglast líka í niðurstöðum undirverkefnis L1 (málrýni).

Skýrsla

1. Yfirlit

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

Þáttunarhluti IceNLP[14], *IceParser*, er hlutaþáttari sem hefur þegar verið notaður við fjölda verkefna, og má helst nefna MerkOr (1, GitHub), og í tilraun með sjálfvirka upplýsingaheimt (1, GitHub). Hann var einnig notaður til þáttunar upprunalegrar útgáfu grunnþáttuðu prófunargagnanna fyrir þetta undirverkefni.

Greynir parser[15] er (full-)þáttari sem hefur meðal annars verið notaður til að leiðrétta málfræðivillur (undirverkefni L7), fyrir þáttun upprunalegrar útgáfu fullþáttaðra prófunargagna innan þessa undirverkefnis, í textavinnslu fyrir vélþýðingar (sjá undirverkefni V3b), sem vélin bakvið raddstutt app/gerviþjón voice assistant app, og einnig fyrir þáttaða málheild (GreynirCorpus).

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

Fyrir M4 var áherslan á endurbætur reglubyggða fullþáttarans. Grunnþáttarinn verður endurbættur í vörðu seinna á öðru ári.

[14] Á GitHub: https://github.com/hrafnl/icenlp
[15] Á GitHub: https://github.com/mideind/GreynirPackage

2. Forgreining

2.1 Gagnakröfur
Þar sem mikið magn af málfræðilegum gögnum var búið til á fyrsta ári, sem gætu hugsanlega verið notuð í þáttunarverkefninu, byrjuðum við annað ár á greiningu gagna.

Við vörpuðum fram eftirfarandi spurningum:

- Hvers konar málfræðileg gögn eru tiltæk og eiga við þetta undirverkefni?
- Hvaða gögn þurfum við?
- Hvernig geta tiltæk gögn hentað þörfum okkar?
- Samræmist gagnaleyfið leyfisþörfum okkar?
- Hvernig öflum við gögnunum og frá hverjum?
- Hvaða skörð þarf að fylla í með því að búa til gögn, ef einhver?

Við þurfum gögn til að (1) bæta í *orðaforða,* (2) víkka *málfræðireglur,* (3) *prófanir* þáttaranna, og (4) *þjálfun* tauganetsþáttara og mállíkan.

Fyrir (1), höfum við nú þegar BÍN fyrir máltækni (undirverkefni G4) sem grunnorðaforða. Til að bæta við hana má nota hvaða gerð textamálheildar sem er, eins og Risamálheildina (undirverkefni G1) og GreynirCorpus, til að grafa eftir nýjum orðmyndum. Við höfum að mestu látið undirverkefni G4 sjá um þetta, og uppfært orðabók þáttarans reglulega í nýjustu útgáfu BÍN; seinast í janúar 2021.

Fyrir (2), höfum við þegar lista íslenskra sagnorða, nafnorða og lýsingarorða ásamt rökliðum þeirra. Við tökum eftir rökliðum þar sem á við. Áherslan hér hefur verið á endurbætur og aukningu færslanna.

Fyrir þetta notum við markaða og/eða þáttaðar málheildir. Við getum leitað að ákveðnum mynstrum, eins og sagnliður + nafnliður-þgf. (eða bara sögn + nafnorð í þágufalli), og bætt viðeigandi sögnum við rökliðaramma (e. argument frames) okkar.
Einnig höfum við svipuð gögn úr öðrum verkefnum sem hægt væri að sameina okkar gögnum. Þessi svipuðu gögn þjóna örlítið öðrum tilgangi og því þarf að yfirfara þau handvirkt og lagfæra áður en þeim er bætt í stillingar þáttarans. Af þessum gögnum ber helst að nefna Íslenska nútímamálsorðabók.

Fyrir (3), viljum við nota handvirkt þáttuð gögn sem eru þegar tiltæk. Nokkrar markaðar (en ekki þáttaðar) málheildir eru til, en þar sem við höfum aðra valkosti reiðum við okkur ekki á þær eins og er. Við notum GreynirCorpus að mestu, eins og tilgreint er í hluta 3.1. IcePaHC er önnur handvirkt þáttuð málheild, en nútímamálshluti hennar er tiltölulega lítill og inniheldur að mestu bókmenntatexta.

Fyrir (4), munum við nota sjálfvirkt þáttaða hluta GreynirCorpus, sem inniheldur 7 milljón setningar, til þjálfunar. (Undirbúningsvinna gefur til kynna að tauganetsþáttari þarf í reynd ekki

fleiri en u.þ.b. 1 milljón setninga til að ná samleitni (e. converge).) Handvirkt þáttaða þróunarsettið, sem inniheldur 4.500 setningar, verður notað fyrir fínstillingu.

2.2 Utanaðkomandi þarfir
Byrjun annars árs felur í sér fjölda undirverkefna sem reiða sig á þáttarana. Því mátum við það mikilvægt að byrja á greiningu þarfa hvers undirverkefnis.

Málrýniverkefnið notar þáttarann. Það þarfnast sérstakra málfræðireglna, normaðrar (e. normalized) myndar úr tilreiðslu fyrir ákveðin greinarmerki, stuðning fyrir rangar skammstafanir, o.s.frv.

Við höfum fundið þörf í öðrum undirverkefnum (eins og G1 og T9) fyrir sérstaklega markað og lemmað úttak. Þetta er ekki hluti af þessum verkþætti en verður skoðað í framtíðinni.

Hvað utanaðkomandi notendur þáttarans varðar, reiðir raddstudda appið (e. voice assistant app) Embla sig á reglubyggðu þáttunina. Til að greina og flokka gildar fyrirspurnir úr raddinntaki þarf Embla sveigjanleika og mikla sértækni skýrra málfræðireglna (e. explicit grammar rules).

3. Prófanir og mat

3.1 Prófunarmálheild
GreynirCorpus inniheldur 5.000 handvirkt þáttaðar setningar úr fréttatextum samkvæmt skema Greynis. Þeim er skipt í þróunarsett, sem inniheldur 4.500 setningar (90%), og prófunarsett, sem inniheldur 500 setningar (10%).

Að vel athuguðu máli ákváðum við að nota GreynirCorpus til mats í stað áður tiltæks, mun smærra prófunarsetts. Þetta tekur á vandamáli þess að prófunarsettið sé of vítt og að hluta utan óðals. Þar sem mikilvægt notagildi þáttarans er málrýni, leiðir það af sér að þróun hans ætti að henta þörfum endanlegra notenda málrýnis, þeirra á meðal blaðafólks, nemenda og annarra höfunda formlegra, staðreyndabyggðra texta.

Önnur staðreynd sem styður þessa ákvörðun er að niðurstöðum fyrir sumar gerðir texta í upprunalega prófunarsettinu var að ýmsu leyti ábótavant, sem og gæðum textans. Við gerum okkur grein fyrir því að þáttarinn nær aldrei góðum árangri á öllum óðulum, til dæmis bókmenntatextum á frjálsu formi, og er ekki við öðru að búast. Þáttararnir eru reglubyggðir, svo þá má alltaf stækka fyrir fleiri óðul ef og þegar notagildi verður fyrir hendi.

Eins og tilgreint var í M3 skýrslunni, inniheldur GreynirCorpus prófunarsettið - sem er aðallega safn af fréttatextum - einhverjar óformlegar og jafnvel ranglega myndaðar setningar, setningar með innsláttar- og málfræðivillum, og setningar með veigamiklum erlendum setningum. Engar villur voru leiðréttar fyrir fram í textunum. Óreglulegu setningarnar hafa verið merktar og innleiddar í gullstaðalinn. Setningar sem innihalda veigamiklar villur sem ekki er hægt að gera ráð fyrir að þáttarinn (og líklega hvaða þáttari sem er) ráði við hafa verið merktar sem slíkar, og

niðurstöður gefnar án þeirra. Inngilding þessara setninga er kostur fyrir málrýniverkefnið, auk þjálfunar tauganetsþáttarans.

Að lokum má nefna að bæði IceParser og Greynir geta gefið aukalega málfræði og merkingar í úttaki sem eru viðbætur almennrar þáttunarskema.

3.2 Pípa
Eins og tilgreint var í M3 skýrslunni hefur prófunarsvíta fyrir innri mælingar verið þróuð. Hún notar hefðbundna matsforritið *evalb* til að mæla F-gildi út frá svigum, meðal annars.

Nýja málheildin hefur verið tengd og matspípan aðlöguð að henni. Ein stærsta breytingin er að GreynirCorpus inniheldur handvirkt þáttuðu skrárnar í skema Greynis, en þáttuðu skrárnar út frá svigum (e. the bracketed parsed files) í almenna skemanu dvelja í prófunarpípunni og má uppfæra fyrir hverja keyrslu.

4. Endurbætur
Í þessum undirkafla útskýrum við framfarir síðan í vörðu M3. Eldri endurbætur hafa verið útskýrðar í skýrslum fyrsta árs.

Nöfn persóna og aðrar nafneiningar: Fyrir M3 bættum við samsetningu nafneininga í einn tóka. Fyrir M4 var þessi samsetning gerð enn ágengari, þar sem það voru mörg viðeigandi tilfelli í þróunarmálheildinni og mjög fá mótdæmi.

Margra orða segðir (e. MWEs): Við rákumst á fleiri rangar skilgreiningar margra orða segða í nýja þróunarsettinu. Þær voru fjarlægðar, þar sem of áköf sameining í margra orða segðir kom í veg fyrir rétta þáttun. Við fundum raunar nokkrar nýjar og bættum þeim við.

Sagnarammar: Við erum að vinna við að uppfæra og bæta sagnaramma okkar. Verkinu er skipt í þrjú þrep:

1. Skilgreining ákveðinna ramma fyrir hverja sögn. Þetta veitir möguleika á því að tengja ákveðna merkingu við hvern ramma og röklið, sem veitir möguleika á merkingarfræðilegri greiningu. Þetta er vel á veg komið.
2. Viðbót nýrra dálka í sagnarammana. Sem dæmi, stefnum við á að bæta við föstum andlögum (e. static objects), til að auðvelda greiningu á orðatiltækjum, undirskipuðum setningum (e. subclauses), nafnháttarsetningum, ögnum, og fleiru.
3. Sameining gagnanna úr þrepi 1 við gögnin sem nefnd eru í hluta 2.1 (2), þar sem það gæti vantað einhverja ramma eða mikilvæga dálka. Þetta mun byrja um leið og þrepi 1 er lokið og gögnin í hluta 2.1 (2) eru tilbúin.

Meðhöndlun skilgreinds forms endurbættra sagnaramma verður innleidd í þáttarann um leið og búið er að ganga frá forminu (þrep 2).

Setningar sem innihalda spor og tóma liði: Sérhæfðum reglum hefur verið bætt við samhengisfrjálsu málfræðina til að taka á algengustu setningagerðunum.

Yfirborðsgerð setninga: Málfræðireglum hefur verið bætt við fyrir óþáttanlegar setningar í þróunarsettinu og fyrir óþáttanlegar setningar úr þróunarsetti villumálheildar í undirverkefni L1 (málrýni), þar sem gera mátti ráð fyrir að þáttarinn gæti þáttað setningarnar. Þetta fækkaði verulega fjölda óþáttaðra setninga í málheildinni.

Samræming atvikssetninga: Þetta var uppfært enn frekar eftir yfirferð þróunarsettsins í undirverkefni L1. Hegðun fyrir hverja undirskipaða setningu var skjöluð, eins og t.d. form hverrar tengingar.

Hugbúnaðarvinna: Við höfum endurbætt afköst og minnisnotkun þáttarans, sem bætir skilvirkni þáttunarferlisins, sérstaklega á löngum og flóknum setningum.

Orðaforði: Nýrri útgáfu BÍN (undirverkefni G4) frá janúar 2021 hefur verið pakkað inn í þáttarann, sem bætir orðaforða hans.

Uppfylling krafa: Eins og tekið var fram í hluta 2.2, þarf að uppfylla kröfur annarra verkefna. Við höfum innleitt betri meðhöndlun tölulegra setningarliða fyrir Emblu, endurbætt greinarmerkjatóka til að innihalda normaðar myndir, og bætt við stuðningi fyrir rangar skammstafanir.

Aðrar endurbætur: Til viðbótar við endurbætur þáttunar, höfum við fínstillt tilreiðarann (undirverkefni I3) til að styðja betur við þáttun, á grundvelli niðurstaðna þróunarsetts.

5. Mat

5.1 Samanburður við grunnlínu
Fyrir vörðu M3 greindum við frá eftirfarandi niðurstöðum:

```
Heimt út frá svigum: 74,29%        Meðalskörun: 2,50
Nákvæmni út frá svigum: 77,85%     Nákvæmni marka: 93,95%
F-gildi út frá svigum: 75,97%
```

Fyrir vörðu M4 greinum við frá eftirfarandi niðurstöðum (á öðru, stærra prófunarsetti):

```
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%
```

Taka skal fram að niðurstöðurnar er ekki hægt að bera beint saman, þar sem prófunarsettið hefur breyst og verið stækkað verulega. Gallaðar setningar (töflugögn, íþróttaniðurstöður, tíst (e. Tweets), o.s.frv.) eru ekki meðtaldar í niðurstöðum.

Eins og sjá má fer heimt hækkandi á meðan nákvæmni fer lækkandi. Við þessu er búist þar sem nýjum reglum fyrir áður óyfirfarnar setningar hefur verið bætt við. Það þýðir að við getum þáttað fleiri setningar, en óþáttaðar setningar eru ekki hafðar með í niðurstöðum *evalb*, og má búast við

að nákvæmni lækki með því að hafa þær með. Nýju reglurnar gætu líka leyft of mikið, sem leiðir þá af sér að setningar sem áður þáttuðust rétt fái nýja uppbyggingu. Þáttun setninga, jafnvel ekki alveg rétt, ætti að teljast mikilvægari en að þátta þær ekki.

Algengustu villurnar á þessu stigi stafa af erfiðleikum við að kortleggja rétt öll fullþáttunartré yfir í einfaldað staðlað skema sem notað er til mælinga, sérstaklega í kringum sagnliði. Algengar þáttunarvillur eru meðal annars viðhenging röklíða (e. argument attachment) og viðhenging forsetningarliða (e. PP attachment). Við búumst við því að nýir sagnarammar taki á mörgum þessara vandamála.

5.2 Dekkun
Eins og tekið var fram í hluta 4 bættum við málfræðireglum við fyrir margar nýjar setningagerðir, sem bætti dekkun. Þar sem prófunarsettið hefur breyst getum við ekki borið tölur saman beint við M3 vörðu, en það er tilfinning okkar að fjöldi óþáttanlegra setninga hafi minnkað umtalsvert.

6. Samstarf við atvinnulífið

Þáttarar eru ekki almennt notaðir beint af endanlegum notendum, en þjóna sínum tilgangi sem hluti af stærri hugbúnaði. Eins og tekið var fram í hluta 1 hefur þáttarinn verið nýttur í ýmis tilraunaverkefni og einnig í notendahugbúnaði.
Við höfum gefið nýja atvinnulífstengli SÍM kynningu á þáttara Greynis, GreynirCorpus og prófunarsvítunni. Kynningin fól í sér ítarlegar upplýsingar á því hvar má finna gögnin og hvernig má hugsanlega nota þau. Við búumst við því að kynningin leiði af sér áhugaverð tækifæri til að vinna með aðilum atvinnulífsins á komandi árum.

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 við nýjustu aðferðir í tauganetsmállíkönum. Slík líkön eru orðin grunnur að mörgum af bestu útfærslunum fyrir ýmis verk málvinnslu (e. NLP, natural language processing).

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 (e. downstream tasks) eru málrýni, vélþýðingar, talgreining, málfræðileg mörkun/lemmun, nafnakennsl og þáttun.

Nokkur BERT-lík líkön af 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 þær séu vel skipulagðar, og þetta þarf að hafa í huga við áætlun mannmánaða.

Varða 4 - lýsing

Nokkur BERT-lík mállíkön hafa verið forþjálfuð og metin samkvæmt viðmiði íslenskra málvinnslugagnasafna / verkefna sem nýta mállíkönin.

Afurðir

Öllum markmiðum verður náð 8. febrúar. Viðmiðunarmælingar hafa verið búnar til og Transformer mállíkan hefur verið þjálfað. Verið er að þjálfa stærra líkan sem verður tilbúið fljótlega.

Afurðir:

- ELECTRA-small: Transformer-líkan þjálfað á Risamálheildinni.
- ELECTRA-base: Transformer-líkan þjálfað á Risamálheildinni.

Skýrsla

Transformer líkön

Frá útgáfu BERT (e. Bidirectional Encoder Representations from Transformers) líkansins 2018, hefur Transformer högunin orðið leiðandi á sviði málvinnslu. Slík líkön eru þjálfuð í tveimur þrepum: fyrst eru þau *forþjálfuð* á viðgjafarlausan (e. unsupervised) hátt á stórri textamálheild á almennum verkum eins og að fylla inn hulin orð (e. masked words), og svo eru þau *fínstillt* á sértækari (og oftast mikið smærri) gagnasettum.

Ferli forþjálfunar er reiknilega dýrt og tekur oftast daga eða vikur að klára. Kostnaður við forþjálfun nýlegs líkans, eins og GPT-3, hefur verið metið á milljónir dollara. Við völdum því ELECTRA líkanið fyrir tilraunir okkar með íslensku, þar sem það hefur virst reiknilega hagkvæmara en flest önnur Transformer líkön. Að auki benda nýlegar rannsóknir á að það sé einnig einstaklega skilvirkt á gögn, þar sem það þarf minna af gögnum til forþjálfunar til að ná sömu niðurstöðum og önnur líkön.

Við höfum forþjálfað ELECTRA-small Transformer líkan með 2018 útgáfu Risamálheildarinnar (RMH) sem inniheldur 1,5 milljarð tóka. Þó að stærð RMH sé minni en helmingur málheildarinnar sem var notuð til að þjálfa upprunalega BERT líkanið (3,2 milljarðar tóka), reyndist hún samt fullnægjandi til að þjálfa líkan sem nær betri árangri en áður útgefin (ekki Transformer-byggð) líkön á gagnasöfnum fyrir málfræðilega mörkun (e. PoS tagging) íslensku auk nafnakennsla.

Umsókn um aðild að TFRC (TensorFlow Research Cloud) var lögð inn snemma í janúar. TFRC er verkefni styrkt af Google sem býður vísindafólki endurgjaldslausan aðgang að þinvinnslueiningum (e. Tensor Processing Units, TPUs) Google Cloud, sem er kraftmikill flýtivélbúnaður (e. hardware accelerators) sem geta þjálfað stór mállíkön. 29. janúar fengum við 90 daga kvóta fyrir TPU, sem verður notaður til að þjálfa fjölda Transformer-byggðra mállíkana fyrir íslensku. Nú stendur yfir þjálfun ELECTRA-grunnlíkans (e. ELECTRA-base model), sem verður tilbúið 8. febrúar. Grunnlíkanið er töluvert stærra, sem fjölgar heildartölu stillinga úr 14M í 110M. Það lengir einnig runulengdina úr 128 í 512, sem veldur því að ekki þarf að skipta jafn mörgum setningum í smærri hluta. ELECTRA-grunnlíkanið er þjálfað á 2019 útgáfu RMH, sem inniheldur 1,55 milljarða tóka.

Viðmiðunarmælingar

Eins og er felast viðmiðunarmælingar okkar í þremur málvinnsluverkefnum: málfræðilegri mörkun, nafnakennslum og vélrænum útdrætti (e. ATS, automatic text summarization). Málfræðileg mörkun er metin á MIM-GOLD málheildinni[16], sem samanstendur af u.þ.b. 1M tókum úr Markaðri íslenskri málheild (MÍM), sem hafa verið handvirkt merktir með málfræðilegum mörkum. Nafnakennsl eru metin með MIM-GOLD-NER[17], sem er útgáfa MIM-GOLD málheildarinnar sem hefur verið handvirkt merkt með nafneiningum (e. named entities).

[16] MIM-GOLD: http://hdl.handle.net/20.500.12537/39
[17] MIM-GOLD-NER: http://hdl.handle.net/20.500.12537/42

Vélrænn útdráttur er metinn með IceSum málheildinni[18], sem samanstendur af 1.000 íslenskum fréttagreinum sem hafa verið handvirkt merktar með útdráttarsamantekt. Fleiri verkum verður hugsanlega bætt við síðar, eftir því sem viðeigandi málheildir verða tiltækar.

Niðurstöður

Hér tilgreinum við niðurstöður fyrir ELECTRA-small og IceBERT, RoBERTa-grunnlíkan (e. RoBERTa-base model) þjálfað af Miðeind[19]. Fyrir málfræðilega mörkun (e. PoS) tilgreinum við nákvæmni mörkunar með tíföldum krossprófunum (e. 10-fold cross-validations). Fyrir nafnakennsl (e. NER) tilgreinum við F1-gildi með tíföldum krossprófunum. Fyrir vélrænan útdrátt notum við ROUGE aðferðina, sem mælir hlutfall skörunar n-stæða milli markmiðs útdráttar og myndaðs útdráttar. Við tilgreinum ROUGE-2 heimt.

Líkan	PoS	NER	ATS
Besta ekki-Transformer	95,6%[20]	83,7%[21]	50,8%
ELECTRA-small	96,7%	90,3%	44,4%
IceBERT	96,8%	91,8%	59,4%

Fyrir bestu ekki-Transformer niðurstöður fyrir ATS, tilgreinum við gildið fyrir Lead-3, einfalda grunnlínuaðferð sem býr til samantekt úr fyrstu þremur línum tiltekins skjals.

[18] IceSum: http://hdl.handle.net/20.500.12537/96
[19] Símonarson, H., Snæbjarnarson V., Ragnarsson, P., Einarsson, H. (2021). Language modeling for morphologically rich languages (IceBERT). Manuscript in preparation, Miðeind and University of Iceland.
[20] ABLTagger: http://hdl.handle.net/20.500.12537/53
[21] http://hdl.handle.net/1946/36562

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 4 - lýsing

Matssett fyrir setningafræðileg líkindi og skyldleika tilbúin, byggð á alþjóðlegum stöðlum og aðlöguð að þörfum íslenskrar orðhlutafræði.

Afurðir

- Tæmandi listi allra merkingareinkunna fyrir MultiSimLex.
- Listi einkunna frá þeim sem unnu merkingarvinnu við MultiSimLex og voru ekki síuð út af APIAA.
- Textaskrár sem innihalda orðapör fyrir hvern flokk íslenska settsins.
- Textaskrár sem innihalda allar mögulegar samsetningar af orðafernum fyrir hvern flokk, til að nota með vigrahliðrunar aðferðinni til að meta orðagreypingar.

Allar afurðir má finna hér: https://github.com/steinunnfridriks/greypingar_profunarsett

Skýrsla

MultiSimLex:
Fyrir setningafræðileg líkindi völdum við að þýða Multi-SimLex[22] (til styttingar munum við nota „OMSL“ þegar við vísum í eiginleika sem eiga aðeins við upprunalega (e. original) útfærslu og aðferðafræði Multi-SimLex, og „MSL“ þegar við vísum annaðhvort í eigin útfærslu, eða eiginleika sem eru óbreyttir úr upprunalegri útfærslu). MSL gagnasettið er nokkuð nýlegt. Það er viðauki eldra gagnasetts sem er vel þekkt og kallast SimLex-999, en það hefur margfalt gagnamagn. Þau gögn hafa verið fengin úr ýmsum öðrum gagnasettum, þeirra á meðal SemEval-17, Card-660, og SimVerb-3500, með það að markmiði að gefa OMSL víðari og jafnari fjölbreytni en SimLex-999 hafði. Höfundar OMSL hafa þegar gefið gagnasettið út á 12 tungumálum, og hafa komið á fót röð prófana og bestu verklagsreglna til að tryggja bestu mögulegu gæði merkinga gagnasettsins.

[22] https://arxiv.org/abs/2003.04866

Byggingu nýs MSL-gagnasetts má skipta í tvo megin áfanga: Þýðing upprunalegra enskra orðapara; og merkingu setningafræðilegra líkinda paranna.

MSL-gagnasett samanstendur af 1.888 orðapörum. Orð geta verið endurtekin milli ólíkra para, en hvert par verður að vera einstakt og bæði orð þess frábrugðin hvoru öðru. Af því leiðir að fyrir hvaða tvö ólíku orð sem er, „A“ og „B“, má aðeins vera til eitt (A::B) eða (B::A) par í öllu settinu, og (A::A) & (B::B) eru ekki leyfð. Þýðingar verða að viðhalda setningafræðilegum skyldleika hvers pars þegar mögulegt er. Eins orðs þýðingar eru ákjósanlegar, en orðasambönd (e. multi-word expressions) má nota ef þau hjálpa í tilteknum aðstæðum. Ef margir möguleikar eru jafnir í gæðum eru þýðendur beðnir að velja einn af handahófi. Taka skal fram - og höfundar OMSL viðurkenna gagngert - að þessar skorður geta stundum leitt til nokkurra vandkvæða við þýðingu heils OMSL, sérstaklega í tilfellum þar sem mörg samheiti eða sambærileg orð eru til staðar fyrir tiltekið hugtak á ensku, en mun færri á markmiðstungumálinu.

Í fyrri þýðingum af OMSL voru tveir aðilar fengnir til að þýða allt MSL gagnasettið: Svokallaður „aðal“ (e. main) þýðandi, hvers þýðing er notuð fyrir endanlega niðurstöðu, og „sjálfstæður“ (e. „independent“) þýðandi, hvers þýðing er notuð til að mæla gæði vinnunnar. Vegna flækjustigs við þýðingu ákveðinna orða, ákváðum við að nota tvo „aðal“ þýðendur frekar en einn.

Fyrir „sjálfstæðu“ mælinguna eru 100 pör valin af handahófi með sama hlutfall málfræðilegra flokka og allt MSL settið. Samræmi milli þýðenda er á reiki milli tungumála, og auk þess oft á milli einstakra málfræðilegra flokka. Heildarmeðaltal fyrir tungumál OMSL er 84,8%, á meðan MSL þýðingar voru með 74,0%. Í íslenska settinu var árangur á nafnorðum góður, með 83,9% samræmi; sagnorð voru nokkuð neðar með 70,8%; og bæði lýsingarorð og atviksorð náðu aðeins 50,0%. Svipuð mynstur ósamræmis, þó ekki jafn öfgafull, má sjá í eistnesku, finnsku og spænsku. Helsta orsök þessara niðurstaðna eru mjög líklega erfiðleikar við þýðingar þeirra orðapara þar sem bæði orðin sjálf og merkingarlegi þráðurinn sem tengir þau eiga sér ekki eina ákveðna hliðstæðu í markmiðstungumálinu. Í þeim tilfellum er mjög líklegt að skilyrði þess að þurfa að velja þýðingu af handahófi útskýri ósamræmi milli þýðinga. Þar að auki er hægt að sjá áhugaverðan viðsnúning þegar þessar einkunnir eru bornar saman við mikilvægari merkingareinkunnir: Þau tungumál sem höfðu hæsta samræmi milli þýðenda - þeirra á meðal velska, hebreska og rússneska - höfðu lægsta samræmi milli merkjara, þar sem íslenska náðu auðveldlega betri árangri.

Í merkingarferlinu eru nokkrir þátttakendur beðnir að gefa líkindum allra para í MSL einkunn. Það eru engin samskipti milli þátttakenda, en þeim er leyfilegt að nota utanaðkomandi hjálpartæki eins og orðabækur eða samheitaorðabækur. Aðeins sett með fullkláruðum einkunnum allra 1.888 para eru nothæf.

