MÁLTÆKNI FYRIR ÍSLENSKU 2018-2022

VERKÁÆTLUN

ANNA BJÖRK NIKULÁSDÓTTIR
JÓN GUÐNASON
STEINÞÓR STEINGRÍMSSON

Mennta- og menningarmálaráðuneytið
Júní 2017

Útgefandi: **Mennta- og menningarmálaráðuneytið**
Sölvhólsgötu 4
101 Reykjavík
Sími: 545-9500
Netfang: postur@mrn.is
Veffang: www.menntamalaraduneyti.is

ISBN 978-9935-436-69-6

EFNISYFIRLIT

SKÝRSLUHÖFUNDAR

Anna Björk Nikulásdóttir lauk M.A.-prófi í máltækni frá Háskólanum í Heidelberg árið 2007. Vann meðfram námi við Fraunhofer stofnunina í Darmstadt og European Media Laboratory í Heidelberg. Tók þátt í verkefninu „Hagkvæm máltækni utan ensku - íslenska tilraunin“ og stundaði doktorsnám í máltækni við Háskóla Íslands á árunum 2009-2012. Meistara- og doktorsverkefni Önnu fjölluðu um sjálfvirka greiningu merkingarvensla í orðabókum og textum. Hún vann síðar við hugbúnaðarþróun og verkefnastjórnun hjá VICO Research & Consulting í Stuttgart, fyrirtæki sem sérhæfir sig í greiningu samfélagsmiðla, en starfar nú við Gervigreindarsetur Háskólans í Reykjavík.

Jón Guðnason er lektor í rafmagnsverkfræði hjá Háskólanum í Reykjavík, forstöðumaður Gervigreindarseturs Háskólans í Reykjavík og sérfræðingur í talmerkjafræði. Hann lauk doktorsprófi frá Imperial College London árið 2007 í merkjafræði og fjallaði verkefnið um auðkenningu á rödd með því að vinna einkenni úr raddlind. Meistaraverkefni Jóns var unnið í Háskóla Íslands árið 2000 og fjallaði um líkanagerð á talmerki með afturvirkum tauganetum. Jón starfaði sem sérfræðingur í talgreiningu hjá fyrirtækinu SpinVox í Bretlandi árið 2006-2008 og var gestafræðimaður í Columbia háskóla í New York 2008-2009. Síðan hann hóf störf hjá Háskólanum í Reykjavík árið 2009 hefur Jón byggt upp rannsóknarhóp í máltækni og raddmerkjavinnslu og stýrir verkefnum meðfram háskólakennslu á sviði merkjafræði og vélræns lærdóms.

Steinþór Steingrímsson er tölvunarfræðingur og íslenskufræðingur frá Háskóla Íslands. Lauk M.Sc. prófi í máltækni (Speech and Language Processing) frá Edinborgarháskóla árið 2005. Hefur unnið að máltækni-verkefnum á Íslandi frá 2011, fyrst í META-SHARE verkefninu og síðar við þróun mállegra gagnasafna hjá Stofnun Árna Magnússonar í íslenskum fræðum, Markaða íslenska málheild, Risamálheild, Málróm og fleiri verkefni. Vann áður við hugbúnaðarþróun í tveimur bönkum, tölvuleikjaþróun, kennslu og fréttamennsku. Er nú verkefnastjóri upplýsingatækni hjá Stofnun Árna Magnússonar í íslenskum fræðum.

Við undirbúning verksins hittu höfundar sérfræðinga víða að úr heiminum að máli og fengu hjá þeim upplýsingar og ráðgjöf um máltækni og/eða máltækniáætlanir. Þeir eru Hynek Hermansky, Julian S. Smith prófessor í rafmagnsverkfræði og forstöðumaður Máltækni- og talvinnsluseturs Johns Hopkins háskóla, Bandaríkjunum. Kadri Vider, sérfræðingur við Stærðfræði- og tölvunarfræðideild Háskólans í Tartu, Eistlandi. Heiki-Jaan Kaalep, fræðimaður við Stærðfræði- og tölvunarfræðideild Háskólans í Tartu. Mark Fišel, deildarforseti og dósent í máltækni við Stærðfræði- og tölvunarfræðideild Háskólans í Tartu. Tanel Alumäe, fræðimaður við tölvunarfræðideild Tækniháskólans í Tallinn. Einar Meister, fræðimaður við tölvunarfræðideild Tækniháskólans í Tallinn. Meelis Mihkla, deildarforseti og fræðimaður við Stofnun Eistneskrar tungu. Tõnis Nurk, deildarforseti og fræðimaður við Stofnun Eistneskrar tungu. Martin Eessalu, sérfræðingur í rannsóknum hjá eistneska mennta- og rannsóknamálaráðuneytinu. Andero Adamson, sérfræðingur í málvísindum hjá eistneska mennta- og rannsóknamálaráðuneytinu. Etienne Roth, sérfræðingur í máltækni hjá EPC Consulting & Software GmbH, Þýskalandi. Markus Foti, verkefnastjóri hjá MT@EC/eTranslation, Directorate-General for Translation (DGT). Andreas Eisele, sérfræðingur hjá MT@EC/eTranslation, Directorate-General for Translation (DGT). Michael Jeelinghaus, sérfræðingur hjá MT@EC/eTranslation, Directorate-General for Translation (DGT). Szymon Kocek, sérfræðingur hjá MT@EC/eTranslation, Directorate-General for Translation (DGT). Oddur Kjartansson, sérfræðingur í máltækni og gagnasöfnun hjá Google, London, Bretlandi. Alexander Gutkin, sérfræðingur í talgervingu hjá Google, London. Hanna Silen, sérfræðingur í talmerkjafræði hjá Google, London. Bente Maegaard, vísindamaður við Máltæknimiðstöð Kaupmannahafnarháskóla og fulltrúi CLARIN í Danmörku. Koenraad De Smedt, prófessor í máltækni við Háskólann í Bergen, fulltrúi CLARIN í Noregi. Jens Edlund, lektor í taltækni við KTH, Stokkhólmi. Eiríkur Rögnvaldsson prófessor í íslensku við Háskóla Íslands. Anton Karl Ingason, lektor við Háskóla Íslands. Sigrún Helgadóttir, sérfræðingur við Stofnun Árna Magnússonar í íslenskum fræðum. Kristín Bjarnadóttir, rannsóknardósent við Stofnun Árna Magnússonar í íslenskum fræðum. Jón Hilmar Jónsson, rannsóknardósent við Stofnun Árna Magnússonar í íslenskum fræðum. Jón Friðrik Daðason, tölvunarfræðingur. Vilhjálmur Þorsteinsson, forritari.

Öllu þessu fólki kunnum við okkar bestu þakkir fyrir ómetanlega aðstoð og ráðgjöf.

AÐDRAGANDI

Illugi Gunnarsson, mennta- og menningarmálaráðherra, skipaði haustið 2016 stýrihóp til að hafa umsjón með kortlagningu á tækni fyrir máltækni, stefnumörkun og vali á tæknilegri útfærslu fyrir íslensku. Nefndinni var einnig falið að gera stöðumat á íslenskum gagnasöfnum og nákvæma fjárhags- og verkáætlun fyrir 5 ára máltækniáætlun. Í tilkynningu um stofnun stýrihópsins segir: „Vaxandi áhrif tölvutækni á daglegt líf munu á næstu árum krefjast aðgerða af hálfu stjórnvalda til að tryggja að íslenskan verði gjaldgeng í samskiptum sem byggja á tölvu- og fjarskiptatækni. Það er mat þeirra sem best til þekkja að íslenskunni stafi hætta af þessari þróun verði ekkert að gert. Jafnframt felast í því mikil tækifæri fyrir íslenskt samfélag ef hægt er að nota tungumálið til fulls í samskiptum við snjalltæki ýmiss konar.“

Þessi skýrsla er niðurstaða vinnuhóps sem stýrihópurinn kallaði saman til að meta stöðu íslenskrar máltækni og gera verkáætlun til fimm ára. Lagt er til að fjórar opnar kjarnalausnir verði til á tímabilinu: talgreinir, talgervill, þýðingarvél og málrýnir. Þróun þeirra og aðlögun fyrir íslensku skal ná nógu langt til að lausnirnar verði gagnlegar og notaðar af almenningi, fyrirtækjum og stofnunum á Íslandi. Grunnforsenda í smíði máltækniverkfæra er að til séu gagnasöfn og stoðtól. Nauðsynlegri vinnu á því sviði er einnig lýst. Auk verkáætlunar eru gerðar tillögur um tilhögun áætlunarinnar og gerð grein fyrir hvernig staðið hefur verið að sambærilegum áætlunum í öðrum löndum.

Anna Björk Nikulásdóttir
Jón Guðnason
Steinþór Steingrímsson

ÁGRIP

Framtíð tölvunotkunar verður samofin máltækni. Með nýrri gervigreindartækni er mögulegt að hagnýta gríðarstór texta-, mál- og upplýsingasöfn með áður óframkvæmanlegum hætti. Því fylgir að tungumálið verður sífellt meira notað í samskiptum við tæki og tæki sífellt meira notuð til að vinna með tungumálið. Breytingarnar fela í sér mikla möguleika – og nýjar kröfur notenda. Sjálfvirk fyrirspurna- og samtalskerfi geta aukið hagkvæmni og bætt þjónustu fyrirtækja og stofnana. Þýðingarvélar geta aukið framleiðni þýðenda og þannig gert það mögulegt að meira efni sé aðgengilegt á tungumálinu. Með áheyrilegum talgervlum verður hægt að gera margfalt fleiri bækur aðgengilegar á hljóðbókaformi en smæð markaðarins byði annars upp á. Hugbúnaður sem gerir fólki kleift að tala eða skrifa, sem annars gæti það ekki vegna fötlunar eða sjúkdóma, gjörbreytir lífsgæðum þess og svo mætti lengi telja. Það er sérstaklega mikilvægt fyrir smá málsamfélög að nýta sér þessa tækni. Hún er ekki aðeins tungumálinu til framdráttar heldur getur hún skipt sköpum um það hvernig málsamfélaginu reiðir af.

Við stöndum á krossgötum. Við þurfum að velja hvort þeir innviðir, sem til þarf til að íslenskan verði nothæf í breyttum tækniheimi, verði þróaðir eða hvort við bíðum átekta og sjáum hverju fram vindur. Ef íslenska er ekki með í nýrri tækni verða tækin notuð á öðrum tungumálum. En þá takmarkast not tækninnar við fá svið máltækninnar, tækifæri á öðrum sviðum glatast og þeir sem ekki eru vel talandi á erlendum tungumálum sitja eftir. Möguleikarnir geta þannig aldrei orðið þeir sömu og ef tæknin er til á móðurmálinu.

Víst er að tækniþróun er ekki ókeypis, en það getur líka verið samfélaginu dýrt að missa af nýjum tækifærum og viðhalda gömlum aðferðum sem úreldast hratt. Raunverulegt val stendur því á milli þess annars vegar að sætta sig við glötuð tækifæri, lakari lífsgæði og kostnað sem hlýst af því að nýta ekki bestu fáanlegu tækni og hins vegar að fjárfesta í tækniþróun sem eykur samkeppnishæfni atvinnulífsins, samfélagsins og tungumálsins.

Til að tryggja að íslenska verði valkostur í tækniheiminum þarf að búa svo um hnútana að fólk, fyrirtæki og stofnanir geti nýtt máltækni án þess að flókin og þung tækni- og innviðaþróun standi þeim fyrir þrifum. Í skýrslunni er lögð áhersla á þrjá meginþætti til að svo geti orðið:

Uppbygging innviða: Gagnasöfn um tungumálið og grunnverkfæri kallast málföng. Ef þau eru ekki til staðar eða eru mjög ófullkomin er vonlaust að þróa máltækni fyrir tungumálið. Gagnasöfnin geta verið stór textasöfn,

hljóðupptökur og orðalistar sem búið hefur verið um þannig að hægt sé að nota þau sem efnivið í máltækni. Hægt er að byggja á þeirri vinnu sem þegar hefur farið fram hér á landi en til að geta nýtt möguleika tækninnar er mikilvægt að setja mun meiri kraft í uppbygginguna en hingað til hefur verið gert.

Grunnverkfæri í máltækni eru til að mynda opnir talgreinar og talgervlar sem vinna með almennt mál en sem hægt er að laga að sérþörfum. Þau geta verið tól til mál- og framburðargreiningar eða verkfæri til mállíkanagerðar, nauðsynleg stoðtæki í máltæknitól fyrir endanotendur. Þau geta verið almennar þýðingarvélar sem hægt er að laga að ákveðnum sviðum og svo mætti áfram telja. Nauðsynlegt er að þau séu opin svo að allir sem vilja þróa máltæknilausnir fyrir íslensku geti samnýtt þau og farið beint í að þróa notendahugbúnað en þurfi ekki að sinna tímafrekum grunnrannsóknum og grunnþróun.

Með góðu skipulagi og samstilltu átaki getur íslenskan orðið hluti af stafrænum heimi framtíðarinnar.

Nýsköpun í máltækni: Nauðsynlegt er að styðja við fyrirtæki sem stunda nýsköpun í máltækni og/eða geta notað máltæknitól til að bæta þjónustu eða framleiðsluþætti og tryggja þátttöku þeirra í þróuninni. Fyrirtækin finna lausnir á þörfum samfélagsins á máltækni og því eiga þau að geta útfært nauðsynleg verkfæri sem byggð eru á þeim innviðum sem verða til í áætluninni. Tryggja þarf þennan þátt með hvatakerfi og góðum samskiptum og samvinnu þeirra sem að verkefninu koma á öllum stigum áætlunarinnar.

Samstarf og klasamyndun: Þátttaka í erlendu samstarfi leikur lykilhlutverk í að tryggja íslenskunni sess í tækjum og tölvum framtíðarinnar. Sækja þarf af festu og harðfylgi að erlend stórfyrirtæki á sviðinu bjóði upp á íslensku í sínum kerfum með sama hætti og önnur tungumál. Leiðin til þess er regluleg samskipti við fyrirtækin en einnig við stofnanir og háskóla erlendis sem vinna að sömu markmiðum fyrir önnur tungumál. Þátttaka í samstarfi um þróun máltækni fyrir málsvæði með rýr málföng er mikilvæg. Við upphaf áætlunarinnar verður myndaður klasi innanlands með öllum helstu áhuga- og hagsmunaaðilum. Með honum verður til góður vettvangur til þess að setja á laggirnar samstarfsverkefni innanlands og taka þátt í verkefnum erlendis sem gagnast íslenskri máltækni.

Með góðu skipulagi og samstilltu átaki getur íslenskan orðið hluti af stafrænum heimi framtíðarinnar. Lifandi þróun íslenskrar máltækni gerir það mögulegt að fella íslensku inn í tækni og þjónustu þannig að hún verði raunverulegur valkostur í öllu viðmóti og upplýsingavinnslu.

ÚTDRÁTTUR

Máltækni felur í sér alla þá tækni sem gerir hugbúnaði kleift að fást við tungumál. Tölvur eiga auðvelt með að framkvæma flókna útreikninga og vinna með mikið magn af gögnum í lokuðum kerfum þar sem reglur eru skýrar. Tungumálið er hins vegar eitt af því sem erfitt er að setja upp í formleg kerfi fyrir tölvur. Hæfileikinn til þess að skilja og tala tungumál er meðfæddur og margt er enn hulið í sambandi við það hvernig málvinnsla fer í raun fram. Tungumál eru lifandi, ný orð verða til, orð breyta um merkingu, og möguleikarnir á að mynda setningar og tengja saman hugtök og hugmyndir eru óendanlegir. Það er því krefjandi og flókið verkefni að fá tölvur til þess að vinna með tungumál á líkan hátt og manneskjur, að skilja málhljóð, orð og setningar og tengja við almenna þekkingu um málið og veröldina í kringum okkur.

Frá því um miðja síðustu öld hafa fjölmargar aðferðir verið þróaðar fyrir mismunandi svið máltækni. Lengi vel var árangurinn takmarkaður en undanfarin ár hefur orðið algjör bylting með tilkomu gríðarmikils gagnamagns, öflugs vélbúnaðar og þróunar í gerð algríma, sem oft byggjast þó á gömlum grunni. Það hefur sýnt sig að frumforsenda fyrir árangri í máltækni að hafa yfir miklu magni gagna að ráða. Þróun í vélbúnaði og algrímum hefur svo gert vinnslu nauðsynlegs gagnamagns mögulega.

Nú er svo komið að máltæknibúnaður er orðinn alltumlykjandi og ljóst orðið að tungumál sem verða fyrir utan þróun á þessu sviði munu eiga erfitt uppdráttar á allra næstu árum. Fólk í tæknivæddum samfélögum venst því að tala við tæki og fá svör við spurningum sem það ber upp. Það venst því að upplýsingaleit einskorðist ekki við einföld leitarorð heldur að hugbúnaður geti unnið upplýsingar upp úr miklu magni mismunandi gagna, tengt saman og dregið ályktanir, að hægt sé að lesa inn texta sem tölva skrifar niður og að tölva lesi upphátt hvaða texta sem er, að hægt sé að átta sig á innihaldi tals og texta á framandi tungumáli með aðstoð vélþýðinga og svo mætti lengi telja.

Máltækni hefur þó ekki einungis gríðarleg áhrif á daglegt líf fólks og samskipti. Við lifum á upplýsingaöld, á tímum þar sem gögn og upplýsingar eru meðal þess verðmætasta sem fyrirtæki ráða yfir. Fyrirtæki og stofnanir sem vinna með gögn á „snjallan“ hátt bæta samkeppnisstöðu sína. Nauðsyn þess að sinna þessari þróun má líkja við það að varla finnst það fyrirtæki eða stofnun í dag sem ekki býður upp á upplýsingar eða þjónustu á netinu, eitthvað sem á sínum tíma virtist ef til vill ekki skipta sköpum.

Innan máltækni eru fjölmörg sérsvið sem fást við sérhæfð verkefni. Vinna við ólík svið hennar, eins og til að mynda greiningu tals eða málfræðigreiningu texta, krefst mismunandi en oft fjölbreyttrar sérþekkingar. Tölvunarfræði, málvísindi, verkfræði, stærðfræði, heimspeki og tölfræði eru dæmi um fræðigreinar sem nýtast innan máltækni. Hefðbundin menntun í máltækni felst fyrst og fremst í að tvinna saman tölvunarfræði og málvísindi. Kjarni máltæknimenntaðra sérfræðinga er nauðsynlegur fyrir sviðið en fjölmargir aðrir sérfræðingar geta tekið þátt í að mynda öflugan þekkingariðnað fyrir máltækni á Íslandi. Þó að þróa þurfi sérhæfða máltækni fyrir íslensku er ekkert sem stendur í vegi fyrir því að velheppnaðar lausnir verði hægt að útfæra fyrir önnur tungumál og þar með stærri markað.

Máltækniáætlun fyrir íslensku 2018–2022 hefur það að markmiði að tryggja að hægt verði að nota íslensku í samskiptum við tæki og í allri upplýsingavinnslu.

Máltækniáætlun fyrir íslensku 2018–2022 hefur það að markmiði að tryggja að hægt verði að nota íslensku í samskiptum við tæki og í allri upplýsingavinnslu. Áður hefur verið gerð grein fyrir því sem þarf að gera til þess að þetta markmið megi nást, fyrst í skýrslu starfshóps um tungutækni árið 1999. Fyrirliggjandi skýrsla fjallar um það sem þarf að gera á tíma áætlunarinnar til að þetta verði hægt, hvernig skipulagi vinnu og verkefna gæti verið háttað og mikilvægi samskipta milli þátttakenda innanlands og ekki síst við erlend stórfyrirtæki og samstarfsaðila.

Hér á eftir fer stutt samantekt um efni skýrslunnar með vísunum í nánari umfjöllun þar sem það á við.

VERKEFNI MÁLTÆKNIÁÆTLUNAR

Forgangsverkefni máltækniáætlunarinnar eru þau verkefni sem mynda nauðsynlegan grunn fyrir áframhaldandi þróun á mismunandi sviðum máltækni fyrir íslensku. Verkefnin eru flokkuð í talgreiningu, talgervingu, vélþýðingar, málrýni og málföng.

Talgreining (Kafli 2.1) snýst um það að breyta töluðu máli í ritmál. Hún er forsenda þess að við getum átt samskipti við tölvur og tæki með þeim hætti sem flestum er eðlilegast: með því að tala. Möguleikinn á raddstýrðum samskiptum er sérstaklega mikilvægur við aðstæður þar sem ekki er hægt að nota hendur eða það er truflandi, t.d. við akstur, eða fyrir þá sem af einhverjum ástæðum hafa ekki gott vald á skrifuðu máli eða eiga í vandræðum með að nota lyklaborð eða snertiskjái.

Talgreining (Kafli 2.1)

Talgreining er notuð til þess að rita upp langan samfelldan upplestur, ræður eða dikteringar; fylgjast með samræðum og taka jafnvel þátt í þeim eða taka við raddskipunum til frekari greiningar. Talgreiningarhugbúnaður er

því eins konar lyklaborð fyrir röddina. Eins og á öðrum sviðum máltækni – og innan gervigreindar almennt – er sérhæfður hugbúnaður yfirleitt raunhæfari lausn en hugbúnaður sem á að virka í öllum aðstæðum. Við þróun talgreiningar þarf að taka tillit til mismunandi atriða. Það skiptir máli hver talar, við hvaða aðstæður og um hvað er talað:

- **Hver talar?** Raddir og framburður eru ólík milli fólks. Ákveðnir hópar eiga þó meira sameiginlegt en aðrir, t.d. eru kvenraddir líkari öðrum kvenröddum heldur en karlröddum, fólk á svipuðum aldri frá sama málsvæði hefur oft svipaðan framburð o.s.frv. Það er nauðsynlegt að hafa yfir hljóðupptökum að ráða sem eru lýsandi fyrir þann hóp sem talgreinirinn á að nýtast fyrir. Hefðbundinn talgreinir virkar t.d. ekki fyrir börn þar sem raddir þeirra eru of ólíkar röddum fullorðinna.

- **Við hvaða aðstæður er talað?** Aðstæður sem geta haft áhrif á talgreiningu eru bakgrunnshávaði (umferðarniður, náttúruhljóð, skvaldur, einstök hljóðmerki eins og bjalla í þingsal), hvort fleiri eru að tala nálægt tækinu, hversu fumlaust talið er (ekki mikið af hikorðum og endurtekningum) og gæði upptöku. Við hönnun talgreinis fyrir leiðsögukerfi bíla þarf til að mynda að gera ráð fyrir bílhljóðum og bakröddum farþega.

- **Um hvað er talað?** Einföld raddstýring getur falist í því að einungis er leyfilegt að nota ákveðin stök orð og verkefni talgreinis er þá einungis að greina á milli leyfilegra orða. Samskipti eru þó almennt flóknari, nýta fjölbreyttan orðaforða og setningaskipan. Þau geta verið afmörkuð við ákveðið efni, t.d. læknisfræði, eða verið alveg opin. Sníða þarf talgreina að því innihaldi sem líklegt er að hann fáist við.

Markmið: Að til verði almennur íslenskur talgreinir aðgengilegur til notkunar í gegnum vefþjónustu. Allar aðferðir og gögn verði einnig aðgengileg sem grunnur fyrir þróun sérhæfðra talgreina. Að íslenskum talgreinum verði komið inn í snjalltæki. Að talgreiningarhluti raddstýringar-, fyrirspurna- og samræðukerfa verði þróaður. Að unnið verði að talgreiningu fyrir börn og unglinga.

