Samstarf um íslenska máltækni
SÍM

Máltækniáætlun fyrir íslensku

Varða 1 – lokaskýrsla

10. febrúar 2020

Efnisyfirlit

Yfirlit yfir aðgengilegar auðlindir

Gefið út til CLARIN:

Tilreiðari fyrir íslenskan texta
https://repository.clarin.is/repository/xmlui/handle/20.500.12537/11

Aðrar aðgengilegar málheildir og hugbúnaður

G1 – Risamálheild (RMH)
Aðgengi að málheildinni: http://malfong.is/?pg=rmh
Vefviðmót:
https://malheildir.arnastofnun.is/#?stats_reduce=word&isCaseInsensitive&searchBy=word&cqp=%5B%5D
Textar á ákveðni sviði fyrir handvirka stjórnun:
https://drive.google.com/drive/folders/12Esg6fNZtM_-1F2QjihE21Zp53jb72-m

G4 – Beygingarlýsing íslensks nútímamáls (BÍN)
Vefviðmót og aðgengi að almennri útgáfu gagnagrunnsins:
https://bin.arnastofnun.is/

G6 – Framburðarorðabók
Yfirlit afbrigða í íslensku hljóðkerfi og framburði, umritunarreglur og viðmið; frumgerð af tóli til yfirferðar orðabóka:
https://github.com/grammatek/pedi

G7 – Íslenskt orðanet
Vefviðmót: http://ordanet.is/

I1 – Mörkunartól
Annotald: https://github.com/mideind/annotald
UDAnnotatrix: https://maryszmary.github.io/ud-annotatrix/standalone/annotator.html
Málfræðilegur markari: https://github.com/stofnun-arna-magnussonar/pos-tagging-tool

I3 – Textatilreiðari
Gefið út til CLARIN. Github veita:
https://github.com/mideind/Tokenizer

V2 – Samhliða málheild
Prófunargögn, samhliða málheild:
https://drive.google.com/drive/folders/1yGZFClbwqyoOE6mSI9rFwFE2YAa_uuul?usp=sharing

V3 – Grunnlína í vélrænum þýðingum
Forritaskil þýðinga (sjá skýrslu bls. 38):
https://velthyding.mideind.is/nn/translate.api
http://nlp.cs.ru.is/moses/translateText
Frumkóði / veitur máltækni:

- *Transformer* og *Bi-LSTM* (tauganet):
    - *Docker compose* (fyrir Docker gáma) skrár: https://github.com/vesteinn/docker-greynir
    - *Docker* skrá: https://hub.docker.com/u/mideind
    - Kóði fyrir vefþjónustu: https://github.com/mideind/nnserver
    - Þjálfuð líkön með stillingum: http://velthyding.mideind.is/data
    - Forvinnsla og þjálfunarskriftur: https://github.com/mideind/nnserver/tree/master/nnserver/scripts og https://github.com/mideind/GreynirTranslate/tree/master/utils
- Moses:
    - Framendi þýðinga: https://hub.docker.com/r/haukurp/moses-lvl
    - Þjálfuð kerfi: https://hub.docker.com/r/haukurp/moses-smt
    - Kóði: https://github.com/cadia-lvl/SMT
    - *Docker compose* skrá: https://github.com/cadia-lvl/SMT/blob/master/docker-compose.yml

V5 – Viðmót á þýðingarvél
Frumgerð vefviðmóts:
http://velthyding.mideind.is/

M6 – Málrýnir fyrir stakorðavillur
Hugbúnaður, skrifleg gögn:
https://github.com/mideind/reynircorrect

H1 – Gagnaupptökur með Samrómi
Vefsíða Samróms:
https://samromur.is/

T1 – Upptaka á einingavalsgögnum
Hugbúnaður upptökuumhverfis:
https://github.com/cadia-lvl/lobe

T7 – Forskriftir að röddum
https://github.com/arnacadia

T12 – Mynstur og setningar
Leslistar og hugbúnaður:
https://github.com/cadia-lvl/tts_data

Aths: fyrirliggjandi skjal er þýðing úr frumútgáfunni, sem er á ensku.

G1 – Risamálheild

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

Varða 1 - lýsing

Skýrsla með lista yfir texta sem skal bæta við frá vefnum og stöðu leyfa. Meta hversu miklum texta skal bæta í málheildina, með tilliti til orðafjölda. Hefja viðræður við útgefendur bóka og leyfishafa útgefinna bóka. Skilgreina matssett fyrir undirmálheildir sem skal fara handvirkt yfir.

Afurðir

Öllum afurðum var skilað. Við kortlögðum vefsíður sem henta málheildinni og erum að tryggja leyfi, þar sem mögulegt er. Viðræður við bókaútgefendur hafnar.

Listi veita og gagnasafna til að bæta í málheildina, ef leyfishafar samþykkja og gefa okkur skriflegt leyfi:

Albumm	albumm.is	óútkljáð
Deiglan	deiglan.is	óútkljáð
Eiðfaxi	eidfaxi.is	óútkljáð
Eyjafréttir	eyjafrettir.is	óútkljáð
Eyjar	eyjar.net	óútkljáð
Fréttatíminn	frettatiminn.is	óútkljáð
Hringbraut	hringbraut.is	óútkljáð
Húni	huni.is	óútkljáð
Kaffið	kaffid.is	CC BY 4.0
Kvennablaðið	kvennabladid.is	óútkljáð
Mannlíf	mannlif.is	óútkljáð
Sunnlenska	sunnlenska.is	CC BY 4.0
Trú	tru.is	óútkljáð
Myllusetur	vb.is	óútkljáð
Myllusetur	fiskifrettir.is	óútkljáð
Vikudagur	vikudagur.is	óútkljáð
Viljinn	viljinn.is	CC BY 4.0

Tafla 1: Mögulegar veitur af vefnum 2020

Skilgreina matssett fyrir undirmálheildir sem skal fara handvirkt yfir

Textar sem nota skal til prófana hafa verið valdir af 9 sviðum: þingræður, lög, dómsúrskurðir, fréttir frá vefmiðlum og blöðum, fréttir úr sjónvarpi og útvarpi, fréttir frá íþróttamiðlum, útgefnar bækur, fræðsluvefsíður, og blogg og skoðanir í deiglunni.

Um það bil 20.000 orð voru valin af hverju sviði og skipt í 10 hluta. Hver hluti inniheldur setningar úr mismunandi málsgreinum, valdar af handahófi fyrir málheildina en dreift jafnt með tímanum. Gögnin eru staðsett á https://drive.google.com/drive/folders/12Esg6fNZtM_-1F2QjihE21Zp53jb72-m
Markaðan og lemmaðan texta má finna í skránum x.lemmad (þar sem x er tala milli 1 og 10). Upplýsingar um gögnin er að finna í skránum x.info. Fyrsti dálkurinn er slóð að XML-skrám sem málgreinin er tekin úr, annar dálkurinn númer málsgreinar og þriðji dálkurinn fjölda tóka í málsgreininni.
Dæmi frá 1.lemmad af sviði frétta (frettir):

```
#info:/textar/sofnun/tei/MIM/mbl/1999/06/G-15-3902398.xml   3
Ríkharður   nken-s      Ríkharður
sagði sfg3eþ      segja
að    c     að
leikmenn    nkfn  leikmaður
væru  svg3fþ      vera
hvergi      aa    hvergi
bangnir     lkfnsf      banginn
fyrir ao    fyrir
leikinn     nkeog leikur
gegn  aþ    gegn
Rússum      nkfþ-s      Rússi
ytra  lheevm      ytri
og    c     og
gerðu sfg3fþ      gera
okkur fp1fo ég
vonir nvfo  von
um    ao    um
að    cn    að
ná    sng   ná
einu  tfheþ einn
stigi nheþ  stig
.     .     .
```

Skýrsla

Skýrsla með lista yfir texta sem skal bæta við af vefnum og stöðu leyfismála

Fyrsta útgáfa *Risamálheildar* (IGC) var gefin út 2018, og inniheldur u.þ.b. 1,3 milljarða runuorða. Haustið 2019 var textum sem innihalda alls 105 milljón orð safnað, mest frá veitum sem við höfðum leyfi frá, en líka frá þrem nýjum veitum sem samþykktu CC BY 4.0 leyfi fyrir texta þeirra.

Við munum halda áfram að bæta við nýjum gögnum frá núverandi veitum á hverju ári, og metum það sem u.þ.b. 100 milljón orð árlega. Við höldum einnig við lista veita sem við erum að ræða við eða ljúka viðræðum við varðandi leyfismál (sjá kafla: Afurðir). Við reiknum með

að tryggja texta frá a.m.k. þriðjungi þessara veita, metið sem 10 milljón orð til viðbótar árið 2020.

Við horfum einnig til annarra veita, eins og blogga, og munum reyna að ná sambandi við virkustu bloggara. Listi yfir þá bloggara er í vinnslu og hafð verður samband við hvern í gegnum tölvupóst. Við vonumst til að safna allt að 5 milljón orðum með þessum hætti.

Það eru fáar stórar vefsíður með virk vefþing (spjallborð) á Íslandi. Við getum ekki haft samband við alla notendur, en munum ráðfæra okkur við lögfræðing ef það telst leyfilegt samkvæmt höfundarréttarlögum að safna þessum gögnum og bæta í málheildina eftir að við höfum skipt þeim í setningar. Við höfum metið að þrjár þessara vefsíðna innihalda u.þ.b. 80 milljón orð alls. Sjá Töflu 2 fyrir mat á heildartölum.

Hefja viðræður við útgefendur bóka og leyfishafa útgefinna.

Viðræðum við bókaútgefendur hafnar. Málheildin var kynnt fyrir formanni Félags íslenskra bókaútgefenda og tillaga til breytingar hefðbundins samnings milli útgefenda og höfundarréttarhafa verður rædd í sambandi við yfirstandandi endurskoðun samningsins. Breytingin mun bæta ákvæði í samninginn, sem segir að höfundarréttarhafi leyfi texta þeirra til notkunar í málheild undir sambærilegum skilmálum eins og voru samþykktur fyrir Markaða íslenska málheild (MÍM), útgefin 2012.

Fundur við Rithöfundasamband Íslands hefur verið bókaður. Hann verður haldinn í enda janúar 2020.

Ef umræddar breytingar eru samþykktar af öllum aðilum mun það gera okkur mögulegt að safna textum úr bókum án þess að eyða tíma í leyfismál, en þetta mun aðeins gilda fyrir bækur sem er ekki búið að gefa út. Fyrir þegar útgefnar bækur munum við þurfa að hafa samband við höfunda til að fá leyfi til að bæta textum þeirra í málheildina. En ef samkomulag er til staðar við útgefendur mun það auðvelda aðgengi að textunum sjálfum þegar höfundarréttarhafar samþykkja leyfið.

Á fundi við Félags íslenskra bókaútgefenda var ákveðið að stefna á lok viðræðna fyrir sumarið.

Þegar við höfum lokið viðræðum vonumst við til að ná samkomulagi við fjölda höfundarréttarhafa og safna textum sem innihalda alls 20 milljón orð 2020.

Flokkar	Fjöldi tóka (mat)
Viðbætur við núverandi veitur	100.000.000
Nýjar vefsíður	10.000.000
Blogg	5.000.000
Vefþing	80.000.000
Útgefnar bækur	20.000.000
Heildarfjöldi viðbættra tóka 2020	**215.000.000**

Tafla 2: Hugsanlegar viðbætur í Risamálheild 2020

G4 – Beygingarlýsing íslensks nútímamáls fyrir máltækni (BÍN)

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

Varða 1 - lýsing

Úrtak hluta úr BÍN, sem máltækniútgáfa gagnagrunnsins skal innihalda. Upplýsingar sem skal innleiða, með viðeigandi merkingum og lýsigögnum, tilbúnar.

Afurðir

Þeir hlutar BÍN sem eru aðgengilegir til niðurhals hafa verið skilgreindir:

1. Orð
    a. Grunnupplýsingar
        i. Nefnimyndir
        ii. Kenni
        iii. Orðflokkur
        iv. Kyn nafnorða
        v. BÍN flokkur [flokkun eftir viðfangsefni: almennur orðaforði; nöfn; önnur sérnöfn, íþróttir, matur, o.s.frv.]
    b. Aðrar upplýsingar (aukalegar)
        i. Flokkur: 0-5 [lýsir notagildi fyrir réttritun og villuleit, aldur, o.s.frv.]
        ii. Málfar [formlegt, óformlegt, sjaldgæft, staðbundið mál, úrelt, niðrandi, o.s.frv.]

iii. Málfræði [fyrir notkun í máltækni: stafsetning (með millivísunum), tökuorð, aðeins notað í ákveðnum orðtökum (takmörkuð notkun), o.s.frv.]
iv. Framburður [óreglulegur framburður eins og -f- í *sófi*, -g- í *sígaretta*, o.s.frv.]

c. Afmörkun of upplýsinga forræðishyggju (til aðstoðar við skrif)
i. Millivísanir
ii. Aukaflettur [þegar hefðbundin nefnimynd er víkjandi fyrir öðru formi orðs, t.d. þegar eintala er óalgeng en fleirtala algeng. Dæmi: *aðför*/*aðfarir*]
iii. Orð sem hafa verið handvirkt borin saman við viðtekið mál er merkt BÍN-kjarni

2. Beygingarform
a. Grunnupplýsingar
i. Beygingarform
ii. Mark [sama og í eldri útgáfu gagnagrunnsins]
b. Greining
i. Flokkur (1. b. i.)
ii. Málfar (1. b. ii.)
iii. Málfræði, flokkun afbrigða: ráðandi og víkjandi beygingarform, takmörk notkunar, aldur, o.s.frv.

Skýrsla

Ný útgáfa BÍN gagnagrunnsins hefur verið sett upp. Hún inniheldur fleiri skýringar en áður og meiri greiningu gagna. Notendur munu geta halað niður gögnum sem innihalda miklu meiri upplýsingar en áður.

Núna erum við að vinna í því að veita aðgang að gögnum samkvæmt skemanu sem lýst er hér að ofan. Niðurhalanleg gögn verða tímastimpluð, með nýrri útgáfu í boði til niðurhals fjórum sinnum á ári, til að gera rannsakendum kleift að endurskapa eldri niðurstöður. Þar til nú hefur bara verið ein útgáfa í boði til niðurhals, sem er dagleg demba nýjustu gagna.

Verið er að setja upp síðu til niðurhals þar sem notendur geta valið hvaða upplýsingar úr BÍN þeir vilja hala niður og hvaða tímastimpluðu útgáfu. Gögnin verða aðgengileg á .csv formi

G6 – Framburðarorðabók

Skýrsla: Grammatek ehf.
Aðilar: Grammatek ehf.

Varða 1 - lýsing

Umritunarreglur fyrir hvert framburðartilbrigði skilgreindar. Orðalisti prófunarsetts tilbúinn. Uppbygging gagnagrunns orðabókar þróuð. Frumútgáfa vinnslutóls tilbúin.

Afurðir

1. Lýsing íslenska hljóðkerfisins, afbrigði framburðar, og hljóðritun.
2. Listi hljóðritunartákna sem nota skal við umritun íslensku, IPA og X-SAMPA útgáfa.
3. Hljóðritunarreglur fyrir valin afbrigði framburðar: hefðbundinn framburð, norðlenskan framburð, norðausturlenskan framburð, og sunnlenskan framburð. Einnig, val n-stæða stafa sem hægt er að bera fram mismunandi eftir stíl/málfari.
4. Tveir orðalistar, prófunarsett og þróunarsett til sjálfvirkra rittákns-til-fónems (g2p) umritana.
5. Lýsing á uppbyggingu gagnagrunns orðabókar.
6. Frumútgáfa vinnslutóls orðabókar, PEDI.

Allar afurðir er hægt að finna á https://github.com/grammatek/pedi

Skýrsla

Lýsing á íslensku hljóðkerfi, framburðartilbrigðum, og hljóðritun skýrir eiginleika íslenskrar hljóðfræði og hljóðkerfisfræði. Einnig er átta mállýskum lýst, og einni almennri breytingu í framburði, þ.e. eitt afbrigði sem talað er einkum af fólki yfir 60, en annað nú talað af miklum meirihluta mælenda. Einn hluti sýnir hvernig samhljóðaklasar eru oft einfaldaðir í framburði. Alls eru 94 n-stæður stafa skráðar, sem margar hverjar er hægt að bera fram skýrt, og einnig á einfaldaðri hátt.

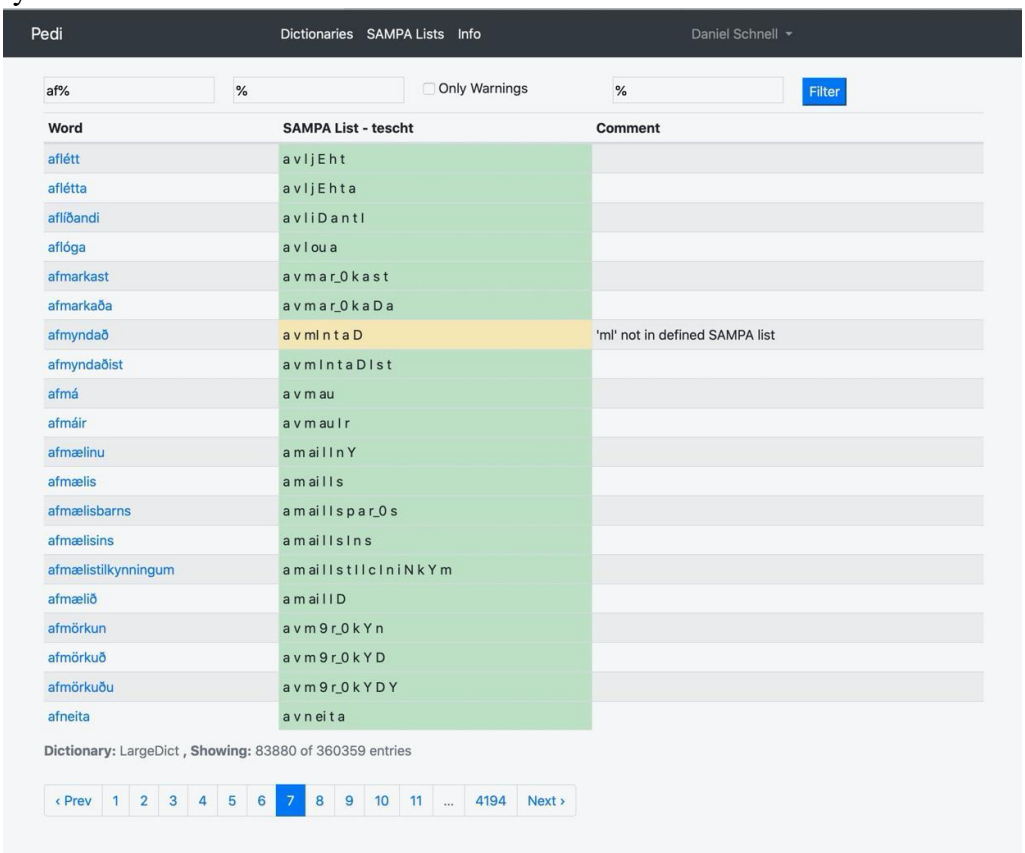
Byggt á þessari lýsingu voru þrjú afbrigði valin til innleiðingar í fyrstu útgáfu framburðarorðabókar: 1) norðlenska afbrigðið með *fráblæstri p/t/k í orðum,* 2) norðausturlenska afbrigðið með *röddduðum nefkveðnum og hliðmæltum hljóðum á undan lokhljóðum,* 3) sunnlenska afbrigðið með *hv-framburði.* Önnur afbrigði eru metin síður mikilvæg fyrir núverandi tilgang orðabókarinnar. Í áframhaldandi þróun verður hinsvegar auðvelt að bæta fleiri afbrigðum við fyrir sértækan tilgang. Fyrir hvert þessara fjögurra valinna afbrigða (hefðbundinn framburður + þrjú afbrigði) verður einfölduð og óeinfölduð útgáfa of 37 n-stæða stafa, sem leiða af sér átta útgáfur orðabókarinnar.

