

Máltækni fyrir íslensku 2019-2023

Yfirlit verkþátta við Vörðu 2, 1. júní 2020

SÍM-hópurinn hefur nú farið yfir stöðu verkþátta gagnvart verkefnisvörðu 2 þann 1. júní 2020. Verkefnahóparnir þurftu, eins og aðrir, að endurskipuleggja vinnu sína og að yfirstíga ýmsar hindranir vegna Covid-19 ástandsins. Flest verkefni eru samt sem áður á áætlun, helst bitnuðu lokanir og samkomubann á verkefnum sem fela í sér upptökur í stúdíói og annars staðar, þar sem starfsmenn þurfa að vera viðstaddir og stýra upptökum.

Eftirfarandi samantekt er ætlað að gefa yfirlit yfir stöðu allra verkþátta samkvæmt tæknilyðingu fyrir fyrsta ár máltækniáætlunar (*M2* merkir *Varða 2*).

Málföng

Verkefni innan málfanga tengjast texta- og orðasöfnum sem nýtast í máltækni. Langflestir verkþættir eru unnir hjá Stofnun Árna Magnússonar að undanskildum verkþætti *Framburðarorðabók*, sem unninn er hjá Grammateki ehf.

Risamálheild

Risamálheildin (RMH) inniheldur mikið magn texta af ýmsum sviðum, til notkunar í máltækni. Því stærri sem málheildin er og því fjölbreyttari sem textarnir eru, því betur nýtist málheildin í þróun almennra máltæknitóla.

Textarnir í málheildinni eru markaðir og lemmaðir, þ.e. hvert orð er merkt með upplýsingum um orðflokk, fall, o.s.frv. og hverju orði fylgir grunnform (*lemma*). Hluti þróunarvinnunnar felst í að uppfæra þessar upplýsingar á sjálfvirknan hátt, eftir því sem gæði sjálfvirkrar mörkunar og lemmunar aukast. Til þess að sannreyna að uppfærsla auki nákvæmni í mörkun, eru prófunargögn útbúin, sem farið er handvirkt yfir.

M2: Málheildin hefur verið mörkuð og lemmuð með nýjustu tólum. Ný útgáfa er komin út, IGC-2019 (IGC: Icelandic Gigaword Corpus). Búið er að fara handvirkt yfir um helming prófunargagna.

Staða: Uppfærð Risamálheild, IGC-2019 hefur verið gefin út á CLARIN og er einnig leitarbær á malheildir.arnastofnun.is. Málheildin inniheldur nú 1.550 milljón orð og hefur verið stækkuð um 115 milljón orð frá síðustu útgáfu. Um 40% prófunarmálheildar hafa verið yfirfarin. – Verkefni á áæltun

Beygingarlýsing íslensks nútímamáls (BÍN)

Beygingarlýsingin inniheldur beygingardæmi íslenskra orða, eins og er hátt í 300 þúsund uppflöttiorða. Verkefni máltækniáætlunar er að gera BÍN enn betur nýtilegri í máltækni, m.a. með því að koma gagnagrunninum á hentugra form og auka við upplýsingar sem fylgja hverju orði/orðmynd.

M2: Skýrsla um margærðar orðmyndir í BÍN tilbúin. Forritunarskil (API) og vefþjónusta uppsett. Gögn útgefin á texta- og/eða pökkuðu formi undir opnu leyfi.

Staða: Tímaröð verkefna var að hluta til breytt: í stað skýrslu um margræðar orðmyndir í BÍN var skýrsla um orðhlutagreiningu unnin fyrir M2. Orðhlutagreiningu átti að skila við M3 og verður skýrslan um margræðar orðmyndir þess í stað skilað í lok verkefnisárs. Að öðru leyti stenst Varða 2.

Orðskiptingar

Töl sem skiptir orðum sjálfvirkt milli lína samkvæmt stafsetningarreglum. Verkefnið er tiltölulega ný hafið og fer að mestu fram á seinni hluta verkefnisárs.

