Endurskoðuð markaskrá fyrir málfræðilega mörkun íslenskra texta

Starkaður Barkarson, Einar Freyr Sigurðsson, Eiríkur Rögnvaldsson, Hildur Hafsteinsdóttir, Hrafn Loftsson, Steinþór Steingrímsson, Þórdís Dröfn Andrésdóttir
29. maí 2020

1 Inngangur

Þetta skjal lýsir endurskoðun á markaskrá fyrir málfræðilega mörkun íslenskra texta sem fram fór veturinn 2019–2020 og gerir grein fyrir því hvernig staðið var að endurskoðun íslensku gullstaðlanna, Orðtíðnibókar og MÍM-GULLs, með tilliti til endurskoðaðrar markaskrár.

Samhliða útgáfu markaskrárinnar eru þessar endurskoðuðu gerðir af þessum tveimur íslensku gullstöðlum gefnar út og hafa útgáfunúmerin MIM-GOLD 20.05; OTB 20.05 og MIM-OTB 20.05 fyrir pakka þar sem gullstaðlarnir hafa verið sameinaðir í einn. Útgáfunúmerin eru í samræmi við leiðbeiningar um útgáfunúmer efnis sem gefið er út innan máltækniáætlunar fyrir íslensku 2019-2023. Gullstaðlarnir eru samtímis gefnir út með skiptingu í tíu hluta, til nota við þjálfun og prófanir á málfræðimörkurum.

2 Fyrri breytingar á markamenginu og vinna við MÍM-GULL

Endurskoðað markamengi byggir á markaskrá sem var gerð fyrir Íslenska orðtíðnibók árið 1991 (Jörgen Pind, Friðrik Magnússon og Stefán Briem, 1991). Þeirri markaskrá hefur áður verið breytt, eins og gerð er grein fyrir hér að neðan. Við þá endurskoðun sem fram fór fyrir útgáfu 20.05 var gengið nokkuð lengra. Til að mynda voru skammstafanir og styttingar settar í sérstakan flokk. Nýr flokkur var búinn til fyrir tákn, sem eru orðin algengari í ritmáli, sérstaklega á netinu. Breytingar voru gerðar á því hvernig erlend orð eru greind o.fl. Nánari grein er gerð fyrir breytingunum í kafla 3.

Mörkuð íslensk málheild (MÍM) var gefin út árið 2013. Málheildin hefur að geyma um 25 milljónir orða af textum sem voru ritaðir á fyrsta áratug 21. aldar.

Á vinnslustigi MÍM var tekið úrtak með um einni milljón lesmálsorða úr 13 mismunandi textaflokkum af 23 textaflokkum í MÍM. Nýja málheildin átti að koma í staðinn fyrir textasafn Íslenskrar orðtíðnibókar sem gullstaðall fyrir þjálfun námfúsra markara fyrir íslensku.

Árið 2013 var veittur aðgangur að útgáfu 0,9 af Gullstaðlinum. 2018 var veittur aðgangur að útgáfu 1,0. Hér er gerð grein fyrir því hvernig Gullstaðallinn var þróaður. Ferlinu er skipt í 5 lotur, tölusettar frá 0 til 4.

Starkaður Barkarson (2017) fjallar í meistarprófsritgerð sinni um áhrif þess að greina erlend sérnöfn sem e og einnig að nauðsynlegt sé að einfalda greiningu á greinarmerkjum.