Við höfum skrifað umritunarreglur fyrir hvert tilbrigði. Þessa reglur munu gegna hlutverki grunngildis fyrir orðabókina, þ.e. allar ákvarðanir umritana munu vera afdráttarlausar, og þannig gera allri þróun í framtíð kleift að finna alltaf skýringar á ákveðnum umritunum í þessu reglusafni, eða kerfisbundið breyta ákveðnum umritunum. Reglusafn þetta verður nýtt á næstu vikum við handvirka umritun, og þannig munum við geta fínstillt reglusafnið eftir þörfum.
Með völdum afbrigðum framburðar, og lista 97 n-stæða stafa sem hafa/geta haft einfaldaðan framburð, var listi 217 n-stæða valinn til sköpunar prófunar- og þróunarsetts til sjálfvirkrar

rittákns-til-fónems (g2p) umritana. Fyrir hverja n-stæðu geta allt að fimm orð verið dregin úr í núverandi útgáfu framburðarorðabókar fyrir hvert sett (prófunar- og þróunar-). 36 n-stæður eru ekki geymdar í orðabókinni, eða eru mjög sjaldgæfar, og því var Beygingarlýsing íslensks nútímamáls (BÍN) notuð til skoðunar á þessum n-stæðum. Alls innihalda prófunar- og þróunarsettin um 1.000 orða hvert.

Núverandi útgáfa Íslensku framburðarorðabókarinnar inniheldur aðeins tvo dálka: orð (framsetning rittákns) og hljóðritun. Þetta þýðir að þrátt fyrir að nokkur afbrigði mállýska séu þegar í orðabókinni, eru þau ekki merkt og ekki nýtt kerfisbundið. Yfirstandandi verkefni stefnir að því að merkja hverja færslu með mikilvægum upplýsingum til sjálfvirkar vinnslu orðabókarinnar. Eftirfarandi skýringar (dálkar í gagnagrunninum) verða innleiddar: orð (framsetning rittákns), IPA-umritun, X-SAMPA umritun, mállýska, málfæri, tungumál, málfræðilegur orðflokkur. Eftirfarandi búlskar (e. *boolean*) merkingar verða einnig innleiddar, þ.e. gildi þessara merkinga er *satt* eða *ósatt*: samsett orð, seinni hluti samsetningar, fyrri hluti samsetningar, yfirfarið (handvirk staðfesting umritana hefur verið gerð).

Starfhæf frumútgáfa vinnslutóls orðabókar hefur verið innleidd. Frumútgáfan gefur kost á handvirkri yfirferð, leiðréttingu og viðbætum við umritanir prófunar- og þróunarsetta, með notkun X-SAMPA stafrófsins. Tólið er vefforrit þar sem notandi skráir sig inn og fær aðgang að öllum orðabókum skráðum á viðkomani notanda. Tólið merkir sjálfvirkt færslur sem innihalda merki umritana utan skilgreinds setts X-SAMPA merkja fyrir íslensku og veitir möguleika á mismunandi aðferðum til röðunar og síunar. Skjáskot PEDI-frumútgáfunnar er sýnd að neðan:



G7 – Íslenskt orðanet

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

Varða 1 - lýsing

Gagnaform skilgreint fyrir Íslenskt orðanet (e. Icelandic Wordnet).

Afurðir

Við höfum skilgreint gagnaform sem við leggjum til að verði notað fyrir Íslenskt orðanet (e. Icelandic Wordnet).

Skýrsla

Formáli
Íslenskt orðanet (skrifað af Jóni Hilmari Jónssyni, http://ordanet.is; lýst í gagnrýni af Rögnvaldsson 2018) er merkingarfræðilegur gagnagrunnur fyrir íslensku. Það sýnir ýmis tengsl milli orða á merkingarfræðilegum grunni. Meðan orðanet Princeton (e. Princeton WordNet) (Miller 1995, Fellbaum 1998; https://wordnet.princeton.edu/) er undirstaða orðaneta, er Íslenskt orðanet frábrugðið því á nokkra vegu.

Skilgreining hnúta og vensla í Íslensku orðaneti
Hnútar (e. *nodes*)
Grundvallaratriði til skilgreiningar hnúta orða byrjar með orðflokki. Íslenskt orðanet notar sambærilega orðaflokka eins og Princeton WordNet, eins og nafnorð, sagnorð, lýsingarorð, atviksorð. Ólíkt WordNet fyrir ensku, þarf íslenska að fást við beygingarmyndir fyrir öll þrjú málfræðileg kyn (karlkyn, kvenkyn, og hvorugkyn), og fjögur málfræðileg föll (nefnifall, þolfall, þágufall, eignarfall). Líkt og enska, hefur íslenska tvö gildi fyrir tölu (eintölu, fleirtölu).

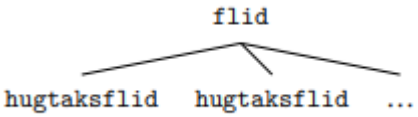
Undirflokkun upplýsingar með fall viðfangs skipta líka máli fyrir íslensku. Til dæmis, sögnin *snúa* með viðfang í þágufalli þýðir ‘snúa <e-u>’, venda eða breyta um átt, en ef viðfang er í þolfalli þýðir það ‘snúa <e-ð>’, t.d. flétta eða vinda <e-ð>. Þessar upplýsingar varða líka orðanet fyrir íslensku.

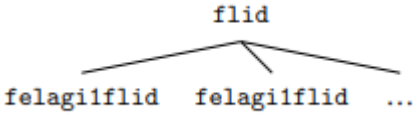
Vensl

Að neðan er yfirlit yfir hugtök (eins og **`flid`**) og tengsl notuð í Íslensku orðaneti, sýnd með viðeigandi tréi (fl (fletta) = lemma; id (kennimerki) = ID).

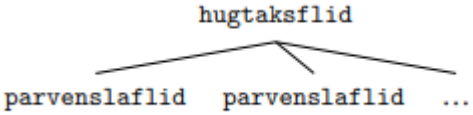
`flid`	Kennimerki nefnimyndar
`hugtaksflid`	Kennimerki hugtaks
`felagi1flid`	Kennimerki nefnimyndar félaga
`parvenslaflid`	Kennimerki parvensla
`frumflid`	Kennimerki frumnefnimyndar
`skyldflid`	Kennimerki skyldra nefnimynda
`tengiflid`	Kennimerki tengdra nefnimynda
`grannflid`	Kennimerki grannnefnimynda

1. Vensl eftir hugtaki
   a. Nefnimyndir eftir hugtaki

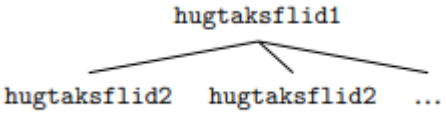




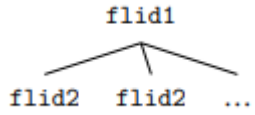
   b. Skyldar nefnimyndir eftir pörum



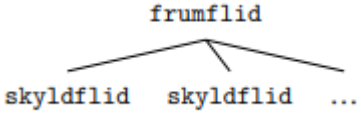
   c. Skyld hugtök



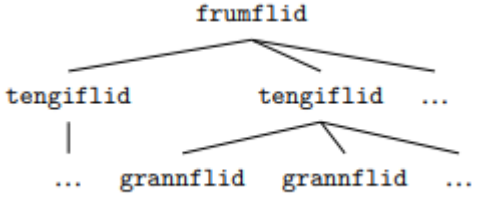
2. Parvensl



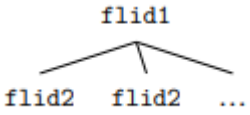
3. Skyldar nefnimyndir



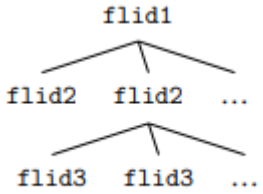
4. Grannheiti



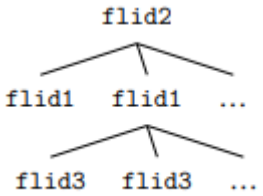
5. Úrskurðuð vensl
   a. Samheiti
   b. Andheiti



6. Samsetningar
   a. Fyrri hluti samsetningar



   b. Seinni hluti samsetningar



Ákvörðun fyrir gagnalíkan Íslensks orðanets

Upprunalega WordNet formið sem hannað er fyrir ensku er einkaleyfisvarið. Þó það hafi verið notað í mörgum verkefnum, telst það slæmur valkostur fyrir núverandi kynslóð orðaneta í þróun. Það þarfnast tóla sem aðeins eru hönnuð fyrir það gagnaform, og nýlegar þróanir sýna framtíð samofinna orðaneta sem eru tengd hvoru öðru og fyrir hugbúnað sem var ekki til við smíði upprunalega orðanetsins (e. WordNet), eins og Merkingarvefinn (e. Semantic Web).

Ef litið er yfir þau fjölmörgu orðanet og regnhlífarverkefni (BalkaNet, EuroWordNet, Open Multilingual Wordnet) sem eru nú til, er XML ráðandi form. Eftir því sem orðanetum hefur fjölgað hefur verið vilji til tryggingar samnýtingar og samvinnu orðaneta. Með það að leiðarljósi reyna Alþjóðleg samtök orðaneta, GWA(e. Global WordNet Association), að búa til samhæfða staðla fyrir orðanet, og stuðla að innleiðingu orðaneta í Alþjóðlegt orðanet (e. Global WordNet Grid) óháð tungumálum. Það eru þrjár skemur studdar af GWA (https://globalwordnet.github.io/schemas/) sem innihalda XML (e. Extensible Markup Language), JSON-LD (e. JavaScript Object Notation for Linked Data), og RDF (e. Resource Description Framework). Bæði JSON-LD og RDF skemur eru stækkaðar með lemon (e. The Lexicon Model for Ontologies, https://lemon-model.net/), líkani sem hannað er fyrir orðasöfn véllæsilegum orðabókum.

Svo virðist sem vaxandi staðall orðaneta sé RDF formið stækkað með lemon. Þar sem RDF er samþykkt form GWA, er skrifað í XML (málið er kallað RDF/XML), og styður samþættingu við Merkingarvefinn (e. *Semantic Web*) (McCrae, Fellbaum & Cimiano 2020), sýnir það fram á marga kosti til byggingar orðanets með þarfir tungumáls eins og íslensku. Viðbót lemon veitir einingum leyfisveitandi kótun hljóðfræði, orð- og merkingarfræðileg vensl, uppbyggingu setningarliða, og fallbeygingu, sem er mjög mikilvæg fyrir íslensku.

Staðfestir og umbreytir milli forma sem studd eru af GWA eru veittir á http://server1.nlp.insight-centre.org/gwn-converter/. Í ljósi þess stuðnings sem GWA veitir, og tól þeirra, eins og staðfestir og umbreytir, er augljósa lágmarkið form sem fylgir viðmiðunarreglum þess. Utan GWA, eru bæði RDF og JSON-LD formin studd innan LLOD (e. Linguistic Linked Open Data; http://www.linguistic-lod.org/) standard.

Heimildir
Rögnvaldsson, Eiríkur. 2018. Íslenskt orðanet: A treasure for writers og word lovers. *LexicoNordica* 25:313–328.
Fellbaum, Christiane (ed.). 1998. *Wordnet: An electronic lexical database*. Cambridge, MA: The MIT Press.
McCrae, John, Christiane Fellbaum & Philipp Cimiano. 2020. *Publishing og Linking WordNet using lemon and RDF*. Semantic Computing Group, University of Bielefeld & Princeton University. https://jmccrae.github.io/wn-rdf-paper/slides/slides.html (2. janúar, 2020).
Miller, George A. 1995. Wordnet: A lexical database for English. *Communications of the ACM* 38(11):39–41.

G10 - Gullstaðall

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

Varða 1 - lýsing

Endurskoðuð orðhlutafræðileg markaskrá fyrir íslensku útgefin. Sjálfvirkri umbreytingu marka Gullstaðals lokið.

Afurðir

Við höfum endurskoðað orðhlutafræðilega markaskrá fyrir íslensku, sjá viðauka A. Sjálfvirkri umbreytingu marka Gullstaðals lokið og handvirkt yfirlit og leiðréttingar byrjaðar fyrir möguleg margræð mörk.

Skýrsla

Formáli
Við höfum endurskoðað orðhlutafræðilega markaskrá fyrir íslensku; sjá viðauka A fyrir nýja markaskrá. Breytingarnar voru til þess gerðar að auðvelda það að forðast ósamræmi í mörkun og gera innviði markaskrárinnar samræmdari (ákvarðanir um mörk eiga að vera utan samhengis að mestu). Minnkun ósamræmis ætti að stuðla að fækkun villna sem sjálfvirkir markarar gera.

Samræmi innviða: Mörk utan samhengis
Þegar mörk eru valin fyrir orðflokka og málfræðileg einkenni er það oft svo að formið sjálft yfirgnæfir raunverulega notkun eða þýðingu. Dæmi um þetta er að á forminu neitun + nafnháttur sagnar hefur sömu þýðingu og sögnin í boðhætti + neitun.

(1) Ekki gera þetta.
not do.inf this

‘Don’t do this.’

(2) Gerðu þetta ekki.
do.imp+you this not
‘Don’t do this’

Gera í (1) hefur engu að síður alltaf verið markað sem nafnháttur, þó það sé notað á svipaðan hátt sem sögn í boðhætti -- það er formið sem ákvarðar markið. Annað dæmi eru “óbeygð” lýsingarorð: þau sýna ekki beygingu en með samhengi má færa rök fyrir því að þau hafi fall, kyn og tölu. Hinsvegar er það formið sem ákvarðar að þeir *einmana* í (3) er óbeygt (en *gamall* beygist greinilega með falli, kyni og tölu, eins og sjá má í (4)):

(3) a. nom.masc.sg. **einmana** maður
‘lonely man’
b. acc.masc.sg. **einmana** mann
c. dat.masc.sg. **einmana** manni
d. gen.masc.sg **einmana** manns

(4) a. nom.masc.sg. **gamall** maður
‘lonely man’
b. acc.masc.sg. **gamlan** mann
c. dat.masc.sg. **gömlum** manni
d. gen.masc.sg **gamals** manns

Með þetta í huga er örlítið sérkennilegt að lýsingarháttur þátíðar er markaður með tilliti til sagnarinnar sem velur hann; þolmyndir, venjulega valdar af sögninni *vera*, hafa verið markaðar sem fallbeygður lýsingarháttur en lýsingarhættir þátíðar valdir með *hafa* markaðir sem sagnbót. Hér er oft um að ræða samfall, eins og sýnt er í (5)-(6), þar sem form sagnbótar er það sama og form lýsingarháttar þátíðar í hvorugkyni eintölu.

(5) Hún **hefur** aldrei **borðað** fisk.
she has never eaten.N.SG fish

(6) Af hverju **var grænmetið** ekki **borðað**?
why was vegetable.the.NOM.N.SG not eaten

Í nýju markaskránni, öfugt við þá gömlu, er enginn munur gerður milli þessara tveggja; samkvæmt nýju markaskránni eru allir slíkir lýsingarhættir markaðir sem lýsingarhættir þátíðar.

Önnur breyting sem við höfum gert í nýju markaskránni varðar forsetningar. Í gömlu markaskránni voru forsetningar markaðar sem a(tviksorð) + fallið sem þær stýra. Upplýsingarnar um að þær stýri falli eru mikilvægar (til að skilgreina þær frá eiginlegum atviksorðum) en upplýsingarnar um hvaða fall þær veita eru óþarfar, þar sem það má lesa úr the object itself (bera má þetta saman við sagnir sem einnig stýra falli, en við setjum ekki upplýsingar um fall í mark sagnarinnar). Í nýju markaskránni mörkum við allar forsetningar á sama hátt, sem atviksorð sem stýra falli.

Nýjir flokkar

Í gömlu markaskránni voru styttingar markaðar sem “as” (“a” fyrir atviksorð, “s” fyrir styttingu). Þetta getur verið nokkuð óheppilegt, sérstaklega þar sem styttingar geta til dæmis staðið fyrir nafnliði. Slíkar styttingar, eða skammstafanir eru sýndar í (7).

(7) ESB = Evrópusambandið
EU = European Union

Því höfum við fært styttingar orðsambanda (7) og orða (8) í nýjan flokk, k.

(8) lögg. = löggiltur; hdl. = héraðsdómslögmaður
certified; solicitor

Að lokum var enginn sérstakur flokkur fyrir greinarmerki og önnur merki í gömlu markaskránni. Hvert greinarmerki var til dæmis markað svona (“.” sem “.”, “;” sem “;”, “...” sem “.”, “.”, “.”, o.s.frv.). Ný markaskrá setur greinarmerki í flokk p og önnur merki (eins og fyrir stærðfræði og tjámyndir (e. emojis)) í flokk m.

Viðauki A

Markaskrá MIM-GULL 2.0		
Dálkur	**Formdeild**	**Greiningartákn-greiningaratriði**
1	Orðflokkur	**n-nafnorð**
2	Kyn	**k**-karlkyn, **v**-kvenkyn, **h**-hvorugkyn
3	Tala	**e**-eintala, **f**-fleirtala
4	Fall	**n**-nefnifall, **o**-þolfall, **þ**-þágufall, **e**-eignarfall
5	Greinir	**g**-með viðskeyttum greini
6	Sérnöfn	**s**-sérnafn
1	Orðflokkur	**l-lýsingarorð**
2	Kyn	**k**-karlkyn, **v**-kvenkyn, **h**-hvorugkyn
3	Tala	**e**-eintala, **f**-fleirtala
4	Fall	**n**-nefnifall, **o**-þolfall, **þ**-þágufall, **e**-eignarfall
5	Beyging	**s**-sterk beyging, **v**-veik beyging, **o**-óbeygt
6	Stig	**f**-frumstig, **m**-miðstig, **e**-efstastig
1	Orðflokkur	**f-fornafn**
2	Flokkur	**a**-ábendingarfornafn, **b**-óákveðið ábendingarfornafn, **e**-eignarfornafn, **o**-óákveðið fornafn, **p**-persónufornafn, **s**-spurnarfornafn, **t**-tilvísunarfornafn
3	Kyn/Persóna	**k**-karlkyn, **v**-kvenkyn, **h**-hvorugkyn/**1**-1. pers., **2**-2. pers.
4	Tala	**e**-eintala, **f**-fleirtala
5	Fall	**n**-nefnifall, **o**-þolfall, **þ**-þágufall, **e**-eignarfall
1	Orðflokkur	**g-greinir**
2	Kyn	**k**-karlkyn, **v**-kvenkyn, **h**-hvorugkyn
3	Tala	**e**-eintala, **f**-fleirtala
4	Fall	**n**-nefnifall, **o**-þolfall, **þ**-þágufall, **e**-eignarfall
1	Orðflokkur	**t-töluorð**
2	Flokkur	**f**-frumtala, **a**-ártöl og fleiri óbeygjanlegar tölur, **p**-prósentutölur, **o**-fjöldatölur framan við hundrað og þúsund
3	Kyn	**k**-karlkyn, **v**-kvenkyn, **h**-hvorugkyn
4	Tala	**e**-eintala, **f**-fleirtala
5	Fall	**n**-nefnifall, **o**-þolfall, **þ**-þágufall, **e**-eignarfall
1	Orðflokkur	**s-sögn** (þó ekki lýsingarháttur þátíðar)
2	Háttur	**n**-nafnh., **b**-boðh., **f**-framsöguh., **v**-viðtengingarh., **l**-lýsingarh. nútíðar
3	Mynd	**g**-germynd, **m**-miðmynd
4	Persóna	**1**-1. persóna, **2**-2. persóna, **3**-3. persóna
5	Tala	**e**-eintala, **f**-fleirtala
6	Tíð	**n**-nútíð, **þ**-þátíð
1	Orðflokkur	**s-sögn** (lýsingarháttur þátíðar)
2	Háttur	**þ**-lýsingarháttur þátíðar
3	Mynd	**g**-germynd, **m**-miðmynd
4	Kyn	**k**-karlkyn, **v**-kvenkyn, **h**-hvorugkyn
5	Tala	**e**-eintala, **f**-fleirtala
6	Fall	**n**-nefnifall, **o**-þolfall
1	Orðflokkur	**a-atviksorð**
2	Flokkur/Fallstjórn	**a**-stýrir ekki falli, **f**-stýrir falli, **u**-upphrópun
3	Stig	**m**-miðstig, **e**-efsta stig
1	Orðflokkur	**c-samtenging**
2	Flokkur	**n**-nafnháttarmerki, **t**-tilvísunartenging
1	Orðflokkur	**k-skammstöfun**
2	Flokkur	**s**-skammstöfun, **t**-stytting
1	Orðflokkur	**e**-erlent orð
1	Orðflokkur	**x**-ógreint orð
1	Orðflokkur	**v**-tölvupóstfang, veffang
1	Orðflokkur	**p-greinarmerki**
2	Flokkur	**l**-lok setningar, **k**-komma, **g**-gæsalappir, **a**-önnur greinarmerki
1	Orðflokkur	**m-tákn**

I1 - Mörkunartól

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

Varða 1 - lýsing

Tilgangur
Að velja og stilla handvirkt mörkunartól sem getur verið notað við að skrifa skýringar fyrir ýmsar íslenskar málheildir með málfræðilegum og merkingarlegur upplýsingum. Þessi verkþáttur felst í því að skilgreina þarfir, velja og stilla viðeigandi opinn (e. open-source) hugbúnað, og semja leiðbeiningar fyrir markara.