Hvað þarf til: Mikið magn talgagna af fjölbreyttum toga. Hugbúnað sem inniheldur vel prófuð algrím fyrir talgreiningu og möguleika á aðlögun fyrir eigin kerfi. Þekking á framburði, mállýskum og málkerfi og tól sem nýta þá þekkingu.

Talgerving (Kafli 2.2)

Talgerving (kafli 2.2) breytir rituðum texta í talað mál. Tvö meginsvið talgervingarhugbúnaðar eru upplestur og (radd)samskipti. Talgervlar eru notaðir til þess að lesa texta, til dæmis af vefsíðum eða jafnvel heilu bækurnar. Fólk sem af einhverjum ástæðum getur ekki lesið sjálft eða á í erfiðleikum með það treystir á talgervlatækni í daglegu lífi. Samskiptakerfi, þar sem talgreinir nemur það sem notandi segir, þurfa á talgervlum að halda til þess að hægt sé að svara með rödd. Svörin eiga að hljóma eðlilega og í samræmi við innihald samskiptanna. Mismunandi kröfur eru gerðar til talgervla eftir því hvaða hlutverki þeir gegna:

- **Markmið hlustenda** Hlustandi sem hefur það að markmiði að komast sem hraðast yfir ákveðið efni, til dæmis í námsbókum, þarf á talgervli að halda sem les hratt og skýrt. Hljómfall og áherslur skipta hins vegar meira máli en talhraði í eðlilegum samskiptum eða við upplestur á skáldsögu.

- **Hlutverk talgervils** Talgervlar sem lesa upp bækur eða blaðagreinar eru einhliða, þeir þurfa ekki að bregðast við einhverju sem notandi segir. Slíkir talgervlar þurfa að vera hannaðir til þess að lesa langa texta á þann hátt sem samræmist þörfum hlustenda. Talgervlar sem eiga í samskiptum við notendur þurfa að bregðast við því sem notandinn segir og eru hannaðir til þess að lesa styttri spurningar, svör og skilaboð.

- **Óskir hlustenda** Fólk er ekki endilega sammála um hvernig raddir er þægilegast að hlusta á. Það þurfa að vera til talgervlar sem tala með mismunandi karl- og kvenröddum og þar sem jafnvel er hægt að stilla atriði eins og raddstyrk og áherslur.

Markmið: Að til verði íslenskur talgervill aðgengilegur í gegnum vefgátt. Að til verði opið umhverfi til þróunar á fjölbreyttum talgervlum sem tala íslensku.

Hvað þarf til: Sérhæfðar upptökur til talgervilsþróunar. Vandaða framburðarorðabók sem inniheldur nægilegan fjölda orða og mismunandi framburð og mállýskur. Kerfi til þess að undirbúa texta fyrir lestur talgervils (textastöðlun) og þekkingu og tól fyrir talanda og ítónun. Hugbúnað fyrir þróun talgervla með þekktum aðferðum.

Vélþýðingar (Kafli 2.3)

Vélþýðingar (kafli 2.3) eru sjálfvirkar þýðingar milli tungumála. Þær eru nú þegar orðnar gagnlegar fyrir ýmis tungumálapör, bæði til þess að hjálpa fólki við að átta sig á innihaldi texta á tungumáli sem það er ekki læst á og

til þess að flýta fyrir vinnu þýðenda við tungumál sem þeir eru sérfræðingar í. Enginn þýðingahugbúnaður ræður hins vegar enn sem komið er við að skila þýðingum sem eru nálægt fullnægjandi gæðum, alltaf þarf að yfirfara texta og laga ef þýðing þarf að vera nákvæm.

Það er einkum þrennt sem einkennir mismunandi áherslur í vélþýðingum:

- **Milli hvaða tungumála á að þýða?** Hvert vélþýðingakerfi er yfirleitt einskorðað við eitt tungumálapar þó að einnig séu til aðferðir sem nýta svokallað millimál til þess að varpa þýðingu eins tungumáls yfir á mörg önnur. Slíkt kerfi kemur ekki til greina um sinn fyrir íslensku þannig að velja þarf tungumál sem leggja á áherslu á.

- **Hvað á að þýða?** Vélþýðingakerfi sem á að geta þýtt hvað sem er þarf að glíma við hluti eins og margræðni í miklu meira mæli en kerfi sem er sérhannað fyrir ákveðin umfjöllunarefni. Þýðingakerfi þurfa að þekkja þá tegund texta sem þau eiga að geta þýtt og því er t.d. kerfi sem hefur verið þróað með aðstoð texta úr stjórnsýslunni ekki líklegt til þess að geta þýtt íþróttafréttir vel.

- **Til hvers á að þýða?** Vélþýðingakerfi munu á næstunni ekki ráða við að skila tilbúnum textum og þar með gera starf þýðandans óþarft. Þau eru hins vegar ómetanlegt hjálpartól sem getur sparað háar fjárhæðir í fyrirtækjum og stofnunum þar sem þýðendur þyrftu annars að vinna mikið magn þýðinga frá grunni. Slík kerfi læra einnig með tímanum í gegnum leiðréttingar þýðenda. Annar mikilvægur tilgangur vélþýðinga er að gefa fólki hugmynd um innihald texta án þess þó að sú krafa sé gerð að allar staðreyndir skili sér.

Markmið: Að smíðuð verði opin þýðingarvél sem þýðir á milli íslensku og ensku. Hún á að gagnast við þýðingar á ákveðnum sviðum svo að þýðendur geti fullunnið texta hraðar.

Hvað þarf til: Stórar samhliða málheildir með íslenskum og enskum textum. Opinn hugbúnað til þróunar á þýðingarvélum með þekktum aðferðum.

Málrýni (Kafli 2.4)

Málrýni (kafli 2.4) aðstoðar við að leiðrétta texta og skrifa rétt. Villur í textum geta verið af ýmsum toga: innsláttarvillur, stafsetningarvillur, málfræðivillur eða villur í orðanotkun. Hugbúnaður til sjálfvirkrar ritvilluleiðréttingar er mikilvægur sem aðstoð við skrif texta, bæði fyrir hinn almenna notanda og í fyrirtækjum og stofnunum. Slík tækni hefur einnig brýnu hlutverki að gegna við þróun annars konar máltæknihugbúnaðar og er nauðsynleg ef

gera á ljóslesna texta að fullu nothæfa í stafrænu umhverfi. Notkunargildi málrýni fer eftir færni þess sem skrifar og aðstæðum. Því minni sem færnin er, tímapressan meiri og kröfurnar um gæði því mikilvægari er sjálfvirk málrýni.

Áherslur við gerð málrýnihugbúnaðar:

- **Hvað á að leiðrétta** Ritvillur eru af ýmsum toga. Nauðsynlegt er að skilgreina hvað málrýnihugbúnaður á að geta leiðrétt til þess að notendur, manneskjur jafnt sem annar hugbúnaður, geti treyst niðurstöðum.

- **Uppruni texta** Málrýni getur þurft að greina ljóslesna texta, texta frá öðrum hugbúnaði eða texta sem skrifaðir eru í ritvinnslukerfi. Laga þarf málrýnihugbúnaðinn sérstaklega að slíkum textagerðum. Auk þess þarf að taka tillit til þess að fólk hefur misgott vald á rituðu máli. Hefðbundin málrýni er þróuð fyrir þá sem hafa gott vald á tungumálinu, hafa notið þjálfunar og eiga ekki í neinum sérstökum erfiðleikum með ritun. Aðrir hópar, t.d. lesblindir, fólk sem ekki hefur viðkomandi tungumál að móðurmáli og börn sem eru að byrja að lesa og skrifa þurfa annars konar stuðning.

- **Tilgangur leiðréttingar** Málrýni, sem er hluti af ritvinnslukerfi, býður notandanum upp á möguleika til leiðréttingar þegar hún finnur villu. Umfangsmeiri aðstoð sýnir einnig af hverju villa er villa, t.d. með því að vísa í reglu. Málrýni sem hluti af öðrum máltæknihugbúnaði þarf hinsvegar að vera samþætt því kerfi og ákvarða hvaða leiðrétting verður fyrir valinu hverju sinni.

Markmið: Að þróa almenna málrýni sem ræður við að finna og leiðrétta algengustu villurnar sem finnast í almennum íslenskum textum. Að til verði þekking á eðli ritvillna hjá mismunandi hópum. Að aðferðir verði þróaðar til þess að laga kerfið að mismunandi þörfum, m.a. með tilliti til þjálfunar og kennslu. Að leggja áherslu á að sem flestir geti nýtt sér málrýnihugbúnaðinn, óháð stýrikerfum og ritvinnslukerfum. Að annar máltæknihugbúnaður geti nýtt sér málrýni lagaða að eigin þörfum.

Hvað þarf til: Safn texta sem innihalda ritvillur og greining á þeim. Áreiðanleg stoðtól til málfræði- og merkingargreiningar.

Málföng (Kafli 2.5)

Málföng (Kafli 2.5) Öll máltækni byggist á málgögnum: textum og/eða hljóðupptökum. Þau eru nauðsynleg við greiningu máls, söfnun orðaforða

og til að finna reglur og mynstur. Út frá málgögnum er þannig hægt að „kenna“ tölvunum það sem máli skiptir fyrir þann hugbúnað sem verið er að þróa. Æ algengara er þó að hugbúnaður sé látinn finna reglur og mynstur af sjálfsdáðum og læra þannig að greina tungumálið að miklu leyti án handgerðra reglna. Slíkar aðferðir, sem skila oft langtum betri árangri en handvirkar aðferðir, krefjast mjög mikils gagnamagns og oft verður að undirbúa gögnin á sérstakan hátt. Mismunandi hugbúnaður getur þurft á mismunandi gögnum að halda. Í verkefninu verður lögð áhersla á að koma upp mikilvægum málheildum fyrir tal og texta. Jafnframt þarf að vinna gögn sem geyma upplýsingar um einstaka þætti tungumálsins eins og orðaforða, framburð og merkingu. Þetta eru svokölluð orðaföng og er þeim nánar lýst í köflum 2.5.1.8–2.5.1.18.

Þó mikið sé um sértækar lausnir í máltækni er ákveðinn grunnhugbúnaður sem nýtist á öllum sviðum hennar. Þetta eru yfirleitt falin tól sem greina grunneiningar í textum, allt frá því að greina hvað teljast orð og hvað ekki, upp í að greina flókið málfræðilegt og merkingarlegt samhengi texta. Öll þessi tól, sem ekki eru tilbúnar hugbúnaðarlausnir í sjálfu sér en nauðsynlegur partur máltæknihugbúnaðar og fyrir gagnavinnslu, kallast stoðtól (kafli 2.5.3). Góð stoðtól, sem eru auðveld í notkun og skila áreiðanlegri greiningu, eru grundvöllur fyrir gæði máltæknilausna.

Nægilegt magn viðeigandi gagna ásamt áreiðanlegum stoðtólum er traustur grunnur fyrir alla áframhaldandi þróun. Þessi atriði skipta gríðarlegu máli fyrir gæði flóknari hugbúnaðar en einnig getur öll þróun orðið hraðari heldur en ef grunnurinn er ótraustur.

Markmið: Að safna gögnum og vinna úr þeim málheildir texta og tals. Áhersla verður lögð á áframhaldandi vinnu við stórar texta- og talmálheildir en einnig sérhæfðar málheildir sem skilgreindar eru fyrir einstök kjarnaverkefni. Að vinnu við mikilvæg orðaföng eins og framburðarorðabók, beygingarlýsingu og orðanet verði haldið áfram. Að vinnu verði haldið áfram við þau stoðtól sem þegar eru til og nauðsynlegum stoðtólum bætt við.

Hvað þarf til: Safna þarf gögnum þar sem þau eru til, á vefnum eða hjá stofnunum og fyrirtækjum, og ganga frá tilskildum leyfum til notkunar. Talgögn þarf að miklu leyti að búa til, þ.e. að taka upp tal fólks. Sérfræðingar þurfa að vinna öll gögn og undirbúa til notkunar í máltækni. Orðfræðigögnum þarf fyrst og fremst að koma á það form að þau nýtist við hugbúnaðargerð. Nokkur stoðtól þarf að þróa frá grunni en ganga þarf

frá leyfum fyrir stoðtól sem til eru og á að þróa áfram. Prófunargögn eru nauðsynleg fyrir öll stoðtól.

Í 3. kafla er fjallað um leyfismál og aðgengi. Öll gögn og tól verða gefin út með eins opnum leyfum og mögulegt er. Þannig verður hægt að nýta innviðina sem víðast. Einnig verður að gæta að stöðlum, aðgengi og viðhaldi allra innviða og annarra verkefna.

Leyfismál og aðgengi málfanga og innviða (Kafli 3)

Þegar uppbygging grunninnviða er komin vel af stað þarf að huga að þróun fleiri grundvallarmáltæknitóla. Í 4. kafla er nokkrum slíkum tólum lýst: upplýsingaútdrætti, álitsgreiningu, upplýsingaheimt, spurningasvörun, samræðukerfum og margmiðlunargreiningu. Þessi verkefni eru ekki hluti máltækniáætlunar fyrst um sinn en ættu að komast á dagskrá eins fljótt og auðið er.

Önnur máltækni-verkefni (Kafli 4)

Nýsköpun þarf að vera í lykilhlutverki í íslenskri máltækni. Innviðirnir sem þróaðir verða innan máltækniáætlunarinnar gera fyrirtækjum kleift að þróa máltæknilausnir og að nýta íslenska máltækni án þess að leggja þurfi í umfangsmikla og sérhæfða grunnþróun. Mikilvægt er að skapa og rækta nýsköpunarumhverfi í kringum máltækni sem er hvetjandi fyrir frumkvöðla jafnt og stærri fyrirtæki.

Nýsköpun í máltækni (Kafli 5)

Fyrirliggjandi skýrsla fjallar fyrst og fremst um verkáætlanir á sviði innviða í máltækni. Við hvert kjarnaverkefni er þó einnig fjallað um mögulega nýtingu viðkomandi innviða í tækniyfirfærslu og í 5. kafla eru tekin dæmi um umfangsmeiri máltæknihugbúnað eins og kennsluforrit, sjálfvirka símsvörun, merkingarbæra leit og fleira. Í nýsköpun mun fyrst um sinn verða lögð aðaláhersla á að koma íslensku í notkun í fjölbreyttum máltæknihugbúnaði. Ógrynni tækifæra felast hins vegar í öflugum máltækniiðnaði á Íslandi: fjölmörg stór tungumál utan ensku þurfa á máltæknihugbúnaði að halda og yfirfærsla lausna og þjónustu gæti opnað raunverulegan markað fyrir máltækni frá Íslandi. Alþjóðlegt samstarf á fjölbreyttum vettvangi er mikilvægt og Ísland gæti jafnvel orðið leiðandi í þróun máltækni fyrir smærri málsamfélög.

SKIPULAG MÁLTÆKNIÁÆTLUNAR

Skipulag áætlunar (Kafli 6)

Til þess að markmið máltækniáætlunarinnar gangi eftir, að koma íslensku og íslenskri máltækni í almenna notkun í tölvum og tækjum, þarf allt skipulag að taka mið af því frá upphafi. Mikilvægt er að skilgreina hlutverk stofnana, háskóla, ríkisins og atvinnulífsins vel og að allt starf sé skipulagt með samvinnu allra hlutaðeigandi í huga (6. kafli).

Lagt er til að sjálfseignarstofnunin Almannarómur verði miðstöð fyrir áætlunina. Aðalmarkmið miðstöðvarinnar er að sjá til þess að verkefni áætlunarinnar verði framkvæmd hjá þeim sérfræðingum, stofnunum og fyrirtækjum sem eru fengin til þess að útfæra þau, sjá um samhæfingu milli verkefna og við atvinnulífið og tryggja góð samskipti aðila verkefnisins við atvinnulífið og við erlend fyrirtæki og stofnanir þannig að þeir innviðir og tækni sem þróuð eru í verkefninu komist í notkun.

VERKÁÆTLUN Í HNOTSKURN

Í skýrslunni verður gerð vandlega grein fyrir þeim verkefnum sem við teljum nauðsynlegt að framkvæma til þess að koma upp innviðum fyrir íslenska máltækni. Gerð er grein fyrir þeim mannauði og þeirri fagþekkingu sem þörf er á í hverju verkefni fyrir sig og umfang í mannmánuðum gefið upp eftir því sem unnt er á þessu stigi. Vinnuhópurinn áætlar ekki aðra kostnaðarliði, til að mynda tækjakost, kerfisstjórn, hýsingu eða þjónustu í skýjaþjónustum. Mikilvægt er að gera ráð fyrir þessum atriðum þegar nákvæmari kostnaðaráætlanir verða gerðar.

Fyrir hvert þeirra fimm kjarnaverkefna sem á að vinna að eru skilgreind kjarnateymi. Þau hafa yfir þeirri þekkingu að ráða sem þarf til þess að vinna að grunnþróun á viðkomandi sviði. Nú þegar hefur talsverð reynsla myndast innan Stofnunar Árna Magnússonar í íslenskum fræðum, Háskólans í Reykjavík og Háskóla Íslands. Önnur teymi vinna þvert á kjarnaverkefnin og sinna verkefnum sem ekki endilega krefjast þekkingar eða reynslu í máltækni.

Mælt er með að kjarnateymin séu hvert um sig öflugir hópar sem vinna að viðkomandi verkefnum allan þann tíma sem áætlunin nær yfir. Þannig byggist upp þekking og reynsla í faginu og hæft fólk fæst til starfa til lengri tíma. Einstök verkefni krefjast þess þó vitanlega að utanaðkomandi starfsfólk og háskólanemar komi inn í teymi tímabundið. Vegna umfangs og fjölbreytni verkefna sem tengjast gögnum, stoðtólum og almennri máltækni þarf að byggja upp nokkur teymi á þessum sviðum. Tillögur um samsetningu teyma og greining á því hvaða þekkingu þau þurfa að búa yfir er sýnd á myndunum á bls. 24.

Tafla á bls. 25 sýnir yfirlit yfir kjarnaverkefnin, teymin sem þurfa að koma að framkvæmd þeirra og samantekt um áætlaða mannmánuði. Nákvæma lýsingu verkþátta innan hvers verkefnis má finna í viðkomandi undirköflum sem vísað er til í töflunni. Hugbúnaður og gagnasöfn þarfnast viðhalds og áframhaldandi þróunar eftir að einstökum verkþáttum er lokið. Sú vinna er utan við svið þessarar skýrslu.

Kjarnateymi

T1: Talgreining

talgreining,
tölvunarfræði,
djúptauganet

T2: Talgerving

talgerving,
merkjafræði,
tölvunarfræði

T3: Vélþýðingar

máltækni/vélþýðingar,
þýðingar,
tölvunarfræði,
djúptauganet

T4: Málrýni

máltækni,
málfræði,
tölvunarfræði,
djúptauganet

T5: Gagnahögun

söfnun og högun,
málgagna,
málfræði,
tölvunarfræði

T6: Máltækni

máltækni,
málfræði,
tölvunarfræði

Önnur teymi

T7: Upptökur og frágangur hljóðgagna

tölvunarfræði, meistaranemar
í tengdum fögum, lesarar/ raddgjafar
fyrir talgervla

T8: Snjallsímateymi

forritun snjallsímastýrikerfa
(Android, iOS, WindowsPhone)

T9: Teymi fyrir vefgáttir og viðmót

tölvunarfræði/forritun

T10: Leyfismál

leyfismál gagna,
lögfræði

Verkefni/ Teymi	Talgreinir H.1-H.16, bls. 43-53	Talgervill T.1-T.13, bls. 61-69	Vélþýðingar V.1-V.5, bls. 81-85	Málrýnir M.1-M.14, bls. 98-106	Gögn G.1-G.9, bls. 107-121	Stoðtól I.1-I.8, bls. 122-132	MM ALLS
Teymi 1	125						**125**
Teymi 2		117					**117**
Teymi 3			114				**114**
Teymi 4				143			**143**
Teymi 5		2	60	18,5	97,5	3	**181**
Teymi 6	12	18		42	38	94	**204**
Teymi 7	80	30			4		**114**
Teymi 8	24	18		12			**54**
Teymi 9	12	33	18	6			**69**
Teymi 10	6	6		1	2		**15**
Mann-mánuðir alls	**259**	**224**	**192**	**222,5**	**141,5**	**97**	**1136**

Yfirlit yfir kjarnaverkefni, fagteymi og umfang máltækniáætlunarinnar.

1 MARKMIÐ MEÐ ÁÆTLUNINNI

1. MARKMIÐ MEÐ ÁÆTLUNINNI

Við leitum upplýsinga, pöntum vörur og þjónustu, skráum okkur á viðburði, lesum, hlustum og horfum á fréttir og skemmtiefni í tölvum og snjallsímum. Þessi samskipti manns og tölvu verða sífellt gagnvirkari og miðlunin margbreytilegri. Tungumálið á stóran þátt í upplifun fólks af tækjunum, sérstaklega á stærri málsvæðum þar sem raddstýring, samræðugreining og talgerving gera fólki kleift að afla sér upplýsinga og koma skilaboðum áleiðis með röddinni.

Markmið með máltækniáætlun fyrir íslensku er að tryggja að hægt sé að nota íslensku sem samskiptamáta í tækniheiminum.

Markmið með máltækniáætlun fyrir íslensku er að tryggja að hægt sé að nota íslensku sem samskiptamáta í tækniheiminum. Nú stendur yfir tæknibylting sem byggist á greiningu mikils magns af gögnum, gagnagnóttar (e. *big data*), og því að vinna úr þeim líkön með hjálp tölvugreindar. Líkönin, gögnin og greiningartæknin eru kjarninn í vitvélum sem geta nýtt sér þann margbreytileika sem felst í gögnunum. Í tilviki máltækninnar gengur þetta út á að beita tölvugreind á mikið magn málgagna (e. *language data*) þannig að vitvélarnar geti unnið með tungumálið eins og fólk gerir, ritað upp talað mál, búið til tal úr ritmáli, greint samræður og tekið ákvarðanir út frá þeim, unnið upplýsingar úr texta og fundið stafsetningar- og málfræðivillur svo að einhver dæmi séu tekin.