M2: Reglur um orðskiptingar yfirfarnar og gloppur skráðar. Nauðsynlegar viðbætur við reglur hafa verið skrifaðar. – *Staða:* Verkefni á áætlun.

Framburðarorðabók

Verkefni framburðarorðabókar er þríþætt: að gefa út endurbætta íslenska framburðarorðabók, að þróa hugbúnað og viðmót til þess að vinna með orðabókina, og að þróa sjálfvirkan hljóðritunarþúnað sem nýtist m.a. í talgervingu.

M2: Fyrsta útgáfa sjálfvirks hljóðritunarþúnaðar hefur verið gefin út. Búið er að skilgreina lokaútgáfu orðabókarfórritsins.

Staða: Síðan í febrúar hefur frumgerð orðabókarfórritsins verið í notkun við yfirferð gagna. Nokkrar aðferðir við sjálfvirka hljóðritun eru klárar, útgefið forrit byggir á handskrifuðum hljóðritunarreglum. Þróun orðabókarfórrits heldur áfram í samvinnu við starfsfólk sem vinnur með það. – Verkefni á áætlun.

Íslenskt orðanet

Íslenskt orðanet inniheldur orð og orðasambönd og tengir þau við hugtök og merkingarflokka. Orðanetið var ekki hannað fyrir notkun í máltækni og felst því vinnan innan máltækniáætlunar að mestu í því að varpa því á það form sem nýtist hugbúnaði, sem og að fylla upp í gloppur sem kunna að vera í netinu.

M2: Orðanetinu hefur verið varpað á gagnaform sem nýtist í máltækni. Skýrsla um sjálfvirkar aðferðir til þess að greina gloppur í gagnagrunninum tilbúin. – Verkefni á áætlun

Gullstaðall

Gullstaðallinn – eða *Gullið* – er mörkuð málheild sem er handyfirfarin og nýtist því til prófunar á mörkunartólum.

M2: Uppfærður Gullstaðall útgefinn. Aðrar málheildir hafa verið markaðar með nýju mörkunarsetti.

Staða: Hluti verkefnisins fólst í því að uppfæra markamengi fyrir íslensku: hvaða tákn á að nota til þess að merkja orðflokka, föll, o.fl. og hvaða eiginleika á að merkja. Gullstaðallinn

hefur verið uppfærður með tilliti til þessa nýja markamengis. – Verkefni á áætlun og **verkefni lokið**

Stoðtöl

Hugbúnaður sem framkvæmir grunngreiningu á texta fellur undir stoðtöl. Greining getur verið allt frá því að skipta texta niður í orð upp í að skila djúpum setningafræðilegum greiningum. Verkefni innan stoðtóla eru unnin af Miðeind ehf. og Stofnun Árna Magnússonar.

Orðtökutól

Hugbúnaður sem fæst við að safna orðum úr textum sem ekki eru fyrir hendi í orðasöfnum eins og BÍN.

M2: Þarfalýsing, úttaksformat og skilgreining á aðferðum tilbúið.

Staða: Vinna við verkefnið ný hafin, verkefni mun ljúka á áætlun við M3

Textatilreiðari (tokenizer)

Textatilreiðari skiptir texta upp í orð eða *tóka* og málsgreinar. Textatilreiðing er forsenda fyrir allri frekari greiningu, en nákvæm tilreiðing þarf til að mynda að geta túlkað greinarmerki rétt, þannig að t.d. punktar í skammstöfunum verði ekki til þess að texta sé skipt í málsgreinar á þeim stöðum.

M2: Samantekt um jaðartilvik tilbúin. Tilreiðarinn ræður við flest jaðartilvik, gerð er grein fyrir þeim sem hann nær ekki utan um. Tilreiðarinn skal ráða við dæmigerða fréttatexta af nokkru öryggi, og er prófaður á slíkum textum. Takmarkið er að tilreiðarinn nái 96%-98% nákvæmni. Búið er að ganga frá hugbúnaði til útgáfu í samræmi við staðla verkefnisins.

Staða: Textatilreiðarinn hefur verið gefinn út á CLARIN í samræmi við markmið. – Verkefni á áætlun og **verkefni lokið**.