Í OMSL eru þátttakendur svo beðnir að yfirfara, og hugsanlega breyta einkunnum sem sýna mest ósamræmi milli einkunna hinna þátttakendanna. Eftir það er meðal samræmi hvers merkjara reiknað á móti meðal samræmi hinna, og þeim settum sem sýna mikið ósamræmi er hent. Í samantekt okkar af þessum gagnasettum reyndist töluverð áskorun að fá þátttakendur til að merkja að fullu heil gagnasett. Þar sem ekki var öruggt að áframhaldandi yfirför einkunna yrði

tekin fyrir af þátttakendum, eða heldur að þau myndu veita okkur betri samheldni gagna, ákváðum við að hætta við yfirferðina og einbeita okkur frekar að því að safna meira magni gagna. Við enduðum því með 20 fullkláruð gagnasett, sem eru fleiri en meðaltalið í mati OMSL sem er 12.

Í MSL eru einkunnasett metin með endurteknum reikningi á því hversu hátt samræmi er í einkunnagjöf þeirra sem merkja, en það er reiknað út á tvenns konar hátt (e. average pairwise inter-annotator agreement (APIAA) og average mean inter-annotator agreement (AMIAA)). Í hverri umferð, ef fjöldi setta sem eru eftir er yfir tíu, og núverandi lægsta APIAA einkunnin er hærri en lægsta APIAA einkunn úr fyrri umferð, er núverandi sett fjarlægt úr skoðun og ferlið er keyrt aftur.

Einkunnin 0,6 eða hærri gefur til kynna „sterkt samræmi“ bæði fyrir APIAA og AMIAA. APIAA hefur verið notað víða, á meðan AMIAA er nýrra en jafnvel nákvæmara. MSL sett okkar fékk tiltölulega háar einkunnir, eins og sjá má í eftirfarandi töflum:

Mál	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE	**ICE**
no.	0,661	0,662	0,659	0,558	0,647	0,698	0,538	0,606	0,524	0,582	0,626	0,727	**0,648**
lo.	0,757	0,698	0,823	0,695	0,721	0,741	0,683	0,699	0,625	0,640	0,658	0,785	**0,800**
so.	0,694	0,604	0,707	0,580	0,644	0,691	0,615	0,593	0,555	0,588	0,631	0,760	**0,715**
ao.	0,699	0,593	0,695	0,579	0,646	0,595	0,561	0,543	0,535	0,563	0,562	0,716	**0,704**
Samtals	0,680	0,619	0,698	0,583	0,646	0,697	0,572	0,609	0,530	0,576	0,623	0,733	**0,690**

Tafla 1: Average pairwise inter-annotator agreement (APIAA).

Mál	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE	**ICE**
no.	0,757	0,747	0,766	0,696	0,766	0,809	0,680	0,717	0,657	0,710	0,725	0,804	**0,771**
lo.	0,800	0,789	0,865	0,790	0,792	0,831	0,754	0,792	0,737	0,743	0,686	0,811	**0,859**
so.	0,774	0,733	0,811	0,715	0,757	0,808	0,720	0,722	0,690	0,710	0,702	0,784	**0,819**
ao.	0,749	0,693	0,777	0,697	0,748	0,729	0,645	0,655	0,608	0,671	0,623	0,716	**0,778**
Samtals	0,764	0,742	0,794	0,715	0,760	0,812	0,699	0,723	0,667	0,703	0,710	0,792	**0,799**

Tafla 2: Average mean inter-annotator agreement (AMIAA):

Íslensku einkunnirnar eru skráðar í dálkinn lengst til hægri. Í hverri línu er hæsta einkunn áherslumerkt, og íslenska nær hæstu einkunn fyrir APIAA á lýsingarorðum og sagnorðum. Yfir heildina er APIAA meðaleinkunn 0,631 á meðan íslenska nær töluvert hærri 0,690. Sömuleiðis er AMIAA meðaleinkunn yfir heildina 0,740, á meðan íslenska nær 0,799.

Að lokum kemur fram í OMSL skýrslunni að meðaleinkunn merkjara (þ.e., ekki samræmi merkjara, heldur einfaldlega meðaleinkunn sem hver merkjari gefur orðapörunum sjálfum) sé 1,61 og miðgildið 1,1, og að dreifingin sé á bilinu 0,137 (rússneska) til 0,160 (pólska). Hér er íslenska aftur nokkuð sambærileg: Meðaltöl hennar (meðaltal 1,77 og miðgildi 1,0) eru mjög svipuð öðrum tungumálum, á meðan dreifingin er 1,99, sem gefur til kynna góða fjölbreytni einkunna sem munu hjálpa við prófanir á greypingum í M5.

Íslenskt beygingar- og afleiðsluprófunarsett:
Gögnum hefur verið safnað og prófunarsettið er fullklárað. Það er byggt á BATS (The Bigger Analogy Test Set) sem var lagt til af Gladkova, Drozd og Matsuoka árið 2016[23], en hefur verið aðlagað sérstaklega að íslenskum þörfum og getur ekki talist þýðing upprunalega settsins. Upprunalega BATS er skipt í fjóra meginflokka, tveir merkingarfræðilegir og tveir orðhlutafræðilegir. Íslenska prófunarsettið heldur orðhlutafræðilegu flokkunum fyrir beygingar- og afleiðsluorðapör, auk lítils alfræðiundirsetts. Upprunalega BATS innihélt einnig orðfræðilegt (e. lexicographic) undirsett sem við slepptum alfarið þar sem við teljum íslensku þýðingu MultiSimLex þjóna sama tilgangi. Að auki, þar sem íslenska er orðhlutafræðilega auðugt mál, er nauðsynlegt að leggja sérstaka áherslu á þá flokka til að ná fyllilega taki á íslenskum orðagreypingum.

Beygingar- og afleiðsluhlutarnir hafa hvor um sig 9 undirflokka. Eftirfarandi er lýsing á hverjum flokki.

Beygingarundirflokkarnir innihalda samanburð á:
- frumstigi og miðstigi lýsingarorða í öllum þremur kynjum
- óákveðnum og ákveðnum greini nafnorða í nefnifalli, eintölu
- nefnifalli og eignarfalli nafnorða í eintölu, án greinis
- nefnifalli nafnorða í eintölu og fleirtölu
- nafnhætti og framsöguhætti sagna í eintölu, þátíð
- nafnhætti sagna og lýsingarhætti þátíðar
- lýsingarhætti þátíðar og framsöguhætti þátíðar í eintölu

Undirflokkar afleiðslu eru:
- lýsingarorð með forskeytið ó- og lýsingarorð/lýsingarháttur nútíðar sem þau eru leidd af
- lýsingarorð með viðskeytið -legur og nafnorðin sem þau eru leidd af
- nafnorð með forskeytið aðal- og nafnorðin sem þau eru leidd af
- nafnorð með viðskeytið -ari og sagnorðin sem þau eru leidd af
- nafnorð með viðskeytið -andi og sagnorðin sem þau eru leidd af
- nafnorð með viðskeytið -ing og sagnorðin sem þau eru leidd af
- nafnorð með viðskeytið -ingur og nafnorðin sem þau eru leidd af
- sagnorð með viðskeytið -a og nafnorðin sem þau eru leidd af
- sagnorð með viðskeytið -ja sá lýsingarháttur nútíðar sem þau eru leidd af

[23] https://www.aclweb.org/anthology/N16-2002.pdf

Alfræðiflokkarnir eru aðeins tveir þar sem þeir eru ekki meginmarkmið vörðunnar. Þeir bera lönd saman við þjóðerni (t.d. Þýskaland - þýskur) og bæ við bæjarbúa (t.d. Reykjavík - Reykvíkingur). Allir flokkarnir innihalda 50 orðapör. Hvert orðapar er notað sem tilvísunarhluti í hliðstæðuspurningu (X er fyrir Y eins og A er fyrir hvað?) fyrir öll önnur pör innan hvers undirflokks. Þessum spurningum er beint að hliðrunarviguraðferðinni sem metur orðhlutafræðileg gæði greypinga. Samtals eru einstakar hliðstæðuspurningar í íslenska prófunarsettinu 49.004 sem er, eins og við var búist, rétt yfir helmingur þess sem upprunalega BATS inniheldur.

Orðapörunum var safnað úr útgáfu 20.05 af Risamálheildinni (RMH) sem inniheldur 1500 milljón lesmálsorð í texta ásamt orðhluta- og setningafræðilegum mörkum og uppflettimyndum þeirra. Með því að skoða tíðni markanna í málheildinni komumst við að því að 78% allra einstakra orða eru nafnorð á meðan sagnorð eru aðeins 2%. Þetta kemur ekki á óvart þar sem nýyrði verða oftast til fyrir nafnorð en sjaldnar fyrir aðra orðflokka. Taka skal fram að RMH er mörkuð vélrænt og ekki yfirfarin af mannlegum sérfræðingum sem þýðir að hún er ekki laus við villur. Samt sem áður gefa þessar niðurstöður hugmynd um vægi nafnorða í tungumálinu og það er af þeirri ástæðu að nafnorðum eru gefnir fleiri flokkar afleiðslu en lýsingarorðum og sagnorðum.

Orðin eru valin eftir orðhluta- og setningarfræðilegum tilgangi þeirra í tungumálinu og eftir tíðni þeirra í RMH. Hið síðarnefnda er gert í tveimur skrefum. Fyrst er framkvæmd leit byggð á málfræðilegum eiginleikum orðsins á 100.000 algengustu orðum RMH (þ.e.a.s. uppflettimyndir fyrir afleiðslu- og alfræðiflokkana og orðmyndir fyrir beygingarflokkana). Eftir var mannlegra sérfræðinga á möguleikum eru 50 orðapör valin fyrir hvern flokk og aftur keyrð á móti tíðni RMH. Endanlegt val samanstendur af orðum sem koma fyrir a.m.k. 1000 sinnum í RMH nema í flokkum nafnorða með forskeytinu aðal-, sagnorðum með viðskeytinu -a, sagnorðum með viðskeytinu -ja og í alfræðiflokknum bæjar/bæjarbúa, sem allir hafa þröskuld af a.m.k. 500 tilfellum í RMH vegna skorts á mögulegum orðum.. Við reynum að velja möguleg orð á þann máta að um 20% séu mjög algeng (koma fyrir oftar en 100.000 sinnum í RMH) á meðan 80% koma sjaldnar fyrir. Hins vegar er þetta ekki alltaf mögulegt þar sem fjöldi mögulegra orða er mismunandi eftir flokkum.

Orðagreypingar eru metnar með fyrrnefndum orðapörum með því að nota vigrahliðrunaraðferðina. Þetta er stöðluð mæliaðferð til að leysa hliðstæðuspurningar (e. analogy questions) í vigurrúmslíkani. Mismunur milli tveggja vigra A og A* er vigurhliðrunin sem samsvarar málfræðilegu sambandi þeirra. Tökum sem dæmi orðin *walk* og *walking*. Með því að draga vigurgildi orðsins *walk* frá vigurgildi orðsins *walking*, ættum við að fá stærðfræðilegt jafngildi málfræðilegu *-ing* endingarinnar. Við setjum svo inn orðið *B* og vörpum fram spurningunni *A er gagnvart A* eins og B er fyrir hvað?* Ef *B* táknar orðið *talk*, ætti svarið við spurningunni að vera vigurinn B* hvers hnit í vigurrúmi samsvarast hnitum orðsins *talking*. Við leitum því að þeim vigra sem hefur nálægustu kósínuslíkindin (e. cosine similarity) við punktinn *B** í vigurrúminu. Þessari aðferð verður lýst nánar í skýrslu M5.

Framtíðar- og áframhaldandi vinna:
Hafin er ytri stika möskvaleit (e. hyperparameter grid search) að oðragreypingum þjálfuðum á RMH. Fyrstu niðurstöður lofa góðu fyrir bæði prófunarsettin.

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 4 - lýsing

- Nýjum EES-textum safnað og bætt við málheildina.

Afurðir

Allar afurðir verða viku á eftir áætlun.

Skýrsla

Nýleg EES-skjöl sem innihalda lagagjörninga (e. legal acts), sem voru ekki þegar hluti af ParIce, var halað niður frá efta.int og eur-lex.europa.eu. Skjöl sem innihalda ákvarðanir sameiginlegu EES-nefndarinnar (e. the EEA joint committee) voru ekki hluti af fyrstu útgáfu ParIce en var nú safnað, alls 9.510 skrár. Viðbæturnar samanstanda af u.þ.b. 500 þúsund setningum.
Öll gögnin, eldri gögnin auk viðbóta, hafa verið síuð og þeim endursamraðað. Í gæðaeftirliti fundum við villur sem komu í pípunni, og höfum ákveðið að lagfæra pípuna og endurtaka ferlið. Þetta þýðir því miður að endanlegur pakki verður ekki tilbúinn fyrir enda almanaksviku 5, fyrstu viku febrúars, og upplýsingar um nákvæma tölu þýðingarbúta verða aðgengilegar eftir að því ferli lýkur.

V4 – Opin þýðingarvél fyrir íslensku

Skýrsla: Miðeind
Aðilar: Miðeind

Tilgangur

Markmiðið á öðru ári er að aðlaga, innleiða og fínstilla valdar viðurkenndar lausnir fyrir taugavélþýðingar milli íslensku og ensku. Þessi vinna byggir á lærdómi frá fyrsta ári, sem sýndi fram á að taugavélþýðingaraðferðir eru umtalsvert betri en hefðbundin tölfræðilíkön. Einkum áformum við að halda áfram með sambland af viðgjafarlausum forþjálfunum á BERT-legum mállíkönum þvert á tungumál, og huldum runu-til-runu þýðingarlíkönum.

Varða 4 - lýsing

- Gögn undirbúin; bráðabirgðaniðurstöður frá forþjálfuðum líkönum aðgengilegar

Afurðir

Öllum afurðum var náð. Einmála gögnum á íslensku og ensku hefur verið safnað, þau hreinsuð og undirbúin fyrir þjálfun. Bráðabirgðafínstillingu forþjálfaðra margmála líkana er lokið.

Skýrsla

Gögn

Einmála íslensk gögn eru fengin frá Risamálheildinni (undirverkefni G1). Ensk gögn eru fengin frá almennu Books3 málheildinni[24], Wikipedia[25] og Newscrawl[26].

Samhliða (raunverulegu) íslensk-ensku þjálfunargögnin eru að mestu þau sömu og notuð voru í fyrri vörðum og byggð á gögnunum í ParIce (undirverkefni V2). Einnig höfum við notað málheild Votta Jehóva (e. Jehovah's Witnesses Corpus) sem inniheldur um 500 þúsund íslensk - ensk setningapör.[27]

Bráðabirgðaniðurstöður

Fyrstu prófanir hafa verið byggðar á því að fínstilla mBART líkanið[28]. mBART er afsuðandi (e. denoising) sjálfvirkur kóðari þjálfaður á einmála gögnum frá 25 ólíkum tungumálum.

[24] https://github.com/soskek/bookcorpus
[25] https://dumps.wikimedia.org/backup-index.html
[26] Newscrawl gagnasettið er gefið út með WMT http://www.statmt.org/wmt19/translation-task.html
[27] http://opus.nlpl.eu/JW300.php
[28] https://arxiv.org/abs/2001.08210

Til að fínstilla mBART fórum við frá því að nota Tensorflow og Tensor2Tensor safnið yfir í Pytorch og Fairseq safnið.

Öll gervibakþýdda málheildin, þróuð á fyrsta ári, var notuð ásamt raunverulegu samhliða gögnunum. Raunverulegu gögnin voru þanin (e. upsampled) til að verða ⅔ gagnanna sem sjást í þjálfun, sem er útfærslan sem reyndist best á fyrsta ári.

Samanburður *BLEU-einkunna* hér fyrir neðan er að sumu leyti ósanngjarn þegar kemur að samanburði á undirliggjandi uppbyggingu kerfanna tveggja. Sanngjarn samanburður arkítektúrs er hins vegar ekki í raun markmið verkefnis okkar; heldur einfaldlega raunhæfar endurbætur í árangri þýðinga eins og endanlegur notandi sér. mBART líkanið sem við þjálfuðum er stærðargráðu stærra en base-Transformer og þarfnaðist meiri reikniafls.

BLEU-einkunn	**is-en**	**en-is**
Transformer-base	24,8	21,4
mBART þýðing	26,8	23,5

Þó að einkunnin hafi batnað umtalsvert og nái þegar markmiðum M6, eru enn nokkur augljós vandamál varðandi þol sem við viljum taka á yfir næstu mánuði með með því að nota meðal annars brottfallsaðferðir og fínstilla kóðun. Sérstaklega geta stafsetningarvillur og há-/lágstafavillur valdið villum í þýðingarúttaki.

Dæmi um þýðingu texta EES-reglugerðar sem var ekki með í þjálfunarsettinu:

Upprunaleg enska: *Where Member States consider it appropriate, experts accredited for the energy performance certification of buildings, or for the inspection of heating, air-conditioning and combined heating or air-conditioning and ventilation systems under Directive 2010 31 EU, or for performing energy audits under Directive 2012 27 EU of the European Parliament and of the Council (4), should be allowed to be considered competent also to assess the smart readiness of buildings or building units.*

Íslensk þýðing: *Ef aðildarríkin telja það viðeigandi ætti að heimila sérfræðingum, sem hafa fengið faggildingu til að votta orkunýtingu bygginga eða til að skoða hitunar-, loftræsti- og sambyggð hitunar- eða loftræstikerfi samkvæmt tilskipun 2010 31 ESB eða til að framkvæma orkuúttektir samkvæmt tilskipun Evrópuþingsins og ráðsins 2012 27 (4), að teljast einnig til þess bærir að meta hvort byggingar eða byggingareiningar séu tilbúnar til notkunar með skynsamlegum hætti.*

Þessi þýðing er fullkomlega ásættanleg, þrátt fyrir lengd og flækjustig upprunalegu setningarinnar, og rétt fyrir utan „snjallhæfni“ (e. „smart readiness“) sem þýðist sem „notkun með skynsamlegum hætti.“

Eftirfarandi þýðingu fréttatexta úr (örlítið slappri) íslensku í ensku er hins vegar ábótavant:

Upprunaleg íslenska: *Boris Johnson hafði ekki síður mikið álit á Trump. Sagði til dæmis að næði Trump að laga Norður-Kóreu og kjarnorkusamninginn við Íran væri hann ekki síðri kandídat fyrir friðarverðlaun Nóbels en Obama, sem fékk verðlaunin áður en hann gerði nokkurn skapaðan hlut, sagði Johnson.*

Johnson hefur látið fleiri hrósyrði falla í garð Trumps en þetta um friðarverðlaunin hefur verið rifjað einna oftast upp undanfarið. Ekki síst af því Johnson lét hrósið falla 2018, þegar það var orðið nokkuð ljóst að það var ekki alltaf mikið að marka yfirlýstar fyrirætlanir Trumps.

Ensk þýðing: *Boris Johnson had no less of an opinion of Trump. For example, said Trump's reach to fix North Korea and the nuclear deal with Iran was "not a less-than-ideal candidate for the Nobel Peace Prize" than Obama, who received the award before making any move, Johnson said.*

Johnson has dropped more compliments on the Trumps, but this one about the Peace Prize has been retweeted one of the most frequent times recently. Not least of all because Johnson dropped the 2018 compliment, when it had become fairly clear that there wasn't always much to mark out about the Trumps' stated intentions.

Vandamálin eru meðal annarra orð sem vantar (*said → he said that*) þar sem vantar samhengi innan setningar, og þýðing eignarfallsins *Trumps* úr íslensku í fleirtöluna *the Trumps* á ensku. Netið vill oft líka þýða íslensk orðasambönd of bókstaflega, eins og *mikið að marka → much to mark out*, þar sem réttari þýðing væri *taken seriously* eða álíka. Þessu er hægt að taka á með meiri þjálfunargögnum úr ólíkum áttum.

Núverandi en-is tauganetsþýðingarvélin er opin á vefnum og hægt að prófa á https://velthyding.is.

V4c – Gervimálheild

Skýrsla: Miðeind
Aðilar: Miðeind

Markmið

Lykiláskorun við þjálfun taugavélþýðingarkerfa fyrir tungumál með litlum og meðalmiklum gögnum (e. low- and medium-resource) eins og íslensku er að þjálfa málheildir á á gjöfum með nógu víðtækum orðaforða, sem t.d. inniheldur einnig sjaldgæf orð og óðalsbundin orð (e. domain-specific words). 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 íslensku 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 4 - lýsing

- Upphafsmati lokið

Afurðir

Afurðunum hefur verið skilað með þeim hætti sem greint er frá í skýrslunni hér að neðan. Við vekjum athygli á fráviki frá þeirri nálgun sem lýst var í Markmiðum að ofan.

Skýrsla

Við höfum rannsakað bæði þá nálgun sem lýst er í upprunalegu skilgreiningu verkefnisins og annars konar nálgun með notkun bakþýðinga og tauganetstaps (e. neural network loss). Við höfum staðfest að upprunalega nálgunin virkar til myndunar tilbúinna þjálfunargagna, sem skapar málfræðilega réttar setningar með tilætlaðan orðaforða. Prófanir hafa hingað til takmarkast við nafnorð, en lýsingarorð og jafnvel sagnorð gætu bæst við í viðbót í framtíðinni.

Upprunaleg nálgun

Prófanir okkar með samhliða íslensk-enskri málheild ágripa ritgerða frá *Skemmunni*[29], gagnagrunn ritgerða háskólanemenda í grunn- og meistaranámi. Þessi texti er fræðilegur, formlegur og nákvæmur að eðli, og hentar vel sem grunnur til að draga út óðalsbundinn orðaforða fyrir tauganet til að læra af.

[29] https://skemman.is/?locale=en

Grunnmálheildin var fyrst keyrð í gegnum síu (með þáttaranum úr undirverkefni I5) sem greindi fyrirfram tilbúið sett nafnorða og samsvarandi ensk orð. Dæmi um par í almenna settinu er *rannsókn -> study*. Úttak síunnar gefur lista samhliða sniðmátasetningar, til dæmis:

Þátttakendur í {þgf_et_gr} voru flokkaðir eftir aldri og tekjutíundum. -> Participants in the {sg} were categorized by age and income deciles.

Sniðmátin innihalda nauðsynlegar upplýsingar innan slaufusviga (fall, tala, ákveðinn/óákveðinn greinir) til að gefa kost á málfræðilega réttum skiptum nýrra nafnorða í stað þeirra upprunalegu.

Tilbúnar sniðmátaskrár eru flokkaðar frekar eftir kyni, svo að karlkyns hugtökum sé aðeins skipt í sniðmátum karlkyns nafnorða, o.s.frv. Þetta er einföld leið til að tryggja samræmi með kynjuðum lýsingarorðum og fornöfnum í samhengi hvers nafnorðs, án þess að þurfa að framkvæma fulla lausn samvísana (e. coreference resolution).

Þegar málheild nokkurra þúsunda sniðmátasetninga hefur verið búin til, er hún notuð til að framkalla sjaldgæf hugtök úr lista hugtakapara (íslensku -> ensku) og búa til tilbúna málheild. Til dæmis, úr stjarnfræðilegu hugtakasafni (*tifstjarna -> pulsar*):

Þátttakendur í tifstjörnunni voru flokkaðir eftir aldri og tekjutíundum. -> Participants in the pulsar were categorized by age and income deciles.

Tilbúnu málheildirnar sem búnar eru til verða að vera nógu stórar til að innihalda hvert hugtak í öllum mögulegum beygingarmyndum, sem í tilfelli nafnorða eru venjulega 16 myndir (4 föll x {eintala, fleirtala} x {óákveðinn/ákveðinn greinir}).

Í stuttu máli, höfum við sýnt fram á að þessi nálgun virkar til að búa til gervi þjálfunarsetningar (e. synthetic training sentences). Eftir stendur að skoða hvaða blöndunarhlutfall þessara gervigagna gefur bestu niðurstöður í þjálfun, þ.e. leyfir netinu að læra tilætlaðan orðaforða án þess að skekkja verulega samhengisskilning. Síðarnefndu áhrifin eru viðbúin þar sem gervisetningarnar, þó þær séu málfræðilega réttar, eru auðvitað frekar óskiljanlegar hvað varðar innihald þeirra.

Frekari vinna mun bæta fyrirfram tilbúin almenn nafnorðapör, sem gefur kost á útdrætti fleiri sniðmátasetninga; að finna notkunardæmi hugtaka og möguleg prófunarsett; og finna besta blöndunarhlutfall gerviþjálfunargagna fyrir hugtök í hin þjálfunargögn þýðingarnetsins.

Annars konar nálgun

Eins og tekið er fram að ofan er sjaldnast hægt að skipta orðum eða orðasamböndum innan setninga án þess að merkingin verði óskiljanleg, og lítil sem engin úrræði eru til fyrir íslensku til að meta hvort slíkar breytingar hafa átt sér stað.

Þar sem bakþýðingar eru þegar notaðar í stórum stíl sem hluti af máltækniverkefninu er allur náttúrulega myndaður íslenskur texti þegar innan ramma þess að vera notaður sem þýðingarmarkmið. Það er enginn skortur á slíkum enskum textum.

Í ljósi þessa höfum við skoðað möguleika annars konar nálgunar þar sem tvímála orðalisti, nánar tiltekið tvímála íðorðabanki, er notaður samhliða bakþýðingum til að taka á skorti á samhliða óðalsbundnum textum til að bæta þýðingar markmiðsóðala. Niðurstaða tilraunarinnar er ný bakþýdd málheild með viðbættum (e. injected) orðaforða á upprunalegu hliðinni. Kostur við þessa nálgun væri svo mældur. Þessa málheild má nota til aðlögunar óðala, fyrir utan hlutverk hennar sem almenn samhliða málheild.

Yfirlit aðferðar

Þarfir

Fyrir gefið óðal þarf þessi gagnasett:

a) Tvímála orðalisti, skyldur óðalinu
b) Einmála texti markmiðshliðar (en) sem inniheldur orð úr orðalistanum
c) Einmála texti upprunahliðar (is) í óðalinu (kostur, en strangt til tekið ekki nauðsynlegt)

Að auki þurfum við þýðingarlíkan úr markmiðshlið í upprunahlið (til að búa til bakþýðingar), auk líkans úr upprunahlið í markmiðshlið sem hægt er að bæta.

Gagnavinnsla

Hér búum við til bakþýddu málheildina með innspýtingu íðorðaforða (e. terminology injection):

a) Texti markmiðshliðar (en) er bakþýddur
b) Þegar orð í orðalistanum er til staðar í markmiðstexta er því skeytt inn í þýðinguna (upprunahlið). Bestu aðferð til að gera þetta þarf að meta en samröðunarmat (e.alignment estimation) eða samröðunarathygli (e. alignment attention) koma helst til greina.

Þjálfun

c) Ef einmála afsuðunarlíkan (e. denoising model) er til fyrir íslensku, er hægt að nota það til að fara betur yfir tillögurnar úr (b), sem sjálfstæðan metil úr (d)
d) Halda nýju framvirku (e. forward) líkani (is->en) til að afsuða (e. denoise) þjálfunardæmin sem búin eru til í (c) og innihalda suð
e) Nota nýja framvirka líkanið (is->en) ásamt afturvirka (e. backward) líkaninu (en->is) til að fága og bæta þýðingarmöguleikana í hluta (b)
f) Halda afturvirku líkani (en->is) til að búa til betrumbætta möguleika (e. candidates) (með orðasafnshömlum). Athugið að hér erum við að þjálfa líkan með gervigögn sem markmið.
g) Ítra skref (c) til (f) eins og þarf

Mat

Bæði staðlað sjálfvirkt þýðingamat og mannlegt mat verður notað. Fyrir sjálfvirkt mat notum við BLEU þegar viðmiðunarþýðing til samanburðar er til. Fyrir mannlegt mat mælum við ákjósanlega þýðingu, málfimi og nægð.

Frummat

Fyrir verkefnið höfum við valið svið stjörnufræði þar sem við höfum góð gögn fyrir fræðasviðið, meðal annars samhliða málheild sem inniheldur texta frá Stjörnustöð Evrópulanda á suðurhveli (e. European Southern Observatory, ESO), um 13.000 setningapör.