2.1 Lota 0

Sumarið 2009 fékkst styrkur frá Nýsköpunarsjóði námsmanna til að ráða stúdent til þess að hefja verkið. Stúdentinn vann á vegum Hrafns Loftssonar í Háskólanum í Reykjavík. Á Stofnun Árna Magnússonar í íslenskum fræðum var tekið úrtak úr MÍM sem stúdentinn síðan vann úr. Textunum var fyrst skipt upp í setningar og lesmálsorð með tilreiðara sem er hluti af IceNLP-hugbúnaðinum. Síðan voru textarnir markaðir með fimm mörkurum: fnTBL, MXPOST, IceTagger, Bidir og TnT (Hrafn Loftsson o.fl., 2010). Tólið CombiTagger var svo nýtt til að kjósa á milli markaranna. Notuð var aðferð þar sem valið var það mark sem flestir markarar velja fyrir hvert orð. Markararnir voru þjálfaðir á textasafni Íslenskrar orðtíðnibókar. Markamengi Orðtíðnibókarinnar liggur því til grundvallar mörkuninni.

Veturinn 2009–2010 hófst leit að kerfisbundnum villum í Gullstaðlinum. Notuð voru villuleitarforrit sem byggðust á því að skoða samræmi í nafnliðum (NP), forsetningarliðum (PP) og sagnliðum (VP) eins og Hrafn Loftsson (2009) hefur lýst. Farið var handvirkt yfir stóran hluta þeirra villna sem forritið benti á og þær leiðréttar ef markið reyndist rangt. Nákvæmni mörkunarinnar var síðan metin með því að skoða um 1% úrtak (hundraðasta hvert orð). Mark var talið rétt ef allir stafir í markinu (allt að 6) voru réttir. Niðurstaðan varð 92,3% nákvæmni að meðaltali en

reyndist á bilinu 87,6–95,5% eftir textaflokkum (Hrafn Loftsson o.fl., 2010). Verkefnið fékk einnig framlag af styrk 0906621123 frá Rannís.

2.2 Lota 1

Sumarið 2010 fékkst annar styrkur frá Nýsköpunarsjóði námsmanna til að ráða stúdent til þess að skoða mörkun á öllum orðum í Gullstaðlinum. Byrjað var á að fara yfir villur sem fundust í lotu 0 en höfðu ekki verið leiðréttar (textar úr Morgunblaðinu). Einnig hófst vinna við að fara yfir texta úr prentuðum bókum. Stúdentinn var síðan ráðinn í hlutastarf á skólatíma og á árunum 2010–2011 var farið handvirkt yfir öll lesmálsorð í Gullstaðlinum og mörk leiðrétt. Útgáfa 0,9 af Gullstaðlinum, sem veittur var aðgangur að 2013, hefur að geyma skrárnar eftir þessa umferð af leiðréttingum. Meðalnákvæmni var metin 96,4% og var á bilinu 89,9–98,5% eftir textaflokkum (Sigrún Helgadóttir, Hrafn Loftsson og Eiríkur Rögnvaldsson, 2014). Verkefnið fékk einnig framlög frá META-NORD5 verkefninu og styrk frá mennta- og menningarmálaráðuneytinu.

2.3 Lota 2

Í lok árs 2012 hófst síðan næsta leiðréttingarlota. Textarnir voru þá markaðir á sjálfvirkan hátt með markaranum IceTagger sem er hluti af IceNLP-hugbúnaðinum. Skrifað var forrit sem bar saman mörk sem IceTagger skilaði og rétt (að því talið var) mörk í málheildinni. Ef ekki var samsvörun voru orðin merkt. Farið var handvirkt yfir þau orð sem þannig voru merkt. Ráðinn var nemandi í fullt starf sumarið 2013 og í hlutastarf á skólatíma til þess að skoða villumerkingarnar. Nemandinn sem fór yfir mörkin fékk fyrirmæli um að (i) velja markið sem var fyrir í málheildinni; (ii) velja markið sem IceTagger lagði til; eða (iii) finna rétt mark þegar bæði markið í málheildinni og markið sem IceTagger lagði til reyndust röng. Þegar farið hafði verið yfir um 80% af textunum var meðalnákvæmni metin 99,6% og var á bilinu 99,5–100,0% (Sigrún Helgadóttir, Hrafn Loftsson og Eiríkur Rögnvaldsson, 2014). Enn einn nemandi var ráðinn seint á árinu 2013 til þess að ljúka yfirferðinni sem síðan lauk árið 2014. Nákvæmni mörkunar var ekki metin með því að skoða úrtak eftir að þessari yfirferð lauk. Leiðréttingavinnan var styrkt að hluta af META-NORD verkefninu og einnig af mennta- og menningarmálaráðuneytinu.