V1
Skýrsla um þarfir. Opinn (e. open-source) hugbúnaður fyrir handvirka mörkun. Leiðbeiningar fyrir skýrendur.

Endanlegar afurðir

Fjögur tól eða tólasett hafa verið valin til handvirkrar mörkunar innan verkefnisins:

Annotald: viðbætt kvísl núverandi Annotald tólsins, fyrir djúpþáttaðar málheildir
UD Annotatrix: núverandi tól til mörkunar djúpþáttaðra málheildir
Tólasett fyrir mörkun villumálheildar: í þróun, sjá lýsingu neðar
Tól til málfræðilegrar mörkunar (e. POS-tagging): viðmót fyrir handvirka leiðréttingu málfræðilegrar mörkunar málheilda, þróað innan verkefnisins.

Skýrsla

Eftirfarandi þarfir voru skoðaðar við val/þróun tóla til handvirkrar mörkunar:

Almennar þarfir

- Sköpun nýrrar markaskrár fyrir ný verkefni verður að vera möguleg
- Tólin þurfa að vera opin (e. open-source)
- Tólin þurfa að vera frí
- Tólin þurfa að hafa vefviðmót og virka í öllum vöfrum og stýrikerfum
- Tólin þurfa að ráða við lýsigögn
- Það þarf að vera auðvelt að aðlaga úttak sjálfvirkra markara í snið mörkunartóls
- Það þarf að vera auðvelt að nota úttak mörkunartóls fyrir inntak annarra tóla
- Tólin þurfa að marka vensl milli markaðra eininga
- Tólin þurfa að vera auðvelt í notkun og mega ekki þarfnast mikillar tölvuþekkingar
- Tólin þurfa að vera öflug og áreiðanleg

Aðrar þarfir

- Villumálheildir
    - Formið verður að taka á villum á víðu sviði
    - Best væri ef fleiri en ein möguleg leiðrétting er fundin
- Gullstaðall og Risamálheild
    - Á TEI-formi eins og er, en ætti að minnsta kosti að geta umbreytt mörkuðum gögnum á það form
- Þáttun
    - Ætti að ráða við flókin mörk og nefnimyndir, orðhlutafræðilega greiningu, þýðingu styttinga og skammstafana, o.s.frv.

Verkefnið breyttist töluvert eftir upprunalegar skilgreiningar varða.

Upprunalegar afurðir voru skilgreiningar þarfa, opinn (e. open-source) hugbúnaður fundinn sem hægt væri að aðlaga þeim þörfum, og leiðbeiningar fyrir.

Upprunalega var gerður listi mögulegra tóla fyrir mörg mismunandi verk ásamt kostum og göllum þeirra. Mismunandi skemu voru útlistuð. Þarfir mismunandi hluta verkefnis SÍM (og umfram það) sem þurfa mörkunartól voru greindar. Þessi gögn eru aðgengileg í sameiginlegu vinnurými allra aðila verkefnisins.

Eftir að þessi gögn voru kynnt fyrir öðrum verkhlutum SÍM verkefnisins varð ljóst að ekki þótti raunhæft að nota eitt mörkunartól fyrir alla hluta. Eftir mikla umhugsun var ákveðið að hver verkhluti myndi velja mörkunartól við hæfi, helst af útbúnum lista tóla. Niðurstaðan er sú að afurðirnar hafa breyst í lýsingu aðgengilegra og hæfra opinna (e. open-source) tóla, ásamt lýsingu á ferli mörkunar fyrir verkhluta sem hafa byrjað.

Valinn og þróaður hugbúnaður
Þáttuð mörkun
Annotald tólið hefur verið valið til mörkunar djúpþáttaðra málheilda. Viðbætt kvísl þess er aðgengileg á GitHub. Leiðbeiningar markara (á íslensku) hafa verið samdar, sem hluti af öðru verkefni (utan sviðs Almannaróms/SÍM). Markaðar málheildir sem eru búnar til í því verkefni verða notaðar að hluta til prófunar þáttara í I5 með þáttunarskemu sem þróuð er þar. Á þeim tíma munu leiðbeiningar markara vera aðlagaðar að valinni skemu. Engar aðrar markaðar málheildir í þessum tilgangi eru á sjóndeildarhringnum eins og er.

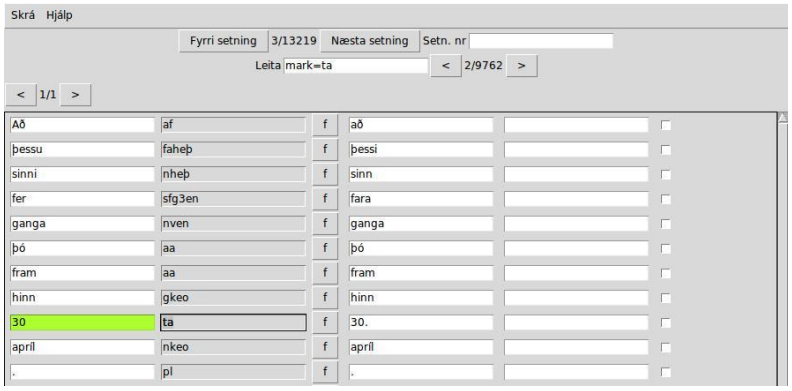
Fyrir möguleg venslaþáttunarverkefni innan SÍM verkefnisins, ætti að nota UD Annotatrix. Engar sérstakar leiðbeiningar fyrir markara eru til, þar sem slík vinna innan verkefnisins byrjar ekki fyrr en miklu seinna.

Mörkun villumálheildar
Til þess að hámarka notkun núverandi yfirlestrargagna og veita kunnulegt viðmót fyrir markara er verkefni villumörkunar hannað til að leyfa Microsoft Word skjöl með breytingarakningum (e. *Track Changes*) sem fyrsta skref í mörkunarferli. Við drögum út upprunalega og leiðrétta útgáfu textans, röðum þeim með sérhæfðri röðunarskriftu, skrifaðri í Python sem nýtir núverandi (og hægt að skipta út) algrím röðunar, sem er nú í BioPython. Við flokkum svo fjölda algengra villna sjálfvirkt með annarri Python skriftu og við erum eins og er að þróa einfalt vefviðmót til endurskoðunar sjálfvirkrar mörkunar og einnig til mörkunar

villna sem þarf að skoða handvirkt. Viðmótið tekur sum gildi hönnunar sem gerðu Annotald kleift að flýta hraða handvirkrar mörkunar í uppbyggingu Sögulega íslenska trjábankanum (IcePaHC), einkum þá notandavænu kosti fyrir markara, sem fást frá litlu og afmörkuðu setti flýtivísa sem þarfnast fárra áslátta og handhreyfinga. Mörkunartólið mun leyfa vistun skráa málheildarinnar í, að minnsta kosti, TEI formi. Aðlögun þessara aðferða fyrir núverandi verk er í vinnslu á meðan mörkun villna stendur yfir. Heilt safn tóla sem notuð verða í þessu ferli verða gerð aðgengileg á Github ásamt leiðbeiningum fyrir notkun.

Málfræðileg mörkun

Sjálfstætt mörkunartól var hannað fyrir leiðréttingu málfræðimarkana og lemmun með Python. Það tekur sem inntak 1) skrá með einu orði í línu, setningar aðskildar með línubili. Í venjulegum ham er hverju orði fylgt með málfræðilegu marki, nefnimynd, athugasemdum og Boole-breytu sem sýnir hvort orðið hefur verið leiðrétt, allt aðskilið með dálkbili; 2) textaskrá með markaskránni; 3) textaskrá með lista orða og mögulegra málfræðilegra marka fyrir hvert þeirra (sem hægt er að búa til úr Gullstaðli) og 4) skrá sem inniheldur python orðabók með mörkum, þar sem hvert mark er aðskilið í stafi og þar sem þýðing hvers stafs er skrifuð (t.d. n = ‘nafnorð’). Það er mögulegt að bæta dálkum í skrá 1 sem innihalda önnur möguleg mörk, t.d. fengin frá öðrum mörkurum. Svo mun einn eða fleiri hnappar birtast bakvið hvert orð þar sem hægt er að smella á til að breyta gildi svæðis sem inniheldur aðal markið.



Tólið sýnir mest 25 orð á síðu. Ef setningin er lengri er hægt að fletta fram og til baka með hnöppum. Einnig er hægt að stökkva frá einu orði yfir í annað. En til að geta stokkið í næstu setningu, er hægt að nota leitaraðgerð. Hægt er að leita að ákveðnum mörkum, nefnimyndum, athugasemdum eða orðum sem hafa verið leiðrétt eða ekki. Tólið færir svo yfir í næstu viðkomandi setningu þar sem viðkomandi orð eru uppljómuð.

Í hvert skipti sem fært er úr einni setningu eða setningarhluta yfir í annan sannreynir tólið mörkin og uppljómar þau ef þau tilheyra ekki markaskránni.

Við leiðréttingu marka eru tveir hamir mögulegir. Í venjulegum ham birtist fellilisti með öllum mögulegum mörkum fyrir hvert orð. Í skriftarham má skrifa mark meðan tólið sannreynir hvern staf og sýnir í fellilista mögulega stafi fyrir hvert skref í myndun marks.

Til þess að notandi þurfi ekki að styðjast um of við músina (sem getur hægt á leiðréttingarferlinu), getur lyklaborðið verið notað til flestra aðgerða.

Tólið er aðgengilegt á GitHub (https://github.com/stofnun-arna-magnussonar/pos-tagging-tool). Nánari leiðbeiningar er að finna í skránni hjalp.html.

I3 - Textatilreiðari

Skýrsla: Miðeind
Aðilar: Miðeind

Varða 1 - lýsing

Tilgangur
Að skila tilreiddum textum til annarra máltæknitóla.
Tilreiddur texti sýnir skil setninga og sýnir hvern tóka aðskilinn með bili.

V1

- Afkastageta greiningar setninga valins tilreiðara skjöluð, með notkun blandaðs prófunarsetts, sem inniheldur jaðartilvik sem eru erfið fyrir greiningu setninga.
- Skilgreining tóka fyrir grunnútgáfu tilreiðara tilbúin, tilreiðari aðlagaður að þeirri skilgreiningu.
- Prófanir á hefðbundnum texta án jaðartilvika ætti að ná nálægt 100% nákvæmni.
- Listi jaðartilvika tilbúinn.
- Listi mismunandi stillinga fyrir þarfir annarra verkhluta verkefnisins tilbúinn, en eftir því sem greining óalgengra tilvika þróast gætu þessir listar breyst.
- Grunnútgáfa tilreiðara afhend eins og mögulegt er samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins; en hugsanlegt er að þær verði ekki fullkláraðar fyrir V1, þar sem leiðbeiningarnar verða aðeins kláraðar um miðjan desember 2019.

Afurðir

Öllum afurðum fyrir vörðuna var náð og tilreiðari fyrir íslensku er útgefinn til CLARIN:
https://repository.clarin.is/repository/xmlui/handle/20.500.12537/11

Skýrsla

Afkastageta greiningar setninga valins tilreiðara skjöluð, með blönduðu prófunarsetti, þ.e. sem inniheldur erfið jaðartilvik fyrir greiningu setninga.
Þegar setning endar með greinarmerki og í kjölfar kemur lágstafstóki, greinir tilreiðari nýja setningu ef seinasta orð fyrri setningar var gilt orð. Ef það var það ekki, er það talið skammstöfun, þekkt eða óþekkt. Í slíkum tilfellum getur tilreiðarinn ekki greint nýja setningu, ef nýja setningin byrjar með tölu eða öðru merki utan stafrófs eða orði án hástafs í byrjun.

Stórt prófunarsett var þróað fyrir tilreiðarann. Það er aðgengilegt á GitHub.
Innbyggðar prófanir eru í skránni `test/test_tokenizer.py` og er hægt að keyra með `pytest`.

Skráin `test/toktest_large.txt` inniheldur prófunarsett 13.075 lína. Línurnar prófa greiningu setninga, greiningu tóka og flokkun tóka. Til greiningar inniheldur `test/toktest_large_gold_perfect.txt` væntanlegt úttak fullkomnar grunntilreiðslu,

og `test/toktest_large_gold_acceptable.txt` inniheldur núverandi úttak grunntilreiðslu. Skráin `test/toktest_large_gold_acceptable.txt` er sama og núverandi úttak. Munurinn milli hennar og `test/toktest_large_gold_perfect.txt` er vel skjalaður og inniheldur aðeins óalgeng tilvik.

Skilgreining tóka fyrir grunnútgáfu tilreiðara, og tilreiðari aðlagaður til að fylgja þeirri skilgreiningu.
Í samstarfi við teymið sem vinnur við G10 - Gullstaðal, endurskoðun og stöðlun markaskrár fyrir íslensku - er skilgreiningu tóka komið á fót.
Skráin `test/toktest_large_gold_perfect.txt` var búin til samkvæmt þeirri skilgreiningu og tilreiðarinn aðlagaður til að fylgja þeirri skilgreiningu eins vel og hægt er.
Skráin `test/toktest_large_gold_acceptable.txt` inniheldur núverandi úttak grunntilreiðslu, þannig að munurinn milli þeirra sýnir hvað er mögulegt og hvað ekki.

Aðal markmiðið var að skipta tókum með hvítbili, með fáum undantekningum. Tókar sem innihalda tákn gjaldmiðla eða styttingar mælieininga (dl, cm) eru meðhöndlaðir eins og sést í grunntilreiðslu en samsettir í fulltilreiðslu. Þetta þýðir að ef það er ekkert hvítbil ($12, 12cm) er tókinn látinn vera, og ef það er hvítbil ($ 12, 12 cm), er það meðhöndlað sem tveir tókar í grunntilreiðslu.
Breytileiki talna flækir málin; tölur með formerkjum (+12, −49) og tölur sem innihalda almenn brot (49¼, 3½) eru hafðar saman en gjaldmiðilstáknum eða styttingum mælieininga er skipt í sundur (−49¼cm → −49¼ cm).

Rómverskar tölur eru látnar vera, sem og blandaðar raðtölur (2svar, 3ji). Háum tölum með brotum er skipt í tvo tóka (5.000¼ → 5.000 ¼).

Einföldum brotum er skipt í þrjá tóka nema þegar hægt er að túlka þau sem dagsetningar (5/49 → 5 / 49; 5/4 → 5/4).

Mörg greinarmerki sem enda setningu eru höfð saman sem einn tóki. Úrfellingarmerki enda setningu ef næsti tóki er hástafur, eftir meginreglu.

Tókar sem innihalda bandstrik eru meðhöndlaðir sem einn tóki (Suður-Hnappadalur, marg-ítreka).

Varðandi flóknari tóka, eru margar gerðir greindar, svo sem gild símanúmer, kennitölur, efnasambönd, Twitter notendanöfn, myllumerki, vefföng, netföng, segðir tíma, o.s.frv. Aðrir tókar eru meðhöndlaðir samkvæmt reglum fyrir þekkta flókna tóka.

Prófanir á hefðbundnum texta án jaðartilvika ætti að ná nálægt 100% nákvæmni.
Skráin `test/toktest_normal.txt` inniheldur runutexta frá nýlegum fréttagreinum, inniheldur 100 setningar og engin jaðartilvik. Stafsetningarvillur voru leiðréttar. Skráin var notuð til prófunar grunntilreiðslu, bæði greiningu setninga og greiningu tóka.

Gullstaðal fyrir þá skrá má finna í skránni `test/toktest_normal_gold_expected.txt`. Tilreiðarinn nær 100% nákvæmni fyrir þessa prófun.

Listi óalgengra tilvika tilbúinn.

Listi óalgengra tilvika var búinn til með því sem við höfum rekist á við þróun, erfið tilvik frá verkhluta skilgreininga tóka, og tillögum frá öðrum rannsakendum í SÍM sem nota tilreiðara fyrir verkefni sín. Þessi óalgengu tilvik voru síðan notuð til sköpunar `test/toktest_large.txt`.
Skráin `test/Overview.txt` (aðeins á íslensku) inniheldur lýsingu prófunarsetts, með númerum lína fyrir hvern hluta í bæði `test/toktest_large.txt` og `test/toktest_large_gold_acceptable.txt`, og mark sem lýsir hvað er verið að prófa í hverjum hluta (greining setninga, greining tóka og flokkun tóka).
Hún inniheldur einnig lýsingu fullkominnar grunntilreiðslu fyrir hvern hluta, ásættanlegri tilreiðingu og núverandi hegðun. Sem slík er lýsingin greining á hvaða jaðartilvik tilreiðarinn ræður við og hver ekki.

Listi mismunandi stillinga fyrir þarfir annarra verkhluta verkefnisins tilbúinn, en eftir því sem greining jaðartilvika þróast gætu þessir listar breyst.
Listi mögulegra stillinga var þróaður á sama hátt og listi jaðartilvika. Eftir því sem verkefnið þróaðist voru tvær mismunandi stillingar tilreiðingar innleiddar, sem kallast fulltilreiðing og grunntilreiðing. Þessar tvær stillingar voru taldar þjóna flestum verkefnum.

Grunntilreiðing gefur einfaldlega hverja setningu sem streng (eða sem línu texta í úttaksskrá), þar sem hver einstakur tóki er aðskilinn með bili. Þetta hæfir vélþýðingu og mörkun, til dæmis. Verkefni I4, mörkun, í verkáætluninni notar grunntilreiðingu sem inntak, ásamt fulltilreiðingu til meðhöndlunar flóknari tóka.

Fulltilreiðing gefur hverjum einstökum hlut tóka sem hefur verið markaður með tegund tóka og öðrum upplýsingum dregnum úr tókanum, til dæmis *(ár, mánuður, dagur)* færslu í tilfelli tóka dagsetninga, og þenslu styttinga. Þetta hæfir þátturum og málfræðilegum leiðrétturum, til dæmis. Verkefni I5, þáttari, og M6 og M7, Málrýnir, í verkáætluninni notar fulltilreiðingu.

Grunnútgáfa tilreiðara afhend eins og mögulegt er samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins, í núverandi mynd. Lagfæringar verða gerðar þegar og ef viðmið breytast í lokaútgáfu.
Tilreiðarinn hefur verið afhendur samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins og er aðgengilegur almenningi á GitHub. Textakóðun er UTF-8 í gegn. Kóði og athugasemdir eru á ensku.
Hugbúnaðurinn er undir MIT leyfi, sem leyfir meira en Apache 2.0 sem var tilskilið sem lágmarksleyfi í verkáætluninni.
Tilkynning höfundarrétta og leyfa er innifalin í hausathugasemd í öllum frumskrám.
Travis CI er notað fyrir endurtekna prófun.
Hugbúnaðurinn er skrifaður í Python 3.6 (en er þó samhæfur eldri Python útgáfum, m.a. 2.7).
Hugbúnaðinum hefur verið skilað til Clarin og er aðgengilegur hér.
Hugbúnaðinn má setja upp sem Python pakka með pip, og er hýstur í Python Package Index (PyPi) sem 'tokenizer'.
Skipanalínutól er fáanlegt.
Skjöl varðandi uppsetningu og notkun eru á GitHub.

I4 - Málfræðilegur markari

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

Varða 1

Málfræðilegir markarar til prófana hafa verið valdir og prófanir til undirbúnings hafa verið gerðar.

Svið texta til frekari prófana hafa verið valin og viðeigandi textar valdir.

Afurðir

Valdir markarar

Fjórir markarar hafa verið valdir til að vinna með. ABL-tagger hefur ennþá gefið bestu niðurstöður og mun því vera aðal markari. Þrír aðrir markarar verða hluti af prófunum okkar: TriTagger, IceTagger og IceStagger.