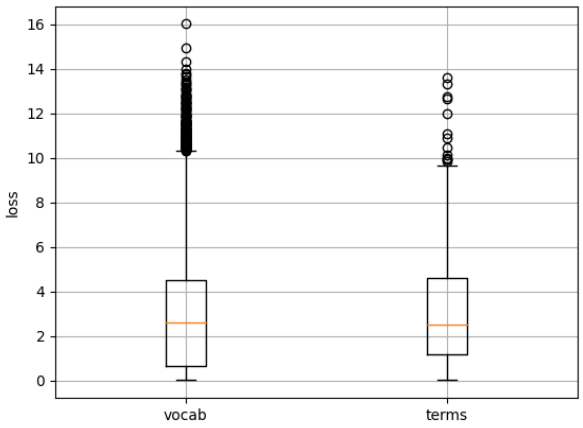
Á markmiðshliðinni notum við ensku Wikipedia sem upprunatexta stjörnufræði.

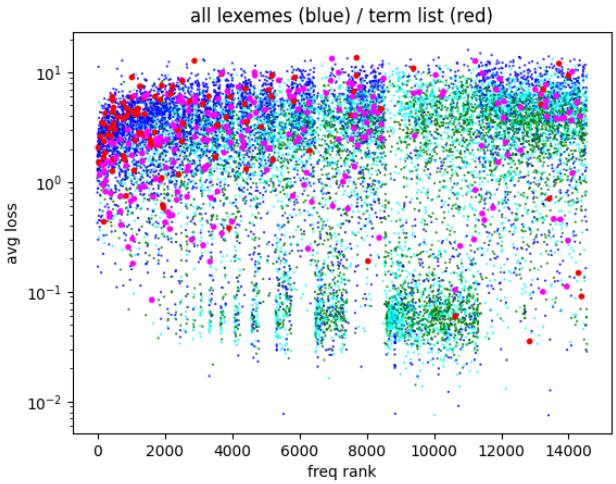
Íðorðabankinn inniheldur 2.400 þýðingar milli íslensku og ensku fyrir stjörnufræði og 4.600 þýðingar fyrir eðlisfræði.

Með þessum gögnum fylgjum við yfirliti aðferðarinnar hér að ofan.

Tapgreining

Við mælum tapið á óðalbundnum, áður óséðum texta, sem tengist orðum í íðorðabankanum sem borinn er saman við samsvarandi orðasett af svipaðri tíðni. Þetta er framkvæmt á þýðingalíkaninu sem ekki hefur verið þjálfað á auknu málheildinni.





Kassateikningin (e. box-plot) að ofan sýnir tap fyrir allan orðaforðann (14.219 án hugtakalistans) samanborið við íðorðin á íðorðalista fyrir stjörnufræði og eðlisfræði (559). Grafið tap yfir tíðni (e. loss over frequency graph) er litað eftir fjölda undirliggjandi orðflísatóka í hverju orði. Samkvæmt þessu getum við ekki komist að neinni niðurstöðu um hvort það er marktækur munur í þýðingartapi fyrir íðorðin í samanburði við almenna orðaforðann í þessari óðalsbundnu, óséðu málheild. Kostir þessarar nálgunar eru því enn óljósir en má rannsaka frekar.

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 handvirkt yfirfarin hönnunar- og prófunarsett. Slík sett búum við til úr textum bæði innan og utan óðala. Með það að markmiði að koma íslensku inn í eitt af sameiginlegu verkefnunum á WMT 2021, verða prófunarsett utan óðals 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 4 - lýsing

- Skilgreina gjafa og stærð þróunar- og prófunarsetta.
- Merkja 50% af þróunar- og prófunargögnum innan óðals.
- Komast að niðurstöðu um það hvort íslenska verði hluti af sameiginlegu verkefni á WMT 2021 og nákvæmlega hvaða gögn þarf.
- Skilgreina umfang vinnu við orðalista

Afurðir

Afurðum hefur verið skilað.

Skýrsla

Skilgreina gjafa og stærð þróunar- og prófunarsetta

Þróunar- og prófunarsett munu innihalda 22.000 gild setningapör frá fjórum mismunandi óðulum auk heilla skjala frá þremur af fimm óðulum.

Óðal	**Pör**
Upplýsingablöð sjúklinga (EMA)	6.000
OpenSubtitles	6.000
EES skjöl	6.000

Fréttatilkynningar Stjörnustöðvar Evrópulanda á suðurhveli (ESO)	2.000
Fréttir frá www.norden.org	2.000
SAMTALS	**22.000**

Frá EMA og OpenSubtitles munum við aðeins gefa út sett einstakra setningapara sem valin hafa verið af handahófi og án samhengis. Frá EES, ESO og Norden munum við gefa út sett setningapara, sett málsgreina og sett skjala.

Merkja 50% þróunar- og prófunargögn innan óðala

Við höfum merkt 6.000 gild setningapör frá bæði EMA og OpenSubtitles, samtals 12.000. Við höfum einnig valið 50 textaskjöl frá EES til að nota í prófunar- og þróunarsetti. Þar sem við sóttum textana úr pdf-skrám þurfum við að hreinsa þá og meðhöndla á ákveðinn hátt til þess að samröðunarferli gangi án hnökra. Við höfum þegar hreinsað og meðhöndlað 20 skjöl sem eru nú tilbúin til samröðunar. Við höfum einnig valið 94 skjöl frá ESO sem verður samraðað.

Skilgreina umfang vinnu við orðalista

Markmið vinnu við orðalista er að búa til handyfirfarna orðalista, með íslenskum orðum og enskum þýðingum þeirra, og öfugt. Möguleg orð fyrir orðalistana verða sótt með ýmsum tvímála aðferðum til útdráttar úr orðasafni (e. lexicon induction methods), m.a. með margmála greypingum (e. cross-lingual and multilingual embeddings), útdrætti mögulegra para með orðasamröðun og með viðgjafarlausum vélþýðingum (e. unsupervised machine translation) til að draga út orðasambandatöflur (e. phrase tables). Hvert par á listum möguleika verður skoðað handvirkt. Greypingar verða svo notaðar til að finna orð sem eru nátengd þeim ásættanlegu til að finna fleiri mögulegar þýðingar fyrir hvert orð. Fyrstu tilraunir hafa verið gerðar fyrir þessar ólíku aðferðir, sem bjuggu til lista möguleika fyrir einstök orð og stutt orðasambönd. Við takmörkuðum orðasamböndin við hámark þriggja orða, en þau eru höfð með til að finna réttar þýðingar fyrir orð sem eiga sér ekki fullkomlega samsvarandi orð í hinum tungumálunum. Dæmi um þetta eru: *rafbíll - electric car, electric vehicle; lánshæfi - credit rating; frá því, síðan - since; rekast saman, rekast á - collide; takast á - clash, fella niður, hætta við - cancel*. Í þessum tilraunum hafa u.þ.b. 6.000 pör verið samþykkt handvirkt sem mögulegar þýðingar. Við stefnum á að búa til lista með minnst 40 þúsund orðum og þýðingum þeirra, yfirfarin handvirkt og samþykkt.

WMT

WMT brást jákvætt við þegar við höfðum samband í lok september 2020 varðandi það að bæta íslensku-ensku í sameiginlega fréttaþýðingaverkefnið (e. the Shared News Translation task) fyrir WMT 2021. Fyrsta febrúar 2021 fengum við endanlega staðfestingu þess að íslenska-enska muni vera hluti af WMT 2021. Íslenska verður fyrsta skandinavíska tungumálið í WMT.

Vinna við þróunar- og prófunarsett fyrir WMT 2021 er yfirstandandi. Alls þarf 4.000 setningar: 2.000 setningar fyrir hvora átt (en->is og is->en), sem er báðum skipt í 1.000 setningar fyrir þróunarsett og 1.000 setningar fyrir prófunarsett. Þýðingu þróunarsettsins verður að vera lokið fyrir lok febrúar. Við skrif þessarar skýrslu hafa 700 setningar verið þýddar í báðar áttir.

V4e – Þýðingarvél fyrir sérsvið

Skýrsla: Miðeind
Aðilar: Miðeind

Markmið

Að tryggja að vélþýðingarkerfi máltækniverkefnisins séu nothæf og viðeigandi fyrir ákveðin óðul, 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 4 - lýsing

- Nákvæm áætlun samstarfsins, prófunarumhverfis og prófunaraðferða tilbúin

Afurðir

Afurðum hefur verið náð. Nákvæm áætlun samstarfsins og prófunarumhverfis er sett fram hér að neðan.

Skýrsla

Inngangur

Samningur um samstarf við þýðingamiðstöð utanríkisráðuneytis Íslands hefur verið undirritaður. Í þýðingarmiðstöðinni starfa tugir þýðenda og starfsfólks og vinna að mestu við að þýða reglugerðir Evrópska efnahagssvæðisins (EES) á íslensku. Þetta er án efa stærsta einstaka íslenska þýðingarverkefnið í heiminum og frábær aðili til að vinna með. Taka skal fram að stærsta og gæðamesta samhliða málheild til þýðinga milli íslensku og ensku er afurð þýðingarmiðstöðvarinnar.

Gögn

Gögn sem gerð eru aðgengileg eru tiltækar þýðingar á TMX-formi, um 2 milljónir setninga. Auk þeirra eru nokkur óþýdd hrá skjöl gefin til prófana innbyggða þýðingarumhverfisins.

Samstarf

Sem hluti af samstarfinu munu rannsakendur hjá Miðeind og þýðendur þýðingarmiðstöðvarinnar vinna saman til að

a) Kortleggja allt vinnuferli þýðenda
b) Skilgreina þarfir fyrir tauganetsþýðingarkerfi (NMT system) sem gæti orðið hluti af þýðingarferli
c) Skilgreina kerfi sem mætir þeim þörfum

d) Prófa og skilgreina prófunaraðferðir fyrir innleiðingu tauganetsþýðingarkerfis (NMT)

Auk þess veitir þýðingarmiðstöðin Miðeind skipulögð gögn (e. structured data) til að nota í þróun lausnarinnar. Endanlegar prófanir og innleiðing veltur á því að Miðeind þrói og innleiði tauganetsþýðingarvél (e. NMT engine) sem má innleiða í vinnuferli þýðenda.

Vinnuferli þýðenda

Þýðendur þýðingarmiðstöðvarinnar nota hugbúnaðinn *Trados*[30] við þýðingu texta. Vinnuferlinu er gróflega skipt í eftirfarandi skref:

a. Verkefni er úthlutað þýðanda úr einum af 4 meginflokkum skjala og 30 undirflokkum
b. Microsoft Word-skjali með EES-reglugerðum sem skal þýða er hlaðið inn í hugbúnaðinn
c. Fyrsta þýðing er gerð með beinum samanburði á setningastigi á móti þýðingarminni. Vanalega finnast á bilinu 20% - 90% setninga í skjalinu.
d. Þýðandinn fer yfir skjalið setningu fyrir setningu
    i. Skoðar Trados-einkunn, setningar með einkunn undir 70% þarf vanalega að endurskrifa
    ii. Notar hugtakasafn til að leita eftir orðum í óþýddum setningum
    iii. Skoðar hvort þýðing er rétt með tilliti til undirflokks
    iv. Réttlætir þýðingar byggt á fyrri þýðingum
    v. Notar venjulega textaleit í samhliða málheild, orðabókum og Google til að skoða málnotkun
e. Ritrýnir fer yfir uppkast þýðingar og leitar eftir villum

Skref d. og e. eru endurtekin eftir þörfum.

Kerfiskröfur/þarfir

Helstu kröfum/þörfum má skipta gróflega í þrjá flokka:

a) Auðveldleiki notkunar
    i) Tólið skal vera innleitt í Trados-kerfið
    ii) Notkun ætti að vera einföld og uppruni upplýsinga skýr, þ.e. tauganets-þýddar setningar ættu að vera greinilega merktar
b) Gæði þýðinga
    i) Þýðingarnar skulu ekki aðeins skrifaðar á góðri íslensku heldur einnig nota orðaforða og stíl sem samræmist fyrri þýðingum
c) Áreiðanleiki þýðinga (e. Translation confidence)
    i) Samröðun þýðinga eða stuðningur úr þýðingarminni
    ii) Frá hvaða undirflokki þjálfunargögnin bakvið þýðinguna koma, og hvernig þýðingin var gerð (þ.e. úr þýðingarminni eða með tauganetsþýðingu)

Taka skal fram að tilgreining uppruna (e. origin and provenance) þýðinga (hluti rannsókna á túlkunarfærni og skýringarhæfni lausna vélræns náms) er yfirstandandi rannsóknarefni og óvíst að hvaða leyti við munum geta uppfyllt þær kröfur/þarfir.

[30] https://www.sdltrados.com/

Prófunarumhverfi og NMT-innleiðing

Flækjustig þess að forrita sérhæfða íbót (e. native plugin) fyrir Trados-kerfið í forritunarmálinu C# er í skoðun en miðað við bráðabirgðamat teljum við það innan ramma verkefnisins.

Meginmarkmið okkar er að nota íbótina þegar stuðningur þýðinga finnst ekki í þýðingarminni en við sjáum áframhaldandi leið handan ramma annars árs:

1. Þegar ekki er hægt að styðja þýðingu með þýðingarminni, leita í forritaskilum þýðinga (e. translation API), og tryggja að þýðingin sé merkt sem slík [**markmið annars árs**]
    a. Undirliggjandi þýðingarlíkanið verður fínstillt eða þjálfað sérstaklega á gefnum samhliða þýðingargögnum
2. Þegar þýðing er aðeins tiltæk að hluta úr þýðingarminni skal fylla inn þá hluta sem vantar með þýðingarþjónustunni. [**Framtíðarvinna**]
3. Hafa með upplýsingar um uppruna þjálfunargagna sem leiða til þýðinga (skýringarmöguleiki þýðingarákvarðana) [**Framtíðarvinna**]

Prófunaraðferðir

Til að meta kosti kerfisins verða nokkur lykilatriði skoðuð byggð á notkun þeirra af starfsfólki þýðingarmiðstöðvarinnar:

1. Fjöldi breytinga sem þarf eftir upprunalega þýðingu með og án notkunar kerfisins (e. post-edit distance)
2. Hversu mörgum þýðingartillögum er hent
3. Könnun meðal þýðenda þýðendamiðstöðvarinnar um hvort þeim finnst kerfið nytsamlegt

V5b – Lýðvirkjun

Skýrsla: Miðeind
Aðilar: Miðeind

Markmið

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, bæði 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 4 - lýsing

- Bakendi fyrir notendur og gagnageymsla tilbúin

Afurðir

Afurðunum hefur verið skilað í formi bakenda sem getur geymt notendaupplýsingar, annast færslubeiðnir byggðar á leyfum, geymslu þýðingargagna og mat þar að lútandi.

Skýrsla

Bakendinn er settur upp með Django[31] og PostgreSQL[32], sem eru bæði víða notuð og frí með opnu hugbúnaðarleyfi. Django hefur gott *ORM-lag* (e. object-relational mapping layer) sem varpar klasa (hugbúnaðarlíkön) í töflur og gefur mikla „sjálfbæra“ virkni fyrir notendur og við meðhöndlun gagna.

Að breyta og hlaða inn gögnum

Hægt er að hlaða upp gögnum í stórum skömmtum og hægt er að gera breytingar jafn óðum í gegnum stjórnendaviðmót ef notandinn hefur slík réttindi.

REST forritaskil

Endapunktarnir verða stilltir þegar kröfur fyrir viðmótið skýrast. Meginhluti vinnunnar felst í uppbyggingu skynsamlegra réttinda byggðri á auðkenndum notanda. Endapunktar fyrir gagnasett aðgengileg almenningi verða opnir til að stuðla að yfirferð þýðinga með sem minnstum óþægindum fyrir notandann.

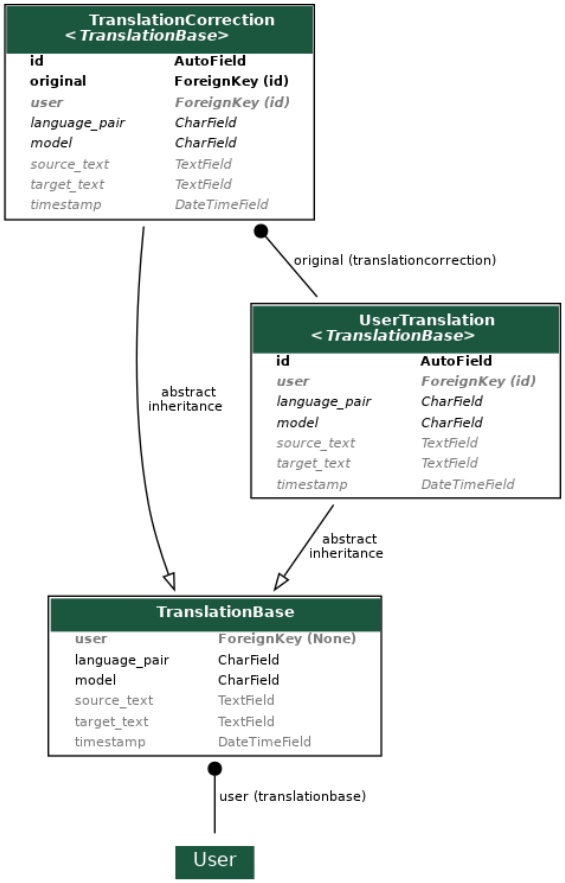
[31] https://www.djangoproject.com/
[32] https://www.postgresql.org/

Django REST-ramminn er notaður sem rammi fyrir endapunktana, sem gefur góða grunnvirkni byggða inn í uppsetningu líkansins.

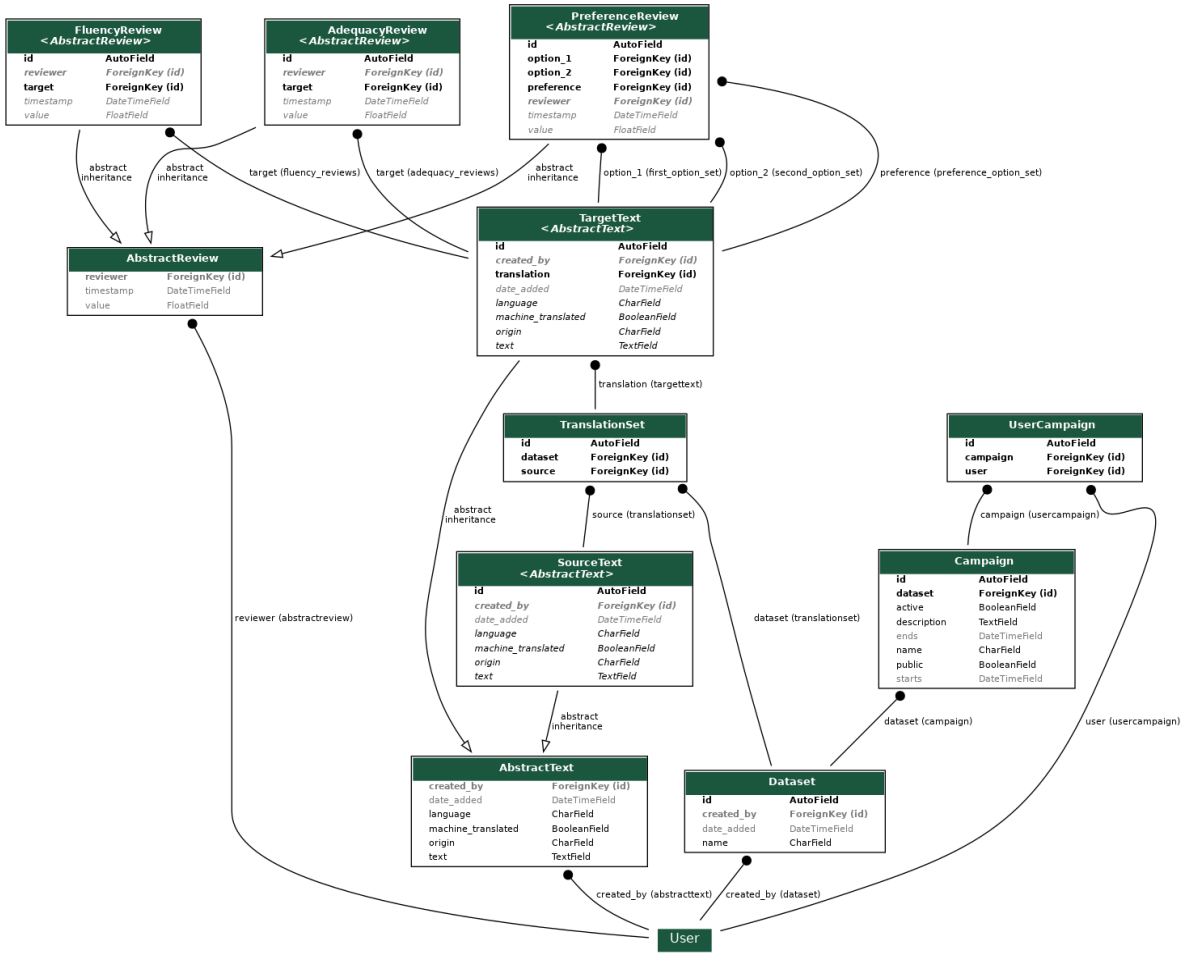
Gagnagrunnsskema

Hér fyrir neðan er skýringarmynd á aðal gagngrunnstöflunum. Búist er við frekari fínstillingum fyrir vinnu fyrir M5 þegar viðmót mats og yfirferðar verður sett upp. Taka skal fram að hugrænir klasar eða töflur eru ekki til staðar í gagnagrunninum.

Við geymum notendaupplýsingar í gagnagrunninum og gefum einstökum notendum hlekki á þýðingar, ritrýni, herferðir og gagnasett. Mismunandi töflur tengja svo markmiðs- og upprunatexta með ritrýnendum og einkunnum þeirra fyrir nægð, málfimi og val.



Gagnagrunnsskýringarmynd fyrir þýðingar notenda



Gagnagrunnsskýringarmynd fyrir ritrýni notenda

L2 - Sérhæfðar villumálheildir

Skýrsla: HÍ
Aðilar: HÍ, 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 4 - lýsing

Verk

- Búa til rafrænt eyðublað um upplýst samþykki til að láta texta af hendi.
- Söfnun texta frá þeim sem læra íslensku sem annað mál.
- Söfnun texta frá málhöfum með lesblindu.
- Leiðrétta villur í annars máls textum og lesblindutextum.
- Flokka villur samkvæmt mörkunarskema sem þróað var fyrir almennu villumálheildina, og bæta við flokkum eftir þörfum fyrir ákveðna samfélagshópa.
- Flytja út á endanlega aðlagað TEI-snið og gefa út á CLARIN.

Söfnun texta frá málhöfum með lesblindu vel á veg komin. A.m.k. 2.000 villur úr þeim textum hafa verið greindar og merktar.

Afurðir

Ekki náðist að klára öll markmið vörðunnar. Færri en 2.000 villur hafa verið greindar úr textum frá málhöfum með lesblindu. Þetta er vegna þess að söfnun texta frá málhöfum með lesblindu hefur reynst erfiðari en búist var við. Áætlun til að ljúka við söfnunina er vel á veg komin, bæði í formi fleiri auglýsinga fyrir verkefnið og með samstarfi við Félag lesblindra á Íslandi. Þessi markmið verða uppfylltar fyrir M5.

Skýrsla

Textar frá málhöfum með lesblindu
Fyrsta útgáfa Íslensku lesblinduvillumálheildarinnar hefur verið gefin út á GitHub https://github.com/antonkarl/iceErrorCorpusSpecialized. Hún samanstendur af 9 textum með 831 endurskoðunarspannir og 1.285 flokkaðar villur. Textarnir eru að mestu stuttar ritgerðir og greinar, allir skrifaðir af aðilum sem hafa íslensku að móðurmáli og hafa verið greindir með lesblindu.

Söfnun þessara texta reyndist erfiðari en búist var við þó að nokkrar mismunandi leiðir söfnunar hafi verið skoðaðar. Námsráðgjöf við Háskóla Íslands gat ekki hjálpað vegna áhyggja varðandi

trúnað, og auglýsingar verkefnisins innan háskólans hafa enn ekki borið árangur, sem fólust meðal annars í samskiptum við nemendafélög varðandi auglýsingar. Við höfðum einnig samband við fyrirtæki sem vinna með lesblindum, Félag lesblindra á Íslandi og TMF[33] Tölvumiðstöð, og erum í samstarfi við framkvæmdastjóra Félags lesblindra á Íslandi til að safna fleiri textum. Þar sem framkvæmdastjórinn er í nánu sambandi við fólk með lesblindu er vonin að áhrif hans muni hjálpa til að ná fleiri textum. Hann vinnur nú að stuttu myndbandi sem kynnir textasöfnunina, sem hvetur fólk vonandi til að senda inn texta.

Við höfum einnig bætt við valmöguleika í innsendingarform þar sem þátttakendur geta skrifað texta beint í textareit í stað þess að senda inn tiltæka skrá. Þessi valmöguleiki veitir áhugasömum þátttakendum möguleika þess að gefa texta. Til að ná til sem flestra höfum auk þess bætt við hljóðbútum í skrá sem veitir upplýsingar um verkefnið, og útskýrir hvers vegna við erum að safna textum og hvernig fólk getur tekið þátt.

Textar frá þeim sem læra íslensku sem annað mál
Þó að vinna við lesblinduvillumálheild hafi gengið hægar en áætlað var, hefur Íslenska L2 villumálheildin stækkað verulega. Útgáfa 0.1 er aðgengileg á GitHub (https://github.com/antonkarl/iceErrorCorpusSpecialized) og samanstendur af 56 textum með 11.113 endurskoðunarspannir og 16.113 flokkaðar villur. Auk þeirra hafa 21 texti verið prófarkalesinn í annað sinn til að tryggja að allar villur hafi verið leiðréttar og að þær séu í raun réttar. Þessi vinna var gerð samhliða söfnun texta, sérstaklega texta lesblindra, og nýjar leiðréttingar gerðar á þeim öllum, sem sannreynir mikilvægi þessarar vinnu.

Merkingarferli og merkingarskema
Merkingarferlið og merkingarskemað úr undirverkefni L1, almennri villumálheild sem var lokið í M3, eru notuð fyrir sérhæfðu villumálheildirnar. Merkingarferlið notar lagskipta nálgun, sem hefst á prófarkalestri, Python-skrifta er notuð til samröðunar og sameiningar réttra og rangra útgáfa, og skilar úttaki á TEI-formi sem inniheldur kortlagningu villna til leiðréttinga. Villurnar sem finnast eru því næst merktar handvirkt og skráin úr því ferli gefin út sem hluti málheildarinnar.

Villukóðarnir úr undirverkefni L1 hafa verið þróaðir eftir því sem leiðréttingar sérhæfðu villumálheildanna þörfnuðust. Þeir hafa einnig verið endurbættir og samanstanda nú af 19 yfirflokkum, sem er öllum skipt frekar í undirflokka. Fjöldi undirflokka er mismunandi eftir flokkum þar sem sumir eru mjög nákvæmir og aðrir víðtækari, 258 alls. Endurbætur villukóðanna reyndist til bóta fyrir undirverkefni L7 sem og fyrir prófarkalesara og merkjara. Vinna prófakalesara varð staðlaðri og afmörkun milli villukóða varð skýrari fyrir merkjurum. Yfirflokkarnir eru (á ensku):

1. Capitalization
2. Collocation
3. Grammar
4. Syntax

[33] TMF stendur fyrir „tækni - miðlun - færni“

5. Nonword
6. Omission
7. Typo
8. Punctuation
9. Spacing
10. Insertion
11. Wording
12. Spelling
13. Foreign
14. Exclusion
15. Numbers
16. Style
17. Other
18. Lexical
19. Unannotated

Tæmandi lista villukóða, ásamt lýsinga og dæma fyrir hvern og einn kóða má finna hér https://github.com/antonkarl/iceErrorCorpusSpecialized/blob/master/IEC_ErrorCodes.pdf.