Markari

Málfræðilegur markari merkir hvert orð með upplýsingum um orðflokk og þeim þáttum sem við eiga, svo sem falli, kyni og tölu fyrir nafnorð. Í verkefninu hafa þrír markarar fyrir íslensku verið skoðaðir og markmiðið er að velja þann besta til áframhaldandi þróunar.

M2: Prófunarsett fyrir mismunandi textategundir/svið hafa verið undirbúin skv. nýju markamengi (sjá *Gullstaðall*). Prófanir hafa verið keyrðar á Gullstaðlinum og villur greindar og flokkaðar. Byrjað er að prófa á mismunandi textategundum.

Staða: Prófunarsett fyrir mismunandi textategundir hafa verið undirbúin og prófanir með þremur mörkurum keyrðar. Ítarlegar villugreiningar hafa verið gerðar til þess að leiða áframhaldandi þróun. Prófanir á mismunandi textategundum verða framkvæmdar í sumar. – Verkefni á áætlun, fyrir utan prófanir á mismunandi textategundum. Prófunarsettin eru klár og prófanir að fara í gang.

Þáttari

Þáttari framkvæmir enn frekari málfræðilega greiningu á mörkuðum texta. Innan verkefnisins er unnið með tvær gerðir þáttara, annars vegar þáttara sem skilar yfirborðsgreiningu, þ.e. greinir hvaða orð eiga saman innan setningar, án þess að greina heildarstrúktur, og hins vegar djúppáttara, sem greinir setningar til fulls og skilar stigveldi.

M2: Búið er að prófa og meta þáttarana og gefa út *grunnlínu* (baseline), þ.e. nákvæmnistölur sem leitast verður við að bæta í áframhaldandi þróun. Forgangsraðaður verkefnalisti fyrir næstu skref í þróun tilbúinn. Byrjað er að bæta þáttarana samkvæmt þessum lista.

Staða: Þáttararnir *IceParser* (hlutapáttari) og *Greynir* (djúppáttari) hafa verið prófaðir og metnir. Grunnlína fyrir *IceParser* er 78,13% *F-gildi* og fyrir *Greyni* 75,17%. Forgangslistar fyrir hvorn þáttara fyrir sig hafa verið settir saman og áframhaldandi þróun samkvæmt þeim hafin. – Verkefni á áætlun

Vélpýðingar

Innan máltækniáætlunar er unnið að sjálfvirkum þýðingum milli íslensku og ensku. Stór hluti verkefnisins, sérstaklega nú í byrjun, felst í gerð *samhliða málheilda*, málheilda þar sem setning á íslensku er við hlið setningar á ensku. Slíkar málheildir gera vélnámshugbúnaði kleift að læra að þýða á milli tungumálanna. Ýmislegt þarf þó að athuga í for- og eftirvinnslu texta sem þýddir eru sjálfvirkt, til að mynda hvernig farið er með eiginnöfn, sem eru beygð í íslensku en ekki í ensku.

Að þessum verkþætti vinna Stofnun Árna Magnússonar, Miðeind ehf. og Háskólinn í Reykjavík

Samhliða málheild EES texta

Þessi málheild mun innihalda þýðingar EES-texta af ensku yfir á íslensku. Fyrsta verkefnisár fer að mestu í að forvinna texta og setja upp sjálfvirkt ferli til samröðunar og síunar. Einnig í að útbúa handyfirfarin prófunargögn.

M2: Búið er að aðlaga „state-of-the-art“ aðferðir til samröðunar og síunar að málheildinni. Prófunargögn eru tilbúin.

Staða: Aðferðir við samröðun og síun hafa verið bætтар, prófunargögn eru útgefin á CLARIN. – Verkefni á áætlun.