Prófanir til undirbúnings

Lokaprófanir verða ekki gerðar fyrr en Gullstaðall, MÍM-Gull, hefur verið breytt yfir í nýju markaskránna (sjá G10). Markararnir fjórir hafa verið settir upp á tölvum í Stofnun Árna Magnússonar í íslenskum fræðum og hafa verið notaðir til mörkunar núverandi Gullstaðals. Python skriftur hafa verið skrifaðar til útbúnings prófunar- og þjálfunarsetta úr Gullstaðli. Skriftur hafa einnig verið skrifaðar sem munu umbreyta öllum orðabókum sem markararnir nota yfir í nýja markaskrá.

Textar til prófana

Völdum sviðum og ferlinu við val setninga úr málheildinni fyrir prófunarsett er lýst í skýrslu fyrir vörðu 1 í G1, Risamálheild.

I5 - Þáttari

Skýrsla: Miðeind
Aðilar: Miðeind

Varða 1 - lýsing

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

Ákvæði
Tilreiðari
Málfræðilegur markari

V1

- Greining þarfa tilbúin.
- Mat aðferða fyrir Greyni og IceNLP tilbúið.
- Stöðluð þáttunarskema skilgreind.

Afurðir

Öllum afurðum fyrir vörðuna var skilað.

Greining þarfa tilbúin.
Hugbúnaður og tól sem þarfnast þáttunar voru yfirfarin og þarfir þeirra greindar. Ekki eru öll tól í boði fyrir íslensku eins og er, og í þeim tilvikum voru hugsanlegar þarfir greindar. Þessar þarfir upplýstu skilgreiningu þáttunarskemu.

Mat aðferða fyrir Greynir og IceNLP tilbúin.
Til þess að bera saman og meta Greynir (sem er liðgerðarþáttari) og IceNLP (sem er hlutaþáttari), þurfti að skilgreina matsaðferðir.
Aðferðirnar lýsa því sem prófa þarf, hvað innihald prófunarmálheilda ætti að vera, grunnlína til samanburðar, og aðferðir til mats mismunandi þáttunarskema.

Stöðluð þáttunarskema skilgreind.
Valin þáttunarskemu, eða tilbúin setningafræðileg mörkunarskemu, fyrir fullþáttun er undir sterkum áhrifum frá IcePaHC, verkefni Sögulega íslenska trjábankans sem var klárað 2011. Hún er því undir óbeinum áhrifum frá skema Penn trjáabankans (e. Penn Treebank).

Til þess að gera skemun eins almenn og hægt er, innihelda þau ekki sértækar markanir og notar aðeins óumdeilda flokka. Þetta stuðlar að því að halda skemanu frá skorðun í einstök fræðisvið og gerir mögulegt að bera saman Greyni og IceNLP. Einnig var jöfn skipting þarfa sjálfvirkrar þáttunar og útdráttar þýðinga (setningafræði á móti merkingarfræði).

Skema fyrir venslaþáttun verður skilgreint er slíkur þáttari verður þróaður sem hluti af verkáætlun á seinna stigi.

Til að skilgreina þáttunarskemu, lýsum við mengi endamarka (e. t*erminals*), mengi markaðra setningahluta og undirgerðum þeirra og málfræðilegum hlutverkum. Að auki gefum við nánari leiðbeiningar þar sem þessar almennu leiðbeiningar eru óljósar.

Skýrsla

Greining þarfa
Eftirfarandi máltækniverkefni geta stutt sig, mismikið, við þáttað inntak.

Málfræðileiðrétting getur notað þáttað inntak til greiningar málfræðivillna. Setningar sem ekki er hægt að þátta samkvæmt málfræðireglum eru líklega rangar. Einnig er hægt að þróa sérhæfðar reglur málfræðivillna til greiningar rangra mynstra og að lokum til aðstoðar í tillögum leiðréttinga. Fullþáttun er líklega hentugust til slíkra verka en venslagreining gæti einnig virkað. Grunnþáttun er síður hentug þar sem hún framkallar minnstar upplýsingar fyrir tilgang þessa verks.

Greining á samvísunum og endurvísunum er ekki tiltæk eins og er fyrir íslensku. Það myndi líklegast þarfnast greiningar á aðalliðum, sérstaklega nafnliðum. Það þyrfti einnig þáttunarskemu sem sýndi spor og færslur til að geta tengt nafnliði rétt, sérstaklega tóma liði.

Merkingarfræðileg greining er tiltæk fyrir íslensku í takmarkaðri mynd. Hún þyrfti minnst greiningu nafnliða og tímaliða. Hún þyrfti fullþáttun eða venslaþáttun. Hún þyrfti aðalliði og sértæka undirflokka nafnliða, svo sem entity, persónu, fyrirtæki, heimilisfang, mælingu, gjaldmiðil, o.s.frv. Nánari undirflokkar fara eftir völdu sviði. Ef merkingarfræðileg viðhenging eru notuð þarf CFG. Það er einnig mögulegt að vinna með fullklárаða þáttunartréið.

Úrdráttur upplýsinga (IE) er tiltækur fyrir íslensku í takmarkaðri mynd. Hann þyrfti aðalliði og sértæka undirflokka nafnliða eins og tekið er fram fyrir merkingarfræðilega greiningu. Hann þyrfti einnig tímaliði og möguleika til að tengja þær við sagnir/atburði.

Upplýsingaheimt (IR) er ekki tiltæk fyrir íslensku eins og er. Þarfir hennar eru svipaðar og IE.

Spurningasvörunarkerfi (QA) eru tiltæk fyrir íslensku. Þau þurfa algenga aðalliði og sértæka undirflokka nafnliða eins og tekið er fram fyrir merkingarfræðilega greiningu. Þau þarfnast einnig phrases denoting temporal expressions og möguleika til að tengja them við sagnir/atburði. Að sjálfsögðu er þáttun spurninga einnig nauðsynleg.

Talþjónar eru tiltækir fyrir íslensku í takmarkaðri mynd. Þeir myndu þurfa aðalliði og að minnsta kosti mark og/eða upplýsingar nefnimyndar. Sértæka liði þyrfti alla vega til að styðja við kerfi sem byggjast á römmum (e. *frame-based*).

Vélþýðingar má bæta með greiningu nafnliða sem ætti ekki að þýða, sem gæti falist í þáttun texta.

Málvísindaleg rannsókn myndi þurfa spor og færslur, aðalliði en líka marga sértæka undirflokka. Þetta myndi þarfnast fullþáttunar, venslaþáttunar og/eða grunnþáttunar.

Mat aðferða fyrir Greynir og IceNLP

Hvað þarf að prófa?
Prófunarmálheildir fyrir fullþáttun þarf til að bera saman Greynir og IceNLP. Prófa þarf bæði liði og mörk (undirflokka og málfræðilegt hlutverk).

Prófunarmálheildir
Trjáabankar eru oftast um 5.000 setningar. Við ættum að stefna að prófunarmálheild 1.000 setninga. Prófunarmálheildin ætti helst að innihalda texta frá síðastliðnum tíu árum. Til þess að halda hámarks jafnvægi og jafnframt einblína á vel skrifaða texta, er eftirfarandi skipting áformuð:

- Fréttatextar - 250 setningar
- Þingræðutextar - 250 setningar
- Vísindavefurinn - 250 setningar
- Bókmenntatextar úr útgefnum bókum - 250 setningar

Grunnlína
Prófunarmálheildin og IcePaHC geta verið notaðar til þjálfunar fullþáttara sem grunnlínu. Úttakinu væri svo hægt að varpa á grunna IceNLP skemað.

Matsaðferðir
Helst skal framkvæma bæði innri og ytri mælingar.
Fyrir innri mælingar er evalb/PARSEVAL algengasta tólið. Handmarkaði hluti prófunarmálheilda verður notaður til mælingar nákvæmni, heimt og F-gildi. Fjöldi yfirliggjandi sviga myndi einnig vera mældur.
LeafAncestor ætti einnig að nota, þar sem það gefur annarskonar en þó gagnlegar mælingar.
Ytri mælingar verða gerðar með leiðréttingu málfræði og vélþýðingum. EPE 2018 notaði viðburðagreiningu, neitunargreiningu og skoðanagreiningu, en þar sem merkingarfræðileg greining er aðeins tiltæk fyrir íslensku í takmarkaðri mynd er það hugsanlega ekki mögulegt.

Stutt yfirlit staðlaðrar þáttunarskemu

Mörk
Mörkin hlíta endurskoðaðri markaskrá í verkefni G10 verkáætlunarinnar.

n - Nafnorð. Sjálfvalinn setningarliður er NP.
s - Sagnorð. Sjálfvalinn setningarliður er VP.
a - Atviksorð. Sjálfvalinn setningarliður er ADVP. Þegar stjórnar falli eftirfarandi nafnliðar, er sjálfvalinn setningarliður PP.
l - Lýsingarorð. Sjálfvalinn setningarliður er ADJP. Þegar við höfum lýsingarorð sem sagnfyllingu án nafnorðs, er sjálfvalinn setningarliður NP.
f - Fornöfn. For ákvæðisfornöfn, er sjálfvalinn setningarliður ADJP. Fyrir sagnfyllingar fornöfn, er sjálfvalinn setningarliður NP. Við greinum undirflokka fornafna; ábendingarfornöfn, óákveðin ábendingarfornöfn, eignarfornöfn, óákveðin fornöfn, persónufornöfn og spurnarfornöfn.
g - Greinir. Sjálfvalinn setningarliður er NP.
t - Töluorð. Sjálfvalinn setningarliður er ADJP. Við greinum undirflokka töluorða; töluorð sem merkja ár og önnur óbeygjanleg töluorð, og prósentur.
c - Samtengingar. Sjálfvalinn setningarliður er C. Athugið að nafnháttarmerki fellur undir þennan flokk.
k – Skammstafanir. Sjálfvalinn setningarliður er NP.

e - Erlend orð. Sjálfvalinn setningarliður er NP.
v - Vefföng og netföng. Sjálfvalinn setningarliður er NP.
p - Greinarmerki (e. punctuation).
m – Merki/tákn. Öll merki sem passa ekki í aðra flokka.

Setningarliðir
NP - Nafnliðir. Við greinum eftirfarandi undirflokka/málfræðileg hlutverk:
NP-SUBJ er frumlag.
NP-OBJ er beint andlag.
NP-IOBJ er óbeint andlag.
NP-PRD er sagnfylling eða andlag tengisagnar.
NP-POSS er nafnliður í eignarfalli.
NP án marks getur staðið innan PPs (forsetningarliða).
Hefðbundinn NP inniheldur nafnorð, mögulega aðra liði eða stakt fornafn. NPs geta innihaldið ADJPs og NP-POSS.

VP - Sagnliðir. Hjálparsagnir eru merktar sem VP-AUX.
Hefðbundinn VP inniheldur sögn og andlög hennar.

ADJP - Lýsingarorðsliðir.
Hefðbundinn ADJP inniheldur lýsingarorð, töluorð eða ákvæðisfornafn. Hann getur einnig innihaldið ADVP sem breytir ADJP. Þessir liðir koma fyrir samkvæmt reglum um orðaröð í íslensku.

ADVP - Atviksliðir.
Hefðbundinn ADVP inniheldur eitt atviksorð eða atviksorðasamband. Hann getur innihaldið annan ADVP, sem breytir hinu atviksorðinu

PP - Forsetningaliðir
Hefðbundinn PP inniheldur forsetningu og nafnliðinn, sem hún stýrir.

IP – Beygingarliðir. Nafnháttarbeygingarliðir eru merktir IP-INF. Hefðbundinn IP inniheldur heilan lið.

CP – Aukasetningar. Eftirfarandi undirflokkar eru greindir:
CP-ADV er atviksliður
CP-QUE er óbein spurning
CP-REL er tilvísunarsetning
CP-THT er fallsetning
Hefðbundinn CP inniheldur samtengingu og IP.

S - Aðalsetningar
Hefðbundinn S inniheldur IP.

S0 - Rót.

V2 – Samhliða málheild

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

Varða 1 – lýsing

Pípu fyrir gagnaöflun, samröðun og síun komið á fót. Viðeigandi skjöl/kóði/skriftur útgefin. Skjöl/skjalahlutar valdir fyrir þróunar-/prófunarsett.

Afurðir

Öllum afurðum var skilað.
Skjöl/skjalahlutar valdir fyrir þróunar-/prófunarsett
Skjöl/skjalahlutar valdir fyrir þróunar-/prófunarsett eru aðgengilegir hér:
https://drive.google.com/drive/folders/1yGZFClbwqyoOE6mSI9rFwFE2YAa_uuul?usp=sharing

Skýrsla

Fyrsta útgáfa málheildarinnar var gefin út um haust 2019. Við köllum hana *ParIce*. Málheildin er byggð upp af samhliða textum af fjölbreyttum uppruna, sá stærsti verandi skjöl EES, mest tilskipanir og reglugerðir. Öll textasöfn eru talin upp í Töflu 1.

Búið er að koma upp ferli til þess að safna textum og undirbúa fyrir samhliða málheild. Þar sem gögnum fjölgar hægt munum við ekki keyra þessa aðferð sjálfvirkt heldur handvirkt árlega.

Pípa **fyrir gagnaöflun, samröðun og síun**

Mismunandi aðferðir voru notaðar fyrir mismunandi uppruna gagna, bæði varðandi söfnun og vinnslu gagnanna. Því þarf pípa að vera aðlöguð fyrir hvern uppruna. Það felur í sér forvinnslu texta, sem í sumum tilvikum felur í sér útdrátt texta úr PDF skjölum, sköpun lýsigagna, röðun texta og síun gallaðra hluta.

Við notum eftirfarandi tól fyrir þessi mismunandi verk:

Söfnun og PDF-útdráttur

Skriftur leita eftir nýjum gögnum í þeim upprunum sem stækka reglulega: EES skjölum, fréttatilkynningum frá Stjörnustöð Evrópulanda á suðurhveli (ESO) og þýðingarhlutum frá Tatoeba. Þessar skriftur verða keyrðar árlega eða þegar pípa fyrir samröðun og síun hefur verið endurskoðuð.

Önnur textasöfn stækka ekki, nema gögn frá EMA. Við notum gögn söfnuðum af TILDE en áætlum að reyna að safna og raða gögnum beint frá EMA og bera niðurstöður okkar saman

við þær frá TILDE.

Í tilvikum þar sem texta þurfti að draga úr PDF-skrá, notuðum við útdráttartólið frá Acrobat.

Samröðun
LFaligner er notaður fyrir sjálfvirka samröðun setninga. Setningaskipti þess (split-sentences.sh) þurfti að breyta lítillega til að taka betur á íslenskum tilvitnunum og til að geta skipt setningum á semíkommum og tvípunktum. LFaligner virkar betur með notkun orðabókar. Við bjuggum til orðabók til bráðabirgða með yfir 12 þúsund lemmum með skrapsöfnun úr hinni íslensku Wiktionary.

Síun
Sérstök pípa til síunar samraðaðra gagna var búin til til þess að meta gæði röðunar og sía út allar línur sem þóttu slæmar. Fyrir utan þau tilvik þar sem engin þörf var á mati vegna eðlis gagnanna (Tatoeba, Biblían), eða þegar röðunin var þegar síuð (KDE4, Ubuntu), voru allar raðanir sendar í gegnum þessa pípu.

Við grófa yfirferð samraðaðra texta kom í ljós að slæmar raðanir komu oftast í böggum. Ef villa kemur í einni röðun er hún líkleg til að hafa áhrif á eina eða fleiri setningar sem koma eftir þar sem LFaligner getur tekið nokkrar línur að finna rétta leið aftur. Sem hluta af síunarferli þýddum við enska texta hvers setningapars yfir á íslensku með OpenNMT og bárum þýðingar svo saman við íslensku setninguna og gáfum hverju pari einkunn eftir þeim samanburði. Öllum böggum setningapara sem fengu slæma einkunn er eytt, en þeim pörum sem ekki er eytt er bætt í orðabókina. Þessi skref þýðinga, einkunnagjafar, síunar og stækkunar orðabókarinnar eru endurtekin nokkrum sinnum.

2. Næstu skref
Við höfum lokið gerð fyrstu útgáfu samhliða málheildar fyrir tungumálaparið enska-íslenska. Til þess að bæta gæði þeirra gagna munum við þurfa að endurskoða aðferðir þessa ferlis og það verk er fyrir vörður 2 og 3. Sérstök áhersla verður á samröðun og síun. LFAligner notar skiptingu setninga og reiðir sig á hunalign fyrir pörun setninga. Við breyttum reglum skiptinga setninga fyrir íslensku, en með útgáfu næsta íslenska tilreiðara (verkefni I3), munum við vilja skipta setningum í aðskildu ferli. Hunalign notar Gale-Church algrímið og orðabók, utanaðkomandi eða afleiddri frá gögnunum. Síðan hunalign og LF Aligner voru gefin út hefur fjöldi annarra aðferða verið notaður. BLEUAlign notar vélþýðingar texta og BLEU sem samanburðareinkunn til að finna áreiðanlega röðun, sem er svo hægt að nota sem markstiklur. Bilin milli markstikla eru svo fyllt með reglum byggðum á BLEU og lengd. Nýlega hafa orðagreypingar þvert á tungumál verið notaðar til útreiknings fjarlægðar milli samsvarana í mismunandi tungumálum, þetta hefur verið innleitt í bivec (https://github.com/lmthang/bivec) og vecmap (https://github.com/artetxem/vecmap). Ný aðferð notar margmála BERT til að finna merkingarleg líkindi milli tungumála og notar það til síunar, með því að sía út minnst líkar setningar, og sýnir einnig að það virkar vel til þess að finna villur í samsvörunum þýðinga (Lo og Simard, 2019). Þó þjálfun góðs BERT fyrir okkur til notkunar í þessu verkefni verði líklega of dýrt, bæði í tíma og öðrum aðföngum, munum við gera tilraunir með aðrar aðferðir, BLEUAlign og notkun orðagreypinga með bivec og vecmap og reyna að aðstoða samröðunarferlið.
ParSentExtract (Grégoire og Langlais, 2018) notar BiRNNs til útdráttar samhliða setninga. Við munum einnig vilja prófa kerfi þeirra á okkar gögn.

Á seinustu árum hefur verið áhersla á síun samhliða gagna, einkum í tveimur samnýttum

verkum hjá WMT, 2018 og 2019. Auk tilrauna með bivec, vecmap og BLEU einkunnir frá máltækni, munum við gera tilraunir með Zipporah aðferðina (Xu og Koehn, 2017), sem blandar reiprennsli og nægð (e. *fluency and adequacy features*) til einkunnargjafar setningapara. BiCleaner er notað í ParaCrawl verkefninu, við munum gera tilraunir með það.

Við munum meta niðurstöður með handvirkt mörkuðum þróunar-/prófunarsettum sem eru í vinnslu.

Heimildir

Francis Grégoire og Philippe Langlais. 2018. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1442–1453, Santa Fe, New Mexico, BNA. Association for Computational Linguistics.

Chi-kiu Lo og Michel Simard. 2019. Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data. Í tíðindum 23. Conference on Computational Natural Language Learning (CoNLL), bls. 206–215, Hong Kong, Kína. Association for Computational Linguistics.

Hainan Xu og Philipp Koehn. 2017. Zipporah: a fast og scalable data cleaning system for noisy web-crawled parallel corpora. Í tíðindum Conference on Empirical Methods in Natural Language Processing frá 2017, bls. 2945–2950, Kaupmannahöfn, Danmörk. Association for Computational Linguistics.

Skjöl/hlutar valdir fyrir þróunar-/prófunarsett.

Lítill hluti málheildarinnar var valinn til þróunar-/prófunarsetta. Við völdum gögn frá þrem undirmálheildum. Heil EES skjöl sem verða röðuð handvirkt og 10.000 setningar frá OpenSubtitles og undirmálheild Lyfjastofnun Evrópu (EMA) sem verða merktar í einum af fjórum flokkum, sem sýna hvort hlutarnir eru rétt raðaðir og þýddir eða ekki. Lítið þróunar-/prófunarsett 2000 setninga frá hverri undirmálheild verður merkt fyrir september 2020 með Keops kerfi, vinna við þetta byrjaði í desember 2019. Hráir ómerktir hlutar eru aðgengilegir hér: https://drive.google.com/drive/folders/1yGZFClbwqyoOE6mSI9rFwFE2YAa_uuul?usp=sharing

Undirmálheildir	Ósíað	Síað
Biblían	32.964	32.964
Bækur	16,976	12.416
EES	2,093,803	1.701.172
EMA	420,297	404.333
ESO	12,900	12.633
KDE4	137,724	49.912
OpenSubtitles	1,620,037	1.305.827
Íslendingasögur	43,113	17.597

Tölfræði Íslands	2,481	2.288
Tatoeba	8,263	8.263
Ubuntu	11,025	10.572
	4,399,583	3.557.977

Tafla 1: Undirmálheildir og fjöldi þýddra hluta í ParIce 1.0, fyrir og eftir síun.