L7 - Kerfisbundin þróun málrýnis

Skýrsla: Miðeind
Aðilar: Miðeind, HÍ, SÁM

Markmið

Að þróa málrýnikerfi fyrir íslensku sem nær yfir ákveðið safn villna, sem eru greindar og þeim forgangsraðað á meðan verkefninu stendur.

Markmið verða skilgreind á ítraðan hátt eftir forgangsröðun villuflokka sem skilgreindir eru í undirverkefni L3 með gögnum frá undirverkefnum L1 og L2. Almenna markmiðið er að greina þessar villur og veita nytsamlegar lýsingar, tillögur og/eða stinga upp á leiðréttingum sem eru viðeigandi fyrir hvert tilfelli.

Ákvæði

- L1 Almenn villumálheild (varða M3)
- L2 Sérhæfðar villumálheildir (varða M6)
- L3 Tölfræðileg úrvinnsla og prófunarmálheildir (varða M3)
- L4 Orðalistar og mállíkön (varða M3)
- L6 Málrýnir fyrir stakorðavillur
- G4 DMII for language technology (allar vörður)
- I3 Tilreiðari (allar vörður)
- I5 Þáttari (fullþáttari (e. full-constituency parser), allar vörður)

Varða 4 - lýsing

- Skipanalínutól málrýnis fyrir stafsetningarvillur getur borið kennsl á og lagt til leiðréttingar fyrir (og valkvætt leiðrétt sjálfvirkt) sett algengustu samhengisháðu stafsetningarvillnanna, auk samhengisfrjálsra villna eins og unnið var að á fyrsta ári. Við stefnum á að heildar F-gildi verði 55-60% fyrir 10 algengustu flokka samhengisháðra stafsetningarvillna, og ættum að ná a.m.k. 75% fyrir 20 algengustu flokka samhengisfrjálsra villna. F-gildi er mælt á villugreiningu.
- Verkáætlun og skilgreiningar varða fyrir notkunardæmi fyrir samstarf við atvinnulíf til reiðu.

Afurðir

Öllum afurðum vörðunnar var náð.

Við náum F1-gildi 71,19% fyrir villugreiningu á 20 algengustu flokkum samhengisháðra villna innan sviðs. F0.5-gildið er 79,93%, sem nær markmiði okkar.

Við náum F1-gildi 41,87% fyrir villugreiningu á 10 algengustu flokkum samhengisháðra villna innan sviðs. F0.5-gildið er 55,20%, sem nær markmiði okkar.

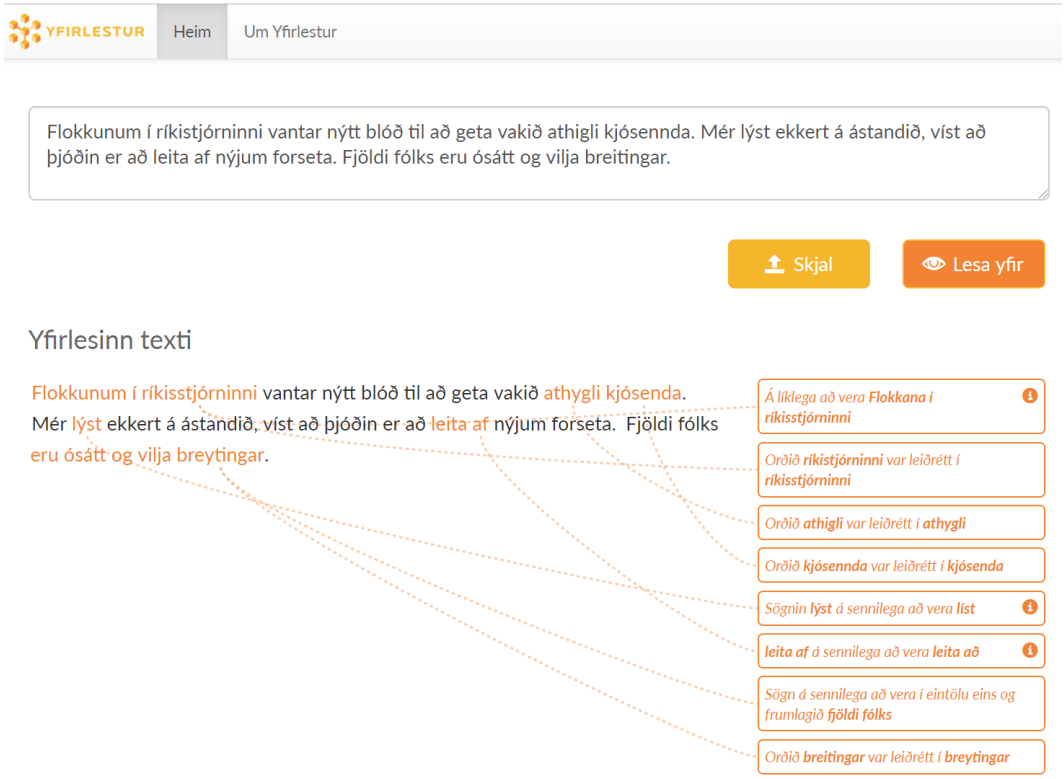
Ný útgáfa GreynirCorrect pakkans er aðgengileg á GitHub.

Skýrsla

1. Yfirlit

Þetta undirverkefni felur í sér þróun málrýnieiningar í Python, auk sýnidæmis í vefviðmóti þar sem notendur geta slegið inn texta og fengið hann sjálfkrafa prófarkalesinn og athugasemdir við stafsetningar- og málfræðivillur. Verkefnið byggir á gögnum sem hefur verið safnað í öðrum undirverkefnum, einna helst mörkuðu villumálheildinni *iceErrorCorpus* (IEC) úr L1, og orðalista og mállíkön úr L4 og L6.

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



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

Verkefnið er í samræmi við kóðunarstaðla SÍM. Frumkóðinn er kóðaður í UTF-8 og er á ensku. Einingaprófanir hafa verið innleiddar með `pytest` og verkefnið sett upp fyrir stöðugar prófanir með GitHub Actions. Frumkóðinn er forritaður í Python 3.6 og fylgir PEP-8. The Black formatter var notaður.

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

2. Forgreining

2.1 Gagnaþarfir

Þar sem við framleiddum margar tegundir gagna á fyrsta ári, og færum okkur yfir í flóknari villur, byrjuðum við annað ár á gagnagreiningu.
Við vörpuðum fram eftirfarandi spurningum:

- Hvers konar málgögn eru tiltæk og viðeigandi fyrir þetta undirverkefni?
- Hvaða gögn þurfum við?
- Hvernig geta tiltæk gögn aðlagast þörfum okkar?
- Fellur gagnaleyfið að gagnaleyfaþörfum okkar?
- Hvernig öflum við gagnanna og frá hverjum?
- Hvaða gögn þurfum við til að fylla í skörð, ef einhver?

Við þurfum gögn fyrir (1) samhengisfrjálsar villur, (2) samhengisháðar villum, og (3) málfræðivillur. Að auki þurfum við (4) opinberar leiðbeiningar til að fylgja til að meta hvað telst villa, og heimildir.

Fyrir (1), höfum við IceSQuEr, safn árangurslausra leitarbeiðna í BÍN (undirverkefni G4) sem hafa verið skoðaðar og tengdar við leiðréttingu. Við höfum einnig öll óorð úr þróunarhluta íslensku villumálheildarinnar.

Fyrir (2), höfum við *Ritmyndir* úr undirverkefni G4, óstaðlaðar ritmyndir einstakra beygingarmynda sem eru tengd uppflettimyndum í BÍN. Þessi gögn eru enn óútgefin en við höfum verið í nánu samstarfi við höfundana, bæði til að deilga gögnum og til að tryggja að flokkunarskemun séu samhæfð.

Fyrir (3), höfum við Málfarsbankann (á íslensku), safn stuttra greina um ýmis málfræðileg umræðuefni og svör við algengum spurningum um málfræðilega notkun.

Auk þessara gagna sem talin eru, höfum við gögnin sem þegar hefur verið safna í verkefninu. Við erum í miðju ferli þess að sameina þau gögn við nýju gögnin í eitt stakt form. Við erum einnig að bæta við upplýsingum, eins og villuflokkum, leiðréttingum, og tengingu við viðeigandi leiðbeiningar og heimildir.

Þróunarhluti íslensku villumálheildarinnar hefur einnig reynst nytsamlegur, sérstaklega fyrir málfræðilegar villur, þar sem aðrar heimildir eru fáar.

Fyrir (4), höfum við *íslenskan málstaðal*, eins og hann birtist í *Ritreglum* og *Stafsetningarorðabókinni* (aðgengileg hjá SÁM).

Eins og lýst er í hluta 4.2 höfum við innleitt samstarfsferli til að hámarka sameiginlegan ávinning verkefnis okkar og yfirstandandi vinnu við *íslenskan málstaðal*.

Fyrir aðrar tilvísunarheimildir höfum við BÍN auk Íslenskrar nútímamálsorðabókar (á íslensku), og *Málfarsbankans*, auk annarra.

2.2 Utanaðkomandi þarfir

Byrjun annars árs felur í sér fjölda undirverkefna sem reiða sig á málrýnirinn. Því mátum við það mikilvægt að byrja á greiningu þarfa hvers undirverkefnis.

Skilgreining þarfa undirverkefna L8 - málrýni í snjalltækjum, L9 - málrýni í ritvinnslukerfum, og L10 - aðlögun að máltæknihugbúnaði er hluti af vörðum þessara verkefna. Gróflega sjáum við að þau þurfa ritspá (e. autocompletion) með mállíkani, villuskynjun, villuleiðréttingu og/eða tillögur, orðabókarskilgreiningar, sérstök villuskilaboð og tengingu við staðla og heimildir, og mismunandi stillingar þegar kemur að hörku (e. strictness) og villuflokkum.

3. Prófanir og mat

3.1 Endurbætur pípu
Pípa prófana og mats er aðgengileg á https://github.com/mideind/GreynirCorrect/blob/master/eval.
Sjálfvirka matssvítan hefur verið verulega endurbætt fyrir M4. Fyrir M3 vörðuna innleiddum við grófgerða matssvítu sem gat mælt nákvæmni, heimt og F1-gildi fyrir villugreiningu á setningarstigi, þ.e. byggt á því hvort setningar eru sagðar innihalda villu eða ekki, réttilega eða ranglega (rangt/rétt neikvætt/jákvætt).

Fyrir M4 bættum við möguleika fyrir mat á tókastigi. Þetta gerði okkur mögulegt að víkka matið og bæta við niðurstöðum fyrir villugreiningu, spannarskynjun, og villuleiðréttingu á tókastigi.

Við bættum einnig við greiningu eftir villuflokk, yfirflokk, og textaflokk. Þetta hefur reynst vera brýnt fyrir endurbætur tólsins, eins og greint er frá í hluta 4, auk villumálheildarinnar. Við gefum niðurstöður fyrir tvær gerðir yfirflokka, þá sem notaðir eru í iceErrorCorpus, og yfirflokka sem nefndir eru í vörðunum.

Við fórum yfir villuflokkana og uppfærðum skilgreiningar eftir því hverjir þeirra voru innan sviðs og hverjir utan (e. in scope/out of scope). Við gefum svo niðurstöður fyrir aðeins þá villuflokka sem eru innan sviðs, og fyrir alla villuflokka. Þetta hefur að mestu áhrif á heildarniðurstöður og niðurstöður yfirflokka. Þessi skilgreining sviða getur breyst á meðan verkefnið stendur yfir.

Við bættum einnig F0.5 einkunn við niðurstöðurnar. Þessi einkunn er gjarnan notuð fyrir málrýni, þar sem hún er talin mikilvægari til að greina réttilega villur frekar en að greina allar villur, þar sem mikilvægt er að halda röngum jákvæðum í lágmarki.

Önnur mikilvæg breyting sem við gerðum á prófunarsvítunni var kortlagning villuflokka GreynirCorrect yfir í villuflokka iEC. Kortlagningin er ekki alltaf einn-á-móti-einum, en yfirflokkarnir eru í flestum tilfellum réttir. Þetta gerði okkur betur kleift að úthluta villuflokk fyrir röng jákvæð; þ.e. þar sem GreynirCorrect greinir ranglega villu.

3.2 Mæligildi
Við byggðum mæligildi okkar á aðferð sem notuð voru í BEA 2019 (B).

Í grófu máli notar hún (1) tókabyggða skynjun, (2) spannarbyggða skynjun, og (3) spannarbyggða leiðréttingu.

Spannarbyggða leiðrétting er erfiðust, þar sem leiðréttingin er aðeins merkt rétt ef spönn hennar er nákvæmlega rétt. Dæmið sem gefið er í BEA 2019 skýrir muninn vel:

Original	`I often look at TV`	**Span-based**	**Span-based**	**Token-based**
Reference	`[2, 4, watch]`	**Correction**	**Detection**	**Detection**
Hypothesis 1	`[2, 4, watch]`	Match	Match	Match
Hypothesis 2	`[2, 4, see]`	No match	Match	Match
Hypothesis 3	`[2, 3, watch]`	No match	No match	Match

Samanburður mismunandi mæligilda[34].

3.3 Prófunarmálheildir
Eins og kemur fram í L3 skýrslu fyrir vörðu M3 hefur *iceErrorCorpus* úr undirverkefni L1 verið skipt í prófunarsett og þróunarsett. Málheildin telst fullkláruð, en verður áfram fínstillt vegna áframhaldandi vinnu við málrýni.

Sem lítið dæmi gerðu niðurstöðurnar okkur kleift að sjá alla flokka sem eru til staðar í villumálheildunum. Við gátum því lagfært stafsetningarvillur villuflokka, yfirfarið flokka sem voru til staðar í viðmiðum en ekki í málheildunum, og bætt við lýsingu villuflokka sem eru til staðar í málheildinni en ekki í viðmiðum. Þetta bætti nákvæmni merkingar og niðurstöður okkar.

Einn eiginleiki hönnunar málheildanna er að faldaðar villur eru ekki leyfðar. Þetta flækir prófanir, þar sem við gætum leiðrétt eina villu en misst af annarri. Gott dæmi um þetta er eftirfarandi upprunaleg og leiðrétt setning í iceErrorCorpus:

Grindvík byrjaði → Leikmenn Grindavíkur byrjuðu.

Grindvík er samhengisfrjáls innsláttarvilla, en jafnvel þó við náum að leiðrétta hana í *Grindavík*, kallar stílvilla (sem er utan ramma þessa undirverkefnis) eftir eignarfallinu *Grindavíkur*. Þetta þýðir að við fáum ekki stig fyrir að leiðrétta innsláttarvilluna. Þetta skal hafa í huga þegar niðurstöðurnar eru skoðaðar.

[34] Sótt af: https://www.cl.cam.ac.uk/research/nl/bea2019st/#eval

Annar flækjandi eiginleiki er að samhliða, óskyldum villum er oft gefin sama villuspönnin, með mörgum villutögum. Þetta gerir rétta greiningu villanna erfiða og nær ómögulegt að gefa rétta leiðréttingu.

4. Greiningarferli

4.1 Forgangsröðun villna
Eftir innleiðingu greiningar á tókastigi í prófunarsvítunni eins og kemur fram í hluta 3.1 gátum við betur forgangsraðar villuflokkum eftir tíðni, yfirflokkum, og hvernig hægt væri að taka á þeim. Fyrir vörðu M4 var lögð áhersla á samhengisháðar villur.

4.2 Endurbótaferli
Við innleiddum ferli til að hámarka ávinning og samstarf. Sem hluta af prófunarsvítunni innleiddum við möguleika þess að sækja öll tilvik villuflokks úr þróunarsettinu.

Fyrir fórum yfir tilvik fyrir hvern villuflokk úr þróunarsettinu. Þetta gaf okkur betri innsýn í villuflokkana og hvað einkennir hvern og einn. Því næst skrifuðum við almenna lýsingu þess flokks. Næst lýstum við meginreglum fyrir atriði í flokknum, skilgreindum jaðartilvik og svipuð tilvik sem eiga heima í öðrum flokki. Dæmi sem áttu heima í öðrum flokkum voru lagfærð. Í einhverjum tilfellum þurfti að skipta flokkum, eða færa hluta atriða í aðra flokka. Viðeigandi reglur úr *íslenskum málstaðli* voru útskýrðar með hjálp höfunda þeirra.

Við fengum ýmis atriði úr þessu.

Í fyrsta lagi, skýra lýsingu á villuflokkunarskemanu. Þetta hafði marga kosti, eins og að hjálpa merkjendum úr L2 við að nota réttan flokk og einnig skýrari skiptingu milli flokka, sem stuðlar að nákvæmari merkingu. Merkjendurnir geta líka borið upp spurningar vegna vandamála við flokkunina ef þeir sjá eitthvað í nýjum textum sem þarf að skoða.

Lýsing skemans er líka í sjálfu sér mikilvægt málfang fyrir komandi verkefni.

Í öðru lagi, skýrari og innihaldsríkari *íslenskan málstaðal*. Þegar við rákumst á óskýrar reglur eða málefni sem hafði ekki verið tekið á, veittum við SÁM gögn til endurbætts efnis, og innsýn í hvar þarf að bæta dekkun. Dæmin úr þróunarsettinu voru þá tilbúin til notkunar í *málstöðlum*.

Í þriðja lagi, drög að viðeigandi villuskilaboðum og viðmiðum fyrir notendur. Þetta mun vera dýrmætt þegar við komum að þeim hluta verkefnisins.

Í fjórða lagi, gott yfirlit til að vinna eftir við villumeðhöndlun. Með öllum þessum gögnum gætum við auðveldlega séð hvort villuflokkurinn ætti að teljast dekkaður eða ef aukinnar meðhöndlunar væri þörf, og hvernig yfirferðartólið virkaði með dæmum úr þróunarsetti fyrir flokkinn.

Vinnu við prófunar- og þróunarsett iceErrorCorpus var haldið aðskilinni. Allir meðlimir hópsins máttu vinna með villuflokkana í þróunarsettinu, en aðeins þeir sem höfðu unnið við

iceErrorCorpus (við Háskóla Íslands) máttu vinna með prófunarsettið. Prófunarsettinu var aðeins hægt að breyta (ef það taldist innihalda villu) eftir að þróunarsettið hafði verið uppfært, og lýsing var tilbúin. Þetta tryggði bestu skiptingu milli setta, og að prófunarsettinu væri aðeins breytt samkvæmt ákvörðunum og reglur byggðar á þróunarsettinu, en ekki öfugt.

5. Endurbætur

Sumir algengir villuflokkar féllu undir fyrirliggjandi gögn og reglur. Fyrir eftirfarandi flokka innan sviðs, lýsum við framvindu síðan í vörðu M3.

5.1 Hástafavillur (lower4upper-initial, upper4lower-noninitial, upper4lower-common, caps4low, lower4upper-proper, lower4upper-acro)
Fyrir M3 skýrsluna hafði meðhöndlun ranglegra há-/lágstafaorða verið innleidd, en aðeins með bráðabirgðavillugögnum. Við höfum aukið gögnin verulega og bætt meðhöndlunina, byggt á rökum úr þróunarsettinu. Reglur um há- og lágstafi í *Ritreglum* voru einnig metnar þegar gögnin voru aukin.

5.2 Samsetningarvillur (compound-collocation, compound-nonword, split-word, split-words, split-compound, missing-dash, merged-words, missing-space)
Við uppfærðum gögn okkar varðandi hvaða orðhlutar geta verið sjálfstæð orð og hvaða orðhlutar þurfa að vera hluti af samsetningu samkvæmt *íslenskum málstaðli*. Þetta bætti niðurstöður okkar fyrir samsvarandi villuflokka.

5.3 Beygingarvillur (nominal-inflection, verb-inflection)
Við erum að vinna að sameiningu listanna sem nefndir eru í M3 skýrslunni. Þar sem samsvarandi vinnu við BÍN er ekki lokið, hefur meðhöndlun ekki verið innleidd.

5.4 Bannorð (taboo-word, gendered)
Gögnunum sem voru undirbúin í undirverkefni L4, sem innihalda víðtæka lista bannorða og fínstillta flokkun þeirra, hefur verið bætt við. Þau juku dekkun umtalsvert.

5.5 Innsláttarvillur (missing-letter, extra-letter, letter-rep, swapped-letters, missing-accent, extra-accent, wrong-accent, y4ý, nonword)
Stækkaða n-stæðumállíkanið hefur bætt niðurstöður fyrir þessa flokka. Það bætir sérstaklega árangur fyrir broddstaafavillur (e. accent mark errors). Við höfum einnig bætt við sameinuðum gögnum úr undirverkefni L4 til að taka á algengustu tilvikum og erum að vinna að viðbót skýringa fyrir villur.

5.6 Stafsetningarvillur (n4nn, nn4n, i4y, y4i, í4ý, hv4kv, pronoun-writing)
Stækkaða n-stæðumállíkanið úr undirverkefni L4 getur skynjað og leiðrétt þessar villur, en þar sem við þurfum að ná í viðeigandi villuskilaboð og viðmið er það ekki alltaf fullnægjandi. Gögnin sem nefnd eru í 5.4 og 5.5 innihalda einnig algengar stafsetningarvillur, og skýringar fyrir þær.

5.7 Skammstafanavillur (abbreviation-period, abbreviation)

Tilreiðingin tekur á röngum skammstöfunum. Við höfum einnig gögn fyrir flóknari rangar skammstafanir.

5.8 Hugbúnaðarvinna
Við höfum bætt afköst og minnisnotkun þáttarans (undirverkefni I5), sem bætti skilvirkni málrýnitólsins, sérstaklega á löngum og flóknum setningum.

5.9 Aðrar endurbætur
Ný útgáfa BÍN frá janúar 2021 hefur verið sett saman sem hluti af undirverkefni G4. Hún hefur verið tengd kerfinu og orðaforðinn uppfærður.

6. Niðurstöður

Niðurstöður eru gefnar fyrir villugreiningu á iceErrorCorpus prófunarsettinu. Taka skal fram að villutíðnin í prófunarsettinu er sýnd, en flokkunum er raðað eftir tíðni í þróunarsettinu, þar sem það sett er mun stærra og gefur betri mynd af óðalinu.

Heildarniðurstöður fyrir villugreiningu á setningastigi eru eftirfarandi:

```
Setningar: 5229
Rétt jákvætt:     777         Rétt neikvætt: 3449
Rangt jákvætt:    577         Rangt neikvætt: 426
Heimt: 65,59%                 Nákvæmni: 57,39%
F1 eink.: 60,77%              F0.5 eink.: 58,69%
```

Síðan í vörðu M3 hefur F1 einkunnin hækkað frá 53,99% í 60,77%. Þetta eru einu niðurstöðurnar sem hægt er að bera beint saman milli varða, þar sem mælingar á tókastigi höfðu ekki verið innleiddar í M3. Samanburðurinn bendir til meiriháttar endurbóta.

Heildarniðurstöður fyrir villugreiningu á tókastigi eru eftirfarandi:

```
Tókar: 103686
Rétt jákvætt:     684         Rétt neikvætt: 99779
Rangt jákvætt:    566         Rangt neikvætt: 2657
Heimt: 20,47%                 Nákvæmni: 54,72%
F1 eink.: 29,80%              F0.5 eink.: 41,00%
```

Við getum ekki borið þessar tölur beint saman við niðurstöður úr M3 skýrslunni (sem voru byggðar á setningum, ekki tókum), og taka skal fram að heildarniðurstöður á tókastigi eru fyrir alla villuflokka, líka þá sem eru utan sviðs (eins og orðalags- og stílvillur). Greiningin að neðan sýnir ítarlegar niðurstöður fyrir flokka innan sviðs.

Niðurstöður á tókastigi fyrir samhengisháða villuflokka eru eftirfarandi:

Villuflokkur	*F1 eink.*	*F0.5 eink.*	*Tíðni í prófunarsetti*

upper4lower-common	35,85	45,67	72
split-compound	42,11	61,54	29
missing-space	62,5	80,65	22
collocation	44,44	66,67	14
lower4upper-initial	0	0	4
repeat-word	88,89	89,33	4
repeat-word-split	0	0	1
collocation-idiom	0	0	1
name-error	-	-	0
split-words	-	-	0

Micro F1 *eink.*: 41,87% Micro F0.5 *eink.*: 55,20%

Eins og sjá má náum við 41,87-55,20% fyrir 10 algengustu samhengisháðu stafsetningarvilluflokkana innan sviðs, sem nær væntingum okkar um 55-60%. Villa (e. bug) í prófunarforritinu þegar enduskoðaðar vörður fyrir M4 voru skilgreindar olli því að þeim var spáð í hærra lagi. Í ljósi þess, og þess að enn vantar mikilvæg gögn, erum við sátt með niðurstöðurnar. Við bendum einnig á að F0.5 einkunnin er mikið hærri en F1 einkunnin, sem samræmist markmiði okkar um að halda röngum jákvæðum (e. false positives) í lágmarki til að tryggja sem besta upplifun fyrir notendur.

Niðurstöður á tókastigi fyrir samhengisfrjálsa villuflokka eru eftirfarandi:

Villuflokkur	*F1 eink.*	*F0.5 eink.*	*Tíðni í prófunarsetti*
missing-letter	81,82	91,84	78
merged-words	87,86	90,69	91
extra-letter	83,12	92,49	45
lower4upper-proper	40,91	63,38	35
letter-rep	75,76	88,65	41
nonword	20,38	14,01	38
missing-accent	83,08	92,47	36
compound-collocation	68,29	84,34	27
split-word	64,00	81,63	17
n4nn	69,39	85,00	32
nn4n	83,33	95,59	14
swapped-letters	92,31	96,77	7
fw	5,66	3,9	13
i4y	50,00	71,43	3
extra-accent	100	100	8
abbreviation-period	83,33	92,59	7
y4i	100	100	4
compound-nonword	66,67	66,67	3
pronun-writing	0	0	1
ý4í	-	-	0

Micro F1 *eink.*: 71,19% Micro F0.5 *eink.*: 79,93%

Eins og sjá má náum við 71,19-79,93% fyrir 20 algengustu samhengisfrjálsu villuflokkana innan sviðs, sem samræmist vonum okkar um 75%.

Flokkarnir 'nonword' og 'fw' fá mikið lægri einkunn en búast mætti við. Þetta er vegna þess að þeir eru sjálfvalin kortlagning (e. the default mapping) fyrir marga villuflokka í GreynirCorrect, sem skýrir hvers vegna þeir fá flest röng jákvæð (e. false positives). Þetta er ekki til vandræða, því þó að þetta valdi ójöfnum villuflokkum er kortlagning milli yfirflokka að mestu rétt svo þetta hefur ekki áhrif á niðurstöður.

7. Samstarf við atvinnulíf

7.1 Kjarninn

Undirpakki innan L7 er notkunardæmi (e. use-case project) í samstarfi við veffréttamiðilinn Kjarnann (https://kjarninn.is/).

Markmið verkþáttarins er að innleiða málrýnitólið í ritstjórnar- og vinnsluferli fréttasíðu. Þetta ætti að hjálpa ritstjórum og fréttafólki síðunnar að skapa og gefa út texta með færri villum og auknum skynjuðum gæðum, sem bætir upplifun lesenda. Á okkar enda fáum við reynslu með notkunardæmi í raunheimi, bæði hvað varðar tæknileg atriði (samþættingu við hugbúnað þriðja aðila, textaform…) og hvað varðar raunverulegar þarfir og forgangsröðun í krefjandi „náttúrulegt“ umhverfi fyrir málrýni.

Eins og komið hefur fram, er meginmarkmið innleiðingar málrýnitóls fyrir Kjarnann að öðlast hagnýta reynslu af raunverulegri notkun. Eins og er felur þetta í sér eigindlega endurgjöf frá þeim notendum sem um ræðir, varðandi árangur tólsins og hagnýta þætti innleiðingarinnar. Þetta mun stýra áframhaldandi vinnu í kjarnaverkefninu og uppfylla mikilvægar kröfur notenda fyrir undirverkefni L8 - málrýni í snjalltækjum - og L9 - málrýni í ritvinnslukerfum.