2.4 Lota 3

Árið 2015 gerðu Steinþór Steingrímsson, Sigrún Helgadóttir og Eiríkur Rögnvaldsson tilraun með að þjálfa markarann Stagger (Östling, 2012) á Orðtíðnibókinni og Gullstaðlinum (Steinþór Steingrímsson, Sigrún Helgadóttir og Eiríkur Rögnvaldsson, 2015). Hrafn Loftsson og Robert Östling gerðu árið 2013 tilraun til þess að þróa íslenskan markara með því að þjálfa og prófa Stagger á Íslenskri orðtíðnibók og náðu 93,84% nákvæmni með því að beita tífaldri krossprófun (Hrafn Loftsson og Östling, 2013). Þar sem þetta var besti árangur sem náðst hafði við mörkun íslensks texta fram að því var ákveðið að prófa forritið á Gullstaðlinum. Samanburður á nákvæmni Staggers þegar hann var þjálfaður annars vegar á Orðtíðnibókinni og hins vegar á Gullstaðlinum leiddi þó í ljós að töluvert var enn af villum og ósamræmi í Gullstaðlinum (Steinþór Steingrímsson, Sigrún Helgadóttir og Eiríkur Rögnvaldsson, 2015). Notuð var útgáfa af Gullstaðlinum eftir að handvirkri leiðréttingu var lokið, þ.e. eftir að lotu 2 var lokið. Tilraun Hrafns Loftssonar og Roberts Östling við að þjálfa og prófa Stagger á Orðtíðnibókinni var endurtekin á Gullstaðlinum. Notaðir voru málþættir fyrir íslensku og forritið IceMorphy (hluti af IceNLP-hugbúnaðinum) sem giskar á mörk óþekktra orða. Einnig var bætt við orðasafni sem byggðist á Beygingarlýsingu íslensks nútímamáls (BÍN). Með tífaldri krossprófun fékkst 92,76% meðalnákvæmni fyrir Gullstaðalinn. Í framhaldi af þessari niðurstöðu var ákveðið að vinna frekar að því að leiðrétta og samræma mörk í Gullstaðlinum. Búnir voru til villulistar yfir ósamræmi og stúdentar fengnir til þess að fara yfir þá handvirkt. Einnig var markamenginu breytt lítillega. Þessari vinnu lauk seint á árinu 2017. Þessi hluti verkefnisins hlaut styrki frá Málvísindastofnun Háskóla Íslands og mennta- og menningarmálaráðuneytinu.

2.5 Lota 4

Starkaður Barkarson fékk gögn Gullstaðalsins þegar lotu 3 lauk og þjálfaði Stagger á textunum (Starkaður Barkarson, 2017). Nákvæmni mörkunar hafði ekki verið metin með því að skoða úrtak orða eins og gert var eftir fyrri leiðréttingalotur. Starkaður endurtók tilraun sem Steinþór Steingrímsson, Sigrún Helgadóttir og Eiríkur Rögnvalds-

son höfðu gert árið 2015. Hann framkvæmdi sambærilega tífalda krossprófun og fékk 92,74% nákvæmni fyrir Gullstaðalinn.