Bakþýðingar

Með bakþýðingum er markmiðið að hafa yfir að ráða stærri samhliða málheildum, sem sýna þýðingar í báðar áttir (ísl.-ens./ens.-ísl.). Þá er þýðingarvél (t.d. ísl.-ens.) notuð til þa þýða texta og hann svo notaður sem þjálfunargögn fyrir líkan sem þýðir til baka í fyrra málið (ens.-ísl.). Þetta skiptir miklu máli, þar sem magn handþýddra texta, sem verkefnið hefur aðgang að, er ekki nægilegt fyrir þjálfun nútíma þýðingarvéla.

M2: Frumgerð að líkani fyrir þýðingar í báðar áttir, þjálfað á síuðum, bakþýddum textum. Byrjað að kanna blöndunarhlutfall handþýddra og sjálfvirkt þýddra texta fyrir valin textasvið.

Staða: Fjölmargar þjálfanir fyrir bakþýðingar hafa verið keyrðar og góð undirstaða að myndast fyrir mat á bestu aðferðum. – Verkefni á áætlun.

Grunnlína í vélpýðingum

Hér er unnið að því að undirbúa þróun hins eiginlega vélpýðingatóls, sem verður afurð áætlunarinnar. Nokkrar mismunandi, þekktar aðferðir eru prófaðar og í lok verkefnisárs verður ein þýðingarvél valin til áframhaldandi þróunar.

M2: Prófunargögn eru tilbúin. Samantekt um nauðsynlega for- og eftirvinnslu texta fyrir hvert kerfi er tilbúin.

Staða: Búið er að gera prófanir með hjálp fólks sem mat gæði þýðinga mismunandi þýðingarvéla innan verkefnisins sem og Google translate. Samantekt um for- og eftirvinnslu tilbúin. – Verkefni á áætlun.

Viðmót á þýðingarvél

Til þess að gefa almenningi kost á að prófa nýjustu útgáfur þýðingarvéla og bera saman, er gagnlegt að bjóða upp á vefviðmót.

M2: Hugbúnaðarpakki sem inniheldur vefviðmót ásamt tengingu við vefþjón og þýðingarvélar hefur verið gefinn út skv. stöðlum verkefnisins. Sjá einnig <https://velthyding.mideind.is> – Verkefni á áætlun.

Lýðvirkjun

Þýðingarvélar skila yfirleitt nokkrum þýðingarmöguleikum og sú þýðing sem hefur hæsta skorið er sýnd notanda. Markmiðið með lýðvirkjunarkerfi er að gefa notendum kost á að velja annan möguleika sem bestu þýðingu en þýðingarvélin sjálf leggur til, eða að skrifa eigin þýðingu (eða að staðfesta að vélin hafi valið bestu þýðinguna). Þessum gögnum verður safnað á skipulegan hátt og þau nýtt til þess að bæta þýðingarvélarnar. Þetta kerfi getur einnig nýst beint við þýðingarvinnu.

M2: Hönnun á endurgjöf tilbúin – Verkefni á áætlun.

Málrýni

Eins og í öðrum kjarnaverkefnum felst vinnan við málrýni í málheildarvinnu og hugbúnaðarþróun. Unnið er að gerð og greiningu málheilda með merktum stafsetningar- og málvillum, til þess að þróun málrýnihugbúnaðar verði sem markvissust.

Miðeind ehf. og Háskóli Íslands vinna að verkefnum innan málrýni, með aðkomu Stofnunar Árna Magnússonar.

Almenn villumálheild

Almenn villumálheild inniheldur texta frá fólki sem almennt hefur ágætt vald á íslensku skrifmáli. Villur sem merktar eru í þessari málheild verða nýttar til greiningar og forgangsroðun verkefna fyrir þróun almenns málrýnis byggir á þessum gögnum.

M2: Búið er að gefa út almenna villumálheild á CLARIN. Gert er ráð fyrir um 5.000 setningum í málheildinni, sem innihalda a.m.k. eina villu. – Verkefni á áætlun

Sérhæfðar villumálheildir

Í sérhæfðum villumálheildum verðuð litið á texta frá fólki sem skrifar íslensku sem annað tungumál og lesblindum, til þess að geta aðlagð málrýnihugbúnað að þeirra þörfum. Byrjað er á því að skoða texta frá erlendum nemendum við HÍ (Íslenska sem annað mál).