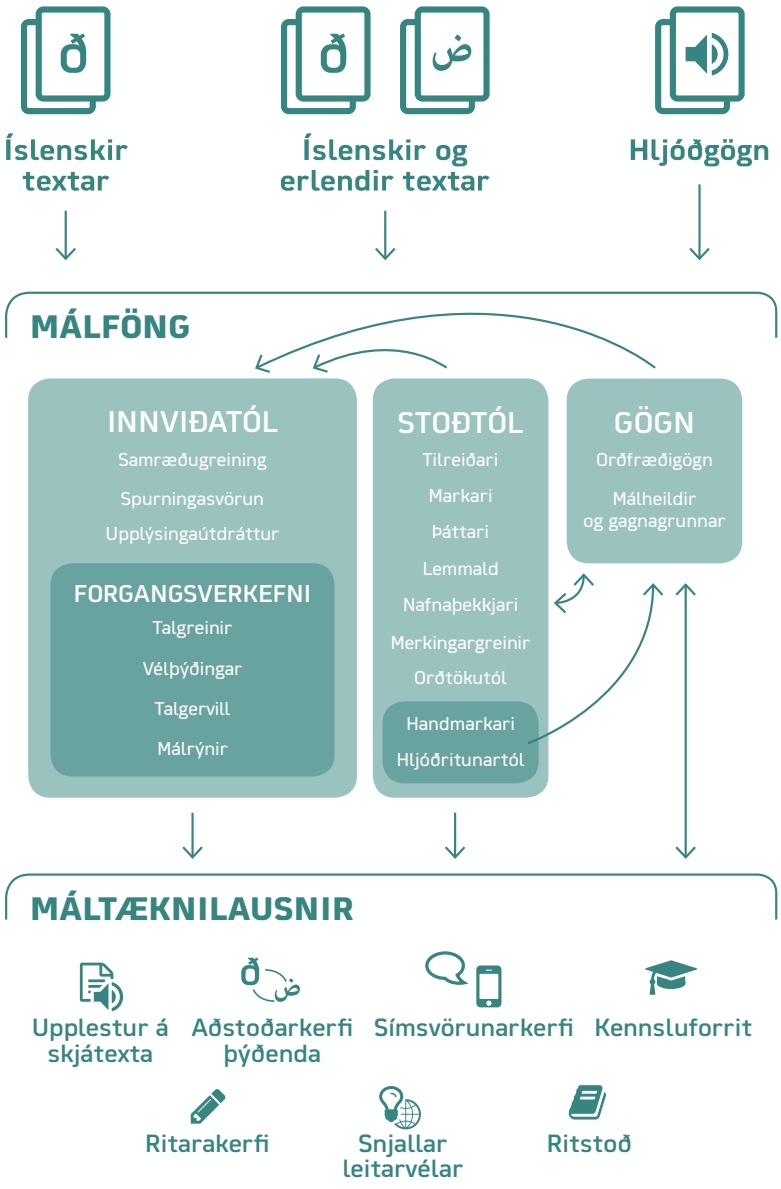
Megináherslan í áætluninni verður því á að byggja upp þá innviði sem þarf til að þróa máltækni – málföng (e. *language resources*), þ.e. gögn og stoðtól sem þarf til að útfæra hugbúnaðarlausnir með máltækni, og grunnverkfæri á borð við málrýni, þýðingarvélar, talgervla og talgreina. Tökum dæmi af talgreini. Helstu gögn sem þarf til að smíða talgreini eru stórt textasafn, framburðarorðabók og hljóðupptökur með samhliða texta. Stoðtól geta verið tengd gagnavinnslunni, eins og upptökubúnaður og innsláttarviðmót, framburðarforrit sem hljóðritar aukaorð sjálfvirkt eða forrit sem eru notuð í beinni líkanagerð fyrir smíði talgreinis eins og forrit sem þjálfa hljóðlíkan og mállíkan. Talgreinirinn sjálfur er síðan grunnverkfæri sem hægt er að samþætta í annan hugbúnað sem styðst við þessa tækni.

Þróun innviða mun lækka þann þröskuld sem almenn hugbúnaðarfyrirtæki þurfa að yfirstíga til að innleiða máltækni í hugbúnaðarlausnir sínar.

Annað markmið með verkefninu er að auka hagnýtingu og tækniþróun í máltækni. Þróun innviða mun lækka þann þröskuld sem almenn hugbúnaðarfyrirtæki þurfa að yfirstíga til að innleiða máltækni í hugbúnaðarlausnir sínar. Sú innleiðing er hins vegar ekki sjálfsögð og er lagt til að sérstaklega verði stutt við þróun á lausnum sem nýta sér máltækni. Það þarf því að leggja áherslu á að styðja við nýsköpun á Íslandi á sviði máltækni með því að styrkja sprotafyrirtæki og tækniþróun hjá stærri fyrirtækjum. Einnig þarf að halda uppi samskiptum og stuðla að samstarfi við erlenda aðila og stórfyrirtæki á sviði máltækni.

Til þess að ná þessu fram er nauðsynlegt að byggja upp þekkingariðnað í kringum máltækni hérlendis. Hlúa þarf að þeirri þekkingu sem fyrir er með því að skipuleggja samstarf og samhæfa framtak þeirra sem stunda greinina en gera einnig fleirum kleift að taka þátt í uppbyggingunni. Samstarf milli háskóla, rannsóknarstofnana og atvinnulífs þarf að samhæfa í gegnum sameiginlegt félag sem sér um að meta og velja verkefnatillögur og mæla árangur. Slíkt félag getur einnig stuðlað að samstarfi við erlend fyrirtæki og háskóla.

„MÁLTÆKNIVISTKERFIÐ“



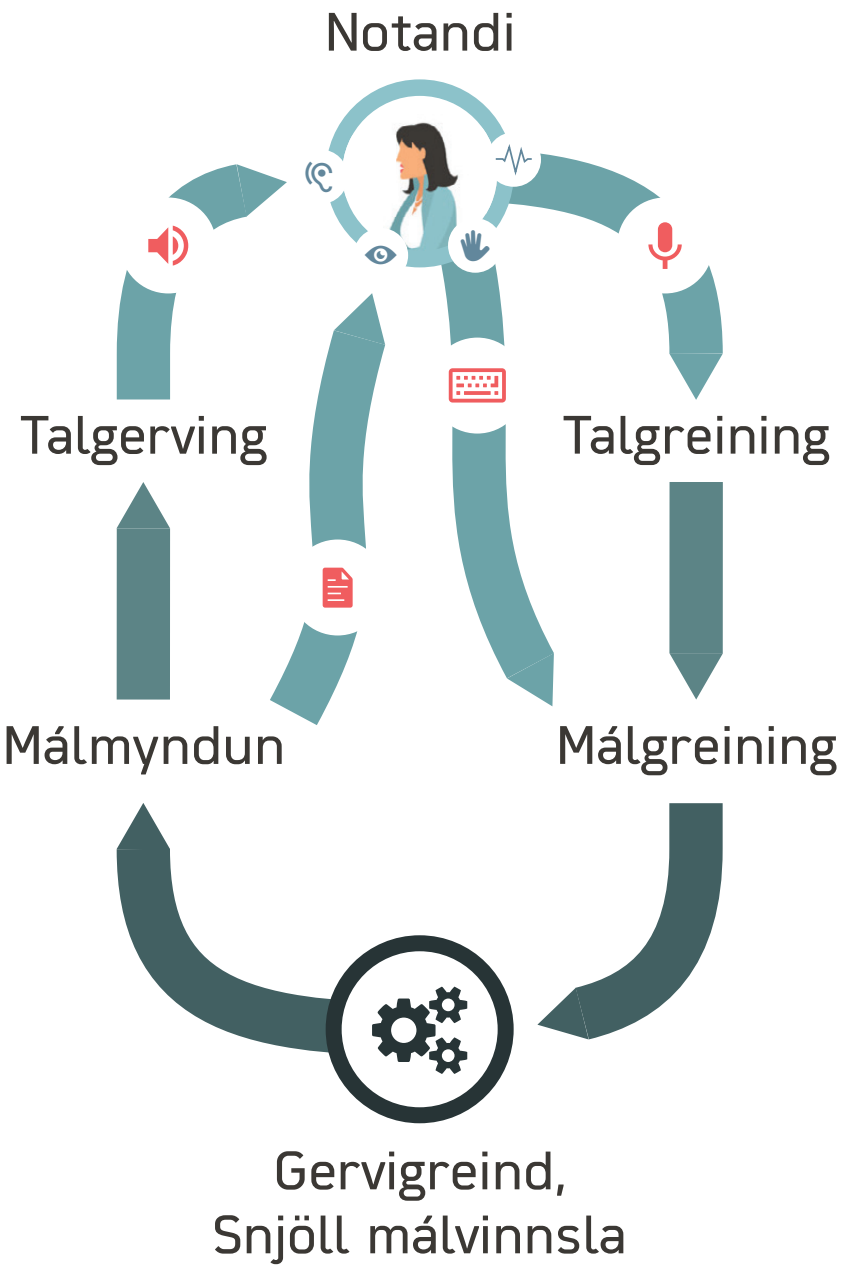
1. MARKMIÐ MEÐ ÁÆTLUNINNI

1.1 UMLYKJANDI TÆKNI

Á Íslandi hefur verið unnin töluverð grunnvinna í máltækni og smíðuð hafa verið fáein tól sem notuð eru af ákveðnum hópum. Smíðaðir hafa verið nokkrir talgervlar sem einkum blindir og sjónskertir hafa notað, stafsetningarleiðréttingatól hafa verið til frá því á níunda áratugnum og síðustu ár hefur staðið yfir þróun á talgreini. Gæði þessara tóla og nýtingarmöguleikar hafa þó hingað til verið takmarkaðir. Víða erlendis hafa sérhæfð eða samsett kerfi náð mikilli útbreiðslu. Símsvörunarkerfi og upplýsingaveitur sem notast við talgervla og talgreina, kerfi sem gera fólki kleift að leita í miklu magni af talupptökum, t.d. sjónvarps- og útvarpsþáttum, talgreinar sem aðstoða við að skrifa upp læknaskýrslur eða þingræður, talgreiningarkerfi sem notuð eru til að bæta auðkenningu, talgervlahugbúnaður sem gerir fólki, sem misst hefur málið vegna tauga-hrörnunarsjúkdóma, mögulegt að tala með augunum, þýðingarvélar sem aðstoða þýðendur og auka framleiðni þeirra og á hamfarasvæðum er farið að nota máltækni við björgunaraðgerðir. Í framtíðinni gætu vitvélar búnar hæfileikum til margmála málnotkunar bjargað mannslífum. Svona mætti lengi telja, máltækni nýtist á flestum sviðum samfélagsins.

Mikið er nú lagt í þróun gerviþjóna hjá stærstu tæknifyrirtækjum í heiminum. Apple Siri, Google Assistant, Microsoft Cortana, Samsung Bixby og Amazon Alexa eru ný gerð hugbúnaðar sem notendur eiga í samskiptum við með talmáli. Gerviþjónarnir geta leitað fyrir notendur að upplýsingum á netinu, aðstoðað þá við kaup á vöru eða þjónustu, stýrt heimilistækjunum og allt annað sem á annað borð er hægt að láta tölvu gera fyrir sig. Lykilþáttur í gagnsemi þessara tækja er geta þeirra til að nota tungumál. Á næstu árum mun talviðmót, eins og í gerviþjónum, verða notað í sífellt fleiri tækjum. Í bílum auka þau öryggi. Með því að stýra staðsetningartækjum, útvarpi og síma með röddinni getur bílstjórinn fylgst með umferðinni án truflana. Þau munu einnig auka þægindi, t.d. með raddstýringu á sjónvarpi, ljósum og öðrum rafbúnaði. Máltæknibúnaður er líka stöðugt meira notaður til að auka lífsgæði þeirra sem búa við fatlanir og sjúkdóma. Máltækni færir ólík málsamfélög nær hvert öðru og getur hjálpað til við að yfirstíga vanda sem fylgir fjölbreytilegu málumhverfi.

Ef við viljum að íslenskan verði með í þróuninni, að hægt verði að nota íslensku í samskiptum við tækin og nota tækin til að hjálpa okkur að tjá okkur á íslensku þarf að bregðast við.

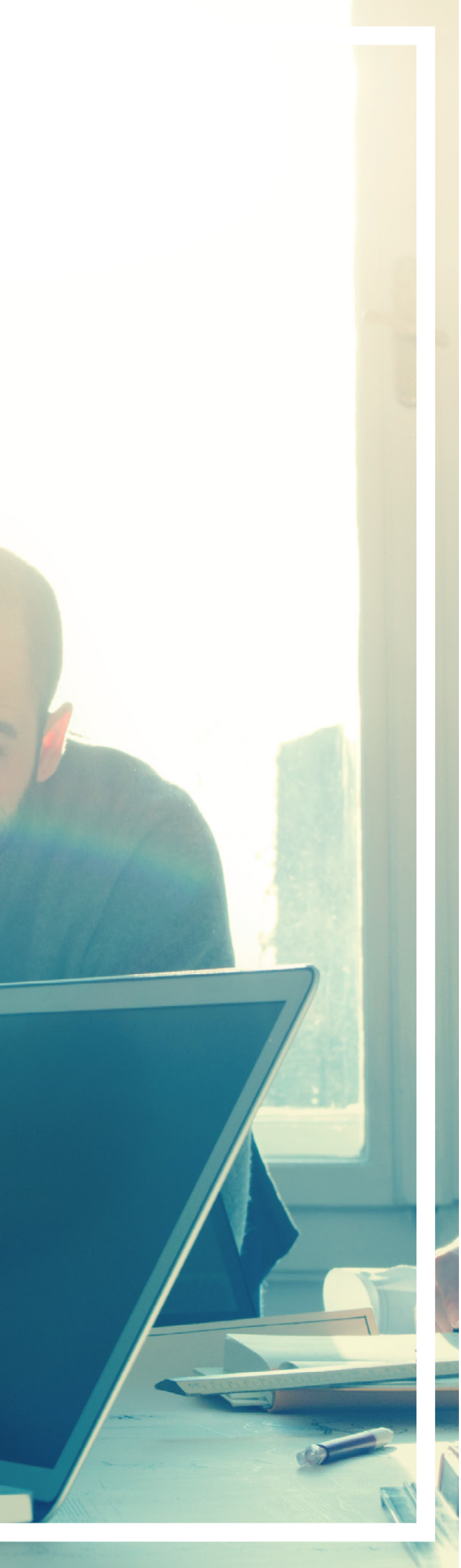


Verkfæri, eins og þau sem talin eru upp hér að framan, er hægt að gera aðgengileg fyrir alla sem tala íslensku samfélaginu til mikilla hagsbóta því að efnahagsleg áhrif máltækninnar geta líka verið umtalsverð. Nefnd hafa verið verkfæri sem nýtast til að auka framleiðni og bæta þjónustu í heilbrigðiskerfinu og stjórnsýslunni. Taltækni og vélþýðingar skapa mikil markaðstækifæri í fjarskiptageiranum og skemmtanaiðnaðinum. Hugbúnaður fyrir tölvustutt tungumálanám, fjarnámsumhverfi og forrit til að uppgötva ritstuld geta aukið gæði náms og aðgengi til náms. Hægt er að nota máltækni

1. MARKMIÐ MEÐ ÁÆTLUNINNI

til að greina umræður á samfélagsmiðlum sem getur nýst í þróun vöru, þjónustu eða til að skilja þjóðfélagsmál betur.

Allt þetta getur haft bein og sýnileg áhrif en telja má að óbeinu áhrifin yrðu enn meiri. Eins og tölvutæknin hefur aukið framleiðni í flestum atvinnugreinum, mun auðveldara aðgengi alls málsamfélagsins að upplýsingatækni og auknir notkunarmöguleikar hennar hafa talsverð áhrif á hag samfélagsins þó að erfitt geti reynst að mæla þau áhrif til fullnustu.



2

KJARNA-VERKEFNI

2. KJARNAVERKEFNI

Innviðauppbygging máltækniáætlunarinnar skiptist í fimm kjarnaverkefni. Fjögur þeirra hafa það markmið að þróa málföng og aðra innviði fyrir talgreiningu, talgervil, vélrænar þýðingar og ritvilluleiðréttingar eða málrýni. Í fimmta forgangsverkefninu er unnið að þróun málfanga almennt, almennar málheildir og orðfræðigögn búin til og nauðsynleg stoðtól þróuð. Í þessum kafla verður fjallað ítarlega um hvert kjarnaverkefni fyrir sig og áætlun sett fram um það hvernig unnið skal að þróun þeirra innan máltækniáætlunar.

2.1 TALGREINING

Þeir sem hanna og þróa raddviðmót fyrir tæki, vefi og upplýsingaveitur geti auðveldlega bætt íslensku við.

Talgreining fyrir íslensku verður þróuð þannig að þeir sem hanna og þróa raddviðmót fyrir tæki, vefi og upplýsingaveitur eða stunda upplýsingavinnslu í talmáli geti auðveldlega bætt íslensku við. Sett verður upp opið umhverfi fyrir smíði talgreina og forskriftir að algengum notkunarmöguleikum verða gerðar aðgengilegar og opnar.

Talgreining breytir töluðu máli í ritmál. Þó menn geti gert það fyrirhafnarlaust krefst mikillar tækni og útsjónarsemi að útfæra talgreiningu í tölvu. Tækninni hefur fleygt fram undanfarin ár og er þróun og notkun á talgreiningu nú mjög viðráðanlegt verkefni ef málföng og þekking eru til staðar. Þrátt fyrir það getur talgreining verið mjög margbreytileg vegna þess hversu fjölbreytilegt talmál getur verið. Talgreiningu má nota til að rita upp langan samfelldan upplestur, ræður eða dikteringar, til að fylgjast með samræðum og taka jafnvel þátt í þeim eða taka við raddskipunum til frekari greininga.

Talgreining er því eins konar lyklaborð fyrir röddina sem gerir fólki kleift að eiga í samskiptum við annað fólk, tölvur eða kerfi.

Notkunargildi og notkunarsvið

Talgreining er forsenda þess að við getum átt samskipti við tölvur og tæki með þeim hætti sem flestum er eðlilegast, þ.e. með tali. Möguleikinn á raddstýrðum samskiptum er sérstaklega mikilvægur við aðstæður þar sem ekki er hægt að nota hendur eða það er truflandi, t.d. við akstur, eða fyrir þá sem af einhverjum ástæðum hafa ekki gott vald á skrifuðu máli eða eiga í vandræðum með að nota lyklaborð eða snertiskjái.

Hvað þarf að greina?

Í talgreiningu þarf að líta til nokkurra atriða sem geta krafist sértækra lausna. Það skiptir máli hver talar, við hvaða aðstæður og um hvað er talað:

- **Hver talar?** Raddir og framburður eru ólík eftir fólki. Ákveðnir hópar eiga þó meira sameiginlegt en aðrir, t.d. eru kvenraddir líkari öðrum kvenröddum en karlröddum, fólk á svipuðum aldri frá sama málsvæði hefur svipaðan framburð o.s.frv. Það er nauðsynlegt að hafa yfir hljóðupptökum að ráða sem eru lýsandi fyrir þann hóp sem talgreinirinn á að nýtast fyrir. Hefðbundinn talgreinir virkar t.d. ekki fyrir börn þar sem raddir þeirra eru of ólíkar röddum fullorðinna.

- **Við hvaða aðstæður er talað?** Aðstæður sem geta haft áhrif á talgreiningu eru bakgrunnshávaði (umferðarniður, náttúruhljóð, skvaldur, einstök hljóðmerki eins og t.d. bjalla í þingsal), hvort fleiri eru að tala nálægt tækinu, hversu fumlaust talið er (ekki mikið af hikorðum og endurtekningum) og gæði upptöku. Við hönnun talgreinis fyrir leiðsögukerfi bíla þarf til að mynda að gera ráð fyrir bílhljóðum og röddum farþega.

- **Um hvað er talað?** Einföld raddstýring getur falist í því að einungis er leyfilegt að nota ákveðin stök orð og verkefni talgreinis er þá einungis að greina á milli leyfilegra orða. Samskipti eru þó almennt flóknari, nýta fjölbreyttan orðaforða og setningaskipan. Þau geta verið afmörkuð við ákveðið efni, t.d. læknisfræði, eða verið alveg opin. Sníða þarf talgreina að því innihaldi sem líklegt er að hann fáist við.

2.1.1 UNDIRLIGGJANDI TÆKNI VIÐ TALGREININGU (STAÐA TÆKNINNAR Í HEIMINUM)

Talgreinar eru þjálfaðir með upptökum af mörgum mismunandi röddum. Algengt er að nokkur hundruð raddir séu í svokölluðu þjálfunarsetti talgreina. Þessi fjöldi radda er nauðsynlegur til þess að hægt sé að búa til almennt hljóðlíkan fyrir talgreininn svo að hann ráði við að greina nýjar raddir. Raddirnar í þjálfunarsettinu verða að vera bæði karl- og kvenraddir og vera frá fólki á mismunandi aldri ef talgreinirinn á að geta greint sem flestar fullorðinsraddir. Ef talgreinirinn á að ráða við sérstakar mállýskur þurfa nægilega mörg dæmi um þær einnig að vera fyrir hendi. Barnaraddir

eru það frábrugðnar fullorðinsröddum að sérstök þjálfunarsöfn þarf fyrir talgreina fyrir börn.

Hefðbundinn talgreinir er þjálfaður þannig að búin eru til pör af hljóð-upptökum og textum í þjálfunarsetti. Samfelldur hljóðstraumur upptakanna er bútaður niður í stutt 25–30 millisekúndna bil sem eru tíðnigreind. Þessi greining er notuð til þess að velja ákveðin einkenni (e. *features*) sem með stöðluðum mynsturgreiningaraðferðum eru sett fram sem vigur af tölum fyrir hvert bil. Tilgangurinn með þessari greiningu er að gera einkennin óháð hvert öðru, staðla einkennavigrana milli mismunandi radda og ýkja muninn á mismunandi hljóðum. Þegar búið er að skipta hljóðmerkinu svona upp og greina það er það tengt við hljóðritun á þeim texta sem fylgir upptökunni. Úr öllu þjálfunarsettinu, upptökum og hljóðritunum, er svo búið til hljóðlíkan. Slík líkön geta verið byggð upp af einstökum hljóðum eða af hljóðasamböndum (e. *triphone models*). Bestu aðferðirnar í dag til þess að þjálfa hljóðlíkön nota afturvirk tauganet með lang-skammtíma-minnis-einingum (e. *recurrent neural networks with long-short term memory units*).

Til þess að talgreinirinn geti skrifað raunveruleg orð og texta þarf að búa til mállíkan með hjálp framburðarorðabókar, sem notast við sama hljóðritunarkerfi og hljóðlíkanið, og með stóru textasafni. Orðabók og textasafn eru valin með tilliti til þess sem talgreininum er ætlað að greina. Einfaldar skipanir eða fyrirspurnir hafa minni orðaforða og einfaldari setningaskipan en opið samskiptakerfi.

Hljóðlíkan og mállíkan eru að lokum sett saman í stöðuvél sem nefnist hulið Markov-líkan. Þar eru líkur á því að ákveðnir einkennavigrar eigi við ákveðin málhljóð reiknaðar saman við fyrirframgefnar líkur á tilteknum hljóðasamböndum, orðum og ákveðinni orðaröð. Þetta stóra líkan (eða hljóðlíkan með litlu mállíkani og stærra mállíkani að auki) notar svo tilbúinn talgreini til þess að greina tal á sama hátt og hann gerði í þjálfunarferlinu. Stundum fer ákveðin forvinnsla fram á hljóðstraumnum til þess að lágmarka bakgrunnssuð og afmarka talmerkið en síðan er hljóðmerkið bútað niður og greint. Markov-líkanið er notað til þess að finna líklegustu textasamsvörunina, orð eða setningu, en einnig geta þeir skilað frá sér grind af líklegum niðurstöðum (e. *lattice*) ef talgreiningin er hluti af stærra kerfi sem bætir upplýsingum við seinna í ferlinu.

2.1.2 KALDI OG ANNAR OPINN HUGBÚNAÐUR

Á síðustu árum hefur orðið bylting í þróun talgreina. Hún tengist að miklu leyti framgangi (djúpu) tauganetanna en þó ekki síst því að opin hugbúnaðarkerfi hafa orðið til sem innihalda öll þau algrím og aðferðir sem mest og best reynsla er komin á.

Útbreiddasta hugbúnaðarumhverfið fyrir þróun á talgreinum í dag nefnist Kaldi. Hugbúnaðurinn er gefinn út með Apache 2.0-leyfi sem er með því opnasta sem gerist. Kaldi er vel prófaður og skjalaður, algrímin sem útfærð eru í honum eru sveigjanleg og bjóða upp á ógrynni stillinga. Á meðan helstu sérfræðingar háskóla og fyrirtækja hafa tekið að sér að útfæra og viðhalda nýjustu og bestu algrímunum sem verið er að rannsaka fyrir talgreiningu hefur stærri hópur um víða veröld sett saman mikið úrval af forskriftum fyrir mismunandi notkun, tungumál og aðferðir sem talgreining býður upp á. Þetta hefur lækkað þröskuldinn töluvert fyrir þá sem hefja þróun í talgreiningu og gert það að verkum að talgreining er mun aðgengilegri tækni en áður.