Þrátt fyrir lagfæringar á Gullstaðlinum virtist nákvæmnin ekki hækka. Til þess að ganga úr skugga um að tilraunirnar hefðu að öllu leyti verið sambærilegar var tilraun Steinþórs og félaga endurtekin, eftir því sem aðstæður leyfðu. Sami Gullstaðall (fyrir síðustu leiðréttingalotu) var notaður og sama skipting í þjálfunar- og prófunartexta. Gögn BÍN voru ekki að öllu leyti sambærileg þar sem nú var notuð nýrri útgáfa. Með því að nota BÍN og IceMorphy náðist aðeins 92,41% nákvæmni í tilraun Starkaðar, í stað 92,76% í tilraunum Steinþórs og félaga. Starkaður telur því að staðhæfa megi að lagfæringar á Gullstaðlinum hafi leitt til 0,30 prósentustiga aukningar á nákvæmni. Hann telur að orsaka á muninum megi e.t.v. leita í því orða- og endingasafni sem IceMorphy hafði aðgang að því mikill munur er á nákvæmni við greiningu á óþekktum orðum (tæp 15%) en lítill á greiningu þekktra orða (0,09%) (Starkaður Barkarson, 2017).

2.6 Breytingar á markamengi

Til þess að auðvelda málfræðigreininguna og ná meira samræmi var markamengi Orðtíðnibókarinnar breytt lítillega í leiðréttingalotum Gullstaðalsins. Þessar breytingar voru gerðar:

- Erlend nöfn voru upphaflega mörkuð sem sérnöfn en í lotu 3 voru þau mörkuð sem erlend orð (**e**) (Steinþór Steingrímsson, Sigrún Helgadóttir og Eiríkur Rögnvaldsson, 2015).
- Í Orðtíðnibókinni voru smáorð sem stóðu á undan *að* greind sem atviksorð (**aa**). Í lotu 2 voru þau hins vegar greind í Gullstaðlinum sem forsetningar ef á eftir fer fallsetning. Þannig er smáorðið *til* í setningunni „Hann hljóp til að komast fyrr heim“ greint sem forsetning sem stýrir eignarfalli (**ae**) (Sigrún Helgadóttir, Hrafn Loftsson og Eiríkur Rögnvaldsson, 2014; Steinþór Steingrímsson, Sigrún Helgadóttir og Eiríkur Rögnvaldsson, 2015; Starkaður Barkarson, 2017).
- Nánari flokkun á sérnöfnum var lögð af í lotu 3 þannig að mörk allra sérnafna fá nú mörkunarþáttinn **-s** í stað **-m** (mannanöfn), **-ö** (örnefni) og **-s** (önnur sérnöfn) eins og áður. Mögulegum mörkum fækkar þannig um 68 (Steinþór Steingrímsson, Sigrún Helgadóttir og Eiríkur Rögnvaldsson, 2015).
- Í lotu 3 var **v** tekið upp fyrir vefföng og tölvupóstföng (Steinþór Steingrímsson, Sigrún Helgadóttir og Eiríkur Rögnvaldsson, 2015).
- Í lotu 3 var **as** tekið upp fyrir skammstafanir, en áður voru skammstafanir brotnar upp og hver stafur greindur eins og um orð væri að ræða (Steinþór Steingrímsson, Sigrún Helgadóttir og Eiríkur Rögnvaldsson, 2015)
- Í lotu 3 var ákveðið að öll tölugildi sem rituð eru með tölustöfum og voru áður greind sem frumtölur (**tf...**) fái markið **ta** og verði ekki greind frekar í kyn, tölu og fall eins og gert er þegar tölugildi eru rituð með bókstöfum (Steinþór Steingrímsson, Sigrún Helgadóttir og Eiríkur Rögnvaldsson, 2015).

3 Breytingar á markamengi í þessari útgáfu

Í þessari útgáfu voru fleiri breytingar gerðar á markamengi, eins og fram kom í inngangi. Gerð er grein fyrir þeim hér.

3.1 Forsetningar

- **ao**, **aþ** og **ae** verða **af** (atviksorð sem stýrir falli).

3.2 Sagnorð

- Það sem var greint sem sagnbót (**ssg** eða **ssm**) er nú greint sem sögn(lh.)-þt.-hk.-et.-nf. (**sþghen** eða **sþmhen**).