V2b – Val og síun gagna fyrir bakþýðingar

Skýrsla: Miðeind
Aðilar: Árnastofnun / Miðeind / Háskólinn í Reykjavík - HR

Tilgangur
Að afhenda stóra samraðaða tilbúna samhliða málheild til beggja átta þýðinga milli íslensku og ensku, til viðbótarþjálfunar vélrænna þýðingarkerfa. Þessi málheild verður byggð á bakþýddum textum, til að vega á móti skorti á samhliða málheilda ensku-íslensku.

Varða 1 - lýsing

- Valdir textar og þjálfunargögn fyrir bakþýðingar og þýðingarlíkön.
- Skýrsla um svið í málheildinni og stærð hvers þeirra.

Afurðir

Öllum afurðum var skilað. Valdir textar og þjálfunargögn eru talin upp að neðan og stærð þeirra kynnt í eftirfarandi undirkafla.

Einmála ensk málheild
Eftirfarandi gagnasöfn voru valin vegna þess að þau eru oft notuð[1] fyrir vélrænt nám, meðal annars fyrir vélþýðingar, tungumálalíkön og líkön til málskilnings. Gagnasöfnin hafa svipuð svið og íslenska Risamálheildin[2], sem er eina stóra eintyngda málheildin fyrir íslensku, og tvítyngda ParIce málheildin[3]. Að lokum eru gögnin misleit í efni, orðanotkun, setningaformi og svo framvegis.

- Acquis - samsafn almennra réttinda og skylda sem eru bindandi í öllum löndum Evrópusambandsins
- Blogg - *The Blog Authorship Corpus*[4]
- BookCorpus - safn bóka á ensku[5]
- Europarl - *European Parliament Proceedings Parallel Corpus*[6]
- NewsCommentary[7] - *News Commentary Parallel Corpus*

[1] Sjá til dæmis gögnin fyrir News Translation Task frá ráðstefnu Workshop on Statistical Machine Translation 2019 http://www.statmt.org/wmt19/translation-task.html
[2] http://malfong.is/?pg=rmh
[3] http://malfong.is/?pg=samhlida
[4] J. Schler et. al. 2006, https://u.cs.biu.ac.il/~koppel/BlogCorpus.htm
[5] Zhu et al. 2015, https://github.com/soskek/bookcorpus
[6] http://www.statmt.org/europarl/
[7] http://www.casmacat.eu/corpus/news-commentary.html

- NewsCrawl [8]- Safn frétta
- PubMed Central[9]
- UN Corpus[10] - Samhliða málheild Sameinuðu Þjóðanna (e. United Nations Parallel Corpus)
- Wikipedia[11] - Síað safn texta frá Wikipedia

Einmála íslensk málheild
Fyrir íslensku er *Risamálheild: A Very Large Icelandic Text Corpus*[12] notuð. Hún er stærsta safn slíkra texta fyrir íslensku. Samkvæmt efniságripi hennar er hún "*a corpus containing more than one billion running words from mostly contemporary texts*". Sjá skýrslu hér að neðan fyrir samantekt innihalds hennar.

Skýrsla

Samhliða málheildir - Tilvitnun
Tvímála ParIce málheildin[13] mun vera notuð til þjálfunar fyrstu þýðingarlíkana. Íslenski hlutinn inniheldur um það bil 47 milljón orð. Hún er samansafn texta, aðallega frá EES, OpenSubtitles og Lyfjastofnun Evrópu (e. EMA).

Einmála ensk málheild

Málheild	Stærð á disk	Stærð í orðum	Svið
Acquis	469MB	34.6m	Lög
Blogg	805MB	140m	Blogg
BookCorpus	4.3GB	985m	Blandaðar bókmenntir
Europarl	322MB	54m	EU þing
NewsComm	53MB	8.5m	Fréttir
NewsCrawl	2.7GB	485m	Fréttir
PubMed	35.7GB (þjappað)	Óþekkt	Vísindi
UN Corpus	3GB	335m	Stjórnmál
Wikipedia	58GB	2.9b	Alfræðilegt

[8] http://data.statmt.org/news-crawl/
[9] https://www.ncbi.nlm.nih.gov/pmc/
[10] https://conferences.unite.un.org/UNCorpus/
[11] https://blog.lateral.io/2015/06/the-unknown-perils-of-mining-wikipedia/
[12] Steingrímsson et al., 2018. http://www.lrec-conf.org/proceedings/lrec2018/summaries/746.html
[13] http://malfong.is/?pg=samhlida

Einmála íslensk málheild
Verkefnið mun nota eina einmála íslenska málheild, *Risamálheildina*, samsett úr mörgum undirmálheildum frá mörgum sviðum.
Stærð *Risamálheildar* (rmh) er 3GB á diski.

Svið	Stærð í orðum
Fréttir	797m
Þingfundir	211m
Dómsúrskurðir	93m
Bókmenntir	5.7m
Innslegnar fréttir	54m
Íþróttafréttir	47m
Reglugerðir	26m
Blogg	10.5m
Greinar	10.8m
Lífstíll	4m

Stærð tilbúinnar málheildar
Í fyrstu mun einmála íslensku málheildinni og hluta einmála ensku málheildanna að ofan vera safnað saman. Þessi tvö einmála söfn verða þýdd yfir á tungumál hvors annars með grunnlínulíkani þýðinga, sem gefa mun stóra tilbúna („gervi“) samhliða málheild.

Rannsóknir hafa sýnt að stærð þessarar tilbúnu málheildar (fyrir hvora átt þýðinga) ætti að vera minnst jafn stór (í fjölda orða eða setninga) núverandi samhliða gögnum. Hinsvegar hafa nýlegar aðferðir sýnt framfarir þegar líkanið sér aðeins hvert tilbúið setningapar einu sinni við þjálfun. Umrædd þjálfun Transformer líkans þarfnast 76 mynstrasafna (endurtekningar yfir allt gagnasafn þjálfunar) til að ná samleitni. Hámarksstærð tilbúins gagnasafn gæti því verið á bilinu 2m til 200m setninga vegna markvissrar sýnatöku (fyrir óalgeng orð) eða síun.

V3 – Grunnlína í vélrænum þýðingum

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

Varða 1 - lýsing

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

V1

- Þýðingarsvið og textar fyrir þróun og prufusett hafa verið valin.
- Öll þrjú kerfin, tölfræðikerfið MT (SMT) Moses, og tauganetin MT (NMT) tvö Tensor2Tensor og OpenNMT, hafa verið yfirfærð/þjálfuð til að þýða frá ensku yfir á íslensku, með því að nota samhliða málheild ParIce.
- Forritaskil í gegnum HTTPS REST eða sambærilegar þjónustur fyrir öll kerfi eru í gangi, aðgengileg með curl, wget eða sambærilegum tólum (ekkert vef- eða myndrænt viðmót).
- Kóði og skjöl eru afhend samkvæmt leiðbeiningum um hugbúnaðarþróun sem verkefnið setur.
- Stutt skýrsla um þjálfunartíma, keyrslutíma og BLEU einkunnir fylgja skjölum.

Afurðir

Öllum afurðum var skilað.

Þýðingarsvið

Þau þýðingarsvið sem valin voru til að meta kerfið eru skjátextar frá Opensubtitles, lyfseðlar frá Lyfjastofnun Evrópu (EMA) og lög tengd EES. Skjátextar og lyfseðlar voru valdir af handahófi frá ParIce á meðan ákveðin heil skjöl voru valin úr skjalasafni EES.[14] Gagnasafn þessara matsgagna var ekki skoðað handvirkt og telst til undirbúnings, eingöngu til að bera saman kerfin til bráðabirgða. Gagnasafnið innihélt einhverjar villur, og til að halda slíkum villum í lágmarki var gagnasafnið síað frekar.[15] Mat á gögnum fyrir öll kerfi í skýrslunni er byggt á þessu síaða gagnasafni.

Kerfi

Eftirfarandi aðferðir voru notaðar, bæði til þýðinga frá ensku yfir í íslensku og öfugt.

[14] 2200 setningapör voru valin frá Opensubtitles og einnig frá EMA. Um það bil 3200 setningapör voru valin frá EES. Gagnasafnið inniheldur 7645 setningapör.
[15] 1930 setningapör frá EES, 1963 frá EMA og 2059 frá Opensubtitles. Samtals 5952 setningapör.

- Transformer (Með Tensor2Tensor[16])
- Tvíátta LSTM (BI-LSTM, með OpenNMT[17])
- Moses[18]

Öll þrjú líkönin voru þróuð með málheild ParIce, samhliða málheild sem inniheldur texta frá fréttum, þingfundum, dómum, bókmenntum, reglugerðum, bloggum og blaðagreinum.

API

Skjöl (v3) varðandi REST API[19] Google þýðingarvélarinnar voru notuð sem ráðfærandi tilgreiningar fyrir HTTP REST API og öll líkön nota JSON fyrir inntak og úttak.[20] Þjálfuðu kerfin eru aðgengileg í gegnum þessa endapunkta með eftirfarandi formi á inntaki.

```
{
  "contents": [
    string
  ],
  "sourceLanguageCode": string,
  "targetLanguageCode": string,
  "model": string,
}
```

- Transformer og BI-LSTM - https://velthyding.mideind.is/nn/translate.api
    - Valkostir fyrir líkönin eru “transformer” fyrir transformer byggðar þýðingar og “bilstm” fyrir tvíátta LSTM líkan fyrir þýðingu.
    - Dæmi: curl 'https://velthyding.mideind.is/nn/translate.api' --data '{"model":"transformer","contents":["Góðan daginn frú mín góð."],"sourceLanguageCode":"is","targetLanguageCode":"en"}' -H "Content-Type: application/json"
- Moses: http://nlp.cs.ru.is/moses/translateText
    - Dæmi: curl -d '{"contents":["Góðan dag frú mín góð."],"sourceLanguageCode":"is","targetLanguageCode":"en","model":"moses"}' -H "Content-Type: application/json" https://nlp.cs.ru.is/moses/translateText

Kóði

Myndrænt viðmót til tilrauna er aðgengilegt á http://velthyding.mideind.is og uppruni kóðans, skjöl, stillingar og líkön eru talin upp hér að neðan.

- Transformer og Bi-LSTM:
    - Docker samsetningarskrár: https://github.com/vesteinn/docker-greynir
    - Docker skrár fyrir þjóna: https://hub.docker.com/u/mideind
    - Kóði fyrir þjóna: https://github.com/mideind/nnserver
    - Þjálfuð líkön með stillingum: http://velthyding.mideind.is/data

[16] https://github.com/tensorflow/tensor2tensor
[17] https://opennmt.net/
[18] http://statmt.org/moses/
[19] https://cloud.google.com/translate/docs/reference/rest/v3/projects/translateText
[20] Sjá Vörðu 1 undir Skýrslu í “V5 – Machine translation interface” fyrir ítarlegri lýsingu.

    - Forvinnsla og þjálfunarskriftur: https://github.com/mideind/nnserver/tree/master/nnserver/scripts og https://github.com/mideind/GreynirTranslate/tree/master/utils
- Moses:
    - Framvinnsluforrit þýðinga: https://hub.docker.com/r/haukurp/moses-lvl
    - Þjálfuð kerfi: https://hub.docker.com/r/haukurp/moses-smt
    - Kóði: https://github.com/cadia-lvl/SMT
    - Docker samsetningarskrár: https://github.com/cadia-lvl/SMT/blob/master/docker-compose.yml

Skýrsla

Forvinnsla

ParIce málheildin þarfnaðist forvinnslu áður en hún var notuð til þjálfunar fyrir kerfin. Þetta var vegna þess að ParIce málheildin innihélt setningar sem voru hvorki á ensku né íslensku, brenglaðar setningar og ósamræmdar setningar. Moses kerfið þarfnaðist frekari síunar, og tekið var á fyrstu tveimur vandamálunum, en ekkert var gert vegna ósamræmdu setninganna. Fyrir Transformer og LSTM kerfin voru síurnar úr GreynirTranslate notaðar.

Þjálfun

Transformer og LSTM líkönin voru þjálfuð á Nvidia 1080 skjákortum í um það bil þrjá til fjóra daga fyrir hvert líkan. Moses líkanið var þjálfað á 10 örgjörvum í um það bil 6 klukkustundir fyrir hvert líkan. Kerfin voru þjálfuð á ParIce málheildinni, að undanskildu litlu hlutasetti sem notað var til að meta líkönin.

Keyrslutími

Afköst hinna mismunandi aðferða eru tiltekin í töflunni hér að neðan ásamt þjálfunartíma þeirra. Tekið skal fram að afköstin ráðast að miklu leyti af lengd setninga fyrir kerfi tauganetsvélþýðinganna (NMT) og uppgefinn hraði takmarkar lengd setninga við ~200 orð.

Líkan	**Meðalafköst [orð/sek]**	**Þjálfunartími [klukkustundir]**
Transformer	~1000 (cpu[21]) / ~8500 (gpu[22])	100
Bi-LSTM	~1500 (cpu[23]) / ~3000 (gpu)	75
Moses[24]	251 (10 þræðir) / 54 (1 þráður)	6 (10 þræðir) / 12 (1 þráður)

[21] Intel(R) Core(TM) i7-7700 CPU @ 3.40GHz
[22] Öll notuð skjákort voru GeForce GTX 1080 8gb
[23] Intel(R) Core(TM) i7-7700 CPU @ 3.40GHz
[24] Tölur fyrir afköst Moses eru byggð á öðru prufusetti en ættu að vera sambærilegar.

Mat

Einkunn á BLEU-kvarðanum var reiknuð milli síaða gagnasafnsins og þýðingarútttaks kerfanna. Tókar voru teknir úr þýðingarúttakinu og BLEU einkunnir allra líkana, óháð há/lágstöfum, á þýðingum í báðar áttir má sjá í töflunni hér að neðan. Einkunnir fyrir NMT kerfin voru reiknaðar með Tensor2Tensor skriftunni *t2t-bleu*[25] og fyrir SMT kerfin með *multi-bleu-detok.perl*[26] skriftunni. Úttök þeirra ættu að vera sambærileg.

Enska → íslenska	EES	EMA	OpenSubtitles	Samanlagt
Transformer	56.31	58.37	34.71	54.71
Bi-LSTM	38.68	41.60	23.32	38.12
Moses	47.71	52.19	25.28	46.98
Google Translate	34.49	37.37	26.53	35.25

Íslenska → enska	EES	EMA	OpenSubtitles	Samanlagt
Transformer	62.48	66.83	36.92	60.20
Bi-LSTM	46.82	50.85	27.00	45.31
Moses	54.55	60.63	27.95	53.92
Google Translate	41.11	48.39	31.04	42.25

Athugið einkunnir fyrir Google Translate er ekki hægt að bera beint saman við líkönin vegna þess að þau gætu haft gríðarlega ólík þjálfunargögn. Þau gögn gætu jafnvel innihaldið hluta prófunarsettsins.

[25] https://github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/bin/t2t-bleu

[26] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu-detok.perl

V5 – Viðmót á þýðingarvél

Skýrsla: Miðeind
Aðilar: Miðeind / Háskólinn í Reykjavík - HR

Tilgangur

Að afhenda vefþjón og notendaviðmót sem gerir þýðendum kleift að skrifa inn texta (beint, eða með því að senda inn hreinan texta (UTF-8) og Microsoft Word skjöl, eða með því að kalla á HTTP/JSON byggð forritaskil) og fá hann þýddan með bakenda þróuðum í undirverkefni V3. Notendaviðmót skal nota samhliða samanburð úttaks kerfanna þriggja, og ætti að geta sýnt hverja upprunalegu setningu á móti þeirri þýddu, í sprettiglugga og/eða með áherslulýsingu. (Í seinna verkefni (V5b) verður bætt við þetta vefviðmót möguleika þess að velja og/eða slá inn aðrar/betri þýðingar fyrir hverja setningu.)

V1

- Forritaskil fyrir öll þrjú vélþýðingarkerfi þróuð og innleidd.
- Kröfur fyrir notendaviðmót skilgreindar.

Afurðir

Öllum afurðum var náð.

Þarfir viðmóts (forritaskil)
Til að ná vel þekktum skilgreiningum og til að forðast það að finna upp hjólið, voru Google Translate REST Resource v3.projects skilgreiningin[27] fyrir *translateText* forritaskilin valin sem viðmið. Nánartiltekið var undirsett skilgreint með eftirfarandi framsetningu JSON samþykktri sem inntaki fyrir öll þrjú kerfi vélþýðinga:

```
{
  "contents": [
    string
  ],
  "sourceLanguageCode": string,
  "targetLanguageCode": string,
  "model": string,
 }
```

Öll þrjú kerfin gefa JSON framsetningu á eftirfarandi formi:

```
{
  "translatedText": [
    string
  ],
 "model": string,
 }
```

Endapunktar forritaskila

[27] https://cloud.google.com/translate/docs/reference/rest/v3/projects/translateText

Eftirfarandi tveir endapunktar eru tiltækir fyrir bakenda vélþýðingalíkananna þriggja:

- NMT - https://velthyding.mideind.is/nn/translate.api
    - Stillingar líkans eru “transformer” fyrir Transformer byggðar þýðingar og “bilstm” fyrir tvíátta LSTM byggðar þýðingar.
- Moses - http://nlp.cs.ru.is/moses/translateText
    - “model” stillinguna ætti að setja á “moses”.

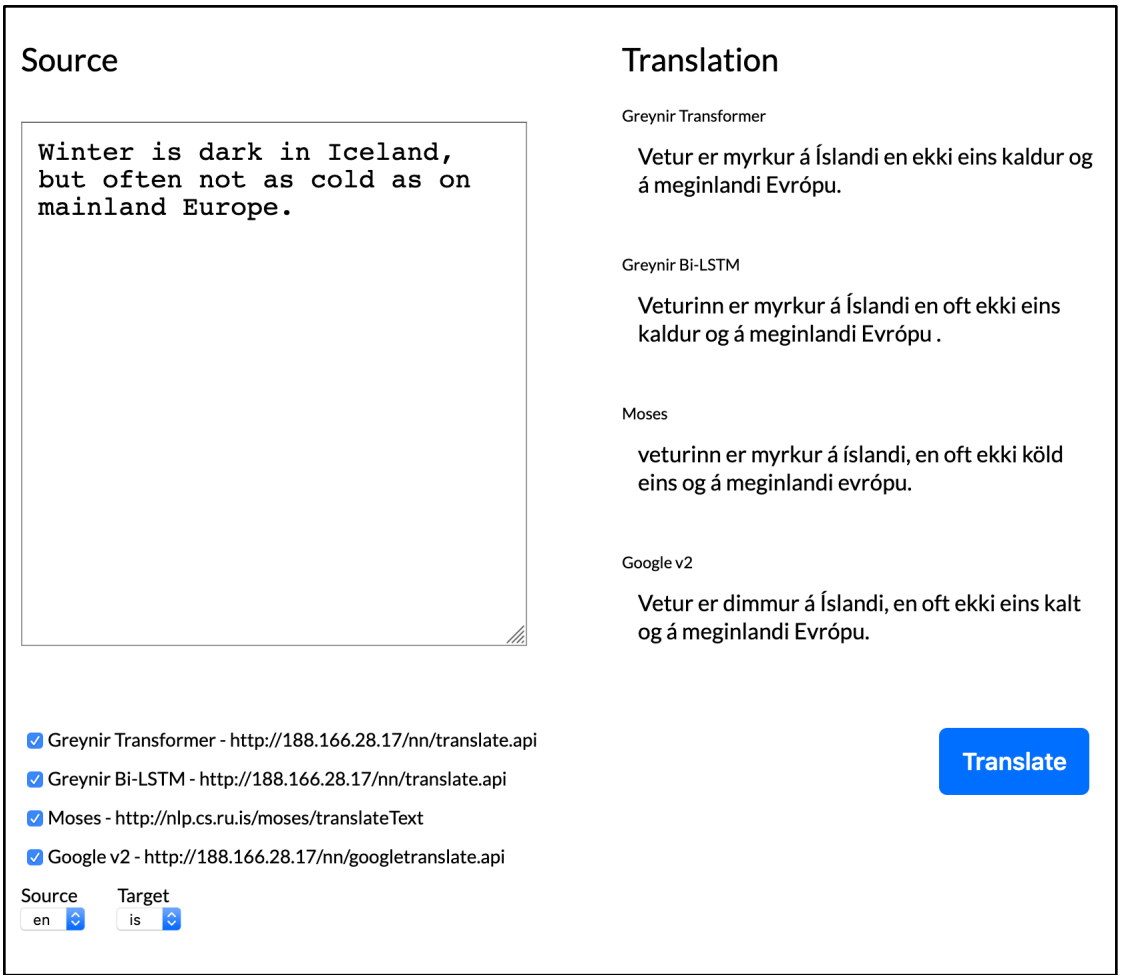
Innleiðing forritaskila NMT
Vefþjónn forritaskila NMT er hluti af verkefninu Greynir sem er opið (e. open-source) á GitHub. Sértækar NMT einingar er að finna í undirmöppu **nn** í meginverkefninu[28]. Vefþjónn forritaskila er skrifaður í Python 3.6 og notar Flask burðargrindina. Tilgangur hans er að endursníða aðkomandi gögn áður en þau eru send áfram til undirliggjandi farvegs tauganets, auk þess að samþykkja endursenda þýðingu frá NN og endursníða áður en hún er send til baka til notanda. Endursníðingin samanstendur meðal annars af umbreytingu inntekins texta í strengi tóka kóðaða sem orðaflísar, og til bara fyrir úttak.