Talgreining er mun aðgengilegri tækni en áður.

Í áætluninni er mælt með að leggja áherslu á að að útbúa forskriftir fyrir talgreiningu með Kaldi-hugbúnaðinum. Til þess þarf að safna gögnum og útbúa íslensk málföng og gera tilraunir með mismunandi forskriftir og stillingar sem Kaldi býður upp á. Annar opinn hugbúnaður ætti einnig að koma til álita. Hugbúnaður eins og HTK frá Cambridge-háskólanum og Sphinx frá Carnegie Mellon-háskólanum gæti einnig nýst áætluninni og ástæða er til þess að hafa þróunina eins fjölbreytta og kostur er. Ef til er einföld og ódýr leið til að búa til sambærilegar forskriftir fyrir íslensku í öðrum hugbúnaðarumhverfum ætti að nýta slík tækifæri líka.

2.1.3 TALGREINING FYRIR ÍSLENSKU OG Á ÍSLANDI

Árið 2012 varð íslenskur talgreinir aðgengilegur hjá Google. Hægt er að leita á Google með því að spyrja á íslensku og nota raddviðmót á Android-snjalltækjum (ekki þó Google Assistant). Einnig er mögulegt að tengja eigin hugbúnað við vefþjónustu Google og fá þannig aðgang að íslenska talgreininum. Gagnasafnið Málrómur var búið til fyrir talgreininn í samstarfi Háskólans í Reykjavík og Google en það er um 150 klukkustundir af upptökum á íslenskum röddum. Þetta gagnasafn er einnig vistað á Íslandi og er opið, en talgreinirinn sjálfur og öll tækni honum tengd er hins vegar

Mikilvægt er að við höfum einnig yfir opinni talgreinistækni að ráða á Íslandi sem hægt er að nýta og aðlaga að mismunandi þörfum..

eign Google. Mikilvægt er að við höfum einnig yfir opinni talgreinistækni að ráða á Íslandi sem hægt er að nýta og aðlaga að mismunandi þörfum.

Grunnþekking á talgreiningu hefur orðið til við Háskólann í Reykjavík undanfarin tvö til þrjú ár. Talgreinir fyrir flugumferðarstjórn var þróaður í samvinnu við Tern Systems ehf. en sá talgreinir byggist á afmarkaðri ensku sem notuð er í samskiptum flugstjóra og flugumferðarstjóra. Samvinnuverkefni Háskólans í Reykjavík, Landspítalans og nýstofnaðs fyrirtækis, Læknaróms, vinnur að gerð talgreinis fyrir röntgenlækna. Þá er verið að vinna verkefni fyrir skrifstofu Alþingis um að rita ræður Alþingismanna sjálfvirkt og auka þannig gæði og hraða á útgáfu þeirra. Þróun opins umhverfis fyrir íslenska talgreiningu er einnig í gangi, en stefnt er að því að gefa út forskriftir og leiðbeiningar fyrir Kalda sem leyfa tæknifólki að hefja þróun sérhæfðra talgreina fyrir íslensku á sem auðveldastan hátt.

Öll þessi vinna er byggð á málföngum sem oft eru háð því viðfangsefni sem talgreinirinn fæst við. Þannig eru málföngin sem notuð eru í talgreiningu fyrir röntgenlækna safn af læknaskýrslum og upptökum (dikteringum) þar sem orðaforði og upplestur er mjög sérhæfður fyrir skýrslutöku lækna. Þau gögn eru einnig mjög viðkvæm og er ekki hægt að nota í almennri þróun. Í öðrum tilfellum, eins og til dæmis Alþingisverkefninu, eru málföngin opnari og möguleiki á að nota þau í frekari talgreiningarvinnu. Nýlega tókst að útbúa um 550 klukkustunda gagnagrunn með Alþingisræðum og er áætlað að hægt sé að stækka hann svo hann verði um það bil tíu sinnum stærri. Hér að ofan var minnst á gagnasafnið Málróm, en gögnin úr honum og minna gagnasafni (um 40 klst) úr Hjal-verkefninu má finna á www.málföng.is.

Önnur málföng sem nauðsynleg eru til að þróa talgreina eru framburðarlýsingar og málheildir. Það er því mikilvægt að þróa slík málföng samhliða því að safna talmálsgögnum

2.1.4 EÐLILEG VILLUTÍÐNI

Einfalt er að mæla gæði talgreinis ef hægt er að bera saman texta talgreinisins við réttan texta. Samanburðurinn býður upp á að meta þrenns konar villur: orð sem hafa verið felld út, orð sem hafa verið sett inn og orð sem hefur verið skipt út úr rétta textanum. Hlutfall þessara villna af fjölda orða í textanum kallast orðvillutíðni (e. *word-error-rate*, *WER*) sem oftast er notuð til þess að mæla árangur í talgreiningu. Orðvillutíðni ræðst ekki bara af gæðum talgreinisins heldur einnig því viðfangsefni sem verið er að fást við. Til dæmis er hægt að ná orðvillutíðni niður fyrir 1% ef orðaforði er

mjög þröngur og/eða viðmælendur kerfisins eru fáir. Aftur á móti ef leyfa á hverjum sem er að ræða við kerfið um hvað sem er þá þarf það að búa yfir mjög stórum orðaforða og geta aðlagað sig að mismunandi röddum. Í slíkum aðstæðum ná allra bestu talgreinarnir sem búnir eru til í dag 5–7% orðvillutíðni, en algengara er að sjá orðvillutíðni á bilinu 10–15%.

2.1.5 INNVIÐIR FYRIR TALGREININGU

Í máltækniáætlun fyrir íslensku 2018–2022 eru sextán verkefni skilgreind fyrir talgreiningu. Þau miða að því að safna gögnum og skrá þau, þróa kjarnainnviði talgreiningar og útbúa stoðtól sem nauðsynleg eru fyrir tækniútfærslur með talgreiningu.

Gagnasöfnun fyrir talgreiningu þarf að vera stöðug og mikil í upphafi áætlunarinnar og taka mið af fjölbreytilegu notkunarsviði. Gagnasöfnun er undirstaða allrar þróunar í talgreiningu og því er lögð áhersla á að setja kraft í þessa vinnu því að án hennar gerist lítið í talgreiningarþróun. Vinnunni er skipt upp í fimm hluta: gagnaupptökur með Eyra, innslegið sjónvarps- og útvarpsefni, upptaka og innsláttur á samræðum og fyrirspurnum, samröðun á stórum málheildum og leyfismál.

Gagnasöfnun er undirstaða allrar þróunar í talgreiningu og því er lögð áhersla á að setja kraft í þessa vinnu því að án hennar gerist lítið.

Þegar skriður kemst á söfnun gagna verður hægt að þróa og hanna talgreiningu á ýmsum sviðum. Með almennum talgreini verður hægt að búa til vefgátt fyrir talgreiningu og koma henni fyrir í helstu stýrikerfum snjallsíma. Þá verður hægt að útbúa forskriftir fyrir talgreiningu fyrir samræður og fyrirspurnir og gera tilraunir með að greina tal barna og unglinga. Smíði á talgreinum fyrir ræður Alþingis og umræður, upplestur og skemmtiefni í útvarpi og sjónvarpi verður einnig að veruleika

Þróun á talgreiningu krefst margra hliðarverkefna sem gera útfærslu og notkun þjálli og betri. Í samfelldu tali er til dæmis ekki augljóst hvar skal setja greinamerki í frálagi talgreinis en án þeirra verður textinn illlæsilegur. Þá getur betri orða- og setningagreining komið að góðum notum fyrir mállíkanagerð fyrir íslensku þar sem er margbrotnara beygingakerfi og virkari orðmyndun en til dæmis í ensku og spænsku. Þekking á almennri hljóðmyndun, dreifingu á styrk og lengd hljóða í íslensku og breytileika eftir mállýskum nýtist við að gera nákvæmari talgreina og við að auðga frálag þeirra. Í samræðukerfum þarf að vera hægt að greina hver er að tala og hvenær ein rödd tekur við af annarri. Þróun hljóðlíkana fyrir barna- og unglingaraddir þarfnast síðan sérstakrar athygli þar sem slík vinna er styttra á veg komin en fyrir fullorðinsraddir. Það mun styrkja íslenska talgreiningu

að útfæra forskriftir fyrir fleiri hugbúnaðarlausnir en Kalda og því munu forskriftir fyrir HTK- og Sphinx-kerfin einnig vera útbúnar.

2.1.5.1 GAGNAUPPTÖKUR MEÐ EYRA

Áætlað er að taka upp 200 þúsund yrðingar fyrir fullorðna og annað eins fyrir unglinga, um 150 klukkustundir af upptökum fyrir hvorn hóp.

Lagt er til að haldið verði áfram með hljóðupptökur á fyrirframskilgreindum texta á sama sniði og því sem safnað var í samstarfi við Google árið 2012. Háskólinn í Reykjavík hefur haldið því samstarfi áfram og þróað hugbúnaðinn Eyra til þess að sjá um slíkar upptökur og gera þær auðveldari. Setningalisti verður útbúinn fyrir fullorðna og unglinga. Þessi söfnun hefst strax á fyrsta ári áætlunarinnar og stendur yfir í þrjú ár. Áætlað er að taka upp 200 þúsund yrðingar fyrir fullorðna og annað eins fyrir unglinga, um 150 klukkustundir af upptökum fyrir hvorn hóp. Huga verður vel að upplýstu samþykki foreldra fyrir seinni hópinn en áætlað er að gefa út báða grunnana með opnu CC BY 4.0-leyfi. Þá verður safnað 100 þúsund yrðingum fyrir fullorðið fólk sem ekki hefur íslensku að móðurmáli. Á seinni árum áætlunarinnar verður hugbúnaðurinn aðlagaður að frjálsara tali þar sem myndum verður lýst og upptökur eru endurteknar. Slá þarf slíkar upptökur inn og krefst það því meiri vinnu en lögð er í upplesturinn. Kosturinn við slíkar upptökur er hins vegar sá að frjálsara flæði næst á talmálinu og hægt verður að fá börn sem kunna ekki að lesa til þess að taka þátt í gagnasöfnuninni. Slík gagnasöfnun krefst auðvitað upplýsts samþykkis foreldra.

H.1 Gagnaupptökur með Eyra

Verkþættir:

- Hljóðupptökur unglinga (200 þúsund yrðingar)
- Hljóðupptökur fullorðinna þar sem tekið er mið af mállýskum (200 þúsund yrðingar)
- Hljóðupptökur fullorðinna sem ekki hafa íslensku að móðurmáli (100 þúsund yrðingar)
- Hljóðupptökur og innsláttur barnaradda (lýsingar á myndum)
- Hljóðupptökur og innsláttur radda fullorðinna (lýsingar á myndum)

Mannauður:

- Gagnasöfnun: 30 mánuðir
- Forritari: 2 mánuðir

Alls: 32 mannmánuðir

2.1.5.2 INNSLEGIÐ SJÓNVARPS- OG ÚTVARPSEFNI

Fjölbreyttari gögn en upplesinn texta þarf til þess að geta beitt talgreiningu á fleiri svið mannlífsins. Byrjað verður að slá inn umræðuþætti í útvarpi og sjónvarpi strax á fyrsta ári áætlunarinnar og tekur þessi innsláttur þrjú ár. Á seinni tveimur árunum verða skemmtiefni og fréttatímar teknir með. Þessi vinna verður gerð í samstarfi við útvarps- og sjónvarpsstöðvar og verður þess gætt að gagnasafnið verði gefið út með eins opnu leyfi og hægt er. Einhver stuðningsforritun þarf að eiga sér stað en gert er ráð fyrir að hægt verði að aðlaga forrit eins og SoundScriber að viðfangsefninu. Annar innsláttur fer fram á seinni hluta áætlunarinnar en þá verða upptökur af umræðufundum, samræðum, fyrirspurnakerfum og sérhæfðum munnlegum lýsingar-verkefnum slegnar inn.

H.2 Innslegið sjónvarps- og útvarpsefni

Verkþættir:

- Innslegnir umræðuþættir útvarp/sjónvarp (250 klst.)
- Útvarps- og sjónvarpsfréttir (innslegið eða samraðað)
- Innslegið skemmtiefni (50 klst.)

Mannauður:

- Ritari: 24 mánuðir
- Forritari: 2 mánuðir

Alls: 26 mannmánuðir

2.1.5.3 INNSLEGNAR SAMRÆÐUR OG FYRIRSPURNIR

Talmál er margbreytilegt og getur í samræðum og fyrirspurnum verið mjög frábrugðið ræðum, upplestri, og útvarps- og sjónvarpsefni. Í þessum verkþætti er ætlunin að taka upp og slá inn samræður á fundum og fyrirspurnir í gegnum síma. Finna þarf þátttakendur sem eru viljugir til að láta taka upp fundi og gefa út. Þetta gæti gerst með samstarfi við fyrirtæki og/eða stofnanir þar sem vinnufundir eða nefndarfundir eru teknir upp. Hægt væri að taka upp fundi hjá fyrirtækjum og gefa þeim tækifæri til að ritskoða þá áður en þeir eru gefnir út og einnig mætti gefa út opna nefndarfundi Alþingis á þessu formi líka.

Þá þarf að setja upp kerfi þar sem þátttakendur hringja inn og biðja um upplýsingar og þjónustu. Fyrirtæki og stofnanir sem gætu mögulega tekið þátt í þessu verkefni eru þau sem hafa stór innhringiver. Sem dæmi má nefna að bankar, flugfélög, orkuveiturnar og ýmsar stofnanir ríkisins halda úti dýrri þjónustu sem sér um að veita viðskiptavinum og skjólstæðingum upplýsingar sem hægt er að gera á sjálfvirkan hátt.

H.3 Innslegnar samræður og fyrirspurnir

Verkþættir:

- Upptökur og innsláttur á fundum
- Upptökur og innsláttur innhringifyrirspurna

Mannauður:

- Gagnasöfnun: 6 mánuðir
- Ritari: 6 mánuðir
- Forritari: 2 mánuðir

Alls: 14 mannmánuðir

2.1.5.4 INNSLEGNIR FYRIRLESTRAR

Í þessum verkþætti er ætlunin að safna upptökum af fyrirlestrum og slá þá inn. Finna þarf fyrirlesara sem samþykkja að fyrirlestrar þeirra séu teknir upp og gerðir aðgengilegir með þessum hætti. Hér getur verið um að ræða kennslufyrirlestra, ráðstefnufyrirlestra eða opna fyrirlestra um hvers kyns efni.

H.4 Innslegnir fyrirlestrar

Verkþættir:

- Upptökur og innsláttur á fyrirlestrum

Mannauður:

- Gagnasöfnun: 3 mánuðir
- Ritari: 3 mánuðir
- Forritari: 2 mánuðir

Alls: 8 mannmánuðir

2.1.5.5 SAMRÖÐUN Á STÓRUM MÁLHEILDUM

Stór gagnasöfn af samhliða upplestri og texta eru þegar til en henta ekki endilega til smíði talgreina á því formi sem þau eru núna. Dæmi um þetta eru Alþingisræður, opin gögn dómstóla og gögn hjá Hljóðbókasafninu.

Nokkur reynsla hefur fengist í að útbúa slík gagnasöfn fyrir talgreiningu og munu afurðir samstarfs Háskólans í Reykjavík og Alþingis verða nýttar í verkefnið. Eftir töluverðu magni af gögnum er að slægjast en þau munu nýtast í smíði talgreina fyrir svipað umhverfi og gögnin koma úr.

H.5 Samröðun á stórum málheildum

Verkþættir:

- Samraðaðar ræður Alþingis (um 5000 klst.)
- Samraðaðar upptökur frá Dómstólaráði
- Samröðun á efni Hljóðbókasafns og geymsla ítónunarprófíla

Mannauður:

- Gagnaforritari: 18 mánuðir

2.1.5.6 LEYFISMÁL

Halda þarf vel á leyfismálum fyrir þau gögn sem búin eru til í verkefninu. Markmiðið er að sem mest af þeim gögnum sem safnað verður verði gefin út með CC BY 4.0-leyfi eða sambærilegu leyfi. Setja þarf ákveðna vinnu í þessi mál sem er annars eðlis en mest af annarri vinnu sem fer fram í áætluninni enda er hún lögfræðileg frekar en tæknileg. Hafa þarf eftirfarandi atriði í huga en fleiri gætu komið upp á verkefnatímanum:

Upptökur á lestri og tali barna og unglinga: Aðalatriðið fyrir gagnasöfnun af þessu tagi er að fá upplýst samþykki foreldra fyrir þátttöku barns eða unglings í söfnun á tali. Ganga þarf frá lýsingu á verkefninu fyrir vísindasiðanefnd og útbúa ferli sem gerir foreldrum auðveldlega kleift að samþykkja (eða hafna) þátttöku barna og unglinga í þeirra forsjá.

Upplestur og texti Hljóðbókasafnsins: Það er augljóslega ekki markmið að endurútgefa efni Hljóðbókasafnsins með opnu leyfi en það ætti að vera hægt að merkja og endurraða gögnunum þannig að þau virki alls ekki eins og bækur en séu samt gagnleg fyrir talgreiningu og aðra máltækni.

Innsláttur á efni fréttamiðla: Semja þarf við fjölmiðla um leyfismál varðandi innslátt á umræðuþáttum, fréttatímum og skemmtiefni.

H.6 Leyfismál talgreinisgagna

Verkþættir:

- Hönnun á ferli fyrir upptökur barna og unglinga (yngri en 18)
- Vinna við leyfismál vegna efnis Hljóðbókasafns
- Vinna við leyfismál vegna efnis fréttamiðla (RÚV, 365 miðla o.fl.)

Mannauður:

- Sérfræðingur í leyfismálum gagna: 6 mánuðir

2.1.5.7 ÞRÓUN Á ALMENNUM TALGREINI

Almennur talgreinir verður þróaður í Kaldi-hugbúnaðinum sem byggist á þeim málföngum sem til eru en þau eru Málrómur, Mörkuð íslensk málheild og framburðarorðabók. Forskrift til þess að þjálfa nýja talgreina á meiri gögnum verður búin til og mun nýtast í öðrum talgreiningarverkefnum áætlunarinnar. Þá verður einnig hægt að byrja á annarri þróun eins og á vefgáttum og snjallsímaílagi sem byggist á þessum talgreini. Hægt verður að gera það þó að nákvæmni talgreinisins sé ekki fullnægjandi.

H.7 Þróun á almennum talgreini

Verkþættir:

- Forskrift byggð á Málrómsgögnum

Mannauður:

- Sérfræðingur í talgreiningu: 12 mánuðir

2.1.5.8 VEFGÁTTIR FYRIR TALGREININGU

Með almennum talgreini verður hægt að setja upp vefþjónustu þar sem talað mál er greint og texta skilað sem niðurstöðu. Þjónustan getur annað hvort tekið við heilum hljóðskrám eða straumi af töluðu máli. Hægt er að setja þessar vefgáttir upp með útgáfu af íslenskum talgreini sem byggist á gögnum sem eru fyrir hendi. Gæði slíks talgreinis verða kannski ekki eins og best verður á kosið í fyrstu en hægt verður að uppfæra hann með tíð og tíma þegar gögn safnast og því munu vefgáttirnar verða betri og betri í að greina tal. Þær munu þjóna mjög mikilvægu hlutverki fyrir verkefnið þar

sem mjög auðveldlega verður hægt að prófa þá talgreina sem verið er að hanna og fylgjast með notkun og árangri.

H.8 Vefgáttir fyrir talgreiningu

Verkþættir:

- Vefgátt fyrir talgreiningu á hljóðskrám
- Vefgátt fyrir talgreiningu á hljóðstraumi

Mannauður:

- Vefforritari: 12 mánuðir

2.1.5.9 TALGREINING SEM ÍLAG FYRIR SNJALLSÍMA

Þrjú helstu snjallsímastýrikerfin, Android, iOS og Windows Phone, bjóða utanaðkomandi aðilum upp á að smíða lyklaborð fyrir kerfin. Fyrir Android-stýrikerfið er þetta auðveld leið til þess að koma íslenskri talgreiningu inn í slíka snjallsíma þar sem talgreiningarhnappur fylgir lyklaborðinu og hægt væri að tengja hann við íslenska talgreiningarþjónustu. Sambærilegar aðferðir mætti nota fyrir hin stýrikerfin en gera þarf ráð fyrir að þetta þurfi að vinna sérstaklega fyrir hvert stýrikerfi.

H.9 Talgreining í snjallsímum

Verkþættir:

- Íslenskt lyklaborð með talgreiningarhnappi fyrir Android-stýrikerfið
- Íslenskt lyklaborð með talgreiningarhnappi fyrir iOS-stýrikerfið
- Íslenskt lyklaborð með talgreiningarhnappi fyrir Windows Phone-stýrikerfið

Mannauður:

- Forritari: 24 mánuðir

2.1.5.10 RADDSTÝRING, FYRIRSPURNIR OG SAMRÆÐUR

Til þess að auðvelda tækniyfirfærslu verða talgreiningarforskriftir fyrir raddstýringu, fyrirspurnir og samræður útbúnar. Hægt verður að hefja þróun á þessum forskriftum áður en gögnum er sérstaklega safnað fyrir þær, en

gert er ráð fyrir að slík gögn auki gæði þess háttar kerfa þannig að þau verði vel nothæf. Hægt er að nota almennt hljóðlíkan en þrengja það síðan með sértækum upptökum. Mesta vinnan mun fara í að útbúa viðeigandi mállíkan sem stýrir því flæði sem þessi notkun býður upp á.

H.10 Raddstýring, fyrirspurnir og samræður

Verkþættir:

- Forskrift fyrir raddstýringar
- Forskrift fyrir fyrirspurnir
- Forskrift fyrir samræður

Mannauður:

- Sérfræðingur í talgreiningu: 24 mánuðir

2.1.5.11 SÉRHÆFÐ TALGREINING

Í áætluninni verður lögð sérstök áhersla á að útbúa forskriftir fyrir talgreiningu unglinga- og barnaradda. Ástæður fyrir því eru einkum tvær: börn og unglingar er sá hópur sem hefur hvað mest áhrif á breytingar í tækninotkun og unglinga- og barnaraddir geta verið það frábrugðnar fullorðinsröddum að þær krefjast sérstakra talgreina.

H.11 Sérhæfð talgreining

Verkþættir:

- Talgreining fyrir unglinga
- Talgreining fyrir börn
- Talgreining fyrir þá sem ekki hafa íslensku að móðurmáli

Mannauður:

- Sérfræðingur í talgreiningu: 32 mánuðir

2.1.5.12 GREINARMERKJASETNING OG ENDAPUNKTASKYNJUN

Mikilvægt er að talgreinir skili frá sér læsilegum texta. Eftir að góðri orðanákvæmni er náð þarf að brjóta upp samfellt talmál með greinarmerkjum. Ekki er einfalt að ákveða hvar skuli setja punkta, kommur, spurningamerki eða önnur greinarmerki. Hægt er að ná ákveðnum árangri með því að búa til greinarmerkjalíkan út frá textamálheildum en betri árangur næst með því að taka tillit til raddarinnar líka. Þá er hlustað eftir lengri þögnum, áherslu og talanda sem fæst með góðri ítónunargreiningu.