3.3 Nafnorð

- Tákn fyrir ókyngreint (**nx**...) var fjarlægt út úr mörkum fyrir nafnorð. Í staðinn notum við - (**n-**...).

3.4 Erlend orð

- Erlend sérnöfn verða greind sem **n----s**, þ.e. nafnorð sem ekki eru greind í kyn, tölu eða fall og hafa ekki greini.
- Erlendar skammstafanir verða greindar með sama hætti ef þær eru ígildi sérnafns (t.d. *CIA*, *NATO*, *LFC* og *KFC*).

3.5 Skammstafanir og styttingar

- Skammstafanir og styttingar eru í sérstökum flokki þar sem greiningarstrengurinn byrjar á **k**.
- Eiginlegar skammstafanir eru markaðar með **ks**. En ef skammstöfun er ígildi nafns (t.d. *KR*, *VR* eða *ÓRG*) þá er það greint sem **n----s**. Það gildir hvort sem skammstöfun er erlend eða íslensk (sjá ofar).
- Styttingar eru markaðar með **kt**. Styttingar eru t.d. *lögg.* (fyrir *löggiltur*) eða *hæstv.* (fyrir *hæstvirtur*), þ.e. þar sem orð er fyrri hluti orðs auk punkts. Einnig er *grunn-* í *grunn- og framhaldsskólar* greint sem **kt**.

3.6 Greinarmerki

- **pl**, lok setninga: .!? (alltaf)
- **pk**, komma: , ; (nema ef , er notað sem gæsalöpp)
- **pg**, gæsalappir: « » „ “ ” ” , ‘ (nema ef , er komma eða ’ er notað sem úrfellingamerki)
- **pa**, önnur greinarmerki: ( ) { } _ : - – — … (ásamt öllum greinarmerkjum sem ekki falla í ofangreinda flokka)

3.7 Mörg greinarmerki í röð

- Tveir eða fleiri punktar (t.d. …), spurningarmerki (t.d. ??) eða upphrópunarmerki (t.d. !!!) í röð eða samsetning upphrópunarmerkja og spurningarmerkja (t.d. !?!) er greint saman sem **pa**.
- Önnur greinarmerki eru slitin í sundur og hvert og eitt greint.

3.8 Tákn

- Öll tákn mörkuð sem **m**
  - stærðfræðitákn: + − × ÷ = < > [ ]
  - tjákn: :) ♥
  - önnur tákn: $ % § © •
- Tákn mætti þá skilgreina sem flest það sem hvorki inniheldur staf né tölu og er ekki greinarmerki – nema þá ef nokkur greinarmerki í röð mynda tjákn.
- Tákni má skipta út fyrir orð (t.d. $ = dollari, + = plús).

4 Ný útgáfa gullstaðla

Eftir endurskoðun markamengisins var farið yfir tvo gullstaðla, Orðtíðnibókina og MÍM-GULL. Í kafla 5 er gerð grein fyrir ákvörðunum sem voru teknar þegar vafaatriði komu upp. Mögulegt er að mynda yfir 600 mismunandi greiningarstrengi með markaskránni, en í MÍM-GULL 20.05 og OTB 20.05, tveimur gullstöðlum sem báðir hafa verið markaðir með þessari endurskoðuðu markaskrá, er 571 mismunandi greiningarstrengur notaður (556 í OTB og 558 í MÍM-GULL).

ABLTagger, tauganetsmarkari sem fram að þessu hefur náð mestri nákvæmni í málfræðilegri mörkun íslensks texta, var þjálfaður á gullstöðlunum eftir að mörkum var breytt til samræmis við nýja markamengið. Mörkunarnákvæmnin er aðeins meiri en þegar eldri markaskrá var notuð. Gerð er grein fyrir mörkunarnákvæmni markarans í kafla 6.