Innleiðing forritaskila Moses
Vefþjónn forritaskila Moses er hluti af verkefninu Moses PBSMT, sem er opið (e. open-source) í undirmöppu **frontend** á GitHub (https://github.com/cadia-lvl/SMT/tree/master/frontend). Forritaskilin eru skrifuð í Python 3.6 og nota Flask burðargrindina. Forritaskilin forvinna gögnin áður en þau eru send til XMLRPC vefþjóns, sem hluti af þjálfuðu Moses kerfi. Forritaskil Moses eftirvinna svo þýðinguna áður en henni er skilað.

Skýrsla

Til viðbótar við afurðir vörðu V1, er einfalt vefviðmót þegar aðgengilegt á http://velthyding.mideind.is/ sem hefur aðgengi að endapunktum sem lýst er að ofan og getur sýnt margar þýðingar á sama tíma til samanburðar. Úttak frá Google Translate er einnig sýnt til viðmiðunar.

[28] https://github.com/mideind/Greynir


Source
Winter is dark in Iceland, but often not as cold as on mainland Europe.
Translation
Greynir Transformer
Vetur er myrkur á Íslandi en ekki eins kaldur og á meginlandi Evrópu.
Greynir Bi-LSTM
Veturinn er myrkur á Íslandi en oft ekki eins kaldur og á meginlandi Evrópu .
Moses
veturinn er myrkur á íslandi, en oft ekki köld eins og á meginlandi evrópu.
Google v2
Vetur er dimmur á Íslandi, en oft ekki eins kalt og á meginlandi Evrópu.
Greynir Transformer - http://188.166.28.17/nn/translate.api
Greynir Bi-LSTM - http://188.166.28.17/nn/translate.api
Moses - http://nlp.cs.ru.is/moses/translateText
Google v2 - http://188.166.28.17/nn/googletranslate.api
Source
en
Target
is
Translate

M1 - Almenn villumálheild

Skýrsla: Háskóli Íslands
Aðilar: Háskóli Íslands, Miðeind

Varða 1 - lýsing

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

V1: Uppruni gagna fundinn, söfnun hafin. Stærð málheildar áætluð, bráðabirgðamat er 5.000 setningar, sem innihalda að minnsta kosti eina villu hver. Mörkunarskema tilbúin og fyrstu mörkunum til prófana lokið.

Afurðir

Öllum afurðum vörðunnar var náð nema mörkunarskema er enn í þróun ásamt áframhaldandi vinnu við mörkun.

Varðandi mörkunarskema: Við höfum þegar valið TEI-formið fyrir úttak markanna og borið kennsl á marga flokka villna sem eru greinanlegir í málheildinni en enn er verið að bæta atriðum í ferlið. Við áformum að fínvinna aðferðir mörkunar og flokkun villna og ljúka mörkunarskemu fyrir 15. Febrúar, 2020. Þessi áframhaldandi vinna við mörkunarskemu hægir ekki á vinnu vegna þess að teymi mörkunar er þegar að leggja fram fyrsta lag mörkunar, sem verður kallað form 1 að neðan.

Skýrsla

Vinna við mörkun Almennu villumálheildarinnar fer fram í Rannsóknarstofunni Mál og tækni í Háskóla Íslands. Áætlun og skipulag þeirrar vinnu felur í sér viðvarandi samræður og reglulega fundi með teyminu sem vinnur að málrýnikerfi hjá Miðeind.

Uppruni gagna fundinn
Við höfum fundið þrjá meginuppruna gagna fyrir villumálheild, 1) Ritgerðir nemenda, 2) Fréttagreinar af netinu, 3) Wikipedia greinar. Við höfum aðgengi að öllum þessum gögnum frá þegar samsettri og útgefinni málheild fyrir íslensku. Við höfum sannreynt að við getum greint nýtilegt magn villna í þessum gögnum.

Bráðabirgðamat lokaútgáfu málheildar
Við höfum keyrt prófanir á hraða mörkunar og gert áætlun um að afhenda minnst 5000 setningar með minnst einni villu hver, skv. upprunalegum áformum.

Mörkunarskema
Okkar ferli mörkunar notar lagskipta aðferð sem endar í XML skjali á TEI-formi með endanlegum villumörkunum. Ferlið hefur einnig birtingarmyndir í miðju ferli. Fyrsta skrefið í ferlinu felur í sér hvers konar tól sem getur leiðrétt villur og einnig viðhaldið upprunalegum texta (form 1). Eins og er notum við *Track Changes* eiginleikann í Microsoft Word fyrir þetta vegna þess að fólk sem vinnur við að merkja villur kann vel á þetta viðmót og gerir þeim kleift að vinna nokkuð hratt. Það er hinsvegar engin þörf fyrir Word, því hvaða textaritil sem

er má nota fyrir þetta skref. Í framhaldinu notum við Python skriftu til samröðunar og samskeytingar réttra og rangra útgáfa í xml-grind sem inniheldur kortlagningu villna til leiðréttinga auk sjálfvirkri mörkun til bráðabirgða af villum sem hægt er að taka á sjálfvirkt (form 2). Við fylgjum því loks með handvirkri leiðréttingu og mörkun eftirstandandi villna og vistum endanlegt úttak til útgáfu í málheildinni (form 3). Eins og kemur fram að ofan er endanleg mörkunarskema fyrir form 3 enn í þróun.

Fyrstu markanir til prófana
Við erum nú að vinna við mörkun fyrstu gerð texta, ritgerðum nemenda, og reynsla okkar af fyrstu prófunum sýnir að þessi aðferð mun leiða til þeirrar útkomu sem stefnt er að. Nú er verkefnisstjóri Rannsóknarstofu Mál og tækni, Lilja Björk Stefánsdóttir, að vinna við mörkun Villumálheildar með aðstoð fjögurra háskólanema í grunnnámi.

M5 - Aðferðir og hugbúnaðarhögun

Skýrsla: Miðeind
Aðilar: Miðeind

Varða 1 - lýsing

Tilgangur
Að hafa hugbúnaðarhögun tilbúna fyrir verkefnið áður en meginverkefni byrja. Umhverfi, verkvangar, aðferðir og forritunarmál hafa verið skoðuð, með tilliti til notagilda málrýna í framtíðinni.

V1
Aðferðafræði og hugbúnaðarhögun fyrir málrýni komið á fót.

Afurðir

Öllum afurðum vörðunnar var náð.

Aðferðafræði og hugbúnaðarhögun fyrir málrýni komið á fót
Tryggt var að tillaga hugbúnaðarhögunar og aðferðafræði fylgdi stöðlum hugbúnaðarþróunar SÍM.
Notendahópar voru skilgreindir og þarfir þeirra greindar.
Algengum aðferðum málrýni var lýst, og besta aðferðin fyrir hverja tegund villna fundin.
Þessum afurðum er lýst nánar að neðan.

Skýrsla

Notendahópar
Tilætlaðir notendur málrýniverkefnanna eru þróendur innan SÍM (DEV), aðrir rannsakendur og þróendur (OTH), einkafyrirtæki og aðrir aðilar (COM) og endanlegir notendur (END). DEV, OTH, og COM hafa sömu þarfir. Þeir þurfa einigar sem hægt er að bæta í forrit, sem veita forritaskil (API) sem geta greint villur í texta og leiðrétt þær.

COM og END þurfa einnig skilmerkilega villugreiningu og upplýsingar um hverja villu, ásamt vísbendingum og hjálp til að forðast villuna.

Fyrirhugaðir endanlegir notendur fyrir þessa fyrstu útgáfu málrýnis hafa íslensku að móðurmáli. Þeir skilja málfræðihugtök og tillögur sem vitna í slík hugtök eru þeim hjálpsamleg.

Endanlegir notendur fyrstu útgáfu eru því ekki hópar eins og lesblindir, fólk með annað móðurmál, börn o.s.frv. Þessir hópar gera ákveðnar villur og gætu þurft annarskonar málfræðilegar vísbendingar og hjálp.

Endanlegir notendur gætu viljað lista mögulegra leiðréttinga frekar en sjálfvirka leiðréttingu með líklegustu staðgöngu.
Undirhópur gæti viljað strangari aðferð - leiðrétta allt mögulegt og allar mögulegar villur greindar, líka samhengisvillur.
Annar undirhópur gæti viljað slakari aðferð - sem gefur aðeins tillögur fyrir villur og bendir ekki á samhengisvillur.

Algengar aðferðir, kostir og gallar
Athugasemd
Ekki er farið yfir djúpar tauganetsaðferðir hér, þar sem aðgreint undirverkefni fer sérstaklega í rannsókn og innleiðingu þeirra. Eins og sjá má er hver aðferð við hæfi mismunandi villuflokka. Sambland þeirra er besta aðferðin.

Greining villna
Orðabókarleit - Þessu er þegar lokið en má alltaf bæta. Nýjum færslum er stöðugt bætt í orðaforða.
Formfræðileg greining - Þegar lokið, fyrir greiningu óþekktra orða. Þetta er nauðsynlegt fyrir tungumál með samsetningar eins og íslenska. Algengum röngum hlutum orða hefur verið safnað og samsett orð með þeim mörkuð sem villur, ef engar aðrar skiptingar eru mögulegar.
N-stæður stafa - Þessar gætu hjálpað að athuga hvort orð fylgir hljóðfræðireglum tungumálsins. Þetta myndi ekki henta íslensku sérstaklega vel, þar sem samsett orð eru mjög algeng, en gæti verið bætt í tilraunaskyni til aðstoðar tillögum leiðréttinga.
Ruglingsmengi fyrir samhengisvillur - Gögnin fyrir sumar slíkra villur eru til, en þessi aðferð hefur ekki verið innleidd. Ítarlegri vinna við söfnun slíkra gagna er í vinnslu í sjálfstæðu verkefni.
Málfræðireglur fyrir málfræðivillur - Þessar hafa verið innleiddar með góðum árangri. Þær henta best vel þekktum algengum villum þar sem einfaldar en nákvæmar reglur virka vel. Dæmi um þetta er notkun á röngu falli með ópersónulegum sögnum (*?Manninum á verkstæðinu vantar skrúfjárn*, þar sem *Manninum* (þágufall) ætti að vera *Manninn* (þolfall)).
Stöðuferjöld - Þessi eru notuð í tilreiðingarlagi til greiningar villna í greinarmerkjasetningu.
Sniðmátun fyrir þáttunartré - Þetta verður innleitt sem hluti af M7 til að grípa villur eins og *Katrín leitaði af kettinum*, þar sem forsetningin *af* ætti að vera *að* í samhengi sagnarinnar *leita - search*.
Flokkari - Oft eru flokkarar þjálfaðir fyrir hverja tegund villna. Þetta hefur ekki verið innleitt fyrir verkefnið (en hefur verið innleitt fyrir íslensku í öðrum verkefnum) en það gæti verið gert. Þetta ætti að skoða í samspili við tauganet, þannig að ef þau vinna betur, ætti vinnan frekar að fara í innleiðingu tauganets. Flokkarar gætu hjálpað við ruglingsmengi, og eru notaðir í verkefninu að ofan.
Sagnaviðmið - Þetta hefur verið innleitt en reglurnar eru of slakar og óljósar eins og er. Nákvæmari reglur villna væri hægt að bæta við, hugsanlega sem hluti af M4.

Listi erlendra orða - Listi erlendra eininga sem fundust við þróun hefur verið safnað saman. Þetta mætti bæta kerfisbundið sem hluti af M4.

Leiðrétting villna
Stafskiptafjöldi - Þetta hefur verið innleitt og er notað til að finna hugsanlegar leiðréttingar og vigta þær.
Hljóðvarps-algrím - Þetta mætti nota við vigtunar tillaga. Skiptingar sem endurspegla almennar hljóðfræðivillur hafa verið skilgreindar en að öðru leyti verður þetta ekki innleitt.
Tækni líkindalykils - Eins og með hljóðvarps-algrím, verður þetta ekki innleitt.
Stafskiptifjölda-algrím og orðabókarleit nær að mestu yfir þetta.
Líkindi stöðuskipta fyrir bókstafi - Þetta mætti nota við vigtunar tillaga leiðréttinga.
Ruglingslíkindi (e. *confusion probability*)- Þetta hefur verið notað fyrir önnur verkefni. Þetta virkar best fyrir villur í ljóslestri (OCR). Þetta gæti verið innleitt fyrir sérstaka OCR útgáfu málrýnis, hugsanlega stjórnaðri með stillingu eða rofa.
Málfræðireglur - Þetta hefur verið innleitt og virkar mjög vel, eins og kemur fram að ofan undir Greining villna. Hjálpleg greining villna og málfræðivísbendingar eru tengdar hverri villureglu og gerðar aðgengilegar í notendaviðmótinu. Reglurnar, og einnig vísbendingarnar, geta verið eins nákvæmar og þarf. Málfræðivísbendingar geta verið stikaðar eftir samhengi.
Flokkari - Eins og greint er frá undir Greining villna, verður þetta ekki innleitt eins og er.
Truflaðar samskiptarásir - Þessi tækni hefur verið innleidd. Athuga þarf að þetta gefur engar upplýsingar um villur til að senda til notanda, svo þetta ætti aðeins að innleiða sem síðasta úrræði.
SMT byggt á setningaliðum - Þetta hefur ekki verið innleitt, en myndi ekki skila neinum upplýsingum um villur til notanda.
Orðatíðni - Þessu ætti að bæta við fyrir röðun tillaga. Söfnun gagna gæti verið hluti af M4.
Tíðni orðflokka/orðmynda - Þetta gæti verið nytsamlegt. Söfnun gagna gæti verið hluti af M4.
N-stæður orða - Núverandi pakkinn Icegrams inniheldur stórt safn þrístæða með upplýsingum um tíðni, og verður nytsamlegt til leiðréttinga stafsetningar. Gagnasafn þrístæðanna mætti og ætti að stækka sem hluta af M4.
Skjalasamhengi - Þetta mætti nota til röðunar tillaga. Orðapoki myndi nægja. Þetta ætti að innleiða.

M6 - Málrýnir fyrir stakorðavillur

Skýrsla: Miðeind
Aðilar: Miðeind, Háskóli Íslands

Varða 1 - lýsing

Tilgangur
Að afhenda sjálfstæðan hubúnaðarpakka og vefþjónustu sem getur þekkt og stungið upp á leiðréttingum einfaldra (samhengisfrjálsra) stafsetningarvillna í dæmigerðum textum, skrifuðum af hæfum riturum íslensku. Á seinni stigum máltækniverkefnisins, mun einingin verða hluti af fullgildum málrýni fyrir íslensku.

Ákvæði
Útgáfa tilreiðara sem hefur viðeigandi stikaðar stillingar til málrýni.

V1

- Skýr verkalisti fyrir málrýni tilbúinn, aðgreindar frá verkum tilreiðara.
- Stikaðar stillingar fyrir tilreiðara skilgreindar.
- Prófunarsett 250 setninga með samhengisfrjálsum orðavillum tilbúið. Málrýnir merkir óþekkt orð, og innleiðing á greiningu óþekktra orða langt komin.
- Hugbúnaðinum er afhent samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins.

Afurðir

Öllum afurðum vörðunnar var náð.

Skýr verkalisti fyrir málrýni tilbúinn, aðgreindar frá verkum tilreiðara.
Þarfir fyrir tilreiðarann voru skilgreindar, og svo innleiddar sem hluti af I3. Niðurstöður prófana voru greindar og verkum forgangsraðað. Þessar eru tilgreindar undir Skýrsla. Athugið að prófunarsett og forgangsröðuð verk voru búin til með þarfir annarra rannsakenda í huga, þannig að prófunarsettin prófa leiðréttingar, ekki villugreiningu, eins og er.

Stikaðar stillingar fyrir tilreiðara skilgreindar.
Eftir því sem verkefni tilreiðara þróaðist urðu stikaðar stillingar sem slíkar úreltar. Þeirra í stað voru tvær stillingar skilgreinar; grunn- og fullþáttun. Málrýnir notar fullþáttun, sem var þróuð með þessar þarfir í huga.

Prófunarsett 250 setninga með samhengisfrjálsum orðavillum tilbúið.
Prófunarsett 325 setningar með 343 samhengisfrjálsum villum var safnað úr ritgerðum nemenda sem partur af verkefni M1.

Villugreining nær F1-gildi 89,2% fyrir grunnt úttak skipanalínu, sem sýnir ekki óþekkt orð sem mörkuð eru sem villur í notendaviðmótinu. Villur eru aðeins leiðréttar, en engar markarnir sýndar. Af 66 óleiðréttum villum voru 29 markaðar sem óþekkt orð í notendaviðmótinu og 12 voru markaðar sem villur og tillögur gefnar en villurnar ekki leiðréttar sjálfvirkt.

Leiðrétting villna nær 86,6% nákvæmni. Aðrar mælingar eru ekki mögulegar fyrir grunnt úttak skipanalínu.

Málrýnir merkir óþekkt orð, og innleiðing á greiningu óþekktra orða langt komin.
Eins og kemur fram að ofan merkir málrýnir óþekkt orð. Áður en á þann stað er komið reynir málrýnir að greina orðið. Ef það er ekki að finna í orðaforðanum eða orðaforða villna, notar málrýnir greiningu samsetninga til að reyna að greina samsett orð. Ef það virkar ekki og engar aðrar aðferðir villugreiningar virka, notar málrýnir stafaskiptifjarlægðar-algrím í `spelling.py` til að reyna að leiðrétta orðið.

Orðalistarnir sem verða þróaðir sem hluti af M4 verða nytsamlegir hér, svo sem listi erlendra orða.

Hugbúnaðurinn er afhentur samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins.
Skipanalínutól málrýnis hefur verið afhent samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins og er aðgengilegt á GitHub. Kóðun texta er UTF-8. Kóði og athugasemdir eru að mesti á ensku en á íslensku þar sem þarf.
Hugbúnaðurinn er undir GPLv3 leyfi og tilkynning höfundarrétta er innifalin í hausathugasemd í öllum frumskrám.
Travis CI er notað fyrir endurtekna prófun. Hugbúnaðurinn er skrifaður í Python 3.6.
Hugbúnaðinn má setja upp sem Python pakka með pip, og er hýstur í PyPi sem 'reynir-correct'.
Skipanalínutól er fáanlegt.
Skjöl varðandi uppsetningu og notkun eru á GitHub.

Skýrsla

Forgangsröðuð verk
Eftir prófanir voru eftirfarandi verk sett í forgang.