M2: Búið er að finna leiðir til þess að útvega gögn og söfnun texta hafin. Búið er að áætla stærð málheilda.

Staða: Jákvæð viðbrögð hafa fengist hjá kennurum og nemendum við HÍ og athuganir varðandi leyfismál og persónuvernd hafa verið gerðar. Byrjað er að yfirfara texta, áætluð stærð málheilda íslenska sem annað mál: 3000 setningar, málheild með textum frá lesblindum: 2000 setningar. – Verkefni á áætlun

Tölfræðileg úrvinnsla og prófunarmálheildir

Hér er villumálheildin nýtt annars vegar til þess að útbúa tölfræði og forgangslista fyrir villutegundir og hins vegar til þess að búa til prófunarmálheildir fyrir málrýni.

M2: Búið er að greina helsut villuflokka. Skilgreining á stærð og innihaldi prófunarmálheilda tilbúin – Verkefni á áætlun

Orðalistar og mállíkön

Málrýnihugbúnaður reiðir sig á ýmis orðasöfn til þess að finna og leiðrétta villur. Mállíkön hjálpa til við að segja fyrir um hvaða orð og orðmyndir eru líkleg til þess að vera rétt í ákveðnu samhengi.

M2: Orðalistar tilbúnir. Undirbúningur þjálfunar mállíkana hafinn.

Staða: Búið er að setja saman mismunandi tegundir af orðalistum, m.a. íslensk heiti erlendra borga, algengar villur, óviðurkvæmileg orð, o.fl. Vinna við orðalista mun hins vegar fylgja kerfisbundinni þróun málrýnis áfram. Einföld gerð mállíkana er tilbúin. – Verkefni á áætlun.

Málrýnir fyrir stakorðavillur

Málrýnir fyrir stakorðavillur finnur villur sem eru ekki háðar samhengi, t.d. *firir* í stað *fyrir*. Hann getur hins vegar ekki leiðrétt villur þar sem rangt stafsetta orðið er í raun til, en á ekki við í viðkomandi samhengi, t.d. *leiti* í stað *leyti*.

M2: Lokaútgáfa stakorðavillugreinis hefur verið gefin út. Vefþjónusta hefur verið sett upp og einfalt vefviðmót er uppi. – Verkefni á áætlun og **verkefni lokið**

Kerfisbundin þróun málrýnis

Þetta er kjarninn í málrýniverkefninu, þar sem samhengisháð ritvilluleit og -leiðrétting verður kerfisbundið haldið áfram öll verkefnisárin.

M2: Búið er að forgangsraða villutegundum til þess að vinna með í næstu skrefum þróunar.

Staða: Tveir listar af villutegundum hafa verið settir saman, sem munu stýra næstu skrefum í þróun. Annars vegar listi sem sýnir villur sem hægt er að greina á *tókastigi*, t.d. endurtekin orð, villur í samsettum orðum, og hins vegar listi sem sýnir villur sem þarfnast setningagreiningar, eins og samræmi innan setningaliða og rangar forsetningar með sögnum (*leita *af*). – Verkefni á áætlun.

Talgreining

Stór hluti vinnunnar innan talgreiningarverkefnisins á fyrsta ári snýr að gagnasöfnun og leyfismálum. Einnig er unnið að rannsóknum og þróun sem munu nýtast eiginlegum talgreiningarhugbúnaði.

Háskólinn í Reykjavík, ásamt Tíro, RÚV og Creditinfo Fjölmíðlavaktin koma að talgreiningarverkefnum.

Gagnasöfnun með Samrómi

Hér er markmiðið að safna sem flestum raddskýnum frá sem flestum Íslendingum og íslenskumælandi fólki og gefa út stóran gagnagrunn til þjálfunar á talgreinum.

M2: Búið er að gefa út gagnagrunn með 100.000 setningum frá fólki 18 ára og eldri. Sá hluti gagnagrunnsins sem kemur frá fólki með íslensku sem móðurmál inniheldur upptökur frá fólki á öllum aldri og af báðum kynjum í góðri dreifingu. Búið er að taka saman aldurs- og kynjadreifingu hjá fólki með íslensku sem annað mál. Búið er að taka upp 80.000 setningar frá börnum og unglingum yngri en 18 ára.