Að auki erum við að skoða möguleika þess að safna eigindlegri endurgjöf, eins og tölfræði villna og tölfræði um hagnýtt notagildi tólsins, t.d. hvort tillaga að leiðréttingu er samþykkt af notandanum. Þetta myndi bæði vera ómetanleg endurgjöf fyrir frekari innleiðingar tólsins og stýra áframhaldandi vinnu við L7 verkefnis sjálft.

Sem hluta af M4 skilgreindum við verkáætlun og vörður.

Vinna við samstarfsverkefnið er þegar hafin, og starfsfólk Kjarnans hefur sýnt velvilja sinn auk þess að veita aðgang að hugbúnaðarbakenda þeirra. Þetta undirverkefni mun þróast samhliða kjarna L7, með það að markmiði að verða full innleitt í ritstýringarumhverfi Kjarnans fyrir vörðu M6, þegar málrýnitólið verður tilbúið fyrir notendaprófanir. Endurgjöf og niðurstöðum verður safnað fyrir vörðu M6.

M4: Á ekki við

M5: Kröfur samþættingar í ritstýringarumhverfi Kjarnans skilgreindar og innleiddar staðbundið.

Umfang mögulegrar eigindlegrar endurgjafar gert skýrt.

M6: Málrýni samþætt í ritstýringarumhverfið. Prófanir hafa verið gerðar. Skýrsla um niðurstöður og kröfur og þarfir notenda tilbúin.

7.2 Annað samstarf

Við höfum gefið hinum nýja atvinnulífstengli SÍM kynningu á GreynirCorrect, iceErrorCorpus, og prófunarsvítunni. Kynningin innihélt ítarlegar upplýsingar um hvar má finna heimildir og hvernig mætti hugsanlega beita þeim til að koma til móts við kröfur notenda. Við búumst við því að kynningin muni leiða til áhugaverðra tækifæra til að vinna með fyrirtækjum á komandi árum.

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. vélþýðing og leit.

Varða 4 - lýsing

Skýrsla um greiningarvinnu tilbúin.

Afurðir

Markmið þessarar vörðu er greiningarskýrsla sem þjónar tilgangi undirbúnings fyrir innleiðingu málrýnikerfis (L7) í annan máltæknihugbúnað. Markmiðum var náð.

Skýrsla

Inngangur

Málrýnirinn GreynirCorrect (https://github.com/mideind/GreynirCorrect) veitir málrýnivirkni SÍM verkefnisins (sjá undirverkefni L7). GreynirCorrect er Python forritasafn og verkfærakista. Markmið L10 er að tryggja að þessi eining sé altæk fyrir ýmsan máltæknihugbúnað og verkvanga.

Núverandi skýrsla lýsir núverandi stöðu málrýnisins (útgáfa v1.1.1) varðandi notagildi fyrir almenna notkun á sviði máltækni hvað varðar eiginleika og viðmót, auk áætlaðrar notkunar hans innan talgreiniverkefnisins (sjá undirverkefni T13).

Kortlagning ólíkra þarfa máltæknihugbúnaðar

Almennt

Við höfum kortlagt almennar þarfir fyrir samþættingu málrýnisins í annan máltæknihugbúnað. Þarfirnar verða endurskoðaðar í innleiðingarferli verkþáttarins, til að stefna að markmiði þess að

a) hafa starfandi framendapípu fyrir talgervil sem inniheldur málrýni, og b) veita almenna skjölun og viðmið fyrir samþættingu málrýnis í annan hugbúnað.

Umhverfi

Máltæknihugbúnaður keyrir á ýmsum ólíkum vélbúnaði og stýrikerfum. Í þessari skýrslu einbeitum við okkur að þörfum almennra borðtölva og þjóna (e. general desktop and server platforms). Verkþáttur L8 tilgreinir þarfir fyrir snjalltæki. Málrýninn ætti að vera hægt að nota á hefðbundnum stýrikerfum fyrir borðtölvur og þjóna eins og Windows, MacOS og Linux. Hann ætti ekki að þurfa keyrslu í sýndarumhverfi eins og Docker, sem er aðeins óbeint í boði á Windows & Mac OS í gegnum sýndarvélar.

Forritunarmál

Þó að máltæknihugbúnaður sé almennt forritaður í ákveðnum forritunarmálum (Python, Java), ætti aðgengi að kjarnalausnum eins og málrýni ekki að vera takmarkað við ákveðin forritunarmál sem sá hluti er forritaður í. Forrit sem sem nota þessa kjarnalausnir ættu að velja hvaða almenna forritunarmál sem er viðeigandi til að leysa hvert einstakt tilfelli. Aðgengi að kjarnalausnum ætti að vera auðvelt og án erfiðleika eða meiriháttar fórna í afköstum til að brúa bil milli ólíkra sérsviða forritunarmála.

Forritaskil

Þarfir fyrir val mögulegra forritunarmála hafa bein áhrif á forritaskil málrýnisins. Það ættu að vera forritaskil sem eru aðgengileg alhliða fyrir mörg mismunandi forritunarmál. Þetta má best leysa með net-forritaskilum eins og gRPC eða REST, sem veita möguleika á því að keyra málrýni staðvært eða fjartengt. Í heildstæðri máltæknihugbúnaðarpípu væri best að hafa aðeins eitt stakt tilvik málrýnis fyrir umsjón stillinga og samræmi útgáfa.
Að auki er hægt að veita sértæk forritaskil fyrir ákveðið forritunarmál fyrir hugbúnað eða aðra máltæknihluta sem nota sama forritunarmálið. Þessi sértæku forritaskil fyrir ákveðin forritunarmál eru þá einnig notuð úr hjúp net-forritaskila (e. networking API wrapper).
Í tilfelli GreynirCorrect, kemur hann með Python forritaskilum, og JSON RPC net-forritaskilum sem eru aðgengileg hér https://github.com/mideind/Yfirlestur, en net-forritaskilin bjóða aðeins upp á undirsett af Python forritaskilunum.

Talgervill (TTS, Text-to-speech)

Fyrsta skref forvinnslu texta fyrir talgervil er sem stendur textanormun (e. text normalization) (sjá undirverkefni T9) og í þessum verkþætti munum við bæta leiðréttingu stafsetningar framan við textanormunareininguna (e. normalization module). Inntakstexti fyrir talgervil þarf ekki endilega að vera málfræðilega réttur, eða heldur að samanstanda af heilum setningum eða mörgum setningum. Almennt séð viljum við að talgervilskerfi lesi texta nákvæmlega eins og hann er skrifaður, þ.e. án mikilla breytinga/leiðréttinga í málfræði eða stíl. Hins vegar geta stafsetningarvillur, eins og ef stafi vantar, stöfum er bætt við eða stöfum víxlað, orsakað úttak talgervils sem er erfitt að skilja. Sjálfvirk leiðrétting slíkra villna er því líkleg til að bæta árangur talgervilskerfis. Þetta krefst þess að málrýni megi fínstilla til að skila eingöngu ákveðnum gerðum mögulegra leiðréttinga.

Þó að talgervilskerfi SÍM noti Python sem forritunarmál framenda (og nokkur önnur forritunarmál/umhverfi í bakenda), ætti það ekki að nota Python málrýniforritaskilin beint fyrir textavinnslu. Þess í stað ætti notandi talgervilsins að hreinsa textann sem talgervillinn fær, þannig að talgervillinn sé ekki beint háður málrýninum. Eftirfarandi þörfum er því beint að notanda talgervilsins, þar sem notandi þýðir ekki endanlegur notandi, heldur hugbúnaður sem notar talgervilskerfið.

Þarfir

- Leiðrétting stafsetningar
- Stillingar til að fínstilla gerðir mögulegra leiðréttingar stafsetningar (t.d. aðeins villur þar sem staf vantar eða er víxlað/bætt við)
- Stuðning við minna en heilar setningar fyrir stafsetningarleiðréttingu
- Málrýnir ætti að læra ákveðin orð / hugtök; bæta annarri orðabók í kerfið?

Spjallmenni / Leitarkerfi / leitarbyggð forrit

Spjallmenni (e. chatbots), leitarkerfi og leitarbyggð forrit eru mest notuð á ákveðnum sviðum og bregðast á ákveðinn hátt við hugtök og lykilorð innan þeirra. Því eru þessi ákveðnu lykilorð og hugtök veigameiri hér miðað við notkun þeirra í almennu tali. Það væri því gott ef þessi lykilorð/hugtök væri hægt að stilla til að a.) greinast sem gild af málrýni / málrýnir ætti að eiga möguleika á því að taka hvaða orð sem er í orðaforða sinn og b.) leiðrétt með miklum líkum. Þessi notkunardæmi gera það nauðsynlegt að bæta textahreinsun beint í hugbúnaðinn, þ.e., þessir hlutar eru bein ákvæði í málrýnirinn.

Inntakstexti sem kemur frá notanda getur samanstaðið af heilum setningum/spurningum en einnig byggður á lykilorðum og getur verið mjög „óhreinn“. Oft er málfræðilega rangur eða óljós texti sleginn inn. Mikilvægara er að skila niðurstöðum en að krefjast málfræðilega réttu inntaki. Því má beita leiðréttingum á málfræði fyrir minniháttar málfræðivillur, en málfræðilega rangur inntakstexti sem ekki er hægt að leiðrétta sjálfvirkt ætti samt sem áður að ná til bakenda.

Þarfir:

- Leiðrétting stafsetningar
- Leiðrétting málfræði
- Stuðning við minna en heilar setningar fyrir leiðréttingu stafsetningar og málfræði
- Taka við öllum mögulegum niðurstöðum leiðréttinga fyrir betri einræðingu lykilorða
- Málrýnir ætti að læra ákveðin orð / hugtök; bæta annarri orðabók í kerfið?
- Málrýnir ætti að raða ákveðnum lykilorðum / hugtökum hærra

Vélþýðing (e. MT, Machine Translation)

Vélþýðing treystir á að stafsetning og málfræði upprunamáls sé að mestu rétt. Stafsetningar- og málfræðivillur geta ruglað kerfið og hugsanlega spillt þýðingu. Því mun forvinnsla í formi leiðréttingar stafsetningar- og málfræðivillna líklega bæta niðurstöður vélþýðinga. Inntakstexti getur samanstaðið af heilum setningum/spurningum og setningaliðum, en einnig stökum eða mörgum lykilorðum, ekki endilega í rökréttri röð.

Þarfir:

- Leiðrétting stafsetningar
- Leiðrétting málfræði
- Stuðning við minna en heilar setningar fyrir leiðréttingu stafsetningar og málfræði

Þarfatafla

	Talgervill	Þýðingarvél	Q/A - Spjallmenni
Leiðrétting stafsetningar	X	X	X
Leiðrétting málfræði	X	X	X
Möguleiki fínstillingar leiðréttingargerða	X	-	X
Stuðning við minna en heilar setningar	X	X	X
Tekur við öllum mögulegum niðurstöðum leiðréttinga	-	-	X
Málrýnir ætti að læra ákveðin orð	X	X	X
Málrýnir ætti að raða ákveðnum orðum	-	-	X

Athugið: reitir merktir grænir eru studdir af GreynirCorrect frá og með v1.1.1. Appelsínugulir og hvítir eru eiginleikar sem eru ekki til staðar.

Núverandi staða GreynirCorrect

Forritaskil málrýnisins (e. The Spell Checker API) miða að þáttun heilla setninga eða málsgreina. Úttak málrýnisins, ætlað endanotendum, annaðhvort á setningarfræðilegu eða málfræðilegu stigi, sem Python hluti (e. Python objects) (í gegnum Python forritaskil) eða JSON (í gegnum net-forritaskilin). Sum þeirra gagna eru einnig ætlum vélvinnslu, t.d. villuflokkar.

Python forritaskilin bjóða eftirfarandi meginföll:

- `tokenize(), check_single(), check(), check_with_stats()`

En net-forritaskilin (e. networking API) hefur fallið:

- `Correct.api`

Hið síðarnefnda samsvarar Python forritaskilunum `check_with_stats()` og skilar gögnum sem JSON streng.

Það er ekki vel stutt eins og er að nota minna en heilar setningar sem inntakstexta. Til dæmis kemur villa ef texti byrjar ekki á hástaf.
Það eru stillingar í Python forritaskilunum til að hafa áhrif á skoðun niðurstaðna, en þær eru frekar ætlaðar til innværar notkunar eða fyrir undirliggjandi tilreiðara.

GreynirCorrect fyrir talgervil

Samkvæmt the þarfatöflunni, þarf að innleiða eftirfarandi eiginleika fyrir talgervil:

- Möguleika til að fínstilla leiðréttingargerðir
- Stuðning við minna en heilar setningar
- Málrýnir (e. spell checker) ætti að læra ákveðin orð

Þessa eiginleika sem vantar má annaðhvort innleiða í tiltækar aðferðir forritaskilanna, eða með því að bæta við nýjum forritaskilum.

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 4 - lýsing

- 200.000 segðir fullorðinna teknar upp
- 180.000 segðir unglinga teknar upp
- 15.000 segðir annars máls mælenda (e. L2 speakers) teknar upp
- Útgáfa af Samrómi sem leyfir mælendum að endurtaka talaðar segðir, annaðhvort upptökur af mennskum mælanda eða segðir myndaðar af talgervli

Afurðir

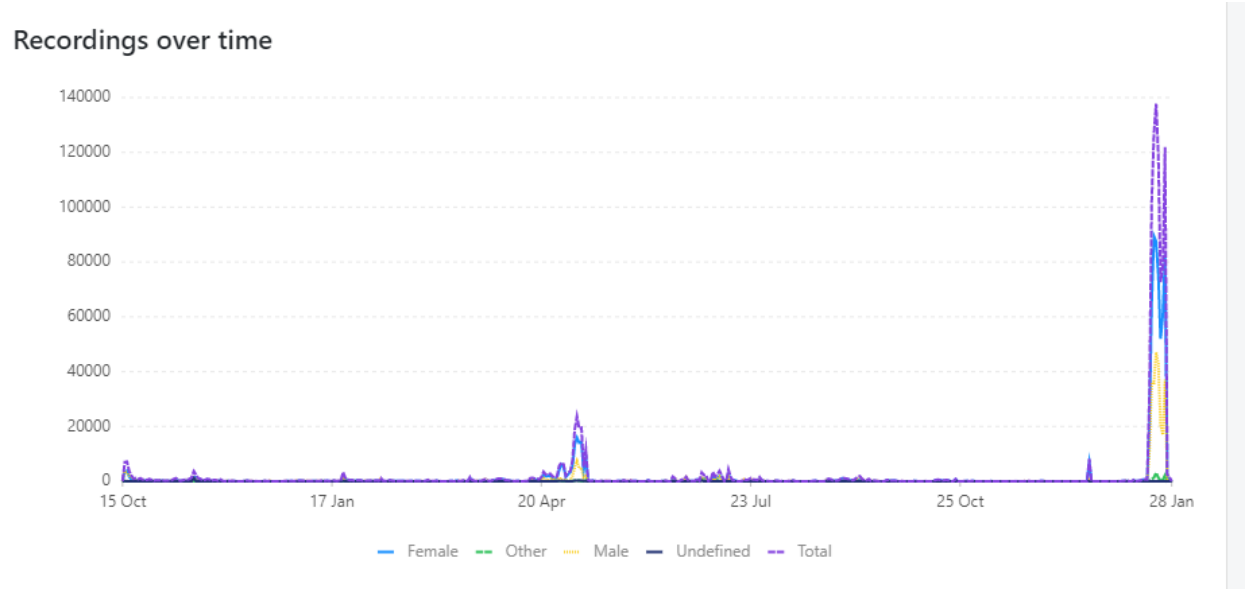
Öllum söfnunarmarkmiðum vörðunnar var náð. Valmöguleika, þar sem notandi getur valið að láta lesa fyrir sig segðir myndaðar af talgervli, hefur verið bætt við Samróm.

Skýrsla

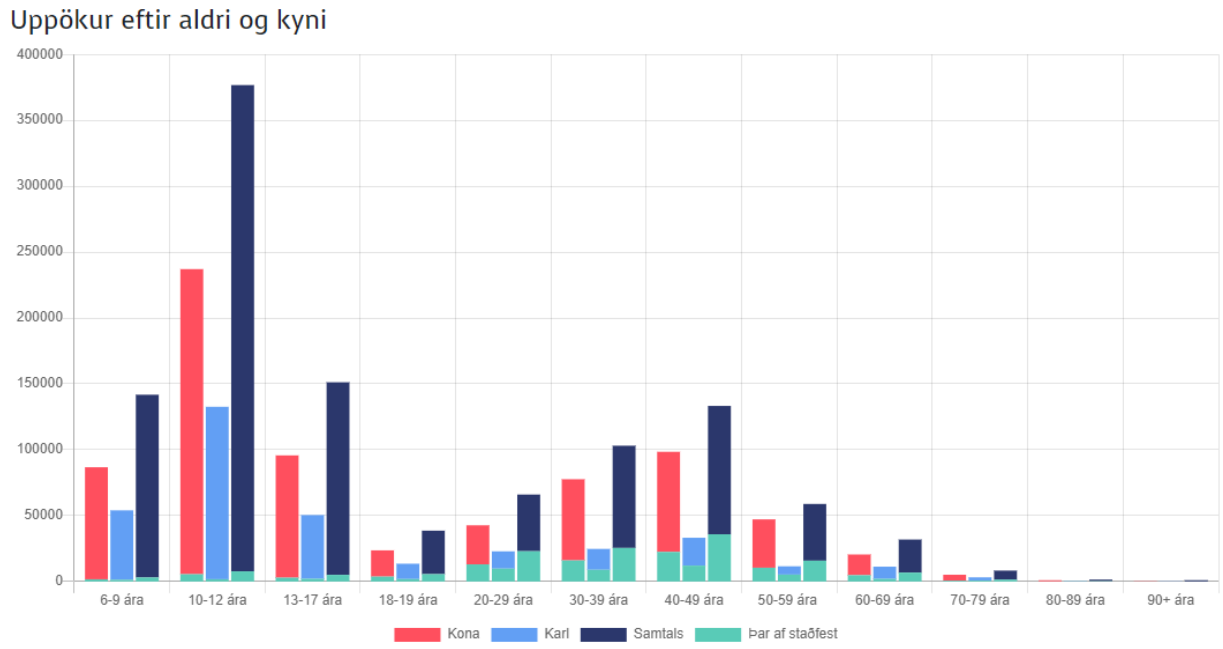
Þegar þetta er ritað er grunnskólakeppninni nýlega lokið og fór þátttakan fram úr okkar björtustu vonum. Samtals fengum við um 790 þúsund segðir frá 6.000 manneskjum. Þetta þýðir að samtals eru segðirnar orðnar fleiri en 1,1 milljón. Keppnin hafði aðra kosti eins og að vekja mikla athygli á verkefninu og vekja vitund á mikilvægi þessara lausna fyrir íslenskt mál.

Lifandi tölfræði er alltaf aðgengileg hér og hér.

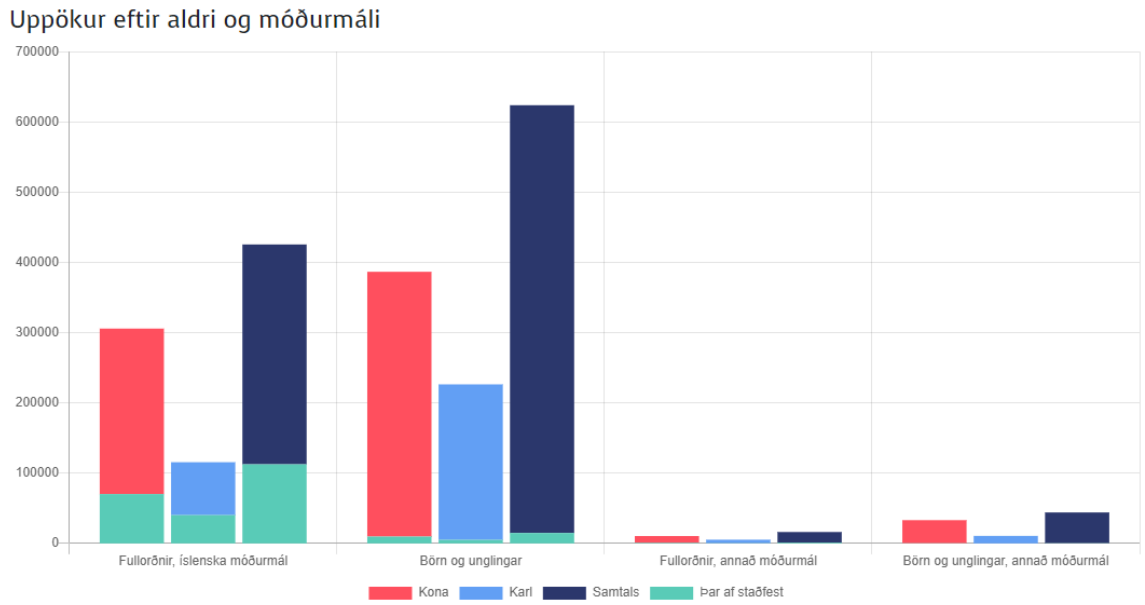
Línuritið hér að neðan sýnir hvernig söfnunin hefur gengið frá upphafi. Bungan í upphafi og bungan í miðjunni voru ótrúlegir dagar, eða það héldum við. Á einum degi í keppninni í ár fengum við fleiri upptökur en söfnuðust samanlagt í keppni síðasta árs.



Línuritið hér að neðan sýnir samanlagðan fjölda upptaka sem við höfum safnað til þessa, flokkaðar eftir aldri og kyni.



Græni litur súlnanna vísar til fjölda klippa sem hafa verið yfirfarnar. Línuritið að neðan sýnir fjölda upptaka, flokkaðar eftir aldri og svo skipt eftir kyni. Minni súlurnar til hægri sýna fullorðna sem hafa íslensku sem annað mál og svo börn einnig. Samanlagt höfum við 60 þúsund upptökur frá fólki sem hefur íslensku sem annað mál.



Frá síðustu vörðu höfum við gert miklar breytingar á söfnunarumhverfi okkar, samrómur.is, sem gerði þessa keppni, eða að minnsta kosti upplifunina af því að taka þátt, meira hvetjandi og áreiðanlegri. Við löguðum fjölmargar villur (e. bugs), bættum við innskráningarmöguleika til að notendur fengju yfirsýn yfir sitt framlag og nú getum við veitt notendum sérstaka heimild, t.d. í þeim tilgangi að yfirfara klippur, sem mun koma sér mjög vel þar sem það verður ærin fyrirhöfn að fara yfir allar klippurnar.

Við höfum farið yfir upptökurnar sem safnað hafði verið fyrir þessa grunnskólakeppni með sjálfvirkri aðferð við sannprófun sem heitir Marosijo[35].

Útgáfa af Samrómi sem leyfir notendum að endurtaka talaðar segðir

Við bjuggum til málheild stuttra setninga með stuttum orðum, sem safnað var úr Risamálheildinni. Þessar segðir voru búnar til (e. synthesised) með polly-talgervilskerfinu frá AWS (við hlökkum til að skipta því út fyrir talgervilinn okkar). Frá og með þessari skýrslu erum við í því ferli að búa til notendaviðmótið fyrir þetta kerfi. Það verður sýnilegt á www.samrómur.is/tala.

Vinnuskilgreining okkar á lýðfræði markhópa þessarar söfnunar er eftirfarandi:

- Fólk á öllum aldri sem er sjónskert
- Börn 7 ára og yngri
- Annars máls mælendur (e. L2 speakers)

[35] https://github.com/cadia-lvl/samromur-tools

H2 - Umritun efnis úr sjónvarpi og útvarpi

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

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 4 - lýsing

Verk:

- Útbúa lista yfir þætti til að umrita á öðru ári, byggt á reynslu fyrsta árs
- Umrita umræðu- og afþreyingarefni
- Kerfi til að samraða útvarps- og sjónvarpsfréttum

Endurskoðaðar vörður:

M4: 150 klukkustundir af útvarpsþáttum umritaðar. Kerfi til að samraða útvarps- og sjónvarpsfréttum hefur verið útbúið.
M5: 200 klukkustundir af útvarpsþáttum umritaðar.
M6: 250 klukkustundir af útvarpsþáttum umritaðar og 250 klukkustundir af samröðuðu útvarps- og sjónvarpsefni tilbúið, sem leiðir af sér 500 klukkustundir af umrituðu og samröðuðu sjónvarps- og útvarpsefni.

Afurðir

7 klukkustundum af sjónvarpsgögnum frá RÚV hefur verið hlaðið upp á Clarin: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/93
Að 145 klukkustundum Creditinfo meðtöldum, eigum við samtals 152 klukkustundir af sjónvarps- og útvarpsefni.

Skýrsla

Umritun efnis úr sjónvarpi og útvarpi

Creditinfo hefur afhent 145 klukkustundir af gögnum. Við höfum verið að fara yfir gögnin til að meta raunverulega útkomu áður en sú vinna sem eftir er við merkingar hjá Creditinfo verður unnin. Það eru vandamál, einkum með tímamerkingar, sem virðist vera erfitt að sækja með núverandi umritunarkerfi Creditinfo. Þetta leiðir til mikillar vinnu við gerð lítilla búta í núverandi umritunarkerfi þess og ónákvæmni tímamerkinga.
Það er til umræðu að nota skilvirkara kerfi við umritanirnar, talgreininn frá Tiro, sem gæti með smá aðlögun uppfyllt allar kröfur sem núverandi kerfi er ekki hannað til að fylgja. Vonandi mun

þetta bæta frammistöðuna þegar kemur að umritunarhraða og því að minnka eftirvinnslu gagnanna.

Samröðun og bútun (e. segmenting) sjónvarpsefnis

Núverandi umritunarkerfi: https://github.com/cadia-lvl/Alignment-and-segmentation
Upphaflega samröðunarkerfið hefur verið útbúið og prófað. Einmitt núna er það að vinna með sjónvarpsþættina Kastljós og Kiljuna frá RÚV. Aftur á móti er það nátengt textanormunarkerfi Alþingistalgreinisins og getur aðeins unnið með gögn sem hefur verið skipt niður í mælendur (e. pre-diarized data) með Gecko merkingatólinu og Kaldi samræðugreindarkerfinu. Frekari skipulagning og vinna verður unnin í vörðu 5 til að hægt verði að vinna með breiðari flóru þátta og aftengja kerfið frá Alþingiskerfinu.
Þar sem Creditinfo hafði ekki aðgang að streymi (e. source stream) sjónvarpsþátta frá RÚV á fyrsta ári reyndist erfitt að útvega réttar tímamerkingar. Þess vegna var ákveðið að útvega sjónvarpsefni með því að samraða textum frá síðu 888 í textavarpinu. Forskriftin krefst upplýsinga úr samræðugreind til að para mælendaauðkenni (e. speaker IDs) við textagögn og skilar þannig úttaksgögnum með mælendaauðkennum, sem henta til að þjálfa talgreina.
Samröðunarkerfið var matað á 86 þáttum (~32 klst.) úr stóra samræðugreindargagnasafni RÚV (e. the extended RÚV Diarization) (ruv-di)[36] og þáttum úr samræðugreindargagnasafni sumarsins (ruv-di v2) að auki. Þættirnir voru Fréttir kl. 19:00, Kiljan og Kastljós. Hins vegar voru textar (e. subtitles) aðeins til reiðu fyrir 36 þætti. Eftir að þáttunum var samraðað, þá eigum við núna ~7 klst. af samröðuðum gögnum úr Kiljunni og Kastljósi. Valdir bútar úr þessum ~7 klst. hafa verið yfirfarnir handvirkt og gagnasafnið gefið út.

[36] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/68

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 lýkur á öðru ári.