Orðalistar og orðaforði. Athugið að þróun hefur að mestu lagt áherslu á meðhöndlun mismunandi villuflokka, kerfisbundnar viðbætur í notaðann orðalista og orðaforða er næsta skref og stór hluti af M4.
Icegrams - Leiðrétt undirsett RMH ætti að bæta við til að endurbæta þrístæður og tillögur í `spelling.py`. Þetta gæti hjálpað við leiðréttingar langra samsetninga.
Gögn frá eldri leitum orða sem ekki eru þegar í orðaforða - Gögnin eru tilbúin og hafa verið flokkuð, þar sem hluti þeirra hefur verið innleiddur en mikið er eftir. Gögnin innihalda samhengisfrjálsar villur, nýyrði, erlend orð o.s.frv.
Færa orð frá [einstakar villur] til [villuform] í ReynirCorrect.conf – Þetta er í vinnslu og tekur tíma. Markmiðið er svo að bæta tilteknum skilaboðum við hverja villu, til að hjálpa notenda sem mest. Seinna atriðið þarfnast flokkun villna frá undirverkefni M1.
Bæta [hástöfunarvillur] í ReynirCorrect.conf - Nöfn íbúa og nöfn landa ætti að bæta við kerfisbundið.
Form sagna sem tákna spurningar með fornafnasníklum (e. *pronoun clitics*) bætt við - Eintöluformum hefur verið bætt við, en fleirtöluformum ætti að bæta við og merkja sem villur.
Bannorð - Þessi listi myndi hjálpa mörgum verkefnum.
Aðrir flokkar villna - Flokkun villna úr undirverkefni M1 geti gefið upplýsingar um aðra flokka villna sem ekki hefur verið tekið á.
Villur í orðaforðanum – BÍN inniheldur mörg form sem eru ekki notuð eða einfaldlega villur. Sjálfstætt verkefni endurskoðunar er í vinnslu og gögn frá því ættu að reynast nytsamleg.

Kóði. Þessi verk marka hvað þarf að einnleiða eða laga í kóðanum.
Tillögur - Úttak skipanalínu ætti að leiðrétta tillögur sjálfkrafa. Þær eru merktar gular í notendaviðmótinu en ekki leiðréttar í grunnu úttaki. Prófanir sýna að þær eru sjaldan rangar og myndu bæta grunnt úttak.
Kerfisbundnar villur í beygingarendingum - Listi þeirra og leiðréttra útgáfa ætti að skapa og meðhöndlun þeirra innleidd. Þetta myndi líklega hjálpa leiðréttingu langra samsetninga, sem stafaskiptifjarlægð nær ekki að taka vel á vegna þess hve sjaldgæf þær eru í orðaforðanum.
Hljóðfræðivillur - Spelling.py inniheldur skilgreiningar skiptinga sem endurspegla hljóðfræðivillur, en gefa ætti þeim hærri einkunn. Dæmi: i/y, fnt/mt/mmt, hv/kv.

Þrístæðuleit - Fyrir þrístæðuna <t1, t2, t3> fyrir t2, ef t1 hefur verið leiðrétt, ætti leiðrétt útgáfa að vera send inn í þrístæðuna fyrir `spelling.py`. Orðaforðinn þar hefur verið leiðréttur og ætti ekki að innihalda nein vitlaus orð, svo að leit með upprunalegri þrístæðu er árangurslaus.

H1 - Gagnaupptökur með Samrómi

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík (og Almannarómur og Deloitte)

Tilgangur

Að útvega stór söfn upptaka stuttra setninga til þjálfunar talgreina.

Varða 1 - lýsing

Kerfi Samróms sett upp með viðeigandi lista setninga. Fullorðnir: 30.000 setningar teknar upp.

Endurskoðuð varða 1:

30.000 setningar fullorðinna mælenda (18+) útgefnar samkvæmt stöðlum CLARIN. Skýrsla um dreifingu aldurs, kyns og íslensku sem móðurmáls, í gagnasafni. Áform tilbúin um hvernig skal ná til hópa mælenda sem eru í of lágu hlutfalli.

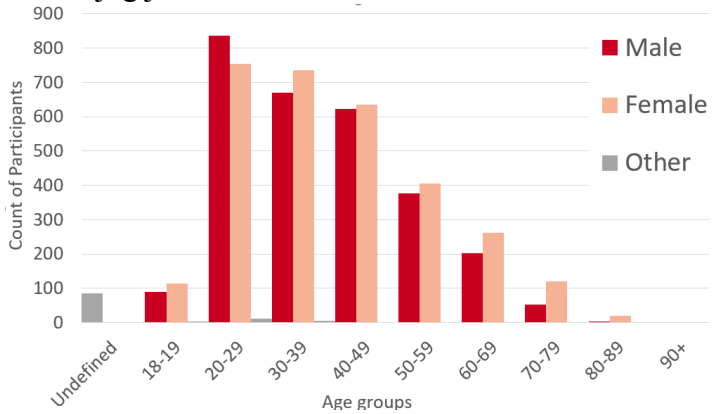
Afurðir

Vinna við gagnaupptökur hefur gengið vel. Vefforritið samrómur.is hefur reynst áhrifaríkt til lýðvirkjunar. Fyrstu vörðu með 30.000 setningar fullorðinna mælenda var náð 8. nóvember 2019. Gagnasafnið hefur ekki enn verið útgefið þar sem staðlar CLARIN hafa ekki enn verið skilgreindir. 15. janúar 2020 voru komnar 53.407 setningar fullorðinna mælenda og það eru 558 setningar fullorðinna með íslensku sem annað mál.

Til að setning teljist gild þurfa tveir mismunandi aðilar að gefa upptökunni jákvætt atkvæði í endurgjöf á Samrómur.is. Við metum um 87% þeirra setninga sem safnast hafa sem fullkomlega nothæfar. Þessi tala gæti hækkað örlítið þegar við skoðum nánar upptökur sem hafa fengið tvö neikvæð atkvæði.

Fáeinar tölfræðilegar upplýsingar um söfnunina eru hér að neðan, en síuppfærðar tölfræðilegar upplýsingar eru aðgengilegar á stats.samromur.is. Eins og staðan er nú sjáum við að gagnasafnið er með of hátt hlutfall ungra einstaklinga, sem kemur ekki á óvart. Þetta

er líklega vegna þess að þessi markhópur er líklegri til að nota samfélagsmiðla. Kynjaskipting í núverandi gagnasetti er mjög jöfn.



Súlurit 1: Dreifing aldurs og kyns. Frá byrjun var kynjaskipting mjög jöfn milli karla og kvenna. 15. janúar er gagnasafnið byggt upp af 52% konum, 47% körlum og 1% öðrum.

Áform um hvernig skal ná til hópa mælenda í of lágu hlutfalli

Fyrstu mánuðirnir gáfu vísbendingu um hvar við þurfum að beina markaðssetningu og sókn. Viðmótið sjálft verður endurbætt á tvo vegu. Fyrri er að hvetja fólk til þess að lesa fleiri en 5 setningar í hverri lotu með því að gefa fólki valkost þess að taka upp 5, 15 eða 30 í hverri lotu. Seinni er að innleiða valkvætt kerfi innskráningar, sem opnar möguleika þess að leikjavæða söfnunarátakið. Þetta myndi, til dæmis, veita okkur möguleika á því að setja upp keppnir milli stofnana, fyrirtækja, eða annarra hópa fólks og möguleika þess að gefa einhverskonar hvata til þátttöku. Þessum þáttum verður lokið við miðjan febrúar.

Úr Súluriti 1, má sjá að upptökur frá einstaklingum allra kynja undir 50 ára aldri eru í of háu hlutfalli. Hópi mælenda í of lágu hlutfalli mælenda 50+ má skipta í tvennt, einstaklingar sem enn eru í vinnu og einstaklingar sem eru á eftirlaunum. Við áformum að taka á þessu mishlutfalli upptaka með beinni markaðssetningu og kynningum. Fyrir einstaklinga á eftirlaunaaldri felur það í sér kynningu verkefnisins á öldrunarheimilum og í samtökum og félögum eldri borgara. Fyrri hópnum mætti ná til með fyrrnefndri leikjavæðingu verkvangsins.

Frá og með þessari skýrslu höfum við opnað afmörkun aldurs fyrir 13-17 ára einstaklingum. Af þeirri ástæðu er sá hópur ekki til staðar í fyrsta súluriti. Í næstu vörðu munum við byrja að leggja áherslu á söfnun gagna frá ungum einstaklingum með samstarfi við skóla og sveitafélög. Eins og er vinnum við með grunnskólum til að afmarka tíma í námsskrá fyrir nemendur þeirra til að taka þátt í Samrómi í kennslustundum íslensku.
Fyrir næstu vörðu munum við byrja að safna setningum einstaklinga með íslensku sem annað mál. Þetta verður gert í samstarfi við samfélagshópa fólks sem deilir sameiginlegum heimalöndum og búa á Íslandi. Sumir þessara samfélagshópa hafa formleg samtök, aðrir aðeins hópa á Facebook.

H2 – Innslegið sjónvarps- og útvarpsefni

Skýrsla: Háskóli Reykjavíkur
Aðilar: Háskóli Reykjavíkur, CreditInfo, RÚV

Tilgangur

Að veita innslegið sjónvarps- og útvarpsefni til þjálfunar talgreina fyrir notkun á sviði miðla.

Varða 1 - lýsing

Matsskýrsla varðandi texta af 888 og hvernig skal nota þá. Skýrsla um aðgengilegt efni og listi yfir efni til innsláttar á fyrsta ári tilbúin, ásamt skilgreiningum forms og lýsigagna. Verkvangur fyrir handvirkan innslátt sett upp. 5 klukkustundir útvarpsefnis innslegnar.

Afurðir

Öllum afurðum vörðu 1 var náð. Við höfum yfir 120 klukkustundir af útvarpsefni og yfir 100 klukkustundir gagna af 888. Þetta er allt aðgengilegt á vefþjóni Mál- og raddtæknistofu Gervigreindarseturs Háskólans í Reykjavík, Terra. Matsskýrsla er innifalin hér fyrir neðan.

Skýrsla

Útvarpsþættir
Við höfum texta úr þættinum *Í ljósi sögunnar* frá byrjun þáttarins (um 80 klukkustundir tals) og texta frá þættinum *Hátalarinn* (40 klukkustundir, eitthvað af því tónlist eða viðtöl) fyrir árið 2019. Textarnir eru nokkuð nákvæmir en höfundarnir segja að sumt af því hafi breyst. Á þessum tímapunkti höfum við ekki mælt nákvæmnina.

Innslegið útvarpsefni

Alls voru fimm klukkustundir efnis innslegið frá Creditinfo. Efninu var skipt í u.þ.b. 2 mínútna búta:

1. Kastljós, 1,5 klukkustundir
2. Samfélagið, (útvarpsþáttur um málefni líðandi stundar), 2 klukkustundir
3. Spegillinn, (útvarpsþáttur um málefni líðandi stundar), 1,5 klukkustundir

Efnið var undirbúið og afhent til HR, þar sem það er geymt á *terra* þjóninum.

Sjónvarpsþættir
Við munum draga út og nota gögn frá eftirfarandi sjónvarpsþáttum:

1. Menningin, (þáttur um menningu)
2. Krakkafréttir, (fréttir barna)
3. Stundin Okkar, (barnaefni)
4. Kastljós, (viðtöl um málefni líðandi stundar)
5. Kiljan, (bókmennta)
6. Fréttir kl. 19:00, (fréttir)
7. Tíufréttir, (fréttir)

1. Menningin
Klukkustundir hljóðs: 44 klst.
Nákvæmni undirtexta: Virðist nokkuð nákvæm
Tímastimplar: Hefur tímastimpla
Merkingar mælenda: Engar
Undirflokkur: Menning

2. Krakkafréttir
Klukkustundir hljóðs: 37 klst.
Nákvæmni undirtexta: Virðist nokkuð nákvæm
Tímastimplar: Hefur tímastimpla
Merkingar mælenda: Engar
Undirflokkur: Fréttir

3. Stundin Okkar
Klukkustundir hljóðs: 25 klst.
Nákvæmni undirtexta: Stundum nákvæmir, en vantar oft og oft ekki á réttum tíma
Tímastimplar: Hefur tímastimpla
Merkingar mælenda: Engar
Undirflokkur: Barnaefni

4. Kastljós
Klukkustundir hljóðs: 20 klst.
Nákvæmni undirtexta: Oft styttir og orðum sleppt
Tímastimplar: Hefur tímastimpla
Merkingar mælenda: Engar
Undirflokkur: Fréttir

5. Kiljan
Klukkustundir hljóðs: 69 klst.
Nákvæmni undirtexta: Nákvæm, einu orðin sem vantar eru fylliefni.
Tímastimplar: Hefur tímastimpla
Merkingar mælenda: Engar
Undirflokkur: Umræður

6. Fréttir kl. 19:00
Klukkustundir hljóðs: 35 klst.
Nákvæmni undirtexta: Oft umritaðir og vantar stundum.
Tímastimplar: Engir
Merkingar mælenda: Engar
Undirflokkur: Fréttir

7. Tíufréttir
Klukkustundir hljóðs: 13 klst.
Nákvæmni undirtexta: Oft umritaðir og vantar stundum.
Tímastimplar: Engir
Merkingar mælenda: Engar
Undirflokkur: Fréttir

Til skýringar eru klukkustundir listaðar lengd þáttanna, við höfum ekki klukkustundir

málgagna sérstaklega.

Millital er almennt merkt með bandstriki.

Við höfðum áhuga á eftirfarandi þáttum, en þeir eru ekki textaðir eins og er:

2. Vikan
3. Silfrið

Ef fleiri klukkustundir þarf, munum við skoða aðra textaða þætti fyrir meira efni.

Niðurhal 888 og myndbönd
Við fengum aðgang að innanhúss forritaskilum RÚV og með hjálp frá hugbúnaðardeild RÚV bjuggum við til skriftu sem dregur sjálfkrafa út myndbands- og textaskrár fyrir þættina Krakkafréttir, Kastljós, Menningin, Kiljan. Fyrir Fréttir kl. 19:00 og Tíufréttir, drógum við sjálfkrafa út myndbönd síðustu mánaða. En forritaskilin höfðu ekki 888 textaskrárnar þannig við þurftum að nálgast þær beint frá tengilið okkar hjá RÚV. Allt þetta efni er nú á okkar vefþjóni.

Merkingar mælenda
Engin gagnanna höfðu merkingar mælenda, sem sagt engar upplýsingar um aldur, kyn eða fjölda mælenda í þáttahluta. Hinsvegar eru þrjár aðferðir til þess að afla nauðsynlegra gagna:
1. RÚV hefur SQL gagnagrunn sem kallast Kistan sem hefur gögn um aldurshópa mælenda og nöfn þeirra, og því kyn (þar sem flestir eru Íslendingar með föðurnafn sem eftirnafn)). Ef einhver frá RÚV eða Háskólanum í Reykjavík getur búið til forritaskil fyrir Kistuna, er hægt að hala niður merkingum mælenda að mestu.
2. Spyrja grafíkdeild hvort hún sé enn með óunnar myndir, en það er ólíklegt eins og er.
3. Það gæti verið mögulegt að nota tesseract-ocr til að draga út nöfn mælenda úr myndböndum ef við vitum tímastimpla byrjunar tals, svo við þyrftum sjálfskapaða tímastimpla úr raðara. En erfiði hlutinn hvernig skal sjálfvirkt ná skjámyndum úr myndböndum á þeim tímastimplum. Handvirk aðferð væri of kostnaðarsöm í tíma og peningum.

Útdráttur og notkun 888 fyrir talgreina
Fyrir þætti með nákvæma texta eins og Kastljós og Krakkafréttir, notum við Kalda og svipaðar aðferðir eins og greint er frá í rannsóknarritgerðinni, "*Building an ASR corpus using Althingi's Parliamentary Speeches*."

Við munum draga út texta úr skrám undirtexta inn í hluta og textaskrár, framkvæma hlutun og röðun, og henda út öllu sem ekki er hægt að raða eða hefur 90%+ villutíðni orða. Við munum nota núverandi hljóðlíkön og normunarlíkön þróuð fyrir Alþingisverkefnið. Núverandi villur sem komu úr þessu líkönum verða leiðréttar til notkunar fyrir öll gögnin. Bjagað (e. *biased*) mállíkan verður að búa til úr RÚV gögnunum. Myndbandaskrám verður að umbreyta yfir í wav skrár áður en þær verða nothæfar.

Tilraunir hafa verið gerðar til hlutunar samkvæmt tímastimplum á 888 textanum. Niðurstöður til bráðabirgða sýna að samröðun við hljóðið er skökk eins oft og hún er nákvæm; Oft bara skökk um eitt orð eða tvö.

Fyrir fréttaþætti eins og kl. 19:00 og Tíufréttir, munum við þurfa að sía út erlenda hluta og CreditInfo mun líklega þurfa að umrita íslenska hluta með tímastimplum. Þá getum við notað

þær upplýsingar og aðeins þjálfað talgreina á þeim hlutum þáttanna eins og við gerum með þætti með mikla nákvæmni 888.

H6 - Leyfismál talgreinisgagna

Skýrsla: Háskóli Reykjavíkur (HR)
Aðilar: HR, Stofnun Árna Magnússonar í íslenskum fræðum, Persónuvernd, RÚV

Tilgangur

Að hafa skýran farveg leyfissamninga fyrir a) veitendur talgreinisgagna, sérstaklega ólögráða einstaklingum, og b) höfundarrétt hljóðgagna. Á fyrsta ári verkþáttarins mun áherslan vera á samkomulög fyrir veitendur talgreinisgagna (fullorðinna og unglinga).

Varða 1 - lýsing

Yfirlýsingar samþykkis fyrir hljóðupptökur tilbúnar fyrir fullorðna. Yfirlýsingar samþykkir fyrir hljóðupptökur tilbúnar fyrir umráðamenn ólögráða unglinga.

Afurðir

Persónuverndaryfirlýsing Háskólans í Reykjavík fyrir raddsöfnun Samróms.

Skilmálar vegna þátttöku í gerð talgagna fyrir talgreini, verktakasamningur tal.

Skilmálar vegna þátttöku í gerð talgagna fyrir talgreini, verktakasamningur, tal & hlust.

Skilmálar vegna þátttöku í gerð Samróms – Ungmenni.

Eyðublað fyrir samþykki forsjáraðila.

Enda-notanda leyfissamningur fyrir RÚV gögn.

Tengiliðir og netföng fyrir söfnun radda ungmenna– Grunn- og framhaldsskólar.

Skýrsla

Til þess að reyna að fá sem flest fólk til að nota gagnasett okkar, höfum við hafið viðræður við ýmsa aðila til að skipta á eða senda þeim þau gagnasöfn máltækni sem við eigum.

Amazon
Við höfum rætt við Amazon í þeirri von að finna tengilið til að nota íslensku gögnin okkar til þess að færa íslensku inn í snjalltæki Amazon. Enn höfum við ekki fundið neinn til að taka það hlutverk.

CLARIN
Við höfum búið til grunnskilyrði fyrir útgáfu talgagna á CLARIN. Það er einfalt að láta lýsigögn fylgja með eins og er skilgreint í CMDI 1.2 vegna sveigjanleika þess. Eitt vandamál sem við höfum rekist á er að venjulega gefur CLARIN hverjum notanda aðeins 10GB. Við munum þurfa að biðja CLARIN tengilið okkar um meira geymslupláss þar sem hljóðskrár okkar eru á wav formi sem hætta til að vera stórar og bara gagnasafnið Samromur gæti verið stærra en 10GB.

Linguistic Data Consortium (LDC) https://www.ldc.upenn.edu
Við höfum samþykkt að senda gagnasett okkar til LDC í skiptum fyrir nokkur gagnasett þeirra sem munu hjálpa okkur í vélþýðingum og talgreiningu. Eins og er sendum við þeim gagnasett þingræða og biðjum í staðinn um ensk gagnasett sem gætu gagnast verkefni vélþýðinga.
Til að leggja gagnasett okkar inn til LDC þurfum við að fylla út eyðublað fyrir innsendingu málheilda. Síðan hafa þau samband við okkur fyrir nánari upplýsingar. Á þessum tímapunkti getum við einnig sent þeim zip-þjappaða útgáfu gagnasettsins með tölvupósti eða hvaða dreifingarkerfi sem er og láta md5sum fylgja með til staðfestingar. Þau munu skoða gagnasettið og senda spurningar til skýringa eða benda á skrár sem þarf að laga, en betra er að gera þetta eins mikið og hægt er áður en gagnasettin eru send.
Þau krefjast ítarlegri og strangari lýsinga en við höfum í upprunalegum gagnasettum hjá Málföngum. Í framtíðinni verðum við að tryggja að gagnasett okkar innihaldi aðeins nytsamleg gögn, ítarlegar lýsingar innihalds og uppbyggingu mappa, og stjórn á öllu misræmi í hljóð- og textagögnum. Við þurfum heilt yfir betri gæðastjórnun gagnasetta okkar.