Staðan: Gríðarlegur árangur í gagnasöfnun náðist með keppni milli grunnskóla. Þegar hafa safnast um 240 þúsund setningar, sem er mun meira ein áætlað var á þessum tímapunkti. Verið er að vinna tölfræðilega úrvinnslu fyrir samantekt um aldurs- og kynjadreifingu. – Verkefni svo gott sem á áætlun.

Innslegið sjónvarps- og útvarpsefni

Að stórum hluta innsláttur á útvarpsþáttum, en einnig samröðuð gögn úr sjónvarpsefni sem er textað í textavarpi.

M2: Kerfi til samröðunar sett upp. Búið er að rita niður eða samræða 60 klst. af efni.

Staða: Hér vantar töluvert upp á magn gagna. Bæði tók meiri tíma en áætlað var, að fastsetja form ritaðra gagna, en einnig setti Covid-19 hér strík í reikninginn. Ráðstafanir hafa verið gerðar til þess að ná lokamarkmiði við M3, með því að ráða inn fleira starfsfólk og skipuleggja vinnuna í samræmi við það. – Verkefni á eftir áætlun, ráðstafanir gerðar fyrir M3.

Innslegnar samræður og fyrirspurnir

Til þess að safna upptökum af „eðlilegu“ máli, verða samræður og fyrirspurnir ritaðar niður.

M2: Búið er að setja upp kerfi til þess að taka upp samræður. Fyrsta samtal hljóðritað.

Staða: Búið er að setja upp kerfi til þessa að taka upp samræður. Upptökur hefjast innan skamms og M3 verður á áætlun. – Verkefni aðeins á eftir áætlun.

Innslegnir fyrirlestrar

Sjálfvirk talgreining fyrirlestra á netinu er eitt af stórum hagnýtingarmöguleikum talgreiningar. Hér verður gögnum fyrir þjálfun slíkra talgreina safnað.

M2: Búið að rita niður 15 klst. af fyrirlestrum. – Verkefni á áætlun.

Leyfismál talgreinisgagna

Leyfismál snúa bæði að rétthöfum efnis og samþykktum þeirra sem gefa raddir.

M2: Leyfi fyrir fréttir frá RÚV eru tilbúin.

Staða: Búið er að ganga frá fjölmörgum leyfum og samþykktaryfirlýsingum á verkefnisárinu, m.a. endanotanda leyfissamning við RÚV (EULA). Frágangur opins leyfis fyrir fréttir er í undirbúningi. – Verkefni svo gott sem á áætlun.

Greinamerkjasetning og endapunktaskynjun

Texti sem talgreinir skilar frá sér, er alveg án greinarmerkja og hástafa. Það er því nauðsynlegt að eftirvinna textann og bæta inn greinamerkjum og hástöfum, til þess að gera textann læsilegri. Endapunktaskynjun snýr aftur á móti að því að greina þá staði í upptöku, þar sem einhver hefur lokið máli sínu, eða þar sem vænlegt er að klippa á lengri upptöku.

M2: Samantekt um hljóðgreiningu sem nýtist í endapunktaskynjun. Fyrstu tilraunir hafa verið framkvæmdar.

Staða: Skipting á áherslum fór fram og tilraunir gerðar með notkun hljóðgreiningar í greinamerkjasetningu, ásamt því að þróa áfram og endurbæta greinamerkjalíkön frá M1. Töl til greinamerkjasetningar er því tilbúið og endapunktaskynjun verður skilað við M3.

Mállýskur, hljóðgreining og samræðugreind

Í þessu verkefni fer fram dýpri greining á tali, sem hægt er að nýta í mismunandi talgreinishugbúnaði. Á fyrri hluta verkefnisársins fór fyrst og frems fram vinna við hljóðgreiningu og samræðugreind, þ.e. greining á því hver talar hvenær í upptökum af samtölum.