Varða 4 - lýsing

Verk:

- Skipuleggja samtöl yfir samræðukerfi (e. dialogue system)
- Skipuleggja lestur fyrirspurna
- Umrita samtöl og fyrirspurnir

H3 - Endurskoðaðar vörður

M4: 15 klst. af gögnum teknar upp og 5 klst. umritaðar.
M5: 5 klst. af gögnum teknar upp og 15 klst. umritaðar. Fyrirspurnamálheild búin til.
M6: 20 klst. af lesnum fyrirspurnum teknar upp.

Afurðir

Á þessum tímapunkti hafa 3 klst. af samræðum verið teknar upp og 1 klst. umrituð, af samtals 15 klst. af upptökum og 5 klst. umritunum. Tólið sem hannað var til að safna samræðunum má nálgast hér. Við höfum aðallega einbeitt okkur að tólinu í þessari vörðu og við erum mjög örugg um að ná 20 klukkustundunum fyrir lok vörðu 5 í staðinn.

Skýrsla

Hugbúnaðurinn sem hannaður var fyrir þessa gangasöfnun gerir það mögulegt að taka upp samtöl á netinu í gegnum vafra. Það hafa komið upp nokkur vandamál og villur í upptökuferlinu, aðallega tengd því að senda upptökuna eftir að upptöku lýkur. Að minnsta kosti 1 klst. af upptökum hefur tapast vegna þessa. Eins og er teljum við mikilvægara að vinna að frekari prófunum á tólinu en að kalla eftir þátttöku og þar með biðja fólk að nota tól með þessum vandamálum. Þess vegna höfum við ákveðið að einbeita okkur að ítarlegri prófunum á tólinu áður en við höldum áfram að safna. Að því sögðu, þá var tól sem safnar samtölum á netinu ekki hluti af upphaflegu hugmyndinni að þessum verkþætti. Hugmyndin var að taka upp samtöl á staðnum, með rannsakanda sem stjórnaði þeim og stillti upp hljóðnemum. Að því leyti hefði verið vel raunhæft að ná markmiðinu um 15 klst. af upptökum, en sökum faraldursins neyddumst við til að leita annarra lausna sem leiddu til þróunar veftólsins. Þess vegna eigum við næstum tilbúið tól sem má nota í þeim tilgangi í framtíðinni og mun auðvelda upptökur þegar það verður tilbúið. Vinnan sem við höfum innt af

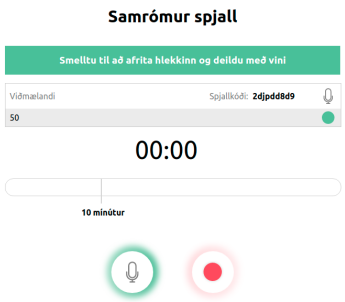
hendi í þessum verkþætti nemur meiri vinnu en við upptökur á 15 klst. með upphaflegu aðferðinni og við erum viss um að geta uppfyllt 20 klst. markmiðið í vörðu 5 í staðinn.

Nokkrar þeirra breytinga sem gerðar voru á eiginleikum og hönnun tólsins innan þessarar vörðu:

- Notendaskilmálum og persónuverndarskilmálum var bætt við þátttökuferlið í samræmi við persónuverndarlög.
- Bætt var við hnappi sem auðveldar notendum að deila hlekk inn á núverandi spjallsvæði.
- Bætt var við kerfi til að stinga upp á umræðuefnum, en því er ætlað að auðvelda þátttakendum að halda samtalinu gangandi. Uppástungurnar birtast í handahófskenndri röð til að hver þátttakandi fái mismunandi umræðupunkta (e. talking points).
- Útbúið var skráningarform til að senda framtíðarþátttakendum.
- Umritunarleiðbeiningar voru skrifaðar og umritarar (e. transcribers) ráðnir til starfsins, sem og til prófarkalesturs.
- Öll umritun og prófarkalestur fer fram í gegnum talgreini TIRO.
- Þar sem flest fólk, að minnsta kosti á Íslandi, deilir hlekkjum sem þessum á Facebook, þá var innbyggð flækja sem þurfti að leysa sem fólst í því að Facebook bætir við upplýsingum til að merkja (e. tagging information).
- Síðasta vandamálið, sem við erum enn að vinna að lausn á, er hvernig á að senda samtöl eftir að þeim lýkur. Við komumst að því að ferlið sumpart ruglingsleg, t.d. þurftu báðir þátttakendur að senda samtalið sitt. Þess vegna gerðum við þann notanda sem átti frumkvæðið að samtalinu ábyrgan fyrir því að senda báðar rásir.
- Viðvörunarskilaboðum til notenda sem ætluðu að vafra í burtu eða loka flipanum var bætt við.

Þekkt vandamál sem verða leyst eins fljótt og mögulegt er:

- Prófun á nýja sendingarferlinu verður forgangsatriði.
- Endurhönnun og clarification á tilgangi græna og rauða stjórnhnappsins í viðbmótinu, sjá mynd 1.



Mynd 1. Skjáskot af viðmóti spjalltólsins (e. chat tool).

Eins og kannski sést á þessum listum þá vonumst við til þess að geta hafið upptökur fljótlega og þar með lokið bæði vörðum M4 og M5 áður en langt um líður.

H7 - Þróun á almennum talgreini

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að þróa almenna Kalda-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 4 - lýsing

Undirbúningur gagna, þar á meðal undirbúningur hreinna þróunar- og prófunarsetta. Sjálfvirkur undirbúningur málfanga (e. resource automation) og einföld HMM-aðferð undirbúin. Stefnan er að ná orðvillutíðni lægri en 20%.

Afurðir:

Þeim áföngum vörðunnar sem lúta að undirbúningi gagna og sjálfvirkningu tilfanga var náð. Einföld GMM-HMM hljóðlíkön byggð á þrístæðum voru þjálfuð og þrí- og fjórstæðu (e. tri-gram and four-gram) mállíkön voru þjálfuð með Samrómstexta og Risamálheildinni. Besta orðvillutíðnin sem náðist var 21,82%. Þetta gildi er hærra en áætlað var, en við komumst að því að þarna reyndust mörk hljóðlíkansins sem innleitt var í M4 liggja, í verki sem þessu með svona miklum gögnum. Búist er við því að orðvillutíðni dragist verulega saman með aðgreiningarþjálfun (M5) og tauganetum (M5), og þegar meira magn af gögnum verður aðgengilegt.

Hlekkur á hirslu verkefnisins:
https://github.com/cadia-lvl/samromur-asr

Skýrsla:

Gerð almenns talgreinis fyrir íslensku er skiptist í þrennt: 1. undirbúningur gagnasafns og sjálfvirkur undirbúningur málfanga, 2. þjálfun hljóðlíkans, og 3. þjálfun mállíkans.

Undirsett 100.000 segða úr Samrómsmálheildinni var valið og yfirfarið, textarnir normaðir, efni yfirfarið. Þessum ferlum var lýst í kafla H1 í M3 skýrslunni. Segðunum var því næst skipt í þjálfunar-, þróunar- og matssett með hlutfallinu 8:1:1. Undirsettin skarast ekki þegar kemur að mælendum og lágmarksskörun er hvað varðar málfræðilegt innihald (e. linguistic content, LC). Hverjum mælanda var úthlutuð skörunareinkunn málfræðilegs innihalds (e. LC overlap score) sem var skilgreint sem summa af vægi setninga normuð eftir fjölda setninga. Vægi setningar samsvaraði tíðni hennar í 100.000 segða málheildinni. Mælendur með lægstu einkunnina voru valdir í matið. Þetta undirsett var erfiðast fyrir talgreininguna, þar sem það inniheldur mikið af

setningum með sjaldgæfum orðum, svo sem örnefnum, eða samhengislausum orðastrengjum, sem lágmarka skilvirkni mállíkansins. Mælendur með hærri einkunn voru valdir við þróunina og afgangurinn við þjálfun. Í töflu 1 má sjá samantekt tölfræði um mælendur og segðir. Skrá með upplýsingum lýsigagnanna var útbúin, sem gerir gagnasafnið tilbúið til útgáfu og skrifta var útbúin til að undirbúa gagnasettin fyrir staðlaða KALDA[37] s5-sniðið.

Tafla 1. Tölfræði um mælendur og segðir í 100.000 segða málheildinni

	Lengd	Segðir	Mælendur	Einstakar segðir.	Mælendur einungis með einstakar (unique) segðir .	Mælendur með einstakar segðir
Mat	3 klst. 48 mín.	10.000	1642	10.000	1642	1642
Þróun	15 klst. 16 mín.	10.000	2109	10.000	2109	2109
Þjálfun	114 klst. 34 mín.	80.000	4642	50.438	54	3856

Við gerð hljóðlíkansins notuðum við ítrunarnálgun til að búa til besta líkanið. Við byrjuðum með minni hljóðlíkön og samröðuðum þeim til að búa til stærri. Fyrst bjuggum við til líkan byggt á stökum málhljóðum, samröðuðum því og bjuggum svo til delta+delta þrístæðulíkan og á eftir fylgdi þrístæðulíkan byggt á línulegri aðgreiningaraðlögun (e. linear discriminative adaption) + hámarkssennileika-línulegri vörpun (e. maximum likelihood linear transformation) (LDA + MLLT), og að lokum þrístæðulíkan byggt á LDA + MLLT með þjálfun með mælendaaðlögun (e. speaker adapted training) (LDA + MLLT + SAT).

Við normuðum texta Risamálheildarinnar[38]. Það felur í sér að skrifa út tölustafi og skammstafanir. Við skiptum henni upp í setningar, með einni setningu í hverri línu. Við flokkuðum setningarnar og fjarlægðum endurtekningar. Endanlega normaða mállíkanaþjálfunarsettið var 44 milljón setningar, eða 797 milljón orð. Við bjuggum til litla þrístæðu (e. tri-gram) fyrir afkóðun og afklipptum (e. unpruned) fjórstæðu til endureinkunnargjafar. Hlutfall orða utan orðaforða er 2,4% og 2,8% í þróunarsettinu annars vegar og prófunarsettinu hins vegar. Framburðarorðabókin var byggð á almennu framborðarorðabókinni fyrir talgreiningu[39], en framburði orða utan orðaforða (e. OOV) var bætt við með Sequitur[40] G2P-líkani sem þjálfað hafði verið á henni. Orðunum sem bætt var við eru íslensk orð sem koma 15 sinnum eða oftar fyrir í mállíkanaþjálfunartextanum, auk orða utan

[37] Kaldi verkfærasett - https://github.com/kaldi-asr/kaldi
[38] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/15
[39] http://www.malfong.is/index.php?lang=en&pg=framb_talgr
[40] https://github.com/sequitur-g2p/sequitur-g2p

orðaforða í þjálfunarsetti Samróms. Með því að nota þrístæðumállíkan (e. tri-gram language model) þjálfað á 100.000 segðum úr gagnasetti Samróms náðum við 45% orðvillutíðni. Með því að nota Risamálheildina náðum við 21,82% og 24,27% á þróunarsettum annars vegar og matssettum hins vegar. Til og með 18. janúar 2021 höfum við safnað meira en 300.000 segðum í gegnum umhverfi Samróms og 200.000 í staðfestingarferli. Þegar ferlinu líkur mun stærð gagnasafnsins hafa þrefaldast, sem gefur okkur traustan grunn að byggja á við betrumbætur hljóðlíkana í framtíðinni, og sér í lagi fyrir tauganetsþjálfun.

H8 - Vefviðmót fyrir talgreiningu

Skýrsla: Tiro
Aðilar: 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 4 - lýsing

Tasks

- Setja upp vefviðmót fyrir talskrár og N-bestu tillögur
- Þróa vinnslu bakenda fyrir textanormun og samræðugreind (e. speaker diarization)
- Þróa framsetningu fyrir auðgað úttak
- Setja upp vefviðmót fyrir auðgað úttak
- Þróa rauntímaframsetningu á úttaki talgreinis
- Setja upp vefþjónustu fyrir talgreiningu í rauntíma

M4
Vefviðmót fyrir talgögn með hráu úttaki.

Afurðir

Öllum markmiðum vörðunnar var náð. Kóðinn fyrir upphaflegu útgáfuna, sem samþykkir stuttar hljóðskrár og skilar N-bestu tillögum, hefur verið gefinn út sem Tiro Speech Core: https://github.com/tiro-is/tiro-speech-core og þjónustan er komin í gagnið á https://speech.tiro.is.

Skýrsla

Talgreinisþjónustan fer eftir stöðlum Google Cloud Speech forritaskila og innleiðir undirsett þeirra, sem eru gRPC[41] forritaskil með REST proxy. Ákveðið var að nothæfa þessi forritaskil, frekar en að hanna ný, til að einfalda yfirfærslu frá talþjónustu Google til þeirrar sem hönnuð var innan verkefnisins. Þetta var metið mikilvægt, þar sem Google er sem stendur eini erlendi aðilinn sem býður upp á talgreini fyrir íslensku. Skilgreining og skjölun undirsettsins er aðgengileg á Github[42].

[41] gRPC: https://grpc.io
[42] Forritaskil talþjónustu: https://github.com/tiro-is/tiro-speech-core/blob/m4/proto/tiro/speech/v1alpha/speech.proto

Staða tækninnar í talgreiningu á netþjónum eru blendingar HMM/tauganetslíkana sem þróaðir voru með verkfærasetti Kalda[43]. Upphaflega útgáfan af Tiro Speech Core styður því aðeins þessi líkön, eða nánar tiltekið líkön þróuð með Kalda Nnet3.
Tiro Speech Core er skrifaður í C++ og notar Bazel[44] til þess að setja saman kerfið (e. build system). Kóðinn fylgir leiðbeiningum Google C++ leiðbeiningunum[45] með þeirri undantekningu að leyfa undantekningar, hann er sniðinn sjálfvirkt með Clang-Format og prófanir eru innleiddar með Catch2. Talgreinirinn sjálfur, þ.e. útdráttarpípa meginþátta (e. main feature extraction pipeline) og kjarnaafkóðari (e. core decoder), er innleiddur með hjálp forritasafns Kalda. FFmpeg[46] forritasafnið er notað til að styðja við afkóðun hljóðskráa og geymt á margs konar sniðum.

Þjónustan er komin í gagnið á https://speech.tiro.is og skjölunin, ásamt tveimur notkunardæmum, er aðgengileg á Github, annað notar REST API[47] forritaskil og hin notar gRPC API sem er skrifað í Python[48]. Þess ber að geta að þar sem talgreinalíkön úr þessu verkefni hafa ekki verið gerð gefin út, þá styðst þjónustan við eldri líkön sem þróuð voru af Tiro með málheild Málróms[49] og gagnasafni BÍN (undirverkefni G4).

[43] Kaldi: https://kaldi-asr.org
[44] Bazel: https://bazel.build
[45] Google C++ Style Guide: https://google.github.io/styleguide/cppguide
[46] FFmpeg: http://ffmpeg.org
[47] REST forritaskil, dæmi: https://github.com/tiro-is/tiro-speech-core/blob/m4/README.md#example-usage-of-the-rest-gateway
[48] gRPC forritaskil, dæmi: https://github.com/tiro-is/tiro-speech-core/blob/m4/examples/python/README.md
[49] Málrómur: http://www.malfong.is/?pg=malromur

H13 - Líkan fyrir orðmyndir og samsett orð

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

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

Varða 4 - lýsing

Frumgerðir þriggja nálgana við gerð orðflísamállíkana prófaðar á texta.

Afurðir

Allar þrjár nálganir við gerð orðflísamállíkana voru innleiddar og afurðum vörðunnar skilað. Við gerðum frumgerðir af Byte Pair Encoding (BPE), Unigram, og Morfesor, og mátum þær hvað varðar flækni (e. perplexity).

Hlekkur á hirslu verkefnisins:
https://github.com/Daviderikmollberg/subword_perplexity_tests

Skýrsla

Íslenska er beygingarmál og ný orð eru stöðugt mynduð með samsetningu. Að finna árangursríka leið til að meðhöndla bæði mismunandi yfirborðsgerðir orða og fjölmargar samsetningar skiptir höfuðmáli til að draga úr tíðni óþekktra orða í verkefnum máltækniáætlunarinnar. Algeng aðferð er að skipta orðunum niður í smærri einingar (oft nefndar orðhlutar eða orðflísar), sem síðan eru notaðar til að þjálfa mállíkan. Þetta ferli fækkar þeim tókum sem mállíkanið þarf að búa til og lágmarkar þar með tíðni utan orðaforða og bætir myndun sjaldgæfra orða.

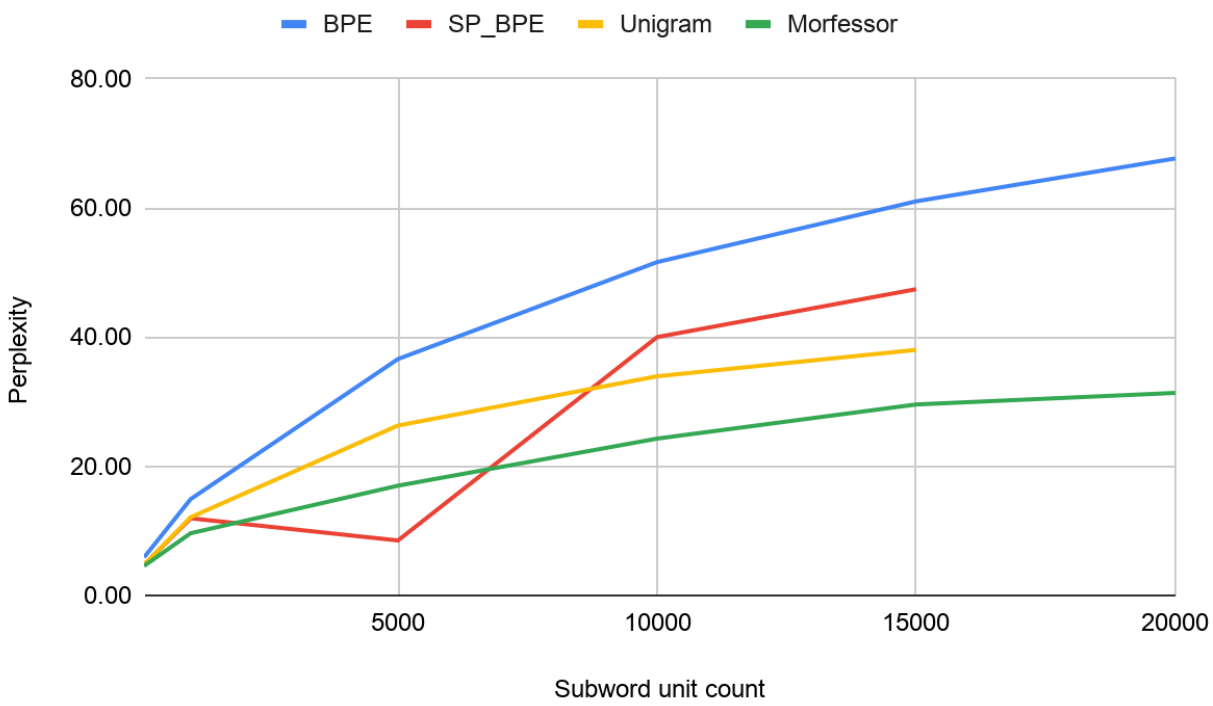
Orðflísar geta verið frá einum staf (sem samsvarar einu fónemi) til eins eða fleiri atkvæða og allt að heilum orðum. Meginreglan við orðflíksamállíkön er að skipta orðunum upp í orðflísar sem eru mismunandi að lengd (einn stafur eða meira). Þetta ferli er líka kallað orðflísatilreiðsla (e. subword tokenization). Tókarnir sem búnir eru til eru síðan notaðir til að þjálfa mállíkan, í staðinn fyrir að nota heilt orð sem grunneiningu við líkanagerðina. Við höfum inleitt tvö Byte Pair Encoding (BPE) algrím, algrím sem er almennt vísað til sem Unigram, og Morfessor.

BPE-algrímið er einfaldasta algrímið af þeim sem við höfum prófað. Algrímið telur tíðni hvers orðs í þjálfunarmálheildinni, sýnir hvert orð sem röð tákna (e. sequence of graphemes), og bætir orðlokatákni (e. end-of-word symbol) við hvert orð. Síðan telur það öll táknapörin og skiptir

hverju tilviki algengasta táknaparsins út fyrir nýtt tákn. Þetta er ítrað þar til tilætluðum fjölda orðflísatóka er náð.

Öfugt við BPE, þá stillir/frumstillir Unigram grunntókana sína á mikinn fjölda tákna og fækkar þeim smám saman til að ná lægri talningu, með því að velja að fjarlægja tákn sem lágmarka lograð sennileikafall (e. log-likelihood) yfir þjálfunargögnin.
Morfessor er ætt vélnámslíkindaaðferða (e. probabilistic machine learning methods) sem leita að málfræðitókum í hráum textagögnum. Með því að nota ytri stika, α fyrir líkindi í kostnaðarfallinu (e. cost function), getum við valið hversu ákaft algrímið tilreiðir (e. tokenizes) orðin. Lágt α gildi eykur áhrif þess sem er fyrirfram gefið (e. the priors) og tekur styttri orðflísar fram yfir lengri. Hátt α gildi eykur áhrif sennileikafalls yfir gögnin (e. data likelihood), og tekur lengri orðflísar fram yfir styttri. Þennan ytri stika má stilla sjálfvirkt að því gefnu að fjöldi orðflísa sé fastur (e. fixed).

Við breyttum skriftu sem er aðgengileg í GALE-arabic forskriftinni í hirslu Kalda fyrir eina af BPE-innleiðingunum og notuðum Google's SentencePiece (SP) forritasafnið fyrir hina BPE-innleiðinguna (SP_BPE) og Unigram-algrímið. Við notuðum líka Morfessor-pakkann.
Málheildinni sem við notuðum er lýst í kafla H7 í skýrslunni, en við þurftum að minnka hana um 10% til að fá SP líkönin til að virka á þjóninum (e. server) okkar, vegna minnishafta (e. memory constraints). Málheildinni var skipt upp í undirsett þjálfunar og prófunar, með hlutfallinu 9:1, með lágmarksskörun setninga. Við notuðum KenLM til að búa til sexstæðu mállíkön (e. 6-gram LMs) með því að nota málheild sem hafði verið tilreidd með öllum fyrrnefndu aðferðunum. BPE og Unigram eru háð ytri stika sem stjórnar fjölda tóka og spurningin um réttan fjölda líkanagerðra (e. modelled) orðflísa er opin spurning. Að lokum mátum við flækni (e. perplexity) mállíkananna á prófunarundirsetti textamálheildarinnar út frá fjölda tóka. Mynd H13.1 sýnir flæknina sem fall af fjölda orðflísa.



Mynd H13.1 Flækni á móti fjölda orðflísa

Hegðun SP_BPE og Unigram líkananna er nánast eins. Í þessu prófi gera þau engar villur utan orðaforða. Hins vegar gerir BPE með GALE-arabic skriftu stöðugt í kringum 5.500 villur utan orðaforða. Þetta er líklega vegna þess að bæði SP_BPE og Unigram innihalda bókstafi upprunamálsins (stafrófið) í lista orðflísa. Þetta þýðir að þau geta alltaf búið til öll orð í prófunarundirsettinu með því að nota þessa bókstafi. Það sama gildir um Morfessor.

H15 - Hljóðlíkön fyrir barna- og unglingaraddir

Skýrsla: Háskólinn í Reykjavík
Aðilar: 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 fullorðinsradda.

Varða 4 - lýsing

Hljóðlíkan fyrir unglingaraddir þjálfað og metið.

Afurðir

Öll skilyrði vörðunnar voru uppfyllt. Undirsett 21 þúsund segða sem innihalda unglingaraddir úr Samrómsmálheildinni voru valdar, yfirfarðar, hreinsaðar og sniðnar á staðlað KALDA s5 snið. Einföld GMM-HMM hljóðlíkön byggð á þrístæðum (e. triphones) voru þjálfuð. Hljóðlíkan var þjálfað með sömu forskrift fyrir Risamálheildina, eins og lýst er í kafla H7 í þessari skýrslu. Bestu orðvillutíðni upp á 23,33% var náð. Sama kóðahirsla er notuð í þessum verkþætti og í H7.

Hlekkur á hirslu verkefnisins:
https://github.com/cadia-lvl/samromur-asr

Skýrsla:

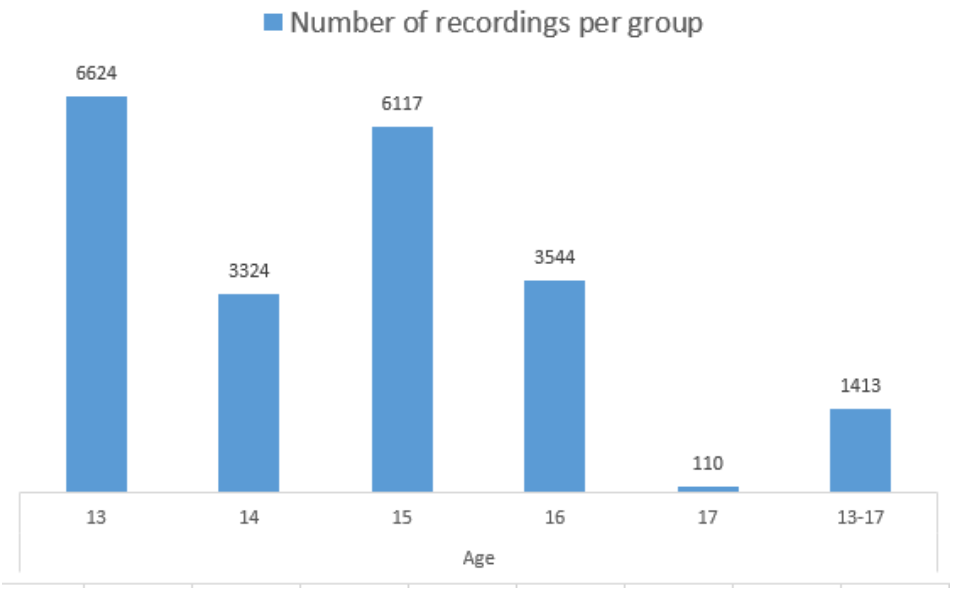
Að búa til talgreini fyrir íslenskar unglingaraddir er tvíþætt: 1. undirbúningur gagnasafns og 2. þjálfun hljóðlíkans.