Endapunktaskynjun er mikilvæg þegar langt samfellt talmál er greint. Talgreiningin þarf þá að geta greint hvar á að stoppa, greiða úr og ljúka orðagrindinni. Orðagrindin stækkar eftir því sem talbúturinn er lengri og fjöldi mögulegra setninga eykst. Ef lengd talbútsins sem greina á er ekki takmörkuð getur talgreiningin auðveldlega fyllt minni og búið til of marga möguleika á setningum sem greina þarf. Því er mikilvægt að búta langt samfellt tal niður og það er gert með endapunktaskynjun. Hljóðgreining fyrir greinarmerkjasetningu nýtist einnig í endapunktaskynjun. Langar þagnir, áherslur og hljómfall gefa til kynna að hægt sé að stöðva talgreininguna á þeim punkti og hefja á ný fyrir næsta bút.

H.12 Greinarmerkjasetning og endapunktaskynjun

Verkþættir:

- Greinarmerkjasetning, endapunktaskynjun og ítónunargreining

Mannauður:

- Sérfræðingur í málvinnslu: 12 mánuðir

2.1.5.13 MÁLLÍKAN FYRIR ORÐMYNDIR OG SAMSETT ORÐ

Stöðluð mállíkanagerð fyrir talgreiningu er að mestu byggð á ensku en tekur minna tillit til beygingarmála eins og íslensku með allar sínar beygingarmyndir, samsettu orð, forskeyti og viðskeyti. Einhver vinna hefur samt verið unnin fyrir slík tungumál, t.d. tékknesku og önnur slavnesk tungumál. Því ætti að skoða hvort hægt sé að búa til flóknara mállíkan sem byggist á þeirri vinnu þannig að talgreiningin verði nákvæmari.

Samsett orð og orð með forskeyti eða viðskeyti eru einnig mun algengari í íslensku en í ensku og því getur verið erfitt að koma öllum mögulegum

orðum fyrir í orðalistanum sem notaður er í talgreiningunni. Þetta gerir talgreiningunni erfiðara fyrir þar sem erfitt er að auðkenna orð sem ekki koma fyrir í orðalistanum. Hins vegar er kosturinn við samsett orð og orð með forskeyti eða viðskeyti sá að hægt er að búa til líkan af slíkum orðum og notast við framburðarlýsingu á smærri orðum og orðhlutum. Máltæknirannsóknir hafa verið gerðar á tungumálum eins og hollensku og þýsku sem hafa þessa eiginleika.

H.13 Mállíkan fyrir orðmyndir og samsett orð

Verkþættir:

- Mállíkan fyrir orðmyndir og samsett orð

Mannauður:

- Sérfræðingur í málvinnslu: 12 mánuðir

2.1.5.14 FRAMBURÐARMÁLLÝSKUR, HLJÓÐGREINING OG SAMRÆÐUGREIND

Betri skilningur á framburðarmállýskum, áherslum, ítónun og tölfræði hljóða í tungumálinu er mjög gagnlegur fyrir talgreiningu. Hægt er að ná meiri nákvæmni með slíkri vitneskju, endapunktaskynjun og greinarmerkjasetning verður auðveldari og ríkari auðkenning fæst á röddina. Þannig gæti frálag talgreinis ekki bara verið texti heldur líka auðkenning á hvers konar tal er í gangi, hver er að tala og hvaða framburðarmállýsku viðkomandi talar.

Samræðugreind í talgreiningu gefur greiningunni tækifæri á að bregðast við kvikum þáttum í talmáli, hröðum samskiptum þar sem tvær eða fleiri raddir skiptast ört á að halda uppi samræðum. Góð ítónunargreining, tölfræði yfir hljóð í tungumálinu og auðkenning á hver er að tala getur gefið lægri orðvillutíðni í talgreiningunni en einnig merkt samræðurnar fyrir frekari greiningu og þátttöku tölvugreindar.

H.14 Framburðarmállýskur, hljóðgreining og samræðugreind

Verkþættir:

- Mállýskur og hljóðgreining og samræðugreind

Mannauður:

- Sérfræðingur í hljóðgreiningu: 12 mánuðir

2.1.5.15 HLJÓÐLÍKÖN FYRIR BARNA- OG UNGLINGARADDIR

Raddir barna og unglinga eru öðruvísi en raddir fullorðinna. Í verkefninu er gert ráð fyrir að safnað verði gögnum og forskriftir þróaðar fyrir talgreiningu slíkra radda en ekki er öruggt að jafn góður árangur náist í því verkefni sökum þess að þróun talgreiningar fyrir barna- og unglingaraddir er ekki komin eins langt og talgreining fyrir fullorðna. Þetta skýrist mögulega af því að ekki hefur verið safnað jafn mikið af gögnum fyrir börn og unglinga eins og fyrir fullorðna. Hér er gert ráð fyrir að hljóðlíkan fyrir börn og unglinga sé þróað sérstaklega með þeim aðferðum sem notaðar eru fyrir erlend tungumál.

H.15 Hljóðlíkön fyrir barna- og unglingaraddir

Verkþættir:

- Aðlögun hljóðlíkana að barna- og unglingaröddum

Mannauður:

- Sérfræðingur í talgreiningu: 12 mánuðir

2.1.5.16 TALGREININGARFORSKRIFTIR Í ÖÐRUM KERFUM

Í áætluninni miðast öll þróun í talgreiningu við að Kaldi-hugbúnaðurinn sé notaður. En við gerum ráð fyrir því að eftir að helstu forskriftir fyrir þann hugbúnað hafi verið gefnar út verði hægt að yfirfæra þær nokkuð auðveldlega á annan opinn talgreiningarhugbúnað sem einnig er notaður alþjóðlega. Þar er einkum horft til Hidden Markov Model Tool Kit (HTK) frá Cambridge-háskólanum og Sphinx frá Carnegie-háskólanum. Kosturinn við að útbúa slíkar forskriftir er að það gerir fólki sem stundar þróun á talgreiningu í þeim kerfum auðveldara að bæta íslensku við rannsóknir og útfærslur.

H.16 Talgreiningarforskriftir í öðrum kerfum

Verkþættir:

- Talgreiningarforskriftir í öðrum kerfum

Mannauður:

- Sérfræðingur í talgreiningu: 3 mánuðir

2.1.6 TÆKNIYFIRFÆRSLA

Þegar innviðir verða til, gögnum hefur verið safnað, forskriftir gefnar út og stuðningsverkfæri og þjónustur tilbúnar verður hægt að yfirfæra þekkinguna og tæknina á viðskiptalausnir og samþætta talgreiningu við kerfi og forrit sem eru þegar til staðar. Eftirfarandi dæmi sýna notagildi þeirra innviða sem smíðaðir verða í verkefninu í hugbúnaði sem smíðaður er til almennrar notkunar.

2.1.6.1 TALGREINING FYRIR ÚTVARP OG SJÓNVARP

Mjög mikilvægt er að geta beitt máltækni á sjónvarps- og útvarpsefni hvort heldur sem er í gegnum hefðbundna miðla eða á vefnum. Talgreining er einn mikilvægasti þátturinn í þessari tækni en hún gerir það mögulegt að skrifa talmál upp sjálfkrafa svo hægt sé að vinna með það eins og texta. Notkunin er margþætt en augljóst dæmi er sjálfvirk textun efnis í rauntíma. Hún leyfir fólki að fylgjast með útsendingum án hljóðs sem hjálpar ekki bara heyrnarlausum heldur einnig áhorfendum sem einhverra hluta vegna geta ekki hlustað á þáttinn. Annað dæmi um notkunarmöguleika talgreiningar fyrir útvarp og sjónvarp er að hægt er að leita í efni eftir lykilorðum og greina efni eftir orðanotkun og málfari.

2.1.6.2 TALGREINING FYRIR ALÞINGI OG DÓMSTÓLA

Hið opinbera stundar innslátt talmáls í stórum stíl. Allar ræður Alþingis eru ritaðar upp og gefnar út. Hjá dómstólum eru vitnaleiðslur og munnlegar skýrslutökur einnig ritaðar upp ef málum er áfrýjað. Talgreining auðveldar þessi umsvif töluvert og býður upp á nýja möguleika á að leita og greina upplýsingar í umræðum og málum.

2.1.6.3 FYRIRSPURNAKERFI FYRIR BANKA OG UPPLÝSINGAVEITUR

Talgreinar munu auka gæði þjónustu þeirra fyrirtækja og stofnana sem þurfa að veita ítarlegar upplýsingar til viðskiptavina og skjólstæðinga sinna. Fyrirspurnakerfi gerir einstaklingum kleift að hringja inn og fá upplýsingar án þess að þurfa að bíða eftir að einhver svari í símann. Þá er fyrirspurn þeirra greind og upplýsingum komið til skila með talgervli. Möguleikar á að þróa slíkar lausnir vaxa samfara innviðauppbyggingu en góður talgreinir með mállíkani sem er smíðað með sams konar málgögnum og kerfið vinnur með er lykilatriði við þróun þeirra.

2.2 TALGERVILL

Talgervlar fyrir íslensku verða þróaðir þannig að hægt verður að framleiða margar mismunandi raddir. Sett verður upp umhverfi og málföng smíðuð og gefin út þannig að hægt verði að smíða gerviraddir á sem auðveldastan hátt. Þannig geta þeir sem vilja bæta sjálfvirkum upplestri eða talsvörun við sín kerfi sampætt talgervingu við sinn hugbúnað.

Talgerving breytir rituðum texta í talað mál. Talgervill er lykilverkfæri í máltækni þegar talmál er annars vegar og veitir tölvu eða samskiptakerfi tækifæri til að koma upplýsingum til skila með rödd. Viðfangsefnin eru mörg og oft þarf að haga hönnun og smíði talgervla eftir því. Þannig hefur talgervill sem les upp langan texta aðra eiginleika en talgervill sem þarf að svara spurningum í stuttum svörum í fyrirspurna- eða samræðukerfi. Það gefur augaleið að talgerving þarf að bjóða upp á fjölbreytta túlkun þar sem miklar upplýsingar sem vantar í texta þurfa að koma fram í rödd. Því þarf að vera hægt að stilla talgervil inn á fyrirfram ákveðna tegund af töluðu máli, halda eiginleikum eins og hljómfalli og raddstyrk svipuðum eða bæta við öðrum upplýsingum eins og skaplyndi eða áherslum í framsögn í rauntíma.

Þróun talgervla á sér langa sögu. Uppfinningamenn fyrr á tímum reyndu að herma eftir mannsrödd með ýmsum aðferðum í mekaník og með rafmagnsrásum. Verulegur árangur í talgervingu náðist samt ekki fyrr en með aðkomu tölvutækninnar. Rödd eðlisfræðingsins Stephen Hawking er frægt dæmi um talgervlarödd sem byggð er á fyrstu kynslóð tölvutalgervla. Tækninni hefur síðan fleygt fram en helstu framfarirnar á þessu sviði um síðustu aldamót áttu sér stað með svokölluðu einingavali (e. *unit selection*). Slíkt kerfi er byggt á stórum lista af svokölluðum dífónseiningum sem notuð eru til að mynda hljóðmerkið. Þessar einingar eru byggðar á einni rödd en í uppbyggingu hljóðmerkisins er besta einingin valin í rununa og hún síðan teygð til og styrkstillt til að mynda besta talið. Talgreinar byggðir á einingavali geta náð mjög góðum árangri en nýlega hefur athyglin beinst að stikuðum kerfum (e. *parametric systems*). Þá eru stikar sem lýsa röddinni ákvarðaðir þannig að eðlilegasta talmerkið sé myndað. Þessir talgreinar voru lengi vel síðri en þeir sem byggðir voru á einingavali en með tilkomu djúptauganeta náðu þeir að komast upp að hlið þeirra og enn er ekki ljóst hvor aðferðin er betri. Kosturinn við nýju stikuðu talgervlana er hins vegar að hægt er að búa til þjálli og fjölbreyttari raddir með sama magni af upptökum og öðrum málföngum. Með einingavalskerfi er ekki hægt að breyta um auðkenni á

röddinni og erfiðara er að stilla af hljómfall og ítónun. Stikuðu raddirnar eru hins vegar ennþá á tilraunastigi.

2.2.1 GÆÐI TALGERVLA

Margar áskoranir felast í að búa til góðan talgervil og eru sumar þeirra nokkuð ólíkar áskorunum í öðrum máltækniverkefnum. Ein aðaláskorunin er að komast yfir svokallaðan uggsdal (e. *the uncanny valley*) gæða og skýrleika talgervla. Fólki finnst yfirleitt ekkert að gæðum talgervla sem hljóma nógu óeðlilega til að þeim sé ekki ruglað saman við mannsraddir. Þeim finnst ekkert að því að hlusta á róbotaraddir enda eru þær þá skynjaðar sem slíkar. En þegar gæði talgervlanna aukast og gerviraddirnar verða líkari mannsröddum þá er truflandi þegar þær gera mistök og bera hljóð fram á mjög framandi hátt. Því er helsta áskorunin í þróun talgervla að komast yfir þennan hjalla og lágmarka framandi framburð.

Setja þarf upp prófunarumhverfi til að meta gæði talgervilsradda. Mat á gæðum talgervils er alltaf að einhverju leyti huglægt því að á endanum byggjast notkunarmöguleikar talgervilsins á því hversu auðvelt er að hlusta á hann til að meðtaka upplýsingar. Þá nota ekki allir talgervla með sama hætti, sumir nota þá til að skima hratt yfir texta og vilja geta hlustað á texta lesinn á margföldum hraða, á meðan aðrir vilja að hann sé áheyrilegur og þægilegur í samræðukerfum, til að hlusta á tilkynningar, blaðagreinar, námsbækur eða jafnvel fagurbókmenntir.

Lagt er til að notast verði við þrjár mismunandi aðferðir til að meta talgervilsraddir, eftir því hvað þykir eiga best við tilgang talgervilsins:

1. Almennir notendur hlusta á texta lesinn af talgervli. Notendurnir geta þrýst á hnapp í hvert skipti sem þeim þykir lesturinn óeðlilegur eða óþægilegur. Með þessum hætti er hægt að sjá hvort ákveðin atriði trufla marga mismunandi notendur og gera þá ráðstafanir til að laga þau. Eins er hægt að bera saman mismunandi raddir með tilliti til þess hversu margar athugasemdir eru gerðar við hverja talgervilsrödd fyrir sig.

2. Almennir notendur hlusta á langan texta lesinn af talgervli. Lesturinn er stöðvaður af og til og notendur beðnir um að svara spurningum úr textanum. Ef þeir geta ekki svarað spurningunum þurfa þeir að fara til baka í upptökunni og hlusta aftur þar til þeir geta svarað öllum spurningunum rétt. Hlustunin er tímamæld og hugmyndin er að sá talgervill sem fljótlegast er að hlusta á og svara öllum spurningum rétt sé áheyrilegastur.

3. Meðalmatseinkunn (e. *mean opinion score*) er notuð til að bera saman mismunandi talgervla. Aðferðin er einföld og er ætlað að meta hversu nálægt talgervillinn kemst eðlilegu tali. Hlustendur eru beðnir um að meta setningar með einkunn á bilinu 0–5, þar sem 0 er óskiljanlegt og 5 er eins og mannsrödd.

Kostirnir við þessar þrjár aðferðir er sá að þær meta fyrst og fremst það hvernig notendum talgervla hugnast þeir. Með þeim er líka auðvelt að bera saman mismunandi talgervla og bera gæði lestrarins saman við gæði lestrar leikara eða annarrar mannlegrar raddar.

2.2.2 UNDIRLIGGJANDI TÆKNI VIÐ TALGERVINGU (STAÐA TÆKNINNAR Í HEIMINUM)

Á stærstu málsvæðum í heiminum þykir nú ekki tiltökumál að smíða nýja rödd fyrir hugbúnaðarlausnir sem þarfnast talgervils. Þau málföng sem tiltæk eru og sá hugbúnaður og þekking sem orðið hefur til í gegnum tíðina gera þetta auðvelt.

Staða talgervla á stærstu málsvæðum í heiminum er sú að ekki þykir mikið tiltökumál að smíða nýja rödd fyrir þá hugbúnaðarlausn sem verið er að þróa og þarfnast talgervils.

Skipta má smíði talgervla upp í málvinnslu og talmyndun. Ílag talgervils er texti sem þarf að forvinna þannig að hægt sé að mynda talmerkið. Helstu skrefin í þessari forvinnu eru tilreiðing, textastöðlun, hljóðgreining, áherslugreining og ítónun. Tilreiðingin afmarkar öll orð í textanum og finnur einingar eins og skammstafanir og tölur þannig að textastöðlunin geti breytt þeim í fullrituð orð. Framburðarorðabók er því næst notuð við hljóðritun, en ef orð í textanum er ekki að finna í henni þarf að búa til hljóðritun fyrir það orð sjálfvirkt. Áherslu- og ítónunargreining er síðan notuð til að fá fram réttar áherslur og hljómfall í raddmerkinu. Frálag málvinnslunar, sem jafnframt er ílag talmyndunarinnar, er runa af hljóðeiningum sem merktar eru með áherslum og ítónun.

Talmyndunin getur verið byggð á einingavals- eða stikuðu kerfi. Einingavalskerfin byggjast á stóru gagnasafni raddmerkja þannig að hver dífónn í tungumálinu kemur að minnsta kosti einu sinni fyrir. Þessum merkjum er skeytt saman samkvæmt hljóðeiningarununni sem verið er að talmynda og við það val eru notuð bestunaralgrím sem taka mið af fyrirframskilgreindum markmiðum og af ítónunar- og áherslumerkingum. Í stikuðu kerfunum er ekki stuðst við hljóðupptökur heldur er talið myndað með hljóðlíkani og raddkóðara (e. *vocoder*). Stikar þessa raddkóðara ráða hvernig talmerki kemur út en þeim er stjórnað af hljóðlíkaninu sem finnur bestu stikana út frá hljóðarununni og áherslu- og ítónunarmerkingum.

Einingavalskerfi komast oft langleiðina yfir uggsdalinn en síðasti spölurinn getur verið erfiður. Ástæðan er helst sú að einingavalskerfi er alla jafna ómeðfærileg. Vænlegasta leiðin til að auka gæði slíkra kerfa er að stækka gagnasafnið þannig að líklegra sé að talgervingin nái yfir alla þá fjölbreytni sem talið og textinn þarfnast. Stikuð kerfi eru aftur á móti mun meðfærilegri þar sem hægt er að líkanagera fjölbreytileikann í textanum betur. Þessi eiginleiki veldur því hins vegar að erfiðara er að finna bestu leiðina til að mynda tal en þróun á þessu sviði er enn í gangi og þykir líkleg til árangurs.

Nýlega hefur nokkur árangur náðst með því að beita djúpum tauganetum og hefur Google til dæmis gefið út talgervil sem nefnist WaveNet. Þessi nálgun er tegund af stikuðum talgervli en í stað þess að nota tölfræðilegt hljóðlíkan og raddkóðara er tauganet notað til að mynda talið beint út frá hljóðarununni. Þessi tækni er enn þá mjög þung í vöfum og glæný en gæðin sem nást með þessari aðferð virðast vera meiri en áður hefur náðst.

2.2.3 OPINN HUGBÚNAÐUR FYRIR TALGERVINGU

Við hönnun og smíði á talgervlum hefur Festival-hugbúnaðurinn lengi verið í almennastri notkun. Honum er haldið við og hann studdur af Carnegie Mellon-háskólanum og Edinborgarháskóla og hafa margir á sviðinu stuðst við þau verkfæri sem þróuð hafa verið innan þess samstarfs. Helstu kostirnir við að hefja þróun á talgervlum sem byggjast á Festival eru þeir að hönnun og útfærsla hugbúnaðarins er mjög góð, hann hefur verið sannreyndur með mikilli notkun og samfélagið í kringum hann býr yfir mikilli reynslu og þekkingu. Gallinn við Festival er að hann er smíðaður í kringum forritunarmálið Scheme sem fátt tæknifólk þekkir núorðið. Því er kostnaðurinn við að kenna nýju fólki á Festival meiri en ef annar hugbúnaður væri valinn.

Aðrar opnar hugbúnaðarlausnir á smíði talgervla hafa verið þróaðar að undanförnu. Idlak er afbrigði af talgreiningarhugbúnaðinum Kaldi sem hefur verið snúið við fyrir talgervingu og byggist á samblandi af C++ og skeljaskipunum. MaryTTS er byggt á Java og er þróað af Þýsku gervigreindarstofnuninni, DFKI, og Saarlandsháskóla og er notað nokkuð víða fyrir önnur tungumál en ensku. Merlin-hugbúnaðurinn er þróaður af Edinborgarháskóla og byggir á vitvélahugbúnaðinum Theano og Python-forritunarmálinu. Merlin er byggður upp þannig að sérfræðingar eiga auðvelt með að búa til sínar eigin forskriftir og deila með öðrum svipað og Kaldi-hugbúnaðurinn gerir fyrir talgreiningu.

Það er ljóst að úr mörgum hugbúnaðarkostum er að velja við þróun og hönnun á íslenskum talgervli. Meta þarf hvert þessara kerfa hentar best fyrir opinn íslenskan talgreini sem getur verið aðgengilegur öllum þeim sem vilja þróa og nota talgervilstækni fyrir íslensku.

2.2.4 TALGERVING Á ÍSLANDI – STAÐAN HÉR Á LANDI

Talgerving fyrir íslensku á sér um þrjátíu ára sögu. Rödd, sem nefnd var Sturla, var þróuð af Háskóla Íslands, Öryrkjabandalaginu og Konunglega tækniháskólanum í Stokkhólmi í lok níunda áratugarins og byggðist á formendatalgervingu (e. *formant synthesis*). Röddin Snorri var þróuð í kringum aldamótin og var byggð á sömu málföngum og Sturla auk þess sem fleiri raddupptökum hafði verið bætt við. Gæði þessarar raddar voru umdeild og því var hafist handa við smíði talgervils sem byggði að fullu á einingavali. Meiri málföngum var safnað og alþjóðlega fyrirtækið Nuance fengið til að smíða talgervil sem hlaut nafnið Ragga. Enn fremur var vefviðmót hannað fyrir röddina og því var röddin oft einnig nefnd Vefþulan. Enn var óánægja með gæði raddarinnar en einnig reyndist erfitt að halda þróun á röddinni áfram þar sem eignarhald á henni var óljóst og þekking á frekari þróun ekki til staðar í landinu. Blindrafélagið fékk pólska fyrirtækið Ivona til að smíða tvær raddir, Karl og Dóru, árið 2010 og eru þær einnig byggðar á einingavalskerfi. Þær þykja mun betri en fyrri raddir og er haldið við af félaginu. Samt er nokkur óánægja með Karl og Dóru, m.a. vegna þess hvernig þau fara með erlend orð, og einnig vegna þess að ekki er hægt að auka hraðann mikið.

Saga talgervingar á Íslandi sýnir svart á hvítu hversu mikilvægt það er að búa til þekkingu og færni í máltækni hérlendis. Ekki er nóg að fela erlendum aðilum að þróa máltæknilausnir fyrir íslensku (þó megi segja að það sé betra en að gera ekkert) heldur verða að vera til aðilar hérlendis sem fara með eignarhald og útvega sérþekkingu á tækninni þegar frekari þróunar er þörf. Hugbúnaður er lifandi tækni. Kerfi, tölvur og hugbúnaður eru oft uppfærð og nýjar samskipta- og reikniaðferðir koma reglulega til sögunnar. Þá þarf þekkingu og getu til þess að aðlaga þá tækni sem er fyrir hendi þannig að dýr fjárfesting úreldist ekki.

Saga talgervingar á Íslandi sýnir svart á hvítu hversu mikilvægt það er að búa til þekkingu og færni í máltækni hérlendis.