Þau hafa viðmið þegar kemur að útvegun gagna:
https://www.ldc.upenn.edu/data-management/providing/technical-guidelines
https://www.ldc.upenn.edu/data-management/providing/documentation-guidelines

Við vonumst til að búa til gagnasett sem virka vel fyrir CLARIN og LDC með minniháttar breytingum, og leggja þau svo inn og gefa út til annarra óbreytt.

Málföng
Við eigum í viðræðum við Málföng um að fá að nýta gagnasafn fyrir talgervingu sem safnað var í samstarfi við Google.

Að setja málheild í Málföng:

1. Við þurfum gögnin
2. Við þurfum að skilgreina leyfið (CC BY 4.0 fyrir þetta verkefni)
3. Við þurfum texta sem lýsa hvað þessi gögn eru er, hvernig þeim var safnað, hvaða ritgerðir og skjöl skal vitna í, o.s.frv. Textinn skal vera bæði á ensku og íslensku. dæmi: http://www.malfong.is/index.php?lang=en&pg=malromur eða þetta http://www.malfong.is/index.php?lang=en&pg=althingisraedur.

Ef tími gefst, myndum við einnig vilja gefa út gagnasett okkar á OpenSLR.

H12 - Greinarmerkjasetning og endapunktaskynjun

Skýrsla: Háskóli Reykjavíkur
Aðilar: Háskóli Reykjavíkur

Tilgangur:

Að afhenda stoðtól sem bæta úttak talgreina sjálfvirkt með greinarmerkjasetningu og endapunktaskynjun.

Varða 1 – lýsing:

Líkan fyrir greinarmerkjasetningu innleitt á íslensku með openFST (textabyggt).

Endurskoðuð varða 1:

Textabyggt líkan greinarmerkjasetningar metið, við stefnum á 80% heildar F-einkunn. Hugbúnaðareining sem inniheldur textabyggð líkön til sjálfvirkrar greinarmerkjasetningar úttaks talgreina samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins.

Afurðir

Öllum afurðum var náð, en F-einkunn gæti verið enn betra með meiri gögnum. Slóð á kóðann notaðan til hreinsunar gagnanna er hér: `https://github.com/cadia-lvl/punctuation-detection`

Líkanið sem notað var til að fá greinarmerki er hér:
`https://github.com/ottokart/punctuator2`

Skýrsla

Eftir rannsókn á nýjustu þróunum í endapunktaskynjun var opið (e. open-source) tól, *Punctuator 2*[1] valið til þjálfunar með íslenskum gögnum. Fyrstu tilraunir voru gerðar með djúpnáms hugbúnaðinum Theano og útfærslu í Python 2.7. Hinsvegar hefur Theano ekki verið stutt síðan 2017 og Python 2.7 missti stuðning í byrjun 2020. Því er þessi tilhögun ekki hæf stöðugri þróun íslensku útgáfu Punctuator 2. Við lentum í vandræðum við umbreytingu Punctuator 2 til að vera samræmanleg Python 3 sem og notkun skjákorts með Theano. Áformað er að endurskrifa kóðann í Keras og Python 3.

Gögnin

Við notuðum hluta Risamálheildar sem inniheldur gögn frá RÚV, 167 MB gagna. Textinn var normaður með *Textahaukur*, frumgerð textanormunarhugbúnaðar fyrir íslenskan texta. Eftir normun, höfðum við 151 MB gagna, 2.047.596 setningar, og af þeim eru 974.183 einstæðar. Við stokkuðum gögnin og gerðum lista yfir 12 milljón tóka (greinarmerki meðtalin), 341.259 einstæð.

Það er vandamál við notkun skjákorts með Theano, þar sem það þarfnast pyGPU, og var ekki árangursríkt. Því var ekki mögulegt að klára þjálfun fyrir meira en 40% upprunalegs gagnasetts frá RÚV (60 MB). Stuðningur við skjákort var svo tiltækur í annarri tölvu sem gat

þjálfað líkanið.

Eftirfarandi tafla sýnir dreifingu greinarmerkja.

Greinarmerki	**Tilvik í gögnum**
Punktur	923.217
Spurningarmerki	10.268
Upphrópunarmerki	1.725
Komma	232.621
Semíkomma	1.542
Tvípunktur	10.380
Bandstrik	3.931

Eins og sjá má úr töflunni eru fá tilfelli allra greinarmerkja nema punkta og komma, sem gerir erfiðara fyrir líkanið til að læra. Við þjálfuð líkanið aðeins með punktum, kommum, spurningarmerkjum og tvípunktum.

Líkanið var þjálfað með þjálfunarsett sem innihélt 80% gagnanna og prófað á hin 20%. Það gaf eftirfarandi niðurstöður:

Greinarmerki	**Nákvæmni [%]**	**Heimt [%]**	**F-gildi [%]**
Punktur	92,0	91,2	91,6
Spurningarmerki	75,2	54,1	62,9
Komma	73,8	56,9	64,3
Tvípunktur	80,5	28,1	41,6
Alls	88,9	83,6	**86,2**

Hlutfall villna orða: 2,64%
Hlutfall villna setninga: 24,5%

Einnig bættum við villum í gögnin og keyrðum þjálfað líkan á þeim. Hlutfall villna var 25% eyðingar, 25% innsetningar, og 50% útskiptingar. Niðurstöðurnar eru eftirfarandi:

Villutíðni orða [%]	**F-gildi [%]**
5	82,9
10	72,8
15	59,8
20	46

Markmið þessarar vörðu var að ná 80% F-gildi, sem náðist, jafnvel með 5% villutíðni orða. Þetta gefur okkur vísbendingar um að við munum getað þróað nákvæman greinarmerkjasetjara fyrir gögn frá talgreinum. Hinsvegar gæti þetta farið eftir gögnum svo að áætlunin fyrir næstu vörðu er bygging nýs greinarmerkjasetjara sem að minnsta kosti eins vel og þjálfa hann á fjölbreyttari gögnum.

Fyrir áframhaldandi vinnu við þetta verkefni þurfum við að nota kóða sem lifir, með því að færa verkefnið yfir í Keras og Python 3. Þá verður stuðningur við skjákort mögulegur og með minni tíma eytt í þjálfun getum við þjálfað á meiri gögnum, sem gefur vonandi betri niðurstöður.

Heimildir

[1] Tilk, O., & Alumäe, T. (2016). Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. *Interspeech 2016.*

T1 - Upptaka á einingavalsgögnum

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

Tilgangur

Að gefa út upptökur mismunandi radda fyrir einingavals-talgervla (e. *Text-to-Speech, TTS*). Markmiðið er að ná fjölbreytni í aldri og mállýskum, með jafn margar karl- og kvenraddir. 8 raddir munu vera teknar upp, 4 karlraddir og 4 kvenraddir, hver þátttakandi ætti að lesa í minnst 20 klukkustundir.

Varða 1 - lýsing

Handrit til lesturs tilbúin (sjá einnig T12). Umhverfi fyrir upptökur uppsett. Upptökum fyrstu tveggja þátttakenda 50% lokið (10 klukkustundir).

Endurskoðuð varða 1:

Umhverfi fyrir upptökur uppsett. Upptökum fyrstu tveggja þátttakenda 50% lokið (10 klukkustundir hver). Gengið frá samkomulagi náð við minnst tvo fleiri mælendur.

Afurðir

Öllum afurðum vörðunnar hefur verið skilað eða eru nálægt því. Nánar:

- Leslistar hafa verið gerðir.
- Efnislegt umhverfi upptaka uppsett.
- Hugbúnaður upptaka er sífellt viðhaldið og hefur verið í notkun síðan í október. (https://github.com/cadia-lvl/lobe) – uppfærð *Readme* skrá á Github til skýringar tilgangs hugbúnaðar.
- Mælandi 1 (Margrét) hefur lokið nær 12 frá skrifum.
- Mælandi 2 (Þorsteinn) hefur lokið nær 11 frá skrifum og er líklegur til að klára fyrir skiladag vörðunnar.
- Við eigum í viðræðum við þrjá aðra mælendur. Minnst einn er mjög líklegur til þátttöku. Þar sem allir 3 mælendur eru kvenkyns höfum við gætt þess að velja ólíkar raddir til að tryggja fjölbreytileika radda.

Skýrsla

Þessi vinna hefur staðið yfir síðan í október og er enn í vinnslu. Upptökur fara fram í hljóðveri 9 hjá RÚV. Mælendur og stjórnendur hittast í tveggja klukkustunda lotum. Við höfum að mestu bókað hljóðverstíma í samfelldum lotum, mælandi 1 09-11 og mælandi 2 11-13. Mælandinn og stjórnandinn fara í gegnum 50 setningar í senn og senda svo til vefþjóns.

Mikil vinna hefur farið í skipulagningu og umsjón þessara tíma, og þarfnast því viðhalds hugbúnaðar upptaka og vinnu við þróun talgervilslíkansins sjálfs. Við munum því endurskoða

dreifingu forða á meðan verkefnið stendur yfir til að tryggja fyllilega afhendingu allra þátta talgervilsverkefnisins.

Gögnin hafa ekki enn verið útgefin og eru hýst á vél hjá HR.

Lesið meira um LOBE í handbókinni hér: https://github.com/cadia-lvl/LOBE/blob/master/other/LOBE_manual.pdf

T7 - Forskriftir að röddum

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Tilgangur

Að gefa út opnar (e. open-source) forskriftir hugbúnaðar til þróunar einingavals- og stikaðra talgervla. Í byrjun munu sérfræðingar vinna með raddupptökur sem eru þegar aðgengilegar (en ekki endilega opnar). Eftir því sem vinnu við upptökur nýrra radda miðar áfram (6.1 og 6.2), munu þær vera notaðar við þróun.

Varða 1 - lýsing

Frumgerð forskriftar fyrir stikaðan talgervil tilbúin.

Afurðir

Frumgerð forskriftar fyrir stikaðan talgervil er tilbúin og hefur verið prófuð á mismunandi gögnum.

Skýrsla

Við höfum notað Ossian (https://github.com/CSTR-Edinburgh/Ossian) til að búa til forskriftirnar sem eru byggðar á Merlin (https://github.com/CSTR-Edinburgh/merlin). Merlin er SPSS tauganetskerfi frá CSTR deildinni við Háskólann í Edinborg (University of Edinburgh). Forskriftin er að stórum hluta byggð á forskrift sem kemur sem hluti af Ossian og notar bókstafi sem einingar hljóðlíkans. Hún notar tauganet fyrir hljóð- og lengdarlíkön. Við höfum bætt við Sequitur G2P líkani fyrir hljóðritun.

Forskriftin hefur verið prófuð öðru hverju, en þar sem gögnin eru enn að streyma inn eru niðurstöður í samræmi við það. Við erum hinsvegar að sjá aukin afköst þegar magn gagna hefur aukist. Vandamál við lendgarlíkanið byrjaði líklega vegna þagna í byrjun hverrar upptöku sem eru mislangar.
Við stefnum á að gefa út einhverskonar prófunartól til að veita öðrum aðilum aðgang að talgervilslíkaninu en það er hýst eins og er á https://github.com/arnacadia.

T12 - Mynstur og setningar

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

Tilgangur

Að afhenda stóra leslista sem ná yfir nægjanlega margar tvístæður til að geta þróað kerfi einingavals talgervla fyrir íslensku. Einnig, að afhenda hugbúnað sem hægt er að nota til að búa til slíka lista og finstilla yfirgrip tvístæða í listum af mismunandi stærðum, t.d. fyrir 1, 4, 10, eða 20 klukkustundir lesturs.

Varða 1 – lýsing

Aðferð til þess að framleiða leslista fyrir einingavalstalgervla með sjálfvirkum hætti og leslisti tilbúinn.

Endurskoðuð varða 1:

Lítill (1 klukkustund), miðlungs (10 klukkustundir), og stór (20 klukkustundir) leslisti útgefinn samkvæmt stöðlum CLARIN. Tól til myndunar sambærilegra lista afhent samkvæmt viðmiðunarreglum hugbúnaðarþróunar verkefnisins.

Afurðir

Öllum afurðum hefur verið skilað nema að gögnin eru útgefin á einföldu textaformi en ekki samkvæmt stöðlum CLARIN, sem enn hafa ekki verið formfestir. Leslisti um það bil 24400 setninga hefur verið útbúinn auk hugbúnaðar til söfnunar gagna frá mörgum stærri gagnasettum og val setninga byggt á umfangi tvístæða.
Hugbúnað og leslista má finna hér: https://github.com/cadia-lvl/tts_data

Skýrsla

Stór textasöfn eru aðgengileg til framleiðslu leshandrita. Við erum hinsvegar að stefna á að finna lítinn fjölda setninga sem henta best fyrir talgervla leslista. Gráðugt algrím er notað við val á setningum. Það er ekki raunhæft að keyra slíkt algrím á öll textagögn sem við höfum aðgang að og einnig óþarft. Því var í byrjun safn um það bil 500K setninga dregið úr eftirfarandi gagnasettum:

Gagnasett	Auðkenni
Hljóðbókasafnið (*)	hlbs
Risamálheild	visir
Risamálheild	silfuregils
Risamálheild	althingi
Risamálheild	school essays
Risamálheild	domstolar
Risamálheild	ras1_og_2
Risamálheild	haestirettur
Risamálheild	wikipedia
Risamálheild	stod2
Sjaldgæfar þrístæður frá Eiríki Rögnvaldssyni (**)	eiki
Gullstaðall	written-to-be-spoken
Gullstaðall	books
Gullstaðall	sjonvarpid
Risamálheild	bylgjan
Risamálheild	ras1
Risamálheild	ruv
Risamálheild	althingislog

(*): Setningar sem eru í TTS gagnasetti fyrir talgervla úr upptökum frá Hljóðbókasafni Íslands. Tengiliðir: annabn@ru.is, jg@ru.is, alfred@hbs.is, einar@hljodbokasafn.is

(**): Listi sjaldgæfra þrístæða var saminn af Eiríki Rögnvaldssyni og var hafður með sem prófunarlisti fyrir gerð þessa gagnasetts.

Að auki, með því að bera kennsl á vantanir eða fátíð tilfelli tvístæða, voru nýjar tilbúnar setningar hafðar með í listanum. Þessar setningar eru merktar með “from_rare” eða “bin_extract”. Fjórum öðrum flokkum var svo bætt við seinna. Flokkurinn “number” inniheldur setningar með mikið af tölum, flokkurinn “short” inniheldur eins orðs setningar, flokkurinn “spell” inniheldur ýmis stafsetningar orða, og að lokum inniheldur flokkurinn “books_long” 15-25 orða setningar frá “books” hluta Gullstaðalsins.

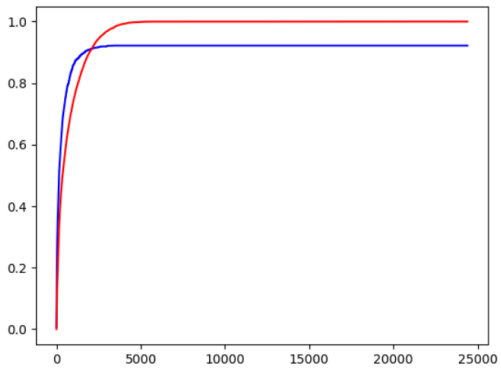
Forþjálfað Sequitur g2p (e. Grapheme-2-Phoneme) líkan var notað til að meta líkindi fóna og umritun er viðhengd hverri setningu í listanum. Hver umritun er byggð úr hvaða fjölda sem er af IPA fónatáknum og tóka þagnar “~” sem er forskeytt og viðskeytt í hverja umritun.

Listinn var hreinsaður skilvirknislega og aðeins orð sem koma upp í BÍN eru leyfð, ein undantekning er listi þrístæða frá Eiríki Rögnvaldssyni sem var óhreyfður og inniheldur orð sem ekki finnast í BÍN. Engir tölustafir eru leyfðir heldur. Allar setningar verða að byrja með hástaf og hver setning er á bilinu 5 til 15, nema úr flokknum “books_long” flokknum “short”.

Listinn er uppraðaður með takmarksaðgerð sem hámarkar umfang tvístæða og jafnframt lágmarkar lengt texta. Takmarksaðgerðinni má lýsa með eftirfarandi formúlu

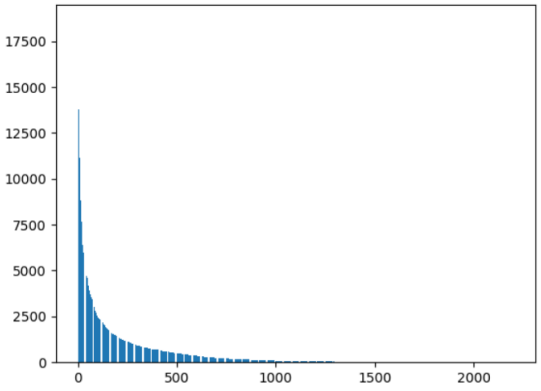
$$R(s) = ||s|| \sum_{i=1}^{k} \frac{1}{\rho(//s_i//)}$$

þar sem s er setning með k bókstöfum og $\rho(//s_i//)$er fjöldi tilvika $//s_i//$ tvístæða í leslistanum enn sem komið er. Vegna þessa má draga marga mislanga leslista úr heila handritinu.

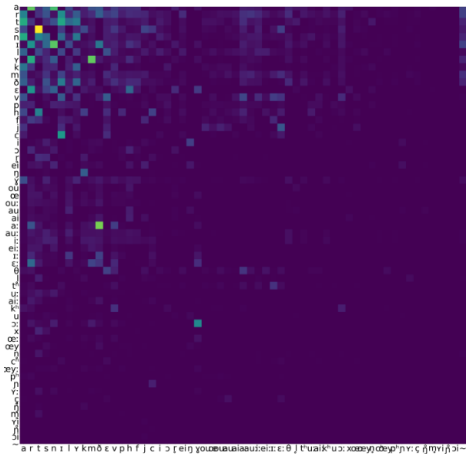


Línuritið að ofan sýnir umfang tvístæða á hverju skrefi innfærslu setninga. Bláa línan sýnir umfang með tilliti til allra mögulega umraðana fóna á meðan rauða línan sýnir umfang aðeins með tilliti til samsetninga tvístæða í leslistanum. Mikilvægt er að taka fram hér að margar tvístæður finnast ekki í íslensku sem útskýrir hvers vegna umfangið nær ekki, og ætti ekki að ná 100%. Það er einnig mikilvægt að taka fram hér að fullt *umfang* í þessu samhengi myndi þýða að minnsta kosti 20 tilvik fyrir hverja tvístæðu. 20 var einfaldlega valin vegna lengdar leslistans, á meðan smærri tala (segjum 5) myndi duga fyrir einingavals talgervla.

Allur leslistinn nær yfir um það bil 81% tvístæða með tilliti til 20 tvístæða af hverri tegund, á meðan hann inniheldur að minnsta kosti eitt tilvik um það bil 87.3% allra tvístæða. Línuritið að neðan sýnir tíðni tvístæða í öllum leslistanum.



Myndin að neðan sýnir fón-fón hitakort fyrir allan leslistann.



Ef við metum sem svo að meðal setning samsvari 5 sekúndum tals þá gætu fyrstu 720 setningarnar myndað 1 klukkustundar gagnasett. Miðstærð gagnasetts 10 klukkustunda myndi innihalda fyrstu 7200 setningarnar og 20 klukkustunda gagnasettið fyrstu 14400. Taka skal fram að fyrstu 11887 setningarnar eru setningar sem eru einhvers virði samkvæmt gráðugu algrími. Allar setningar eftir það bæta ekki tvístæðu í listann sem hafa ekki birst að minnsta kosti 20 sinnum áður. Þessir listar eru aðgengilegir hér. Nánari skjöl er einnig að finna í veitunni. Það er líklegt að einhver vinna verði gerð í framtíðinni úr þessari sömu veitu.