Búið var til undirsett 21 þúsund upptaka úr söfnun Samróms, sem inniheldur raddir unglinga frá 13 til 17 ára að aldri. Þessar klippur voru yfirfarnar handvirkt með Marosjio[50] tólinu sem vinnur eftir meginreglunni um blöndu talgreiningar og samröðunar. Flestar upptökurnar voru yfirfarnar af Marosjio, eins og lýst er í skýrslunni um undirverkefni H1. Öll þessi vinna fór fram áður en grunnskólakeppnin, sem einnig er getið um í H1, var haldin. Ásamt áframhaldandi vinnu við að staðfesta fleiri upptökur handvirkt er búist við því að þetta verði til þess að stækka málheildina verulega. Aldursdreifingu í þessu undirsetti má sjá á mynd H15.1. Upphaflega var ekki greint á milli aldurs unglinganna í söfnunarferlinu, heldur fengu allar upptökur aldursmerkinguna ‘13-17’. Við höfðum safnað 1413 taldæmum með þessum hætti og þessi staðreynd er sýnd með síðustu

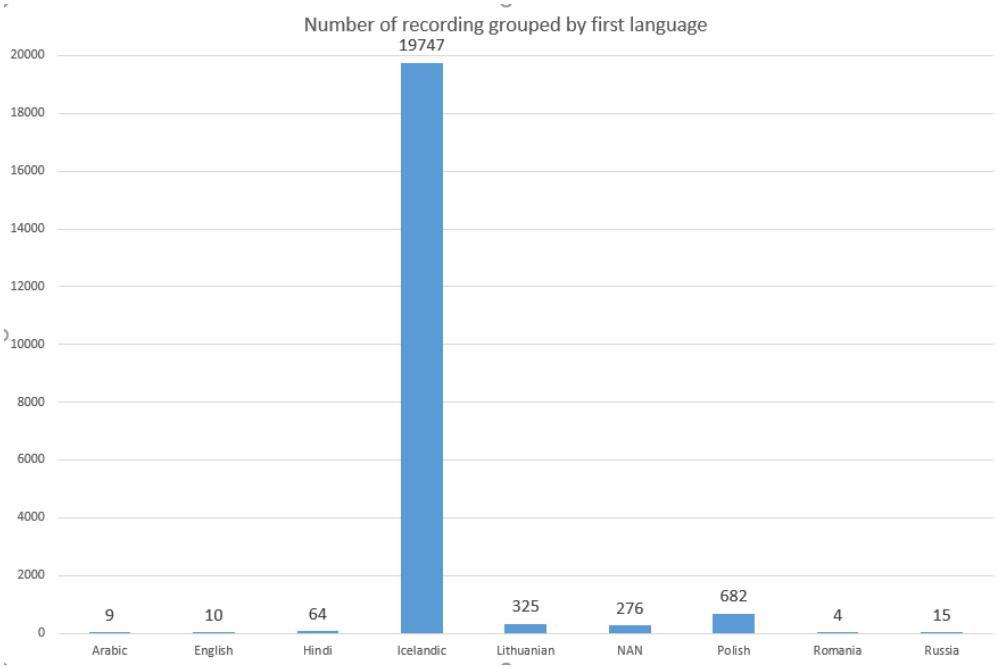
[50] https://github.com/cadia-lvl/Eyra/tree/samromur_qc

súlunni á myndinni, sem einnig er merkt ‘13-17’. Ferlið var gert ítarlegra síðar, sem er ástæðan fyrir einstökum súlum fyrir hvern aldur frá 13 til 17 á mynd H15.1.

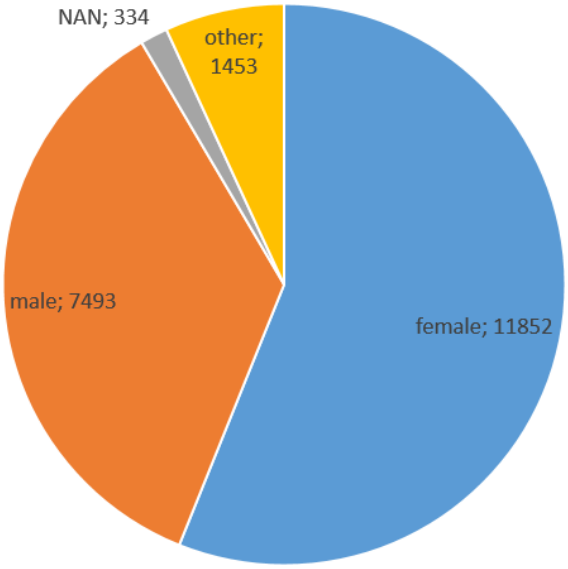
Frekari tölfræði um kynjahlutfall og móðurmál (e. primary language) má sjá á mynd H15.2 annars vegar og H15.3 hins vegar. Þjálfunar-, þróunar- og matsundirsettin voru búin til með sömu aðferð og lýst er í H7 í M4 skýrslunni. Viðeigandi tölfræði er dregin saman í töflu 1. Þjálfunar-, þróunar- og matsundirsettin voru búin til með sömu aðferð og lýst er í H7 í M4 skýrslunni.



Mynd H15.1 Aldursdreifingin í unglingataldæmunum sem safnað var.



Mynd H15.2 Móðurmálsdreifing í unglingataldæmunum sem safnað var.



Mynd H15.3 Kynjadreifing í unglingataldæmunum sem safnað var.

Tafla H1. Tölfræði mælenda og segða í unglingamálheildinni.

	Lengd	Segðir	Mælendur	Einstakar segðir	Mælendur **eingöngu** með einstakar segðir	Mælendur með einstakar segðir
Eval	2 klst. 44 mín.	2250	204	2228	200	203
Dev	2klst. 27 mín.	2049	67	1826	0	67
Train	19 klst. 45 mín.	16.834	215	13.111	0	211

Hljóð- og mállíkönin voru þjálfuð með sömu aðferð og lýst er í H7 í M4 skýrslunni. Með því að nota risamálheildina náðum við 23,33% orðvillutíðni og 30,17% orðvillutíðni á þróunar- og matssettum.

T1 - Upptökur 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. 8 raddir verða teknar upp, 4 karlraddir og 4 kvenraddir, hver þátttakandi skal lesa í a.m.k. 20 klukkustundir. Í þessum verkþætti verður unnið náið með verkþætti 6.5, við undirbúning handrita til upplesturs.

Varða 4 - lýsing

Verk:

- Ljúka upptökum á síðustu tveimur röddunum
- Eftirvinna gögnin
- Gefa gögnin út

Átta þátttakendur hafa tekið upp 20 klukkustundir tals. Eftirvinnslu gagnanna er lokið og þau hafa verið gefin út.

Afurðir

Allar raddirnar átta hafa verið gefnar út sem málheildin Talrómur. Heildarstærð hvers setts og fjöldi klukkustunda á hverja rödd er útlistaður í töflu 1. Við höfum ekki enn fundið stað við hæfi þar sem hægt er að geyma gögnin þannig að almenningur geti nálgast þau. Minni sýnisútgáfa málheildarinnar (e. demo corpus) (22MB) sem inniheldur 10 dæmi frá hverri upptekinni rödd væri hægt að hlaða upp á CLARIN ásamt hlekkjum á heildarupptökurnar þegar hýsisvæði (e. storage location) finnst.
Hljóðið verður gefið út sem 22050hz, 16-bita, einnar rása wave-skrár þó upphaflegu upptökurnar hafi verið 44100hz. Þetta er bæði gert til að minnka stærð málheildarinnar og fylgja sömu stöðlum og mikið notaðar málheildir af sömu gerð.

Auðkenni	Nafn	# upptaka[51]	Heildartímalengd[52]	Stærð
a	Rósa	9.899	16klst32m12s	2,6GB
b	Bjartur	12.048	25klst43m05s	3,9GB
c	Diljá	13.691	27klst57m33s	4,3GB

[51] Fjöldi segða sem ekki eru endurteknar
[52] Tímalengd er reiknuð sem heildarlengd hljóðupptökunnar (að þögnum meðtöldum)

d	Búi	12.357	22klst32m58s	3,5GB
e	Ugla	20.050	31klst28m04s	4,8GB
f	Álfur	19.849	29klst07m18s	4,5GB
g	Salka	16.886	30klst09m38s	4,6GB
h	Steinn	17.637	29klst49m01s	4,6GB
Samtals		**122.417**	**213klst19m49s**	**32,8GB**

Skýrsla

Upptökum var lokið áður en vinna við þessa vörðu hófst. Á meðan á eftirvinnslu stóð uppgötvaðist að einkum tvær raddir (þær fyrstu í upptökuverinu) innihéldu mikið af endurteknum segðum, sem stytti nýtilegt efni af upptökunum um nokkrar klukkustundir.

Vinna við að klippa út þagnir í miðjum segðum og við upphaf og endi segða hófst og lauk. Þegar kemur að útgefnum gögnum, hins vegar, var ákveðið halda upptökunum í sinni upphaflegu mynd, bæði vegna þess að ef einhver mistök hefðu verið gerð í klippingu myndi það valda varanlegum *skaða* á málheildinni, og vegna þess að upphaflegu upptökurnar eru *eðlilegri.* Enn fremur höfum við komist að því, í gegnum reynslu okkar í T6 og T13, að í reynd er þess ekki endilega þörf.

Hluti af klippiferlinu fólst í því að þjálfa þvingunarsamraðara (e. forced aligner) og nota hann til að samraða hljóði og umritun bæði á orðastigi (e. word level) og hljóðastigi (e. phonetic level). Samraðarinn var þjálfaður með öllum átta röddum með Montreal Forced Aligner tólinu[53]. Þetta nýttist ekki afurðum vörðunnar en reyndist nytsamlegt við önnur verk (T6 og T13) svo við leggjum til að samröðunin verði gefin út samhliða málheildinni.

[53] MFA: https://montreal-forced-aligner.readthedocs.io/en/latest/

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 4 - lýsing

30 þátttakendur hafa verið teknir upp.

Afurðir

Markmiðum vörðunnar hefur ekki verið náð. Takmarkanir vegna Covid-19 hafa enn og aftur komið í veg fyrir að við getum hafið upptökur í þessum verkþætti.
Úr úrvalshópi 122 mælenda, höfum við valið og staðfest þátttöku fjögurra mælenda, sem og tilgreint 36 til viðbótar til að velja úr. Við erum reiðubúin að hefja upptökur um leið og við fáum aðgang að upptökuverinu hjá RÚV, en áformað er að það verði 1. febrúar.

Skýrsla

Verkþátturinn fór vel af stað. Við höfðum sett upp nýliðunar- (e. recruitment) og áheyrnarprufukerfi á netinu til að safna radddæmum og velja raddir í málheildina. Prufurnar gengu vel og við söfnuðum dæmum frá 122 mismunandi röddum, 62 karlröddum og 60 kvenröddum á aldrinum 15 til 73 ára.

Við ákváðum að stefna að 4 hópum með 10 manns í hverjum. Við völdum að skipuleggja hópana sem karla- og kvenna-, há og lág F0-gildi, með það einnig að markmiði að raddirnar innan hvers hóps væru eins líkar og hægt væri. Við byrjuðum á því að velja fjórar „kjarnaraddir“ til að vera miðpunktur hvers hóps. Þessir þátttakendur hafa staðfest þátttöku sína í næstu umferð. Næst drógum við i-vigra (e. i-vectors) út úr dæmunum sem hver og einn mælandi lagði til og röðuðum öllum röddum eftir PLDA-einkunn, samanborið við hverja af kjarnaröddunum frá hárri til lágrar. Að lokum fór fram huglægt mat á efstu 20 röddunum fyrir hvern hóp. Það var nauðsynlegt þar sem uppröðun eftir líkindum radda fengnum úr i-vigrum innihélt nokkur frávik. Þrír starfsmenn gáfu öllum 80 samhliða pörunum einkunn á bilinu 0-5. Einkunnirnar voru síðan sameinaðar og 9 raddir með hæstu einkunnina valdar fyrir hvern hóp en reynt að forðast skörun milli hópa.

Sú ákvörðun að nota sömu skriftu fyrir þessa málheild og fyrir hljóðið í T1 er enn í gildi, en með örfáum breytingun þó. Við bættum við tveimur lengri málsgreinum sem ætlað er að bæta

greiningu á hljómfalli og framburðartilbrigðum (e. accent). Málsgreinarnar voru þannig úr garði gerðar að þær innihalda hvert dæmi um tilbrigði í framburði og flækjur (e. intricacies) sem getið er um á https://mallyskur.is/. Málsgreinarnar má sjá í viðauka A.
Einni lengri (1-2 mínútna) málsgrein verður einnig bætt við skriftuna til að auðvelda hljómfallsgreiningu.

Viðauki A - Skriftuviðbætur (e. script additions)

Tveimur lengri setningum var bætt við skriftuna:

Nepjan blés um rjóða vanga stúlkunnar.
Hver hviðan á fætur annarri, beint framan í hana, þar sem hún stóð í þykkri úlpu við lítið sker.
Fuglar flugu yfir og hvalir blésu þegar englar komu að henni siglandi á skötum. Þeir færðu henni gult ljós og einn þeirra sagði: „Sekur er sá, sem lögin hefur brotið. Logið hefur fanturinn hann Bjarni, seint mun annar eins maðkur skríða um þessa akra. Nú hef ég mælt.“ Það var löng þögn.
„Hvað skal gera?“ spurði stúlkan loks með lútandi höfði.
„Hans dómur er ei harður. En hefna skalt þú aumt hjarta.
Í skaut hans skalt þú letra þessi orð, og með því vekja illan anda. Varla mun hann kveljast þessa vegna, en hugur hans mun hverfa.“

Hún hafði lagt rakt handklæðið upp á ofninn hliðina á buxunum.
Lítill pili lék sér með svamp í faðmi hennar.
Sykur, skyr, einn peli af mjólk og hálf kringla var allt sem var eftir.
Hún hefði fundið út úr þessu á endanum, en varla mun henni svo vegna vel í bankanum í dag.

T6 - Veflesari

Skýrsla: Háskólinn í Reykjavík
Aðilar: 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 4 - lýsing

Frumgerð veflesaratólsins.

Afurðir

Öllum markmiðum hefur verið náð. Frumgerðin hefur verið gefin út: https://github.com/cadia-lvl/WebRICE/releases/tag/1.0-alpha. Github-hrislan fyrir WebRICE hefur verið gerð aðgengileg á https://github.com/cadia-lvl/WebRICE. Hún inniheldur einfalda prufuútgáfu (e. demo) af núverandi stöðu verkefnisins.

Skýrsla

Eins og staðan er í dag hermir WebRICE eftir virkni talgervla, hefur einfalt notendaviðmót og uppfyllir kröfu um aðgengi skv. WCAG 2.0 stig AA: stöðlum er fylgt á viðunandi hátt (WCAG 2.0 Level AA: Acceptable compliance).

Innan notendaviðmótsins geta notendur hlustað á lestur vefsíðu, gert hlé á lestrinum, stöðvað hann og hafið hann að nýju, sem og stillt lestrarhraðann. Lyklaborðsviðmót hefur einnig verið innleitt fyrir WebRICE, þannig að frumgerðin uppfyllir kröfur um aðgengi. Notendur geta opnað stillingavalmynd og lesið um hvað veflesari er og hvað hann gerir. Aðgerð til að loka stillingunum auðveldlega var einnig bætt við.

Hönnuðir og stjórnendur vefsíðna geta breytt útliti hnappa á vefsíðum sínum innan kóðans.

Á bakvið tjöldin hefur möguleika á að vista stillingar notenda (e. client storage) verið bætt við til að vista hraðastillingar notanda milli skipta (e. sessions). Þessi virkni verður seinna útvíkkuð til að innihalda frekari stillingar, þegar þeim hefur verið bætt við.

Sem stendur hermir lesarinn eftir virkni talgervils og notar til þess talgervilsraddirnar okkar, sem er í samræmi við þær kröfur sem gerðar eru til frumgerðarinnar. Vefsíðutextanum er enn safnað svo hann verði tilbúinn til sendingar til bakenda, þegar því verður bætt við. Við höfum lokið við

inn- og úttak fyrir forritaskil talgervilsins. Tiro hefur hannað fyrstu forritaskil talgervils fyrir veflesarann sem voru tilbúin seint í janúar.

Með þeim þáttum sem sagt er frá að ofan höfum við mætt þeim kröfum sem gerðar eru til frumgerðarinnar okkar. Þess vegna stefnum við að því á fyrstu vikum M5 að undirbúa WebRICE fyrir notendapróf með Félagi lesblindra og Félagi blindra og sjónskertra. Markmiðið er að þeim verði lokið í febrúar og nægur tími til að innleiða mikilvægar breytingar fyrir næstu útgáfu WebRICE þar sem þróuninni verður haldið áfram.

T7 - Forskriftir að röddum

Skýrsla: Mál- og raddtæknistofa Gervigreindarseturs Háskólans í Reykjavík (LVL)
Aðilar: Mál- og raddtæknistofa Gervigreindarseturs Háskólans í Reykjavík (LVL)

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.

Ætlaðir notendur: DEV, OTH
Afhendingarform: APP
Hugbúnaðarleyfi: Apache 2.0 eða sambærilegt

Varða 4 - lýsing

Fyrsta útgáfa forskrifta einingavalstalgervla gefin út með öllum nauðsynlegum málföngum.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.
Fyrsta útgáfa forskriftar einingavalstalgervis fyrir The Festival Speech Synthesis System er tilbúin fyrir allar átta raddirnar sem gefnar voru út í T1. Forskriftin er aðgengileg undir MIT-leyfi sem hirsla á Github-síðu Mál- og raddtæknistofu Gervigreindarseturs Háskólans í Reykjavík (LVL): unit-selection-festival.

Skýrsla

Nokkrar minniháttar breytingar hafa verið gerðar á forskriftinni síðan frumgerðin var gefin út, færsluskriftan og Docker-skrifturnar voru einfaldaðar eins og hægt er og aðlagaðar til að keyra á síðustu útgáfu talgervilsgagna úr T1 *eins og þær koma fyrir* (e. out of the box*)*, þ.e. án nokkurra breytinga eða stillinga (e. tinkering). Docker-gámurinn fyrir talgervingu (e. synthesis) hefur verið bestaður (e. optimized) til að innihalda aðeins nauðsynlegar skrár fyrir talgervingu, í því skyni að minnka hann.
Nákvæm kembing var framkvæmd á röddunum sem voru búnar til. Þetta fól í sér að hlusta á tilbúin hljóð (e. synthesized audio), orðhluta og einstök hljóð eða einingar og skrifa hjá sér villur á meðan. Niðurstöðurnar sýndu að helsti gallinn við raddirnar eru vitlaust klipptar einingar, enn frekar svo þegar hljóðeining (e. phone) hefur fáa ramma. Þetta var að einhverju leyti lagað með því að einfalda val á einingum. Að öðru leyti virðast raddirnar virka vel, ytri stikar í tengslum við tengingar og sléttun hafa verið bestaðir fyrir einstakar raddir. Þetta býður frekara mats, sem hluti af T13, en í bili hafa þessir ytri stikar verið látnir vera í sjálfgefnum stillingum.

SPSS-forskriftin er á áætlun. Auk þess höfum við bætt við forskrift fyrirFastSpeech2[54], sem inniheldur Docker-skrá til að keyra þjálfað líkan. Bæði forskriftin og líkanið eru enn í þróun en aðgengileg á Github. FastSpeech2-forskriftin keyrir einnig á T1 gögnunum eins og þau koma fyrir og býr til raddir hraðar (~48 klukkustundir á 4 GPU) en aðferðir sem áður höfðu verið innleiddar, s.s. Tacotron2, en þarfnast auka raddkóðara.

Þróun slíkra forskrifta fyrir auka raddkóðara stendur einnig yfir, sem og pökkun þeirra í Docker-gáma. Þetta nær til algengustu raddkóðaranna, WORLD og STRAIGHT. Forskrift fyrir STRAIGHT-raddkóðarann hefur verið gefin út, en hún inniheldur Docker-mynd sem keyrir Matlab-forrit á tvíundasniði í Linux-umhverfi, sem og Docker-skrá til að keyra STRAIGHT afritunartalgervingu (e. copy-synthesis). Bæði forskriftin og forritið á tvíundasniði, ásamt notendaleiðbeiningum, er aðgengilegur á Github.

[54] FastSpeech 2: Fast and High-Quality End-to-End Text to Speech - https://arxiv.org/abs/2006.04558

T8 - Mat á gæðum talgervla

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

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. Prófin eru svo skipulögð með því að undirbúa raddirnar, tímasetja prófin og fá hlustendur. Niðurstöðurnar eru einnig metnar og greint frá þeim (e. reported).

Varða 4 - lýsing

Skýrsla um þróun talgervla. Hönnun og innleiðing hlutlægs mats. Hugbúnaðarþróun á umhverfi huglægs mats.

Afurðir

Skýrsla um huglægt mat talgervla hefur verið rituð.
Engin sérsniðin útfærsla á mælistikum fyrir hlutlægt mat talgervla var hönnuð eða innleidd. Við ætlum okkur að nota núverandi útfærslu á stöðluðum aðferðir á borð við þessa PESQ útfærslu fyrir Python.
Frumgerð á huglægu matsumhverfi hefur verið þróuð sem hluti af LOBE-kerfinu, en frumkóðinn er aðgengilegur á https://github.com/cadia-lvl/LOBE. Eins og er styður þetta umhverfi Mean Opinion Score (MOS) prófanir, sem eru mest notuðu aðferðirnar við rannsóknir á talgervlum.

Athugið að frumgerð matsumhverfisins er markmið innan M5. Þessu var forgangsraðað til að nýta umhverfið til að meta útkomu T1, T2, T7 og T13 á komandi mánuðum.

Skýrsla

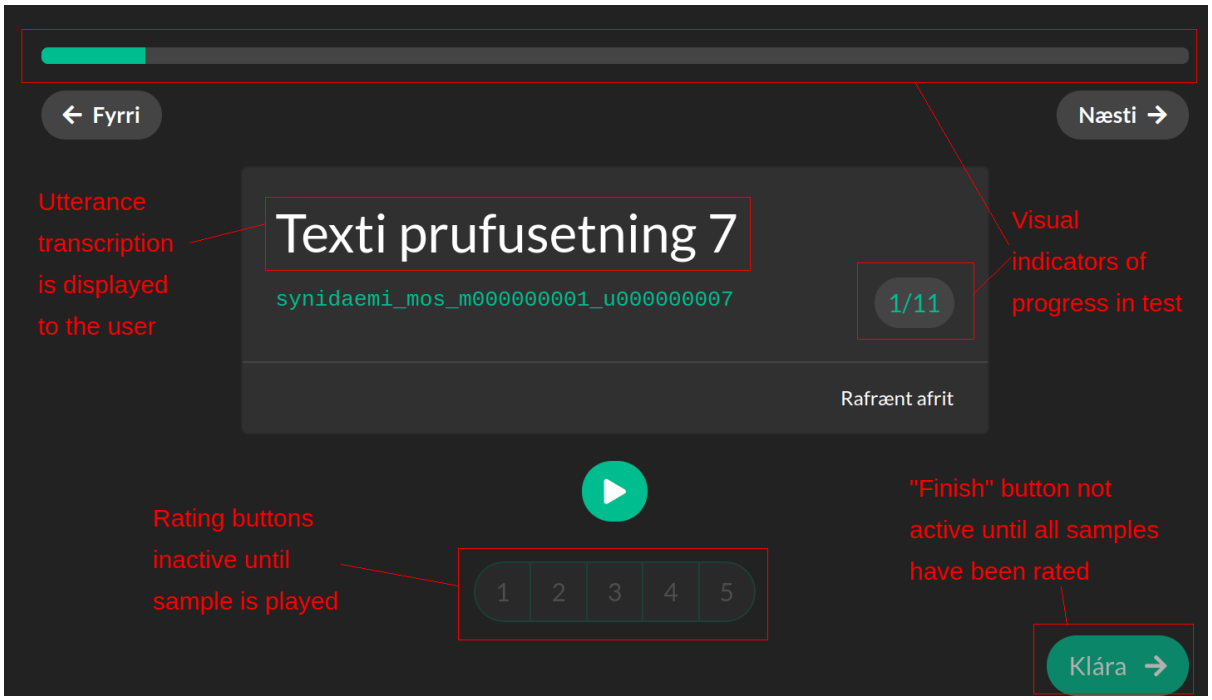
Eftir að hafa farið yfir fræðigreinar drögum við þá ályktun að MOS-virkni væri eðlilegasti upphafspunkturinn fyrir umhverfi huglægs mats, og að innleiðing á okkar eigin hlutlægu matssvítu (e. objective evaluation suite) væri ekki forgangsatriði.

Í MOS-prófi er hlustandinn beðinn að gefa segðum einkunn í heilum tölum frá 1 til 5, byggt á huglægum viðmiðum, t.d. hversu *eðlilega* (e. *natural)* röddin hljómar. Ein segð er kynnt hlustendum í einu, án þess gagngert að uppfræða þá um hvaða líkan þeir eru að hlusta á, eða hvort þeir eru að hlusta á mannsrödd eða tilbúna rödd. Þessar álitseinkunnir (e. opinion scores) eru síðan sameinaðar í einn mælikvarða fyrir hvert líkan, nefnilega meðalálitseinkunn (e. Mean

Opinion Score). Þessi virkni var innleidd sem hluti af LOBE-forritinu. Rannsakendur geta hlaðið upp taldæmum ásamt umritunum og öðrum lýsigögnum. Próf er þá útbúið sjálfvirkt og rannsakendur geta sent hlekk á það próf til þátttakenda. Mynd 1 sýnir eitt þessara prófa í notkun, ásamt nokkrum athugasemdum sem lýsa ýmsum atriðum í viðmóti prófsins. Eftir að einkunnum hefur verið safnað saman reiknar umhverfið út og sýnir lykilatriði á borð við meðaltöl, staðalfrávik, fjölda einkunna á hverja segð og fleira.

Sem stendur erum við að skipuleggja mat á talgervilslíkönunum sem þróuð voru með notkun þessara gagna. Það mun reynast gott próf á umhverfið og við stefnum að því að nota útkomuna og endurgjöfina úr því mati til að bæta og auka við þessa frumgerð.

Núverandi áætlanir um viðbætur fela í sér virkni A/B prófa, sjálfvirkar stillingar rómversks fernings (e. Latin square) og DMOS stuðning.



Mynd 1. Skjáskot af matsumhverfinu í notkun.

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 4 – lýsing

Flokkun óstaðlaðra orða og kennsl borin á margræða flokka. Umfang verkefnisins skilgreint, t.d. með því að skilgreina þær gerðir texta (e. text domains) sem kerfið ætti að ráða við. Prófunargögn tilbúin, vinna við gerð normunarreglna hafin.

Afurðir

Öllum markmiðum vörðunnar var náð. Táknfræðiflokkar (flokkar óstaðlaðra orða, e. Semiotic classes) í íslensku skilgreindir. 40.000 setningar merktar handvirkt. Kerfið sem búið var til ræður við alls konar fréttatexta. Helsta margræðnin lýtur að því hvort bandstrik (e. dash) merki hlé eða úrslit íþróttaleiks. Ef textinn er um íþróttir, þá er óðalið (e. domain) skilgreint sem slíkt.

Óðul sem ekki voru skoðuð sérstaklega voru stærðfræði- og lagatextar, þó núverandi reglubyggða kerfi ætti að ráða við stærstan hluta þeirra. Framtíðarkerfi byggt á tauganetum mun taka þessi óðul með í reikninginn.

Skýrsla

Táknfræðiflokkar

Táknfræðiflokkar fyrir íslensku eru skilgreindir eins og sjá má í töflunni fyrir neðan.