2. KJARNAVERKEFNI

2.2.5 INNVIÐIR FYRIR TALGERVLA

Markmið talgervilsþáttar áætlunarinnar er að búa til fjölbreytt umhverfi fyrir þróun, hönnun og útfærslu á tækninni þannig að notkun hennar geti orðið eins víðtæk og hægt er. Ólíkt fyrri áætlunum um máltækni mun áhersla vera lögð á að innlendir aðilar geti látið útbúa sínar eigin raddir og að mögulegar aðferðir við að búa til talgervla verði sem flestar. Áhersla verður lögð á að safna málföngum fyrir bæði einingavals- og stikuð talgervilskerfi og að utanumhald og eignarhald á þeim málföngum sé vel skilgreint og tryggt.

Í lok áætlunarinnar verða til forskriftir sem sérfræðingar geta notað á auðveldan hátt til að útfæra annaðhvort staðlaðar raddir eða sérraddir fyrir sinn hugbúnað. Raunverulegur möguleiki verður á að búa til fjölbreytt úrval af talgervlum fyrir íslensku og ef ein rödd virkar ekki sem skyldi er líklegt að næsta rödd geri það.* Þannig gætu til dæmis fréttamiðlar og upplýsingaveitur hver fyrir sig stuðst við sitt eigið mengi af röddum við gerð talgervla.

Byggja þarf upp tvenns konar raddupptökusöfn, fyrir einingavalskerfi annars vegar og stikuð kerfi hins vegar.

Fjölbreytni í talgervilsröddum byggist fyrst og fremst á því að margs konar raddir séu teknar upp. Byggja þarf upp tvenns konar raddupptökusöfn, fyrir einingavalskerfi annars vegar og stikuð kerfi hins vegar. Fyrir einingavalskerfi fæst ein talgervilsrödd fyrir hvern upplesara og þó svo að ætlunin sé að ná fjölbreytni hvað varðar aldur og mállýskur þá mun fjöldi þátttakenda vera lágur þar sem hver og einn þarf að lesa í meira en 20 klukkustundir. Fyrir stikuðu kerfin þarf hins vegar að taka upp mun fleiri raddir en upptökur á hverri um sig þurfa ekki að vera nema ein til tvær klukkustundir. Þessar raddir þurfa að vera nógu líkar innbyrðis til að hægt sé að blanda saman upptökum og búa til talgervilsrödd úr þeim öllum. Í báðum tilvikunum þarf að passa upp á að hafa jafnmargar karl- og kvenmannsraddir.

Megináherslan í þróun innviða fyrir talgervla verður á að útbúa forskriftir fyrir einingavalskerfi og stikaða talgervla og að útbúa vefgáttir þar sem almenningur og sérfræðingar geta prófað staðlaðar raddir sem gefnar verða út af verkefninu.

Mælingar á talmáli, meðferð ritmáls og þróun á ýmsum stuðningsforritum er nauðsynleg fyrir innviðauppbyggingu og útfærslu á talgervlum. Unnið verður að sjálfvirkri hljóðritun orðmynda, hljóðrunumyndun, textastöðlun og greiningu á hljómfalli og ítónun þannig að það þjóni talgervlasmíði áætlunarinnar. Enn fremur verður stikuð talgerving skoðuð sérstaklega en hún styðst við raddkóðara sem aðlaga þarf að íslensku.

* Út úr öllum talgervlaverkefnum fyrir íslensku hafa komið ein eða tvær raddir. Það eru aldrei allir ánægðir með niðurstöðuna enda eru þarfir misjafnar og smekkur líka. Með breyttri aðferðafræði verður þetta ekki vandamál lengur.

2.2.5.1 UPPTAKA Á EININGAVALSGÖGNUM

Fyrir einingavalskerfi fæst ein talgervilsrödd fyrir hvern upplesara. Ætlunin er að taka upp átta raddir og ná fjölbreytni í aldri og framburðarmállýskum og gæta þess að hafa jafnmargar karl- og kvenmannsraddir. Ákjósanlegt er að hver þátttakandi lesi upp texta samtals í 20 klukkustundir eða meira til að ná góðri dreifingu á dífónum sem notaðir eru í einingavalskerfinu. Útbúa þarf handrit fyrir upplesturinn, ráða þátttakendur og ganga frá gögnunum. Ein aðalafurð þessa verkþáttar er nákvæm forskrift og ferli sem hægt er að fara eftir eftir að verkþættinum lýkur. Þannig verður hægt að halda áfram að taka upp raddir og búa til talgervla fyrir eigin þarfir eftir lok verkefnisins.

T.1 Upptaka á einingavalsgögnum

Verkþættir:

- Útbúa handrit til upplestrar
- Upptaka radda
- Frágangur gagna

Mannauður:

- Sérfræðingur í máltækni: 3 mánuðir
- Gagnasérfræðingur: 6 mánuðir
- Þátttakendur: 6 mánuðir
- Forritari: 3 mánuðir

Alls: 18 mannmánuðir

Athugasemdir: Þessar upptökur þurfa að fara fram í hljóðeinangruðu rými, t.d. útvarpshljóðveri eða sambærilegu.

2.2.5.2 GÖGN FYRIR RADDBLÖNDUN

Fyrir stikuð talgervilskerfi er ætlunin að safna röddum frá 40 þátttakendum sem hver um sig les upp í allt að tvær klukkustundir. Hér þarf 20 raddir fyrir hvort kyn en að öðru leyti þurfa upplesararnir að hafa áþekkar raddir þannig að þær passi vel saman við blöndun í stikuðu röddunum. Leitast verður við að fá góða og skýra upplesara.

T.2 Gögn fyrir raddblöndun

Verkþættir:

- Útbúa handrit til upplestrar
- Upptaka radda
- Frágangur gagna

Mannauður:

- Sérfræðingur í máltækni: 3 mánuðir
- Gagnasérfræðingur: 6 mánuðir
- Þátttakendur: 6 mánuðir
- Forritari: 3 mánuðir

Alls: 18 mannmánuðir

2.2.5.3 NÝTING ANNARRA GAGNA Í GERÐ TALGERVLA

Upptökur Hljóðbókasafnsins, Alþingis og ljósvakamiðla má nýta við að búa til gögn fyrir talgervla. Hér þarf að vinna í leyfismálum og koma gögnum á rétt snið þannig að upptökur og texti passi saman. Þessar upptökur nýtast ekki bara í að búa til talgervla heldur er hægt að greina ítónun og hljómfall og búa til hljóðprófíl fyrir mismunandi notkunarsvið.

T.3 Nýting annarra gagna

Verkþættir:

- Vinna í leyfismálum
- Högun nýrra gagna
- Notkun á eldri talgervilsgögnum

Mannauður:

- Sérfræðingur í leyfismálum gagna: 6 mánuðir
- Gagnasérfræðingur: 6 mánuðir

Alls: 12 mannmánuðir

2.2.5.4 VEFGÁTTIR FYRIR TALGERVLA

Þrjár vefgáttir verða settar upp þar sem hægt verður að breyta texta í tal. Fyrsta vefgáttin verður fyrir stutta texta sem slegnir eru inn í vefglugga og hægt verður að spila hljóðið beint eða fá það sent sem hljóðskrá. Vefgátt fyrir lengri texta verður einnig sett upp en þá gæti framleiðslan tekið lengri tíma þar sem ítónun og hljómfall er einnig búið til eftir aðstæðum og samhengi í texta.

Þrjár vefgáttir verða settar upp þar sem hægt verður að breyta texta í tal.

T.4 Vefgáttir fyrir talgervla

Verkþættir:

- Vefgátt fyrir stuttan texta
- Vefgátt fyrir langan texta

Mannauður:

- Vefforritari: 6 mánuðir
- Forritari: 9 mánuðir

Alls: 15 mannmánuðir

2.2.5.5 TALGERVLAR SEM FRÁLAG FYRIR SNJALLSÍMA

Flest stýrikerfi bjóða upp á forritunarskil fyrir talgervla. Talgervill verður settur upp fyrir íslensku með lítið spor (e. *footprint*) þannig að hægt verði að setja hann upp á snjallsímum sem nota Android-, iOS- og Windows Phone-stýrikerfin.

T.5 Talgervlar sem frálag fyrir snjallsíma

Verkþættir:

- Talgerving fyrir Android-stýrikerfið
- Talgerving fyrir iOS-stýrikerfið
- Talgerving fyrir Windows Phone-stýrikerfið

Mannauður:

- Forritari: 18 mánuðir

2.2.5.6 VEFLESARI

Viðmót verður sett upp þannig að hægt verði að bæta við íslenskum talgervli á vefsíður. Vefhönnuðir geta þá boðið þeim sem heimsækja síðurnar upp á að texti sé lesinn upp. Þetta er mikilvægt aðgengismál fyrir þá sem ekki eiga þess kost að lesa innihald vefsíðna og mikilvægt að gera vefhönnuðum kleift að bæta við slíkri þjónustu á vefsíðum sínum.

T.6 Veflesari

Verkþættir:

- Hönnun og uppsetning á miðlægum veflesara
- Uppsetning á veflesara fyrir vefþjón
- Uppsetning á veflesaraíbót (e. *web reader plug-in*) fyrir vafra

Mannauður:

- Forritari: 18 mánuðir

2.2.5.7 FORSKRIFTIR AÐ RÖDDUM

Hægt er að búa til forskrift að smíði einingavalstalgervla þegar textastöðlun, gerð hljóðeiningaruna og framburðarlýsingar og talupptökum er lokið. Ákveða þarf í hvaða talgervilshugbúnaði þróa á forskriftina en þegar þessi skýrsla er skrifuð koma Festvox, Merlin og MaryTTS vel til greina. Forskriftin verður gefin út á vefnum ásamt öllum málföngum sem henni tilheyra þannig að hver sem er getur endurframleitt þær raddir sem upptökurnar bjóða upp á og haldið svo áfram að þróa og betrumbæta smíðina, til dæmis með nákvæmari framburðarlýsingu, betra ítónunarlíkani og uppfærðum aðferðum sem talgervilshugbúnaðurinn býður upp á.

Raddir sem byggðar eru á gögnum frá fleiri en einum þátttakanda er hægt að smíða með stikuðum talgervlum.

Raddir sem byggðar eru á gögnum frá fleiri en einum þátttakanda er hægt að smíða með stikuðum talgervlum. Stikaða hljóðlíkanið er annaðhvort útfært með tölfræðilegum aðferðum eða tauganetum. Forskrift fyrir slíkan talgervil verður gefin út fyrir þann hugbúnað sem álitlegastur er á þeim tíma sem verkefnið er unnið. Þessa forskrift má síðan nota til þess að halda áfram að þróa raddir með betri aðferðum, auknum málföngum og fleiri upptökum.

T.7 Forskriftir að röddum

Verkþættir:

- Útgáfa á forskrift fyrir einingavalsraddir
- Útgáfa á forskrift fyrir stikaðar raddir

Mannauður:

- Sérfræðingur í talgervlum: 18 mánuðir

2.2.5.8 MAT Á GÆÐUM TALGERVLA

Eins og fram kemur í Kafla 2.2.1 þá er mat á gæðum talgervla mikilvægt en vandasamt. Notendur geta verið mjög gagnrýnir á gæði talgervla og við smíði þeirra þarf að setja upp notendaprófanir til þess að bera saman og meta árangur. Í þessum verkþætti verður sett upp kerfi til að meta talgervla og notendur fengnir til þess að hlusta á raddir.

T.8 Mat á gæðum talgervla

Verkþættir:

- Uppsetning hugbúnaðar til gæðamats á talgervlum
- Skipulagning notendaprófana

Mannauður:

- Forritari: 9 mánuðir
- Sérfræðingur í talgervlum: 9 mánuðir
- Gagnasérfræðingur: 6 mánuðir

Alls: 24 mannmánuðir

2.2.5.9 UNDIRBÚNINGUR TEXTA OG STÖÐLUN OG ÁHERSLUGREINING

Forvinnsla í talgervingu miðar að því að breyta rituðum texta í hljóðritaða runu tákna. Fyrstu skrefin í forvinnslunni eru tilreiðing (e. *tokenization*) og textastöðlun (sjá 2.5.3.3). Textastöðlun breytir öllum einingum í textanum sem ekki eru orð í orð, eins og til dæmis skammstöfunum og tölustöfum. Þannig er til dæmis setningunni *„Þjóðhátíðardagur Íslands er 17. júní.“* breytt í *„Þjóðhátíðardagur Íslands er sautjándi júní.“*

Forvinnsla í talgervingu miðar að því að breyta rituðum texta í hljóðritaða runu tákna.

2. KJARNAVERKEFNI

Áður en texti er hljóðritaður getur verið gott að ákveða hvar í orði áherslan lendir. Þetta liggur nokkuð beint við í flestum tilfellum en fyrir samsett orð verða áherslureglur flóknari. Forritun slíkra reglna gæti haft mjög góð áhrif á gæði talgervla fyrir íslensku.

T.9 Undirbúningur texta, stöðlun og áherslugreining

Verkþættir:

- Textastöðlun fyrir íslensku
- Áherslugreining

Mannauður:

- Sérfræðingur í máltækni: 18 mánuðir

2.2.5.10 SJÁLFVIRK HLJÓÐRITUN

Talgervill þarf að geta breytt texta í málhljóð. Til þess þarf hann að geta hljóðritað allar mögulegar orðmyndir og lagað hljóðritunina að samfelldu tali ef ekki er um stök orð að ræða. Góð og vel skilgreind framburðarlýsing er nauðsynleg til að hægt sé að þjálfa hugbúnað sem getur hljóðritað óþekktar orðmyndir. Eftir að einstök orð eru hljóðrituð þarf búnaðurinn að skoða síðustu og fyrstu hljóð samliggjandi orða og beita samlögunarreglum þar sem það á við.

Við þróun á sjálfvirkri hljóðritun er nauðsynlegt að fylgjast vel með villutíðni og bæta framburðarlýsinguna ef finnast gloppur í henni sem valda aukinni villutíðni í kerfi þjálfuðu á gögnum hennar.

Til að talgervill geti notað ólíkar framburðarmállýskur þarf framburðar-orðabók að hafa nægilega mörg dæmi um öll helstu tilbrigði í hverri fram-burðarmállýsku fyrir sig til að sjálfvirkt hljóðritunartól geti lært reglurnar.

Hljóðritunartólið yrði notað í talgervlum til að hljóðrita óþekkt orð og orðmyndir en einnig þarf að vera hægt að nota það eitt og sér, til að mynda þegar smíðaðir eru talgervlar fyrir afmörkuð svið sem hafa takmarkaðan og sérhæfðan orðaforða. Þá er hægt að keyra út framburðarorðabók fyrir talgervilinn, annars vegar með þeim framburði sem skráður er í framburðar-orðabókina og hins vegar með framburði sem forritið býr til eftir þeim reglum sem það hefur lært. Þann framburð geta þeir sem smíða talgervilinn farið yfir til að ganga úr skugga um að hann sé réttur.

Hér þarf því annars vegar að gefa út, með skjöluðum árangri, forrit sem notar ákvarðanatré eða aðrar vélnámsaðferðir til hljóðritunar. Hins vegar þarf að smíða vefviðmót ofan á virknina til að auðvelda notkun á tólinu einu og sér.

T.10 Sjálfvirk hljóðritun

Verkþættir:

- Þjálfun og prófun á hljóðritunartólum með framburðarlýsingu
- Hljóðritunarbúnaður
- Vefviðmót

Mannauður:

- Sérfræðingur í máltækni: 6 mánuðir
- Hljóðfræðingur: 9 mánuðir
- Forritari: 3 mánuðir

Alls: 18 mannmánuðir

2.2.5.11 GREINING Á TALANDA OG ÍTÓNUNARGREINING

Bera þarf kennsl á hljóðprófíl upplestrar (e. *speaking style*) og aðlögun að þeirri raddbeitingu sem talgervillinn krefst. Hljómfallið þarf að vera mismunandi eftir því hvort talgervillinn er að halda fyrirlestur, lesa auglýsingar, taka þátt í samræðum, lesa upp úr bókum eða halda ræðu. Talgervillinn þarf einnig að geta breytt ítónun og áherslum eftir viðfangsefni og eftir því hvernig textinn er. Því þarf að koma til sambland af textagreiningu og greiningu á þeim hljóðeiningum sem standa til boða í talgervlinum ef um einingavalskerfi er að ræða. Í talgervlum með stikuðum hljóðlíkönum þarf greiningin að geta haft áhrif á stikana þannig að réttar áherslur og ítónun komi fram.

T.11 Greining á hljómfalli og ítónunargreining

Verkþættir:

- Smíði á ítónunargreini

Mannauður:

- Merkjafræðingur: 12 mánuðir

2.2.5.12 MYNSTUR OG SETNINGAR

Mynstur og setningar er listi sem inniheldur sjaldgæf stafamynstur í íslensku og setningar sem hafa að geyma orð með þessum mynstrum. Setningarnar voru teknar úr skáldsögum frá árunum í kringum 2000. Tilgangur listans var að tryggja að þegar hljóðupptökum hefði verið safnað í Hjal-verkefninu væru þar dæmi um öll stafamynstur í íslensku. Í gagnasafninu eru 1433 setningar og þær eru aðgengilegar með CC BY 3.0-leyfi.

Við smíði nútíma talgervla þarf að safna talsvert fleiri setningum en eru í þessu gagnasafni. Því er þörf á því að skilgreina aðferðir til að smíða stærri leslista sem innihalda öll möguleg mynstur í íslensku. Þær aðferðir yrðu notaðar til að búa til sem besta leslista fyrir upptökur miðað við fjölda setninga sem ætlunin er að taka upp.

T.12 Mynstur og setningar

Verkþættir:

- Skilgreina aðferðir til að búa til leslista
- Búa til tilbúna leslista.

Mannauður:

- Sérfræðingur í máltækni: 2 mánuðir

2.2.5.13 STIKAÐIR TALGERVLAR FYRIR ÍSLENSKU

Stikaðir talgervlar (e. *parametric synthesizers*) er ný tækni sem virðist vera að sanna sig þegar þessi skýrsla er skrifuð. Því er gert ráð fyrir að þessi tækni muni ná bestu raddgæðunum og að með henni verði búnir til þjálustu talgervlarnir strax á verkefnatímanum. Í þessum verkþætti er ætlunin að laga stikaða talgervla að íslensku. Í upphafi verður megináherslan lögð á gerð tölfræðilegs stikaðs hljóðlíkans sem notað verður til að stýra raddkóðara sem býr til tal. Tölfræðilegt stikalíkan fyrir íslensku hefur ekki verið búið til áður og því er mikilvægt að gera ráð fyrir góðum tíma í þessa vinnu. Hægt er að nota almennan raddkóðara til þess að búa til talið en þeir hafa venjulega verið þróaðir sérstaklega fyrir ensku. Aðlögun slíkra raddkóðara að íslensku er því mikilvæg.

Nýjasta tæknin notar djúptauganet í stað tölfræðilegs hljóðlíkans, en einnig er hægt að nota tauganetin í skrefin sem koma bæði á undan og eftir hljóðlíkaninu, þ.e. í sjálfvirka hljóðritun og í raddkóðun. Gert er ráð fyrir að nokkur vinna fari í að aðlaga þessa nýjustu tækni að íslensku.

T.13 Stikaðir talgervlar fyrir íslensku

Verkþættir:

- Tölfræðilegt stikað hljóðlíkan
- Raddkóðari fyrir íslensku
- Stikað hljóðlíkan byggt á tauganeti

Mannauður:

- Sérfræðingur í talgervlum: 21 mánuður
- Merkjafræðingur: 12 mánuðir

Alls: 33 mannmánuðir

2.2.6 TÆKNIYFIRFÆRSLA

Bein nýting á talgervingu snýst fyrst og fremst um að gera ritmál aðgengilegt með upplestri. Þessi tækni er mikilvæg fyrir blinda og sjónskerta en getur einnig nýst fólki sem kemur því ekki við að lesa texta en getur hlustað á upptökur. Talgerving er einnig mjög gagnleg sem hluti af stærri máltæknilausnum þar sem er fléttað saman talgreiningu og textagreiningu. Dæmi um slík verkefni má finna í kafla 5.2.

2.2.6.1 VEFLESARI

Talgervlar eru mikilvægir í að gera ritmál aðgengilegt þeim sem ekki koma því við að lesa efnið en hafa samt tök á að hlusta. Vinsælt er að bæta við upplestrarhnappi á vefsíður þannig að hægt sé að hlusta á innihald þeirra. Þessa tækni má útfæra á ýmsan hátt og býður hún upp á mörg tækifæri til nýsköpunar.

2.2.6.2 LESARI FYRIR RAFBÆKUR

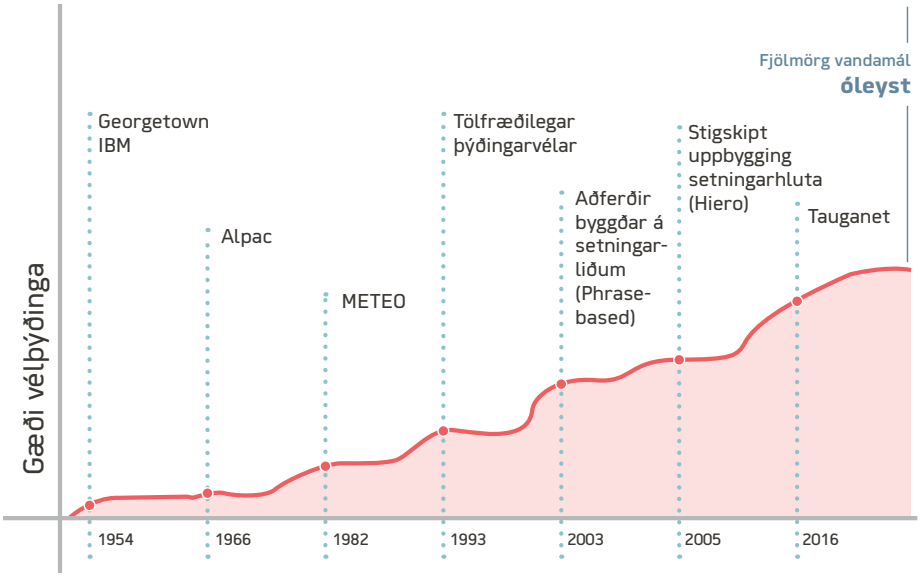
Með velútfærðum talgervli er hægt að láta lesa upp rafbækur og skjátexta sjálfvirkt. Fólk getur til dæmis notað upplestur talgervils til að njóta skáldskapar eða til að hlusta á námsefni. Í öllum tilvikum þarf að huga vel að ítónun og áherslum og passa upp á að hafa gott úrval af röddum.

2.3 VÉLÞÝÐINGAR

Smíðuð verður opin þýðingarvél sem þýðir á milli íslensku og ensku. Gæði þýðinga skulu vera nægilega mikil til að gagnast við þýðingar á ákveðnum sviðum svo að þýðendur geti fullunnið texta hraðar en ef þeir þýddu frá grunni.

Fyrstu hugmyndirnar um hvernig nota mætti tölvur til að þýða texta voru settar fram seint á fimmta áratug síðustu aldar. Þær byggðust á aðferðum sem notaðar voru við að leysa dulkóðun í síðari heimsstyrjöldinni og kenningum um almennar grundvallarreglur tungumála. Fáum árum síðar hófust rannsóknir að einhverju marki í háskólum víðs vegar um Bandaríkin. Þýðingarbúnaðurinn notaðist við reglur og talsverðar vonir voru bundnar við hann. Nú, næstum því 70 árum síðar, eigum við hins vegar enn mjög langt í land með að smíða þýðingarvél sem jafnast á við mennskan þýðanda.