5 Leiðbeiningar við mörkun gullstaðla

Í þessum kafla er því lýst hvaða ákvarðanir voru teknar við mörkun gullstaðlanna. Til að gæta samræmis í mörkun voru allar ákvarðanir skráðar. Með því að fylgja þeim ákvörðunum geta aðrir sem koma hugsanlega til með að marka aðra texta síðar verið í sem mestu samræmi við málheildirnar sem koma út núna.

5.1 Skammstafanir og styttingar

5.1.1 Skammstöfun eða stytting?

- Ef orðmynd stendur fyrir aðeins eitt orð og er fyrri hluti þess (t.d. *lögg.* = *löggiltur*, *hæstv.* = *hæstvirtur*), eða ef orð er samsett og það stafsett með byrjun tveggja eða fleiri hluta (t.d. *lög.stj.* = *lögreglustjóri*, *framkv.stj.* = *framkvæmdastjóri*) og orðmynd samanstendur af þremur eða fleiri stöfum, þá er um styttingu að ræða en ekki skammstöfun.
- Þannig eru *lögg.*, *hæstv.*, *lög.stj.* og *framkv.stj.* styttingar.

5.1.2 Erlendar skammstafanir sem ekki eru sérnöfn

- Erlendar styttingar og skammstafanir (sem standa ekki fyrir sérnöfn) eru **e**.

5.2 Íslensk sérnöfn

- Í fleiryrtum sérnöfnum er aðeins fyrsta orðið merkt sem sérnafn. Dæmi um þetta er *Bóksala stúdenta*:

  Bóksala **nven-s**
  stúdenta **nkfe**

5.3 Erlend sérnöfn

5.3.1 Almennar reglur

- **Nöfn persóna:** eru alltaf greind sem **n----s** (t.d. eru allir hlutar í nafni knattspyrnumannsins Alessandro Del Piero markaðir með **n----s**).
- **Nöfn staða:** eru alltaf greind sem **n----s**.
- **Einyrt heiti stofnana og fyrirtækja:** eru greind sem **n----s**.
- **Löng, fleiryrt sérnöfn:** Almennt gildir sú regla að fyrsta orðið er **n----s** og rest **e**, nema þau sem eru sérnöfn að eigin verðleikum. Nöfn persóna og staða falla ekki undir þetta, sbr. hér að ofan.

5.3.2 Sérstök tilvik

- Heiti undirtegunda bíla (t.d. *Skoda* ***Suberb***, *VW* ***Passat***, *Renault* ***Megane***) eru greind sem **n----s**, enda eru þau mikið notuð ein og sér og eru þar með sérnöfn að eigin verðleikum.
  - Ef eitthvað bætist á eftir undirheitinu (t.d. *Renault Megane* ***Saloon***) er það greint sem **e**.
- Nöfn erlendra íþróttaliða sem eru kennd við borgir eða staði (t.d. *Los Angeles* ***Lakers***, *New York* ***Knicks***, *Utah* ***Jazz***) eru mörkuð með *n----s*, enda eru þau notuð mikið ein og sér og þar með sérnöfn að eigin verðleikum.
- Erlendir titlar (t.d. ***Dame*** *Judi Dench*, ***Mr*** *Feather*, ***Major*** *Miriam Óskarsdóttir*) eru greindir sem **e** þar sem þeir eru ekki beinlínis hluti af sérnafninu sem fylgir á eftir. Hins vegar eru titlar af sama tagi **n----s** þegar þeir eru fyrsti hluti sérnafns, eins og kvikmyndatitils (t.d. ***Mr*** *Deeds*, ***Mrs*** *Doubtfire*).
- Erlend, fleiryrt nöfn fyrirtækja, hljómsveita, bókatitlar, kvikmyndatitlar, titlar á listaverkum, ráðstefnum og hátíðum eru mörkuð á þennan hátt:

Y **n----s**

tu **e**

mamá **e**

también **e**

- Ef eiginleg sérnöfn eru inni í runum af þessu tagi fá þau samt greininguna **n----s**.

5.4 Íslenskt eða erlent?