FLOKKUR	ÚTSKÝRING	DÆMI	NORMAÐ

PLAIN	orðmyndir sem á ekki að breyta	dæmum	dæmum
PUNCT	greinarmerki	.,?!:;”„“…·|	.,?!:;”„“…·|
CARDINAL	höfuðtölur	• 86.761 • 337.429	• áttatíu og sex þúsund sjö hundruð sextíu og einn/ein/eitt/eina/einum/einni/einu/eins/einnar • þrjú hundruð þrjátíu og sjö þúsund fjögur hundruð tuttugu og níu
ORDINAL	raðtölur	• 86.761. • 337.429.	• áttatíu og sex þúsund sjö hundruð sextugasti og fyrsti/sextugasta og fyrsta/sextugustu og fyrstu/ • þrjú hundruð þrjátíu og sjö þúsund fjögur hundruð tuttugasti og níundi/tuttugasta og níunda/tuttugustu og níundu
LETTERS	skammstafanir lesnar sem bókstafir	• KR • ehf • FH • KFUM	• K R (ká err) • E H F (e há eff) • F H (eff há) • K F U M (ká eff u emm)
DATE	dagsetningar, geta verið mánuður-ár, dagur-mánuður-ár, ár, o.s.frv.	• 1919 • 29. september 1928 • 14. mars • september 2008	• nítján hundruð og nítján • tuttugasta/i og níunda/i september nítján hundruð tuttugu og átta • fjórtándi/a mars • september tvö þúsund og átta
TIME	tími dagsins, byrjar oft á kl./klukkan	• kl. 20:00 • 10:30 • kl. 11 • 05:27 • klukkan 11.15	• klukkan tuttugu núll núll • tíu þrjátíu • klukkan ellefu • núll fimm tuttugu og sjö • klukkan ellefu fimmtán
MEASURE	prósentur, metrar, kílógrömm, o.s.frv. Geta haft + eða - á undan.	• 120 kw • 5 km • -39,5 kg • 44,5%	• hundrað og tuttugu kíló(wött/wöttum/watta) • fimm kílómetr(ar/um/a/i) • undir þrjátíu og níu komma fimm kíló(i/s/um/a) • fjörutíu og (fjögur komma fimm prósent/fjórum komma fimm prósentum/fjögurra komma fimm prósenta)

		• 120 g • 180°C • 320 MW	• hundrað og tuttugu (grömm/grömmum/gramma) • hundrað og áttatíu gráður selsíus • þrjú hundrað og tuttugu mega(wött/wöttum/watta)
SYMB	tákn	• + • - • @ • © • 3 **x** 3	• plús • mínus • hjá? • höfundarréttur • þrír **sinnum** þrír
ABBR	styttingar	• a.m.k. • SV-átt • þ.e.a.s.	• að minnsta kosti • suðvestanátt • það er að segja
WLINK	hlekkir og tölvupóstföng	• helgasvala@gmail.com • @BarackObama • https://en.wikipedia.org/wiki/List_of_most-followed_Twitter_accounts • #ljosanott2014	• h e l g a s v a l a hjá g m a i l punktur c o m • hjá b a r a c k o b a m a • h t t p s tvípunktur skástrik skástrik e n punktur w i k i p e d i a punktur o r g skástrik w i k i skástrik l i s t undirstrik o f undirstrik m o s t bandstrik f o l l o w e d undirstrik t w i t t e r undirstrik a c c o u n t s • myllumerki l j o s a n o t t tveir núll einn fjórir
DECIMAL	tugabrot	• 0,45 • 55,63 • 902.543.542,1	• núll komma fjórir/fjóra/fjórum/fjögurra/fjóra r/fjögur fimm • fimmtíu og fimm komma sex þrír/þrjá/þremur/þriggja/þrjár/ þrjú • níu hundruð og tvær milljónir fimm hundruð fjörutíu og þrjú þúsund fimm hundruð fjörutíu og tveir/tvo/tveimur/tveggja/tvær/ tvö komma einn/einum/eins/ein/eina/einni /einnar/eitt/einu
SPORT	úrslit íþróttaleikja	• 2-1 • 2:0 • 16/5 (fráköst)	• tvö eitt • tvö núll • sextán <sil> fimm (fráköst)
RNUM	rómverskar tölur	• XII • X	• tólf/tólfti/tólfta/tólftu • tíu/tíundi/tíunda/tíundu

		• XXIV • III	• tuttugu og fjögur/tuttugasti og fjórði/tuttugasta og fjórða/tuttugustu og fjórðu • þrjú/þriðji/þriðja/þriðju
FRACTION	almenn brot	• ½ • 2/6 • 1⅓	• hálfur/hálf/hálft/hálfan/hálfa/hálfum/hálfri/hálfu/hálfs/hálfrar • tveir/tvær/tvö/tvo/tveimur/tveggja sjöttu • einn og einn þriðji/ein og ein þriðja/eitt og eitt þriðja/einn og einn þriðja/eina og eina þriðju/einum og einum þriðja/einni og einni þriðju/einu og einu þriðja/eins og eins þriðja/einnar og einnar þriðju
DIGIT	tölur lesnar sem tölustafir	• 1109-05-420000 • 5543659 • 083	• einn einn núll níu <sil> núll fimm <sil> fjórir tveir núll núll núll núll • fimm fimm fjórir þrír sex fimm níu • núll átta þrír
MONEY	peningaupphæðir	3000 kr. kr. 4000 $40 €32 32 m.kr.	• þrjú þúsund krónur • fjögur þúsund krónur • fjörutíu dollarar • þrjátíu og tvær evrur • þrjátíu og tvær milljónir króna

40.000 setningar (770.000 orð) voru tekin úr útgáfu Risamálheildarinnar frá 2017 og merkt handvirkt eftir þessum flokkum. Gagnaramminn sem til varð lítur svona út:

Auðkenni setningar	**Auðkenni tóka**	**Fyrir**	**Eftir**	**Táknfræði**
1	1	13.	Þrettánda	ORDINAL
1	2	-	til	SYMB
1	3	30. nóvember 2012	þrítugasta nóvember tvö þúsund og tólf	DATE
1	4	/	skástrik	SYMB
1	5	Reykjanesbær	Reykjanesbær	PLAIN

1	6	88-81	áttatíu og átta áttatíu og eitt	SPORT
1	7	Mosfellsbær	Mosfellsbær	PLAIN
1	8	.	.	PUNCT
1	9	<EOS>	<EOS>	<EOS>

Margræðni

Eins og áður hefur komið fram er það helst bandstrikið sem er margrætt. Það getur gefið til kynna:

1. bandstrik (í hlekkjum),
2. úrslit íþróttaleikja,
3. bil á milli alls kyns talna,
4. þögn,
5. sérstakan tóka (inni í orði) - fyrir þá hluta sem lesa á með viðeigandi áherslu

Dæmi:

1. **kef-fc@keflavik.is**: kef *bandstrik* f c hjá keflavik punktur is
2. **Leikurinn fór 2-1 fyrir KR**: Leikurinn fór *tvö eitt* fyrir K R
3. **Austan 3-8 m/s:** Austan þrír *til* átta metrar á sekúndu
4. **Athugið - á morgun breytast reglurnar:** Athugið <sil> á morgun breytast reglurnar
5. **Austur-Skaftafellssýsla**: Austur-Skaftafellssýsla

Skástrik getur gefið til kynna:

1. skástrik
2. og/eða
3. brot (e. fraction)
4. ekkert

Dæmi:

1. **mbl.is/innlent:** m b l punktur is skástrik innlent
2. **hljóðbækur/rafbækur:** hljóðbækur og eða rafbækur
3. **⅓:** einn þriðji
4. **Grótta/KR:** Grótta K R

Ritun normunarreglna

Fyrra reglukerfi fyrir tetanormun í íslensku var kerfi sem búið var til með OpenFST's *Thrax Development Tools.* Þessi tól taka saman einskonar málfræðireglur á formi reglulegra segða og samhengisháðar umritunarreglur (e. rewrite rules) og varpa þeim á vegin endanleg stöðuferjöld (e. weighted finite state transducers) og eru almennt notuð til að búa til reglubyggð textanormunarkerfi. Nýtt kerfi var búið til með því að nota reglulegar segðir og málfræðireglur,

fengnar með málfræðilegri mörkun. Tala tekur á sig form með form næsta orðs að leiðarljósi, hvort það er nafnorð eða lýsingarorð. Með sama hætti tekur skammstöfun á sig form með forsetninguna á undan að leiðarljósi.

Dæmi um þetta:

Við teljum sem svarar til 3 sek.
|
`(notkun styttingarreglna sem byggja á forsetningunni á undan)`
|
Við teljum sem svarar til 3 sekúndna
|
`(notkun talnareglna sem byggja á nafnorðinu á eftir)`
|
Við teljum sem svarar til þriggja sekúndna

Ef engar málfræðireglur hefðu verið notaðar hefði setningin verið normuð sem:

Við teljum sem svarar til þrjú sekúndur

Með sambærilegum hætti höndlar kerfið muninn á milli ein- og fleirtöluorða, byggt á tölunni sem kemur á undan::

Við teljum sem svarar til 1 sek.
|
`(notkun styttingarreglna sem byggja á forsetningunni á undan)`
|
Við teljum sem svarar til 1 sekúndu
|
`(notkun talnareglna sem byggja á nafnorðinu á eftir)`
|
Við teljum sem svarar til einnar sekúndu

Ef engar málfræðireglur hefðu verið notaðar hefði setningin verið normuð sem:

Við teljum sem svarar til eitt sekúndur

Ályktanir

Niðurstöður nýja kerfisins verða bornar saman við handvirkt merktu gögnin til að ákvarða um gæðin. Þessi varða var einkum notuð til að búa til reglur til að útbúa grunngögnin. Til þess að prófa nýjar aðferðir, eins og tauganet, þurfum við viðunandi gögn sem kerfið getur lært af. Mörg kerfi, eins og vélþýðingarkerfi, búa yfir gögnum sem verða til með náttúrulegum hætti, þar sem manneskja hefur beinlínis þýtt texta af einu tungumáli á annað. Þar sem fólk fólk getur jafnan

lesið tölur, styttingar og önnur óstöðluð orð án þess að það krefjist neinnar skýringar, þá er ekki til nein „þýðing“ á ónormuðum texta til normaðs. Í ljósi þess að ekki er hægt að afla normaðra gagna með náttúrulegum hætti er nauðsynlegt að búa þau til.

T10 - Sjálfvirk hljóðritun

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

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 einingunа til að útbúa nýjar orðabækur, t.d. ef of mörg orð vantar frá sérhæfðu óðali í kjarnaorðabók. Þessi verkþáttur er mjög nátengdur G6, þróun framburðarorðabókar.

Varða 4 - lýsing

Endurbætt g2p-líkön innleidd í bæði talgervilskerfin (einingaval og SPSS).

Afurðir

Öllum markmiðum vörðunnar var náð. Sjálfstæða g2p-vefþjónustan er aðgengileg hér: https://github.com/cadia-lvl/g2p-service/commits/master
Endurbætta g2p-líkanið var innleitt í einingavalskerfið, aðgengilegt hér: https://github.com/cadia-lvl/unit-selection-festival/
Endurbætta g2p-líkanið var innleitt í SPSS fastspeech2 kerfið, aðgengilegt hér: https://github.com/cadia-lvl/FastSpeech2

Skýrsla

Innleiðing á endurbætta g2p-líkaninu var fjórþætt. Í fyrsta lagi skrifuðum við tilvísun með skrá yfir mismunandi g2p-líkön fyrir íslensku. Næst var besta g2p-líkaninu bætt við g2p-vefþjónustuna. Seinna skiptum við núverandi g2p-líkani í einingavalskerfinu út fyrir staðlað g2p-líkan úr G6. Að lokum skiptum við einnig út núverandi g2p-líkani fyrir líkanið úr G6.

Fyrst ber að geta þess að leiðbeiningarnar okkar lýsa í grófum dráttum núverandi g2p-líkönum fyrir íslensku og viðbúnu inn- og úttaki þeirra. Nánari lýsingu má finna hér: https://github.com/cadia-lvl/icelandic-NLP-resources/blob/main/g2p-reference.md

Hirsla	Hugbúnaður	IPA eða SAMPA	PER*	Details
g2p-lstm	fairseq	SAMPA	4,07% (standard)	grammatek/g2p-lstm
tts-data	sequitur	IPA	N/A	cadia-lvl/tts-data

g2p	sequitur	IPA	N/A	atliSig/g2p
g2p-service	sequitur	NA/IPA	N/A	rkjaran/g2p-service
g2p-thrax	thrax	SAMPA	6,49%	grammatek/g2p-thrax
althingi	sequitur	IPA	N/A	kaldi/egs/althingi/s5

Að hafa allar þessar upplýsingar á einum stað ætti að auðvelda hverjum sem er að nota þessi g2p-líkön eða uppfæra í nýrri í sjálfvirkum talgreinum eða talgervilsforritum. Með greinargerð okkar um g2p-líkön höfum við orðið nokkuð góða mynd af núverandi líkönum.

Þannig að við hófumst handa við að setja upp nýjustu g2p-líkönin fyrir LVL (g2p-lstm), sem er mjög einfalt.

Nú þegar við erum komin með virka rödd og virk g2p-líkön, þá innleiddum við fairseq g2p-líkönin í g2p-þjónustuna. G2p-lstm (fairseq g2p) kallar á að conda-umhverfi (e. conda environment) verði búið til. Hins vegar erum við að nota Docker og Docker-skrár virka ekki vel með conda. Þannig að við settum fairseq g2p-líkönin upp án conda. Síðan bættum við pökkunum við pip-kröfur (e. requirements). Sem stendur styður þjónustan núverandi sequitur g2p-líkan og fairseq mállýsku-g2p-líkönin fjögur. Docker-skráin sem inniheldur þessa þjónustu til uppsetningar á þjónum er aðgengileg hér: https://github.com/cadia-lvl/g2p-service

Næst settum við upp og notuðum unit-selection-festival/ til að búa til virka íslenska rödd í Festival[55]. Með því að nota Python-skriftur sem búnar voru til fyrir g2p-þjónustuna, flytjum við framburðinn út á tsv-skrá, aðlögum skeljaskriftu einingavalsins að því að nota fairseq-x-sampa í staðinn fyrir núverandi sequitur-ipa líkön, og breytum scheme skrám þannig að þær noti nýju g2p-lstm líkönin.
Að lokum innleiddum við endurbætta g2p-líkanið með fastspeech2. Við breyttum g2p_is.py til að nota X-SAMPA í staðinn fyrir IPA.

Talgervilsþjónusta fyrir hljóðritaðan texta

Við unnum líka að vefþjónustu fyrir PEDI hljóðritunarforritið í G6, sem tekur inn hljóðritaðan texta í stað venjulegs texta, í samstarfi við Tiro. Í stað þess að taka inn texta þá tekur vefþjónustan inn fónem (X-SAMPA) og úttakið er venjuleg hljóðskrá fyrir hvert orð.

Byrjunarútgáfa forritaskila hefur verið útbúin (fyrri útgáfur virkuðu með espeak og polly). Aftur á móti mun lokaútgáfan vinna með nýju talgervilsröddina okkar.
Á þessari vefsíðu má sjá hvernig á að nota hana: https://tts.tiro.is/

[55] http://www.cstr.ed.ac.uk/projects/festival/

Hér eru dæmi:
orð: https://tts.tiro.is/v0/speak?q=OrD
derp: https://tts.tiro.is/v0/speak?q=tEr_0p
hlaupa: https://tts.tiro.is/v0/speak?q=l_09i%3Apa
https://tts.tiro.is/v0/speak?q=sIltI

Að endingu, afrakstur T10 er ferns konar: g2p-þjónusta, einingavals-festival (e. unit-selection-festival), FastSpeech2, og forritaskil fyrir hljóðritaðan texta.

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: (Upptökur á íslenskum röddum í T1 og T2, frammistaða talgervils í T8 og forvinnsla texta í T9).

Varða 4 - lýsing

M4: Uppsetning málfanga fyrir íslensku og ensku fyrir SPSS. Innleiðing nýjustu SPSS-aðferðar fyrir ensku (frumgerð). Tilraunaniðurstöður bestu gerðar SPSS bornar saman við nýjar (e.novel) með því að nota nýja raddkóðara.

Afurðir

Ekki hefur öllum markmiðum vörðunnar verið náð.
Við þjálfuðum og gerðum tilraunir með nýlegt SPSS-kerfi fyrir ensku. Við innleiddum og gerðum tilraunir með sama kerfi fyrir íslenskar raddir úr Talrómi (T1) Engar niðurstöður hafa verið bornar saman við nýja raddkóðara. Við áformum að framkvæma viðeigandi mat á öllum kerfum sem við höfum gert tilraunir með og þróað á íslenskum gögnum. Þetta er háð bæði T1-M3, hvað varðar gögnin, og T8-M6, hvað varðar huglægan matsumhverfi. Eining fyrir hlutlægt mat og uppsetning fyrir nokkra almenna raddkóðara er tilbúin og hlutlægt mat fer fram samhliða huglægu mati í T8-M6, ef ekki fyrr.

Skýrsla

Farið var yfir nýjustu aðferðir í aðhvarfslíkanagerð (e. regression modelling). Tvær nálganir sem stóðu upp úr sem bestu kerfin sem völ er á í dag voru Tacotron2[56] og FastSpeech2[57]. Innleiðing þeirra beggja er aðgengileg á Github en hvorug er hluti af stærri þróunarumhverfi fyrir talgervla á borð við Merlin[58] eða Ossian[59]. Nokkur árangur hefur þegar náðst með notkun Tacotron2 eins og sjá má í fyrri skýrslum, þannig að meiri áhersla var lögð á FastSpeech2 innan þessarar vörðu.

[56] Tacotron2: https://arxiv.org/abs/1712.05884
[57] FastSpeech2: https://arxiv.org/abs/2006.04558
[58] Merlin TTS: https://github.com/CSTR-Edinburgh/merlin
[59] Ossian TTS: https://github.com/CSTR-Edinburgh/Ossian

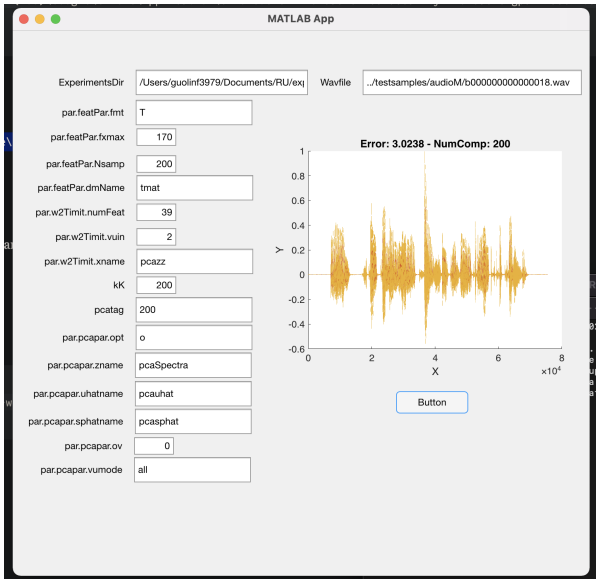
Við byrjuðum á því að endurgera niðurstöðurnar í greininni með því að nota óopinbera innleiðingu með opnu leyfi á Github[60]. Enska útgáfan styðst við LJSpeech[61] gagnasafnið. Endurgerða líkanið gaf niðurstöður sem *virtust*[62] sambærilegar þeim dæmum sem sjá má í upprunalegu greininni.
Við héldum áfram með því að útbúa kvísl (e. fork) af hirslunni og aðlaga kóðann að talaðri íslensku, eða nánar til tekið Talrómsmálheildinni úr T1.

Við þjálfuðum tvær íslenskar raddir, eina kvenrödd og eina karlrödd. Karlröddina með FastSpeech2 og MelGAN raddkóðara en kvenröddina með viðbótarraddkóðara (e. supplementary vocoder) þjálfuðum á kvenröddinni í LJSpeech. Ástæðan fyrir því að við byrjuðum með MelGAN raddkóðara í þessari útgáfu er vegna þess hve fjölhæfur hann er við að aðlagast nýjum röddum. Að þjálfa eina rödd tekur í kringum 48 klukkustundir með því að nota 4 GPU, en þjálfun raddkóðarans tók 2 vikur með 4 GPU. Hins vegar er hægt að þjálfa MelGAN og nota á allar átta raddir samtímis.
Hér má sjá dæmi. Karlröddin er sérstakur raddkóðari þjálfaður eingöngu fyrir hana. Kvenröddin notar raddkóðara sem er þjálfaður á enskri kvenrödd. Til samanburðar má finna Tacotron2 sýnisútgáfu úr M3 hér.

Í þessari vinnu voru afurðir hagnýttar sem búnar voru til í T1:
- Talrómsmálheildin (raddgögn)
- Samröðun fyrir Talrómsmálheildina

Verið er að aðlaga og prófa nýja (e. novel) raddkóðara (gci vocoder) á tveimur íslenskum röddum, einni kvenrödd og einni karlrödd. Hann hefur ekki enn verið metinn samkvæmt hlutlægu mati á borð við Pesq. Til að einfalda prófunarferlið var notendaviðmót útbúið til að ganga úr skugga um að allar slíkar breytur væru á einni síðu og auðvelt að breyta þeim.

[60] FastSpeech2 innleiðing: https://github.com/ming024/FastSpeech2
[61] LJSpeech: https://keithito.com/LJ-Speech-Dataset/
[62] Ekki metið formlega



Mynd 1. GUI fyrir gci-raddkóðara, með inntaksstikum og úttaki hans

Samstarf við atvinnulífið

Skýrsla: Grammatek
Aðilar: Grammatek

Skýrsla

Eins og fram kemur í upprunalega skjalinu *Tæknilýsingu* fyrir annað ár mun SÍM hefja undirbúning að samstarfi við atvinnulífið á öðru ári verkefnisins. Helsta markmið þessarar vinnu er í upphafi að kortleggja umhverfi máltækninnar, þ.e.a.s. hvaða máltæknilausnir eru þegar í notkun og hvaða fyrirtæki eru að þróa máltækni. Einnig að greina hvaða afurðir úr Máltækniáætlun fyrir íslensku má nota við skilgreindar lausnir og skilgreina möguleg skörð í íslenskri máltækni þegar kemur að þörfum atvinnulífsins.

Verkefni nýs starfsmann sem leiðir vinnu um samstarf við atvinnulífið hafa verið fjórþætt: a) kynna sér fyrirkomulag SÍM, verkefnin og teymin, b) að öðlast mikla yfirsýn yfir svið máltækninnar, c) gera athugun á fyrirtækjum og stofnunum, bæði á Íslandi og erlendis, sem gætu hugsanlega orðið samstarfsaðilar að verkefninu, d) stofna til fjölda tengsla, bæði við þá sem eru reyndir á sviði (mál)tækni og hugsanlegra notenda endahugbúaðar og afurða SÍM.

Að neðan er könnun á þeim sviðum sem til álita komu, næstu skref munu fela í sér samræður og ákvarðanir, í samstarfi við SÍM-hópinn og Almannaróm, um það hvaða leiðir skulu farnar og hvaða forgangsatriði skuli skilgreina. Þessar umræður ættu einnig að leiða til ályktana um það hvort SÍM þurfi að aðlaga verkáætlanir, sér í lagi fyrir þau verkefnisár sem eftir eru. Niðurstöður þeirrar vinnu, ásamt nákvæmri áætlun fyrir það sem eftir er af verkefnisárinu, verða afhentar fyrir 1. apríl 2021.

Alþjóðlegir aðilar
Eitt af markmiðum Máltækniáætlunar fyrir íslensku er að koma á fót tengslum við stóra alþjóðlega aðila: Apple, Google, Amazon og Microsoft, með það fyrir augum að gera það mögulegt að íslenska verði innifalinn í þeim hugbúnaðartólum og -umhverfi sem er mest notað. Sumar af vörum þeirra eru þegar mikið notaðar á Íslandi, en á ensku, t.d. stafrænu hjálparhellurnar (e. smart assistants) Siri, Alexa og Google Home. Mjög fá slík forrit og þjónustur eru í boði á íslensku: Google ASR og Google Translate bjóða upp á íslensku, textaútgáfa Microsoft Translator og Polly, talgervilsþjónusta Amazon. Sambandi hefur verið komið á við öll fjögur fyrirtækin, en næsta skref er að útbúa samskiptaáætlun með Almannarómi til að hefjast handa við að koma á raunverulegu samstarfi.

Eitt verkefni á *alþjóðamarkaði* (e. global player field) er að koma á tengingum og stefna að því að íslenska verði innifalin í máltækniþjónustum og -lausnum, þ.e. þar sem framlag SÍM myndi að öllum líkindum felast í gögnum fremur en hugbúnaði. Annað verkefni er að líta til tækja og umhverfis sem er í víðtækri notkun á Íslandi, þar sem þörf er á íbótum (e. plugins) fyrir íslenska

máltækni, t.d. málrýnir, lyklaborðsvirkni sem afmarkast við tungumálið (e. language specific keyboard functionalities) og talforrit í snjalltækjum. Þetta felur í sér að afhenda heildstæðari lausn, í samræmi við almenna staðla og forritaskil fyrir forritin sem um ræðir. Á stigi stýrikerfa verðum við að líta til Windows og OSx fyrir einkatölvur og Android og iOS fyrir snjallsíma/spjaldtölvur. Þegar kemur að vöfrum verðum við að líta til Google Chrome, Microsoft Edge, og Safari, og að lokum til þess að vinnustaða- og textavinnsforrit á borð við MS Office, Google Docs, Google Keep, eru mikið notuð á Íslandi og þurfa góðan stuðning íslenskrar máltækni.

Sérhæfð máltæknifyrirtæki
Við höfum kynnt okkur nokkur sérhæfð máltæknifyrirtæki sem starfa á alþjóðavísu. Nú þegar hefur verið komið á samstarfi við Lingsoft (https://www.lingsoft.fi/en) og Tilde https://tilde.com/. Síðastliðið vor tók Háskólinn í Reykjavík þátt í umsóknum í CEF Telekom áætlunina, Lingsoft leiddi aðra umsóknina og Tilde hina. Báðar umsóknirnar voru samþykktar og markmið þeirra evrópskri málvinnslutækni (e. European NLP technology). Enn fremur er málstofa með Lingstof á dagskránni, þar sem málgreining og máltækni í framleiðslu er meginmarkmiðið.

Dictus í Danmörku (https://dictus.dk/) er upplýst um verkefnið en frekara samstarfi hefur ekki verið komið á fót.

Önnur fyrirtæki sem við höfum verið að kynna okkur, einkum til að rannsaka forrit á þessu sviði, eru Duolingo (https://www.duolingo.com/), Speechmatics (https://www.speechmatics.com/) og Orcam (https://www.orcam.com/en/).

Notkunardæmi og samstarfsaðilar á Íslandi
Við höfum haft samband við fjölda fyrirtækja og stofnana á Íslandi með það fyrir augum að byrja að útfæra hugsanleg notkunardæmi og tækifæri sem innihalda afurðir SÍM. Þetta hefur einnig falist í því að atvinnulífsfulltrúinn (e. industry collaborator) aðstoði fyrirtæki innan SÍM við samskiptin, með það að markmiði að kynna afurðir SÍM og hugsanleg notkunardæmi. Sem stendur teljum við verkefnið Stafrænt Ísland (https://island.is/stafraent-island) vera mikilvægasta samstarfsaðilann. Markmið verkefnisins er að gera samskipti fyrirtækja og einstaklinga við allar stofnanir ríkisins stafrænar, alls staðar þar sem það er mögulegt. Þetta hefur í för með sér risavaxið samhæfingar- og samstarfsferli sem teygir sig til fjölda stofnana og fyrirtækja, og við teljum að við ættum að koma á samstarfi milli SÍM/Almannaróms og Stafræns íslands eins fljótt og mögulegt er. Tengslum hefur þegar verið komið á og umræður hafnar, og þessari vinnu ætti að halda áfram með það fyrir augum að skilgreina sameiginleg markmið innan þessa verkefnisárs.

Sértækir notendahópar
Hvað varðar lífvænleika (e. viability) íslenskra máltæknilausna, þá hefur oft verið litið svo á að smæð markaðarins sé helsta hindrunin í þróun slíkra lausna. Þetta á við um forrit sem ætluð eru almenningi, en jafnvel enn frekar hópum með sérþarfir: börn á tilteknum aldri, eldra fólk, lesblindir, heyrnarskertir eða heyrnarlausir, sjónskertir, svo aðeins fáeinir séu nefndir. Hins vegar hefðu þessir hópar gríðarlegan ávinning af sérsniðnum máltæknilausnum þegar kemur að

persónulegum þroska, réttinum til að hafa aðgengi að upplýsingum, auðveldari samskiptum og út frá heilbrigðis- og öryggissjónarmiðum.

Á heildina litið eru *sértækir notendahópar* mjög fjölbreyttir og samanstanda af fólki með mismunandi þarfir, sem allar ætti að taka til greina við frekari þróun máltækniáætlunar sem fjármögnuð er með almannafé. Fyrstu samningar hafa verið gerðir við forsvarsmenn margra fyrrnefndra hópa, en enn er þörf á frekari vinnu við að skilgreina a) raunverulegar þarfir mismunandi notendahópa, þekkt forrit o.s.frv., b) handhæga þróunarmöguleika (e. low-hanging-fruits‘ for development) c) kostnað og hagkvæmnisgreiningu á þeim forritum sem talin eru mjög mikilvæg / mikilvægust. Þetta felur einnig í sér að komast í samband við hönnuði forrita sem þegar eru í notkun á þessu sviði og fá þá til að greina frá möguleikunum á því að þau taki einnig til íslensku.