Nú, næstum 70 árum eftir fyrstu tilraunir til að nota tölvur í þýðingum, eigum við enn langt í land með að smíða þýðingarvél sem jafnast á við mennskan þýðanda.



Helstu nýjungar í vélþýðingum frá upphafi tölvualdar

Á allra síðustu árum hafa vélþýðingar náð því marki að verða gagnlegar bæði fyrir þá sem vilja átta sig á innihaldi texta á tungumálum sem þeir eru ekki læsir á og til að flýta fyrir vinnu þýðenda við tungumál sem þeir eru sérfræðingar í. En vélþýðingar eru ekki leyst vandamál. Þær hafa ákveðnar takmarkanir sem eru ólíkar eftir því hvaða tækni er beitt.

Þegar nútímaþýðingarvélar eru notaðar til að þýða texta sem er ætlaður til birtingar þarf alltaf að yfirfara textann og laga. Þetta á við um öll tungumálapör. Í mörgum tilvikum er tæknin þó orðin það góð að hægt er að vinna

umtalsvert hraðar með aðstoð þýðingarvéla, enda eru vélþýðingar notaðar af þýðendum víða um heim til að auka framleiðni og auka gæði þýðinga.

Þýðingaiðnaðurinn í heiminum veltir sem nemur um 40 milljörðum bandaríkjadala á ári, að stórum hluta í Evrópu. Þörfin fyrir tækni sem getur auðveldað þýðendum störf sín og aukið framleiðni er mikil og vaxandi. Evrópusambandið hefur lagt umtalsverða vinnu í að þróa þýðingarvélar fyrir öll opinber tungumál sambandsins. Þær eru notaðar til að aðstoða þýðendur í stjórnsýslu, t.d. við þýðingar á lögum, reglugerðum, samningum og öðrum opinberum skjölum. Reynslan þar er sú að þýðingarvélar nýtist í þýðingum á u.þ.b. 50% af skjölum sambandsins og framleiðniaukningin sé um 20–30%. Gagnsemin er þó mjög misjöfn eftir tungumálum.

2.3.1 STAÐA TÆKNINNAR OG HELSTU AÐFERÐIR

Fyrstu vélþýðingarkerfin voru byggð á reglum sem smíðaðar voru fyrir hvert tungumálapar fyrir sig. Undanfarin ár hafa slíkar aðferðir verið á miklu undanhaldi og frá því fyrir aldamót hafa tölfræðilegar aðferðir sem byggjast á miklu magni af gögnum verið ráðandi. Hefðbundin tölfræðileg vélþýðingarkerfi nota forsagnarlíkön (e. *predictive model*) til að kenna tölvu hvernig á að þýða texta og nota til þess samhliða málheildir á tveimur tungumálum. Yfirleitt er þá um að ræða texta sem hefur verið skrifaður á einu tungumáli og þýddur á annað og samsvarandi setningum á tungumálunum tveimur raðað saman. Með samhliða málheild má reikna út líklegustu þýðingarnar en til að úttakið sé gott og gagn sé að því þurfa málheildirnar að vera nógu stórar því að ef ekki er dæmi um orð eða orðasamband í málheildinni þá getur þýðingarvél sem byggist á svona gögnum augsýnilega ekki þýtt það. Ýmsar útgáfur eru til af þýðingarvélum sem byggjast á tölfræðilegum líkönum: þær geta notast við stök orð, setningar eða setningarhluta, setningarfræðilega greiningu og fleiri aðferðir. Kosturinn við þessa tækni er að þýðingarvélarnar sjálfar geta unnið með mörg mismunandi þýðingarpör án þess að aðlaga þurfi þær sérstaklega því að aðlögunarvinnan hefur að mestu leyti farið fram í gagnasöfnunum sjálfum. Til að laga kerfisbundnar villur og bæta þýðingarnar eru stundum einnig notaðar reglur sem keyrðar eru á inntaks- eða úttakstexta. Ókosturinn við þessa tækni er að það getur verið erfitt og kostnaðarsamt að setja saman nógu stórar samhliða málheildir, sérstaklega fyrir minni tungumál.

Þýðingarvélar sem byggjast á tauganetum er nýjasta aðferðin í vélþýðingum. Þróun þeirra hefur aðeins staðið í örfá ár og Google hóf t.d. ekki að nota þær fyrr en síðla árs 2016. Þær eru sagðar geta náð betri árangri en hefðbundnar

tölfræðilegar aðferðir. Tæknin á að líkja eftir tauganetum í mannsheilanum. Við þjálfun slíkra véla eru þær mataðar á samhliða gögnum og við hvert lag sem gögnin fara í gegnum lærir vélin mynstur eða leynda strúktúra sem hún finnur í gögnunum. Þegar tauganet er notað í þýðingarvél geta þessi mynstur verið varpanir á milli orðmynda, orða eða orðasambanda á einu tungumáli yfir í annað, málfræði tungumálanna sem unnið er með og undantekningar frá málfræðireglum. Með því að afkóða þessi mynstur „lærir“ þýðingarvélin tungumálin og það sem þarf til að færa texta úr öðru tungumálinu yfir á hitt. Auk samhliða tvímála málheildar nýtir þýðingarvélin sér einmála málheildir til að læra betur hvernig setningar eru myndaðar á markmálinu. Allt þetta nýtir hún sér svo þegar henni er sagt að þýða úr einu tungumáli á annað.

Gæði vélþýðinga ráðast af ýmsum þáttum. Eiginleikar tungumálanna sem verið er að þýða á milli hafa mikið að segja, t.d. hvort beygingarkerfi sé flókið og hversu virk orðmyndun er í málunum. Ef beygingarkerfið er flókið og orðmyndun mjög virk rekst kerfið oftar á sjaldgæfar orðmyndir sem það hefur ekki séð áður. Þess vegna skiptir magn og gæði þjálfunargagna miklu máli. Því færri óþekkt orð því betra. Vísbendingar eru um að tauganetin gagnist tungumálum með flókna málfræðilega byggingu mun betur en eldri aðferðir. Í nýrri grein sem sérfræðingar MT@EC (vélþýðingardeildar Þýðingamiðstöðvar Evrópusambandsins) skrifuðu um vélþýðingar fyrir ungversku kemur t.d. í ljós að með hjálp tauganeta hafi verið hægt að smíða þýðingarvél sem gagnast þýðendum og er notuð. Sú þýðingarvél sem stofnunin hafði áður smíðað fyrir ungversku og sem byggðist á tölfræðilegum aðferðum skilaði hins vegar ekki nægilega góðum árangri til að vera notuð. Gagnsemi fyrir þýðendur er kannski mikilvægasti mælikvarðinn á gæði þýðingarvéla. Ef þýðingarvélarnar gagnast þeim ekki er hagnýtt gildi þeirra ekki mikið.

Tauganetsþýðingarvélar eru hraðvirkar og fyrstu tilraunir hafa gefið góða raun. Þær skila eðlilegri texta en tölfræðilegu þýðingarvélarnar, þ.e. þær eru líklegri til að skila af sér texta sem líkist mannlegu máli. Það getur þó mögulega skapað vanda því að því eðlilegra yfirbragð sem þýddi textinn hefur, því erfiðara getur verið fyrir þýðendurna sem nota kerfin að sjá þýðingarvillurnar, sem eru auðvitað áfram fyrir hendi þó að þær séu líklega færri en í eldri kerfum. Það gæti haft neikvæð áhrif á afköst þýðenda sem nota slíkar vélar sér til fulltingis. Þetta þarf að meta og mæla. Þess vegna eru þýðendur mikilvægustu samstarfsaðilarnir við þróun þýðingakerfa. Þeir búa ekki aðeins til gögn sem notuð eru við þjálfun kerfanna heldur geta þeir veitt nauðsynlega endurgjöf á þróunarstigi kerfis.

2.3.2 OPINN HUGBÚNAÐUR FYRIR VÉLÞÝÐINGAR

2.3.2.1 MOSES

Moses er tölfræðileg þýðingarvél, gefin út undir opnu LGPL-leyfi (sjá kafla 3.1.1). Þróun við Moses hófst árið 2005 og hún hefur verið mjög virk síðan. Moses er bæði notuð við rannsóknir og í kerfum sem rekin eru í viðskiptaskyni, t.d. hafa bæði Google og Microsoft notað Moses í sínum kerfum.

Í Moses og öðrum tölfræðilegum þýðingarvélum eru kerfin þjálfuð á miklu magni samhliða gagna. Moses notar þau til að læra hvernig þýða skuli setningar eða setningarhluta. Einnig notar kerfið einmála gögn, yfirleitt margfalt stærra gagnasafn en samhliða gagnasafnið, til að læra hvernig markmálið, tungumálið sem þýtt er á, skuli líta út.

Þjálfunarferlið í Moses tekur inn samhliða gögn og finnur dæmi um það þegar sömu orð og orðasambönd koma fyrir í samhliða setningum eða setningarhlutum. Kerfið dregur ályktanir um merkingu út frá því og út frá líkindatöflum sem það býr til með hlutfallslegri tíðni slíkrar pörunar í gögnunum. Kerfið er einnig hægt að þjálfa á ítarlegri greiningu á textunum og þá getur það nýtt sér upplýsingar um setningarlega stöðu eða aðra málfræðilega þætti til að setja upp líkindatöflur.

Segja má að kerfið sé tvískipt. Annars vegar er um að ræða þjálfunartól, sem notuð eru til að búa til líkindatöflurnar, þýðingarlíkönin og mállíkan fyrir markmálið. Hins vegar er það vélþýðandinn sem nýtir þýðingarlíkönin og mállíkanið til að þýða texta úr upprunamáli yfir á markmál.

Í þessu ferli er hægt að stilla kerfið með ýmsum hætti og gefa ólíkum þýðingarlíkönum mismikið vægi. Eins er hægt að velja á milli mismunandi þýðingaralgríma og mismunandi tegunda af tungumálalíkönum. Það er breytilegt eftir tungumálum hvað hentar best og því getur þurft töluverðar tilraunir til að ná sem mestri nákvæmni.

2.3.2.2 NEMATUS OG OPENNMT

Nematus og OpenNMT eru þýðingarvélar sem notast við tauganetsaðferðir. Þær eru báðar í mikilli þróun um þessar mundir. OpenNMT er gefin út með MIT-leyfi af Harvard-háskólanum í Bandaríkjunum og SYSTRAN, vélþýðingafyrirtæki sem hefur verið starfandi um langt skeið. Nematus er samvinnuverkefni vísindamanna við háskóla í Bandaríkjunum, Þýskalandi og Skotlandi, gefið út með BSD-leyfi. Bæði leyfin eru mjög opin og bæði

kerfin eru glæný, kynnt í ritrýndum greinum snemma árs 2017. Um miðjan maí 2017 gaf rannsóknardeild Facebook út þýðingarvél sína, *fairseq*, með opnu leyfi og ritrýndri grein. Vinnuhópnum gafst ekki tími til að kynna sér hana til hlítar og því verður ekki fjallað um hana hér.

Ástæða þess að þessi kerfi koma fram á sama tíma er mjög aukinn áhugi á notkun tauganeta við að leysa vandamál sem tengjast máltækni. Áhuginn kemur til vegna þess að tilraunir hafa gefið góða raun. Vélþýðingar með tauganetum fengu litla athygli og voru rannsakaðar af sárafáum allt til 2014 en þá urðu kaflaskil og tveimur árum síðar hafði orðið bylting, staðan var gjörbreytt og vísindamenn sem vinna í þessum geira beina nú langflestir athyglinni að tauganetstækninni.

Aðferðirnar sem notast er við byggjast á grein gervigreindar sem fengið hefur mikinn byr allra síðustu ár og kallast djúp tauganet. Grunnhugmyndin er að allt vélþýðingarferlið er sett upp í eitt tauganetslíkan, en nokkrar mismunandi tegundir tauganetslíkana eru til.

Í upphafi kaflans er gerð stuttlega grein fyrir grunnhugmyndinni í virkni tauganeta. Kerfin tvö sem hér eru nefnd, Nematus og OpenNMT, byggja í grunninn á sömu tækni og eru um margt lík. Ekki þykir ástæða til að greina nákvæmlega muninn á kerfunum hér en kanna þarf hvort kerfið nýtist betur við smíði á þýðingarvél sem vinnur með íslensku.

2.3.3 VÉLÞÝÐINGAR FYRIR ÍSLENSKU

Tvö þýðingarkerfi hafa verið búin til fyrir íslensku. Bæði byggjast á reglugrunni. Annað þeirra er Tungutorg sem þýðir í báðar áttir á milli íslensku og ensku, úr íslensku á dönsku og úr esperantó á íslensku. Þetta kerfi var smíðað af Stefáni Briem. Það er lokað og kóðinn hefur ekki verið gefinn út. Hitt kerfið er Apertium-is-en. Það er frumgerð, sem byggir á Apertium þýðingarkerfinu. Það var þróað af nemendum í HR á árunum 2009–2010. Hvorugt þessara kerfa er í almennri notkun og tæknin sem þau byggja á er ólíkleg til að ná betri árangri en nýjar aðferðir á borð við tauganetsvélþýðingar (NMT).

Google Translate þýðir einnig á milli íslensku og annarra tungumála. Sú þýðingarvél er ónákvæm enda vinnur hún eins með alla texta og gerir engan mun á efni texta eftir mismunandi sérsviðum og er þess vegna líkleg til að gera villur við þýðingar á margræðum orðum, orðasamböndum og hugtökum. Google Translate er lokaður hugbúnaður og því getur enginn annar en Google lagað hann að sérþörfum eða þróað áfram.

2. KJARNAVERKEFNI

Erlend ráðgjöf um vélþýðingar fyrir íslensku

Höfundar skýrslunnar leituðu ráðgjafar hjá erlendum sérfræðingum í vélþýðingum um það hvernig best væri að móta stefnu um íslenskar vélþýðingar..

Höfundar skýrslunnar leituðu ráðgjafar hjá erlendum sérfræðingum í vélþýðingum um það hvernig best væri að móta stefnu um íslenskar vélþýðingar. Annars vegar var fundað með sérfræðingum MT@EC, vélþýðingardeildar Þýðingamiðstöðvar Evrópusambandsins, og hins vegar eistneskum sérfræðingum sem unnið hafa að þróun vélþýðinga innan eistnesku máltækniáætlunarinnar.

MT@EC hefur það hlutverk að þróa þýðingarvélar fyrir opinberar stofnanir innan Evrópusambandsins, í Noregi og á Íslandi. Samtals hefur hópurinn skyldur gagnvart 26 tungumálum, þ.e. öllum opinberum tungumálum ESB auk norsku og íslensku. Vegna skorts á íslenskum þjálfunargögnum hefur hópurinn þó ekki getað sinnt íslensku og hann vinnur nú með 24 tungumál í 78 tungumálapörum. Þjónustan sem hópurinn veitir stjórnsýslunni er ókeypis og verður það til ársins 2020. Eftir það er stefnt á að stærra verkefni taki við, eTranslation, sem einskorðast ekki við þýðingar á opinberum skjölum.

Fyrsta útgáfa þýðingarkerfis MT@EC var gefin út árið 2013 og frá upphafi voru útvaldir þýðendur fengnir til þess að prófa kerfið og gefa ábendingar. MT@EC hefur notast við tölfræðileg líkön og opna tölfræðilega þýðingarkerfið Moses. Á síðari hluta árs 2016 hóf hópurinn fyrst tilraunir með tauganetskerfi.

Flókin málfræðileg uppbygging íslenskunnar og mikið af óþekktum orðum, fyrst og fremst samsettum orðum, er líkleg til að gera hefðbundnum tölfræðilegum þýðingarvélum erfitt fyrir.

Markmið MT@EC er fyrst og fremst að þýða skjöl fyrir stjórnsýsluna og þýðingarvélar þeirra eru sérsniðnar að þeim. Þýðingarkerfi MT@EC er hægt að nota í gegnum vefviðmót eða með því að nýta vefþjónustu. Hönnun kerfisins miðar að því að þýða skjöl í heild sinni fremur en einstakar setningar. Venjuleg notkunarleið er að hlaða upp skjali til þýðingar og svo er þýðingin send aftur með tölvupósti þegar hún er tilbúin. Hópurinn stefnir á að reka tvöfalt kerfi – eitt sem leggur áherslu á hraða (á kostnað gæða) og annað sem leggur áherslu á gæði (á kostnað hraða).

Gæði vélþýðinga eru mjög misjöfn eftir því milli hvaða tungumála er þýtt. Tilraunir hafa verið gerðar með íslensku hjá MT@EC en þær hafa ekki skilað nothæfum niðurstöðum enda höfðu þeir afar lítið þjálfunarsafn, um 400 þúsund samhliða setningar eða setningarhluta. Almennt telja sérfræðingar MT@EC að 25–35 milljón samhliða setningar eða setningarhlutar sé góður grunnur. Stærsti grunnur sem þeir hafa er með um 50 milljón samhliða setningum eða setningarhlutum, það er ensk–spænsk málheild.

Flókin málfræðileg uppbygging íslenskunnar og mikið af óþekktum orðum, fyrst og fremst samsettum orðum, er líkleg til að gera hefðbundnum tölfræðilegum þýðingarvélum erfitt fyrir. Það á auðvitað við um fleiri tungumál þar sem við sambærilegan vanda er að etja. Hjá MT@EC hafa tilraunir með tauganet gefið góða raun fyrir ungversku, sem áður var í svipaðri stöðu og íslenskan, og tilraunir með tauganetsvélar standa yfir á fleiri tungumálum sem reynst hafa erfið. Kostir tauganetanna liggja fyrst og fremst í getu þeirra til að skilja flókna málfræði.

Sérfræðingar MT@EC segja að fyrir íslensku megi gera ráð fyrir að setja þurfi saman töluvert stór söfn þjálfunargagna.

MT@EC notar opinn hugbúnað, Moses, fyrir tölfræðilegar aðferðir og NEMATUS og OpenNMT fyrir tauganet. Sérfræðingar þar segja að fyrir íslensku megi gera ráð fyrir að setja þurfi saman töluvert stór söfn þjálfunargagna til að þýðingarvélar skili góðum árangri. Fyrir fyrstu tilraunir telja þeir þó að um 2 milljónir samhliða setninga og setningarhluta ættu að nægja til að setja grunnlínu hvað varðar lágmarksgæði og nákvæmni. Ómögulegt sé þó að segja til um það fyrirfram hversu mikið af gögnum þurfi til að fá gagnlegar niðurstöður, til að svara því þurfi rannsóknir og tilraunir.

Hjá MT@EC hafa verið gerðar tilraunir með að nýta málfræðilega mörkun og setningagreiningu í vélþýðingar. Það virðist geta aukið gæði þýðinga smávægilega en það er tímafrekt og þess vegna ekki gert að staðaldri. Þýðingarvélar þeirra forvinna gögn, staðla texta, svo sem greinarmerki og tölur, gæta þess að notaðir séu bæði hástafir og lágstafir við þjálfun og þýðingar og laga önnur smávægileg atriði. Fyrir tiltekin tungumál geta ákveðin atriði verið líkleg til að þýðast vitlaust. Eftirvinnsla úttakstexta, sem er löguð að hverju tungumáli fyrir sig, getur lagað það og hjálpað þýðanda með því að draga athygli hans að hugsanlegum villum. Þessa þætti má ekki vanmeta. Fyrir- og eftirvinnsla getur bætt þýðingar umtalsvert, ekki síst þegar tölur eru þýddar eða annað sem hlítir skýrum kerfisbundnum reglum en er breytilegt á milli tungumála.

Til að vélþýðingar fyrir tungumál eins og íslensku, sem fáir tala, verði gagnlegar þarf gott samstarf við sem flesta þýðendur sem vinna með íslensku, bæði til að búa til samhliða gögn og til að veita mikilvæga endurgjöf við þróunina. Sérfræðingar MT@EC leggja til að unnið sé að því að fá sem flesta þýðendur til að leggja sínar þýðingar inn í sameiginlegt kerfi gegn því að geta notað það gjaldfrjálst.

Til að vélþýðingar fyrir tungumál eins og íslensku, sem fáir tala, verði gagnlegar þarf gott samstarf við sem flesta þýðendur sem vinna með íslensku.

Í eistnesku máltækniáætluninni hefur verið unnið að vélþýðingum frá árinu 2015. Meginmarkmiðið er að gæði vélþýðinga verði næg til að kerfin gagnist við þýðingar á ákveðnum sviðum og að þýðendur geti fullunnið

texta hraðar en ef þeir þýddu frá grunni. Meginvinnan hefur annars vegar legið í því að laga aðferðir sem gagnast hafa í vinnu með önnur tungumál að eistnesku og að gera tilraunir með þýðingarfyrirtækjum til að smíða búnað til að þýða texta á afmörkuðum sviðum. Á fyrstu stigum hefur talsverð vinna farið í að taka saman nýtanleg gögn og setja upp samhliða málheildir. Meðal vandamála sem lögð hefur verið áhersla á í máltæknirannsóknum fyrir eistnesku er hvernig best sé að vinna með samsett orð, en í eistnesku er orðmyndun mjög virk, rétt eins og í íslensku. Þá er reynt að brjóta orð upp í orðhluta fyrir þjálfun og við forvinnslu texta. Sé það gert rétt skilar það betri þýðingum. Erfitt er fyrir þýðingarvélar að mynda setningar í eðlilegri orðaröð á markmálinu, sérstaklega þegar markmálið er eistneska (en ekki enska) og þýtt er á milli ensku og eistnesku. Bestu niðurstöðurnar fást, eins og við má búast, þegar setningar eru stuttar. Svo virðist sem nægilega góður árangur hafi þegar náðst til að gagnlegt sé að nýta vélþýddan texta við þýðingar en ítarlegar rannsóknir hafa ekki verið gerðar á því. Nú er verið að rannsaka hvort þýðingarvélar sem nota tauganet skili betri niðurstöðum og fyrstu niðurstöður benda til þess. Orðaröð og málfræði er að miklu leyti rétt en í um fjórðungi þýddra setninga er einhverju sleppt í þýðingunni eða hluti af þýðingunni er efni sem ekki er í upprunalega textanum. Svona villur getur verið erfitt að greina en gera þarf rannsóknir á því hversu mikil áhrif þær hafa á eftirvinnslu þýðenda. Líklegt má telja að í vinnu með íslensku þurfi að glíma við mörg sömu vandamálin og við er að eiga í vinnu með eistnesku þannig að áhugavert er fyrir íslenska sérfræðinga í vélþýðingum að fylgjast með framvindu mála í Eistlandi.

Áherslur í þróun þýðingarvéla fyrir íslensku

Leggja þarf áherslu á að þróa íslenska þýðingarvél sem er gagnleg..

Leggja þarf áherslu á að þróa íslenska þýðingarvél sem er gagnleg. Einfaldast er að meta það út frá því hvort þýðendur noti þýðingarvélina sér til aðstoðar við vinnu sína. Ekki er hægt að gera ráð fyrir því að þýðingarvélar nýtist á öllum sviðum fyrst um sinn heldur þarf að skilgreina hvar ávinningurinn er mestur og hvar líklegast er að hægt sé að ná árangri. Fyrstu skrefin eru að setja saman gagnasöfn og finna út hvaða þarfir þarf að uppfylla fyrir gagnlega íslenska þýðingarvél og hvaða tækni hentar íslenskunni best.