M2: Hljóðgreining og samræðugreind

Staða: Töl til hljóðgreiningar og samræðugreindar hafa verið unnin, þjálfuð á íslenskum gögnum. – Verkefni á áætlun

Talgerving

Í talgervisverkefninu, eins og í talgreiningunni, er aðaláherlsan á gagnasöfnun nú í upphafi verkefnis. Þjálfun hefðbundinna talgervla krefjast ekki eins mikils magns af gögnum eins og þjálfun talgreina. Á móti kemur að gögn fyrir talgervingu þarf að taka upp í stúdíóumhverfi, upptökurnar þurfa að vera hnökralausar og velja þarf sérstaklega fólk til þess að lesa inn. Sökum þess að hljóðver þurftu að loka vegna Covid-19, urðu verulegar tafir á upptökum fyrir talgervla. Þess í stað var aukinn kraftur settur í þróun og þökkun talgervilsfrumgerðar. Háskólinn í Reykjavík leiðir vinnu við talgervla, upptökur fara fram í hljóðveri RÚV.

Upptökur á einingavalsgögnum

Fyrir hvern einingavalstalgervil eru upptökur frá einni manneskju klipptar niður í hljóðeiningar, sem er svo aftur skeytt saman þegar talgervillinn les texta. Talgervillinn les því í raun með rödd þess sem las inn.

M2: Búið er að semja við alla átta lesara. Upptökum sex lesara er lokið, 20 klst. af tali hver.

Staða: Búið er að taka upp 2 kvenmannsraddir og 1 karlmannsrödd, 20 klst. af tali hver. Upptökur eru hafnar á næstu tveimur karlmannsröddum og verið að ganga frá samningum við þriðju konuna. Að lokum er verið að undirbúa upptökur einnar karlmannsraddar og einnar kvenmannsraddar á Akureyri, en þeir lesarar munu leggja til raddir með norðlenskum framburði. Þar sem upptökur geta farið fram samtímis á Akureyri og í Reykjavík mun verkefnið standa skil á M3 eftir áætlun. – Verkefni á eftir áætlun, ráðstafanir gerðar fyrir M3.

Gögn fyrir raddblöndun

Talgervlar sem beita raddblöndun nýta upptökur frá mismunandi lesurum og búa til nýja rödd / nýjar raddir. Hér þarf því að taka upp mun fleiri raddir en fyrir einingavalstalgervil, en ekki nema um 2 klst. hverja rödd.

M2: Leslistar tilbúnir. Upptökuumhverfi tilbúið, búið að taka upp 12 þátttakendur.

Staðan: Upptökur eru ekki hafnar. Búið er að leggja niður nákvæmt skipulag til þess að ná tilsettum fjölda radda við M3. Það er þó ákveðin óvissa fyrir hendi, þar sem mikilvægt er að kvenmannsraddirnar annars vegar og karlmannsraddirnar hins vegar séu mjög áþekkar, til þess að raddblöndun gangi vel. Leitast verður við að ná til hópa eins og t.d. kóra, til þess að ná að safna röddum frá mörgum á einum stað. Þegar réttu raddirnar eru fundnar má gera ráð fyrir að upptökum fyrir M3 (20 þátttakendur) sé hægt að ljúka á 6-8 vikum með því setja auka kraft í upptökur samhliða upptökum á einingavalsgögnum. – Verkefni á eftir áætlun, ráðstafanir gerðar fyrir M3.

Forskriftir að röddum

Samhliða upptökum radda er unnið að forskriftum að talgervilsröddum. Hér eru þær upptökur sem koma inn jafn óðum nýttar í þróun.

M2: Frumgerð að einingavalstalgervli tilbúin

Staða: Búið er að þróa frumgerðir af einingavals- og stikuðum talgervlum. Að auki er búið að setja frumgerðirnar upp í Docker, sem gerir það mun auðveldara fyrir alla áhugasama að prófa sig áfram með talgervlana, en handvirk uppsetning á því kerfi er flókin. – Verkefni á áætlun.