- Erlend orð sem eru löguð að íslenskri beygingu (t.d. *Steinwayinum* og *Bösendorferinn*) eru greind sem íslensk orð:

  Steinwayinum **nkeþgs**

  Bösendorferinn **nkengs**

5.5 Annað

- Tákn: **m** eða **pa**
  - Ef hefðbundið greinarmerki stendur fyrir orð er það **m**.
    Dæmi: '-' í *frá klukkan 13-15*, '/' í *km/klst.*

6 Sjálfvirk mörkun með nýju markamengi

Við þjálfuðum ABLTagger (Steinþór Steingrímsson, Örvar Kárason og Hrafn Loftsson, 2019) með gullstöðlunum tveimur, hvorn í sínu lagi og saman, eftir að mörkin voru uppfærð. Tafla 6 sýnir nákvæmnina sem markarinn náði fyrir hverja þjálfunarmálheild fyrir sig, við tífalda krossprófun.

Gullstaðall	Nákvæmni	Þekkt orð	Óþekkt orð
OTB	95.54%	95.86%	53.95%
MÍM-GULL	94.47%	95.58%	66.64%
MÍM-GULL + OTB	95.09%	95.93%	63.96%

Meiri ónákvæmni í óþekktum orðum í Orðtíðnibókinni skýrist af því að markarinn notar BÍN og flest orð í Orðtíðnibókinni eru í BÍN. Því eru fá óþekkt orð þar og þau sem eru þar líklega þess eðlis að erfitt getur verið fyrir markara að átta sig á því í hvaða flokk þau falla.

Tilvísanir

Hrafn Loftsson. 2009. Correcting a POS-tagged corpus using three complementary methods. *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, bls. 523–531. Association for Computational Linguistics, Athens, Greece. https://www.aclweb.org/anthology/E09-1060.

Hrafn Loftsson og Robert Östling. 2013. Tagging a morphologically complex language using an averaged perceptron tagger: The case of Icelandic. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, bls. 105–119. Linköping University Electronic Press, Sweden, Oslo, Norway. https://www.aclweb.org/anthology/W13-5613.

Hrafn Loftsson, Jökull H. Yngvason, Sigrún Helgadóttir og Eiríkur Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*. LREC 2010, Valetta, Malta. https://www.isca-speech.org/archive_open/saltmil/SALTMIL2010_Proceedings.pdf#page=57.

Jörgen Pind, Friðrik Magnússon og Stefán Briem (ritstj.). 1991. *Íslensk orðtíðnibók*. Orðabók Háskólans, Reykjavík.

Sigrún Helgadóttir, Hrafn Loftsson og Eiríkur Rögnvaldsson. 2014. Correcting Errors in a New Gold Standard for Tagging Icelandic Text. *Proceedings of the* 9^{th} *International Conference on Language Resources and Evaluation*. LREC 2014, Reykjavík, Iceland. http://www.lrec-conf.org/proceedings/lrec2014/summaries/677.html.

Starkaður Barkarson. 2017. Þjálfun málfræðimarkarans *Stagger* með nýjum gullstaðli. MA-ritgerð, Háskóla Íslands, Reykjavík. http://hdl.handle.net/1946/29474.

Steinþór Steingrímsson, Sigrún Helgadóttir og Eiríkur Rögnvaldsson. 2015. Analysing Inconsistencies and Errors in PoS Tagging in two Icelandic Gold Standards. *Proceedings of the* 20^{th} *Nordic Conference of Computational Linguistics*, bls. 287–291. NODALIDA 2015, Vilnius, Lithuania. https://www.aclweb.org/anthology/W15-1838.

Steinþór Steingrímsson, Örvar Kárason og Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. *Proceedings of the International Conference Recent Advances in Natural Language Processing*. RANLP 2019, Varna, Bulgaria.

Östling, Robert. 2012. Stagger: A modern POS tagger for Swedish. *Proceedings of the Swedish Language Technology Conference, SLTC*. Lund, Sweden.

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**