Reynsla MT@EC er að gæði þýðinga eru mjög misjöfn eftir því milli hvaða tungumála er þýtt. Flókin málfræðileg uppbygging íslenskunnar og virk orðmyndun kallar líklega á talsvert stór gagnasöfn. Vísbendingar eru um að tauganet séu líkleg til að skila betri árangri fyrir íslensku en tölfræðilegar þýðingarvélar, a.m.k. ef tekið er mið af rannsóknum á öðrum tungumálum þar sem við sambærilegar flækjur er að eiga. Í því samhengi

má nefna slavnesk tungumál, finnsk-úgrísk mál og þýsku. Því ætti að leggja áherslu á tauganetskerfi en gagnlegt væri að gera tilraunir með hefðbundnar tölfræðilegar aðferðir líka til að fá samanburð og tryggja að verið sé að velja rétta leið.

Vísbendingar eru um að tauganet séu líkleg til að skila betri árangri fyrir íslensku en tölfræðilegar þýðingarvélar.

2.3.4 GÆÐAMAT

Hægt er að meta gæði vélþýðinga með ýmsum hætti. Matsaðferðirnar eru mismunandi, eftir því hvað nákvæmlega er verið að mæla. Nota má sjálfvirkar aðferðir til að meta hvort þýðingarvél hafi batnað eða versnað við kerfislægar breytingar á hugbúnaðinum eða viðbætur. Aðrar aðferðir eru hins vegar gagnlegri til að meta hvort hafa megi not af þýðingarvélinni, hvort hún flýti fyrir vinnu þýðenda við að ganga frá þýðingum til útgáfu.

Við gerum grein fyrir fjórum aðferðum til að meta gæði þýðingarvéla. Þessar aðferðir eru ýmist notaðar hjá MT@EC, vélþýðingadeild Þýðingamiðstöðvar Evrópusambandsins, eða voru notaðar í SUMAT-verkefninu sem gekk út á að þróa þýðingarvélar til að þýða skjátexta. Tvær aðferðanna voru notaðar bæði hjá MT@EC og í SUMAT-verkefninu.

1. Hjá MT@EC er fylgst með því hversu stórt hlutfall þýðenda í hverju tungumálapari (t.d. enska -> þýska eða rúmenska -> enska) skoðar vélþýðingar eða hefur þær til hliðsjónar við þýðingarvinnuna. Þar hefur komið í ljós að talsverður munur er á þessu hlutfalli milli tungumálapara. Munurinn stafar líklega að mestu leyti af því að vélþýðingarnar eru misgóðar eftir tungumálapörum. Þar sem vélþýðingarnar eru bestar eru þýðendur líklegastir til að hafa þær til hliðsjónar enda geta þær þá gagnast þeim. Að jafnaði eru vélþýðingar skoðaðar í um 50% tilvika þegar öll tungumálapör eru skoðuð. Þýðendur Þýðingamiðstöðvarinnar hafa alltaf aðgang að vélþýðingum og því má nota þennan mælikvarða til að skoða hvort þýðingarvélarnar batni yfir lengri tíma og fyrir hvaða tungumál þær séu gagnlegastar.

2. TER (*Translation error rate*) mælir hversu margar breytingar þarf að gera á vélþýðingu til að fá þá þýðingu sem þýðandinn skilar af sér. Þá er hlutfall orða sem þarf að breyta til að fá endanlega þýðingu reiknað út. Þessi mælikvarði sýnir hversu mikla eftirvinnslu þarf fyrir hvert tungumálapar fyrir sig og þannig er hægt að bera saman nákvæmni vélþýðingarinnar. Þessi aðferð er notuð bæði hjá MT@EC og í SUMAT-verkefninu.

3. BLEU-kvarðinn er þekktur í rannsóknum og þróun á þýðingarvélum. Hann er notaður hjá MT@EC og í SUMAT-verkefninu. Hann hentar vel til að meta hvort breytingar sem gerðar eru á þýðingarvél séu líklegar til að skila betri árangri. Hann er ekki alltaf mjög gagnlegur til að meta endanleg gæði þýðinga og heldur ekki til að bera saman kerfi sem byggjast á ólíkum undirliggjandi aðferðum.

4. Í SUMAT-verkefninu voru þýðendur tímamældir til að meta gagnsemi vélþýðinga í hverju tungumálapari fyrir sig. Þá eru þeir annars vegar látnir þýða án þess að nota þýðingarvél og hins vegar með aðstoð þýðingarvélar. Niðurstöður verkefnisins voru breytilegar á milli tungumálapara en að meðaltali jókst framleiðni þýðenda, mæld í skjátextum á mínútu, um 35,5%.

Við þróun á þýðingarvél fyrir íslensku er gagnlegt að nýta sem flestar aðferðir til að meta gæði. Hvaða aðferðir eru notaðar ræðst af því á hvaða stigi þróunar er verið að prófa aðferðirnar. TER skyldi nota á öllum stigum þegar hægt er að fá þýðendur til að meta gæði vélþýðinga og þegar prófa skal gæði þýðingarvéla eru tímamælingar líklega besta leiðin til að meta gagnsemi.

2.3.5 ÞRÓUN INNVIÐA FYRIR VÉLÞÝÐINGAR

Meginmarkmið með þróun vélþýðinga fyrir íslensku innan máltækniáætlunar er að smíða þýðingarvélar sem gagnast þeim sem starfa við þýðingar og eykur framleiðni þeirra. Innan áætlunarinnar er skynsamlegast að einbeita sér að einu tungumálapari, ensku/íslensku. Til að ná markmiðinu þarf að setja saman nægilega góð gagnasöfn, smíða opið almennt þýðingarkerfi og stuðningskerfi fyrir það og setja upp verkferla til að aðlaga það að ákveðnum sérsviðum.

2.3.5.1 GAGNASÖFNUN OG MÁLHEILDIR

Engar opnar samhliða málheildir eru til fyrir íslensku en þær eru lykillinn að því að hægt sé að þróa þýðingarvélar. Mikilvægt er að hefjast handa sem allra fyrst við að setja saman slíkar málheildir. Miða ætti við að á tímabilinu yrði komið upp samhliða málheild með 25–30 milljón pörum, þ.e. setningum eða setningarhlutum.

Fara skal í tvö verkefni í uppbyggingu málheilda með opnum leyfum fyrir vélþýðingar. Fyrra verkefnið gengur út á að búa til samhliða málheildir með sjálfvirkum aðferðum úr efni sem aðgengilegt er á vefnum. Textar á

Wikipediu og OpenSubtitles eru aðgengilegir öllum og hluti þeirra er á mörgum tungumálum. Þar að auki er mikið magn af veftextum aðgengilegur í CommonCrawl-gagnagrunninum. Þessi gögn hafa verið notuð til að smíða samhliða málheildir fyrir önnur tungumál. Skoða þarf hvernig það hefur verið gert annars staðar og nýta þær aðferðir til að byggja upp sambærilegar samhliða málheildir fyrir íslensku.

V.1 Samhliða málheild með vefefni

Verkþættir:

- Samhliða Wikipediu-málheild.
- Samhliða OpenSubtitles-málheild.
- Samhliða CommonCrawl-málheild.

Mannauður:

- Tölvunarfræðingur: 24 mánuðir
- Málfræðingur: 6 mánuðir

Alls: 30 mannmánuðir

Athugasemdir: Miða skal við að þessi málheild sé byggð upp með sjálfvirkum aðferðum svo að hægt sé að stækka hana með einföldum hætti eftir því sem aðgengilegt efni á vefnum eykst.

Í seinna verkefninu verður búin til samhliða málheild úr skjölum sem Þýðingamiðstöð utanríkisráðuneytisins hefur þýtt vegna EES-samningsins. Um 7000 skjöl eru tiltæk á íslensku og samsvarandi skjöl eru til á ensku og í mörgum tilvikum á fleiri Evrópumálum. Skjöl og setningar í textum skjalanna verða pöruð saman til að búa til samhliða málheild úr þessum gögnum.

2. KJARNAVERKEFNI

V.2 Samhliða málheild með EES-þýðingum

Verkþættir:

- Skjöl á íslensku og ensku pöruð saman.
- Setningar paraðar saman.

Mannauður:

- Tölvunarfræðingur: 12 mánuðir
- Málfræðingur: 18 mánuðir

Alls: 30 mannmánuðir

Athugasemdir: Miða skal við að þessi málheild sé byggð upp með sjálfvirkum aðferðum en gögnin yfirfarin, a.m.k. að hluta til, svo að til verði samhliða málheild sem hægt er að treysta á að sé rétt.

Í þessari vinnu, hvort sem um er ræða EES-gögnin eða önnur gögn, er notaður hugbúnaður á borð við *hunalign* sem parar saman setningar úr textum á tveimur tungumálum. Ef samhliða málheildin á að vera alveg rétt þarf að yfirfara pörunina og laga hana þar sem hugbúnaðurinn gerir villur. Þar sem markmiðið er að ganga úr skugga um að til séu gögn sem treysta má að séu rétt þarf að meta hversu stór hluti textanna þurfi að uppfylla þau skilyrði. Þá þarf að kanna hvaða villuhlutfall er þolanlegt til að óyfirfarin málheild skili tilætluðum árangri.

Aðrir möguleikar á samhliða málheildum eru einnig fyrir hendi. Sjónvarpsstöðvar og útgefendur myndefnis búa yfir skjátextaþýðingum og bókaútgefendur þýðingum á erlendum bókum. Það efni er hins vegar að langmestu leyti verndað með höfundarrétti og þyrfti annaðhvort lagabreytingar eða mikla samningavinnu til að fá leyfi til að nýta slíkt efni í þýðingarvélar. Við gerum því ekki ráð fyrir nýtingu á slíku efni innan kjarnaverkefna máltækniáætlunar.

Þegar þýðingarvélar eru sniðnar að tilteknum sérsviðum þarf að vera fyrir hendi samhliða málheild með textum á því sérsviði. Þýðingar á EES-reglugerðum gæti verið fyrsta tilraun í þá átt en kortleggja þarf hvort og þá á hvaða sérsviðum hægt væri að byggja upp samhliða málheildir fyrir íslensku.

ELRC er verkefni á vegum Evrópusambandsins sem gengur út á að safna samhliða málheildum og öðrum gögnum sem nýtast í þýðingarvélum.

Verkefnið hefur skyldu til að sinna 30 Evrópumálum, þar á meðal íslensku. Kanna þarf möguleika á samstarfi við verkefnið eða fá aðstoð þaðan til að koma á fót samhliða málheildum fyrir íslensku.

Stór einmála málheild með íslenskum textum er nauðsynleg til að hægt sé að búa til mállíkan sem þýðingarvélar nota til að læra að setja saman setningar á íslensku. Risamálheildin uppfyllir þau skilyrði. Fjallað er um hana í kafla 2.5.1.3. Úr henni verður gagnlegt að smíða mállíkön til nota í þýðingarvélum.

Tvímála orðabækur og orðalistar gagnast þar sem gloppur eru í samhliða málheildum og til að velja réttar þýðingar á ákveðnum sérsviðum. Hluti orðasafna í íðorðabanka Árnastofnunar er aðgengilegur með CC BY-SA-leyfi, en önnur tvímála orðasöfn ekki. Mikilvægt er að kanna hvort hægt sé að ráða á því bót. Engin ensk-íslensk orðabók hefur verið í smíðum hjá opinberum aðilum og því gæti það verið að einhverju leyti takmarkandi á þessu sviði. Hugtakasafn Þýðingamiðstöðvar utanríkisráðuneytisins er aðgengilegt á vefnum. Það geymir þýðingar á hugtökum sem notuð hafa verið í þeim skjölum sem Þýðingamiðstöðin hefur þýtt en ekki almennan orðaforða í ensku. Það hefur ekki verið gefið út með leyfi sem gerir notkun þess mögulega í máltæknihugbúnaði.

2.3.5.2 INNVIÐIR

Gera þarf tilraunir með þýðingarvélar, helst bæði tölfræðilega þýðingarvél og tauganetsvélar. Í fyrstu er hægt að notast við þýðingarminni Þýðingamiðstöðvar utanríkisráðuneytisins (sjá 2.5.1.19 Þýðingarminni Þýðingamiðstöðvar), um 1,2 milljón setningapör, til að þjálfa vélarnar en bæta svo við gögnum úr samhliða málheildum eftir því sem þau verða til. Við lok tilraunaverkefnis þarf að velja þann opna hugbúnað sem er líklegastur til að skila bestum árangri fyrir íslensku og halda áfram þróun á því kerfi. Í þessu verkefni verða prófuð nokkur opin tól: Moses, OpenNMT og NEMATUS. Notuð verða þýðingarminni Þýðingamiðstöðvar og önnur samhliða gögn sem kunna að verða aðgengileg á verkefnistímanum.

2. KJARNAVERKEFNI

V.3 Grunnlína í vélþýðingum

Verkþættir:

- Setja upp valdar opnar þýðingarvélar og laga tiltæk gögn að hverri fyrir sig.
- Gera athuganir á gæðum úttaks hverrar þýðingarvélar fyrir sig.
- Greina þarfir fyrir forvinnslu og eftirvinnslu texta.
- Velja bestu þýðingarvélina og áætla gagnaþörf til að hún nái tilætluðum árangri.

Mannauður:

- Tölvunarfræðingar: 18 mánuðir
- Þýðandi: 6 mánuðir

Alls: 24 mannmánuðir

Gefa þarf út þýðingarvél og öll opin gögn sem henni fylgja þegar ákveðnum vörðum er náð. Fyrsta varðan er einfaldlega sú að þýðingarvélin skili setningu á íslensku þegar hún er mötuð með setningu á ensku, óháð gæðum. Unnið skal með kerfið sem valið var í V.3.

V.4 Opin íslensk þýðingarvél

Verkþættir:

- Þróa þýðingarvélina, setja upp forvinnslu- og eftirvinnslureglur og annað sem þarf til að ná sem lægsta TER (sjá 2.3.4 Gæðamat).
- Setja upp leiðbeiningar og þau forrit sem þarf til að hægt sé að setja upp sértækar þýðingarvélar með einföldum hætti.

Mannauður:

- Tölvunarfræðingar: 72 mánuðir
- Þýðandi: 18 mánuðir

Alls: 90 mannmánuðir

Athugasemd: Verkefnið er til þriggja ára og settar vörður sem gera þarf grein fyrir á sex mánaða fresti.

2.3.5.3 AÐGANGSSTUÐNINGUR

Setja þarf upp forritaskil (API) til að veita öllum sem áhuga hafa aðgang að kerfum sem eru í þróun og setja upp prófunarumhverfi þar sem almenningi gefst kostur á að setja inn sinn eigin texta, fá niðurstöður úr mismunandi þýðingarvélum og velja þá niðurstöðu sem þykir best. Hvort tveggja verður rekið meðfram þróun þýðingarvélarinnar til að veita sem flestum aðgang að henni á þróunarstigi og fá sem fyrst endurgjöf á vinnuna. Með þessum hætti má til dæmis með einföldum hætti bera saman afurðir máltækniáætlunar við lokuð verkfæri á borð við Google Translate.

V.5 Viðmót á þýðingarvél

Verkþættir:

- Sett upp forritaskil (API).
- Smíðað prófunarumhverfi.
- Forritaskil og prófunarumhverfi sett í rekstur.

Mannauður:

- Tölvunarfræðingar: 18 mánuðir

2.3.6 TÆKNIYFIRFÆRSLA

Nákvæmni og gæði vélþýðinga ræðst af því hversu vel þýðingarvél tekst til við að færa innihald rétt á milli tungumála og hversu læsileg þýðingin er á markmálinu. Með því að afmarka vélþýðinguna við tiltekið sérsvið er hægt að auka nákvæmni umtalsvert. Þannig er hægt að draga úr líkum á tvíræðni í þýðingum á milli mála. Þýðingarvélin er líklegri til að velja rétta merkingu á markmálinu þegar orð eða orðasamband á einu máli getur haft ólíka merkingu á öðru eftir samhengi orðanna og efni texta.

2.3.6.1 SKJÁTEXTAÞÝÐINGAR

Smíðuð hafa verið kerfi til að flýta fyrir vinnu þýðenda sjónvarpsefnis og kvikmynda. SUMAT-verkefnið, var rannsóknarverkefni á vegum Evrópusambandsins á árunum 2011–2013. Markmið þess var að setja saman þýðingarkerfi fyrir skjátexta til að þýða á milli evrópskra tungumála. Kerfið var þróað fyrir ellefu tungumálapör og með notkun kerfisins jókst framleiðni hjá þýðendum sem þýddu á milli allra tungumálapara nema tveggja. Meðalafköst þýðenda sem notuðu kerfið jukust um 40%.

Ef til væri þýðingarkerfi fyrir skjátexta sem þýðir á milli ensku og íslensku sem yki framleiðni þýðenda gætu þýðingar bæði orðið ódýrari og betri. En til að slíkt kerfi verði öflugt þarf mikið af þýðingargögnum. Æskilegt væri að íslensk fyrirtæki, sem láta þýða fyrir sig skjáefni, tækju höndum saman um slíkt verkefni.

2.3.6.2 ÞÝÐINGAR Á OPINBERUM SKJÖLUM

Mikil vinna fer fram hjá Þýðingamiðstöð utanríkisráðuneytisins við að þýða opinber skjöl. Á vegum MT@EC hefur verið þróað þýðingarkerfi til að aðstoða við þýðingar á slíkum skjölum á milli flestra opinberra mála í Evrópusambandinu og ef sérfræðingar verkefnisins hefðu aðgang að íslenskum gögnum yrði einnig reynt að þróa vélþýðingar fyrir íslensku, Íslendingum að kostnaðarlausu. Þegar MT@EC-verkefninu lýkur væri hægt að nýta sér reynslu þaðan af vinnu með íslenskar vélþýðingar. Þá þyrfti að tryggja að slíkt kerfi yrði aðgengilegt fyrir íslensku og halda áfram þróun þess með það í huga að kerfið gagnist Þýðingamiðstöðinni og öðrum sem vinna að þýðingum fyrir hið opinbera.

2.4 MÁLRÝNI

Íslenskur málrýnir verður þróaður til þess að bera kennsl á og leiðrétta ritvillur. Hann mun greina málfræðilegt og merkingarlegt samhengi og þannig ráða við mun fleiri villur en nú er mögulegt. Opinn málrýni verður hægt að tengja við ritvinnslu, snjalltæki og annan máltæknihugbúnað.

Hugbúnaður til þess að leiðrétta ritvillur er mikilvægur til aðstoðar við skrif en einnig sem hluti annars máltæknihugbúnaðar. Fyrir íslensku eru til nokkur forrit sem leiðrétta villur í einstökum orðum ef villuorð finnst ekki í orðabók. Nauðsynlegt er að þróa opinn hugbúnað sem getur greint málfræðilegt og merkingarlegt samhengi til þess að finna samhengisháðar villur. Til þess að gera þróunina markvissa þarf að safna gögnum um algengar ritvillur og skilgreina markmið fyrir hvert skref í þróun leiðréttingarforritsins.

Máltæknihugbúnaður til þess að finna villur í texta og leiðrétta þær er orðinn staðalbúnaður fyrir mörg tungumál. Villur í texta geta verið af ýmsum toga: innsláttarvillur, stafsetningarvillur, málfræðivillur eða villur í orðanotkun. Hér verða slíkar villur nefndar ritvillur (hefðbundin merking þess orðs á þó helst við um stafsetningarvillur) og hugbúnaður til þess að þekkja og leiðrétta þær kallaður málrýnir. Hugbúnaður getur einnig aðstoðað að öðru leyti við skrif texta, til dæmis með því að fara yfir stíl og málsnið og með tilkomu íslensks málrýnis verður þróun slíkra tóla möguleg.

Við textaskrif verða fólki á mistök, t.d. sökum fljótfærni, ónógrar þjálfunar eða lesblindu. Það getur verið erfitt að lesa texta sem í eru margar ritvillur og merking hans getur jafnvel riðlast eða orðið óljós. Í stafrænu umhverfi reiðum við okkur einnig á að finna texta og upplýsingar með leitarvélum. Þær þurfa að geta fundið viðeigandi upplýsingar þrátt fyrir að fyrirspurn eða texti sem leitað er í innihaldi ritvillur. Annar máltæknihugbúnaður, t.d. talgervlar sem lesa texta, treystir sömuleiðis á að textarnir séu því sem næst villulausir. Málrýnir er því nauðsynlegur grunnhugbúnaður í hvers konar textavinnslu, sjálfvirkri og handvirkri.

Það getur verið erfitt að lesa texta sem í eru margar ritvillur og merking hans getur jafnvel riðlast eða orðið óljós.

Margir kannast við sjálfvirka málrýni fyrir ensku og önnur tungumál í Microsoft Word-ritvinnslukerfinu. Stuðningur við textaskrif í slíkum forritum, við tölvupóstskrif o.s.frv. er það notkunarsvið sem flestir þekkja í sambandi við leiðréttingarhugbúnað. Notkunargildi slíks stuðnings má skilgreina frá mismunandi sjónarhornum: út frá því hver skrifar og við hvaða aðstæður og út frá því hvaða gildi vandaðir textar hafa á mismunandi sviðum.

- **Notkunargildi út frá færni þess sem skrifar:** Fólk hefur misgott vald á rituðu máli. Hér skiptir ekki aðeins menntun og þjálfun máli heldur einnig hvort ritað er á fyrsta máli, aldur og atriði eins og t.d. lesblinda.

- **Notkunargildi út frá aðstæðum sem skrifað er við:** Ef skrifa þarf texta undir tímapressu vinnst oft lítill tími til þess að fara yfir textann áður en honum er skilað eða hann jafnvel birtur. Þannig geta innsláttarvillur og ýmsar fljótfærnisvillur slæðst með þó að sá sem skrifar hafi annars gott vald á málinu.

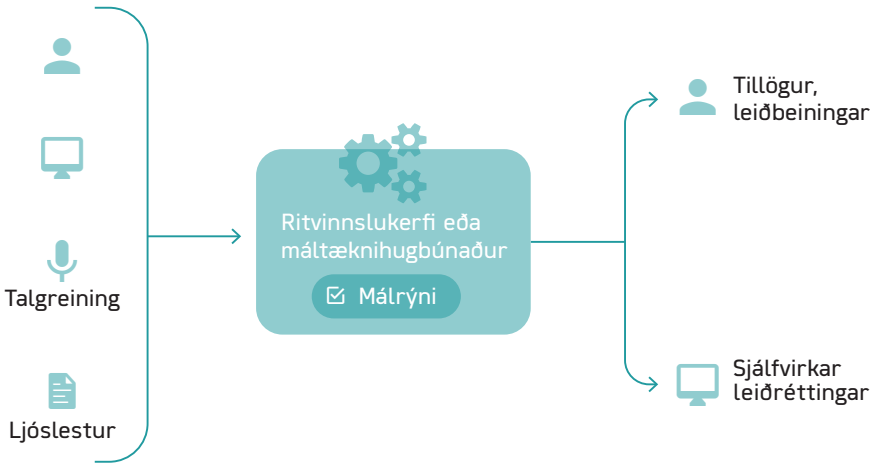
- **Notkunargildi út frá mikilvægi gæða:** Kröfur til gæða texta eru mismunandi. Útgefið efni, bækur, dagblöð o.fl. innihalda alla jafna texta sem gera ætti miklar gæðakröfur til, textinn er sjálf varan sem seld er. Víða er mikilvægt, þar sem mikið magn texta verður til, að þeir séu skrifaðir eftir ákveðnum stöðlum til þess að auðvelt sé að finna það sem leitað er að og til að ekki komi upp óþarfa vafamál vegna villna í texta. Þetta getur t.d. átt við hjá heilbrigðisstofnunum, í stjórnsýslunni og í dómskerfinu. Fyrir mörg fyrirtæki hafa textar einnig áhrif á ásýnd fyrirtækisins, t.d. í gegnum skrifleg samskipti við viðskiptavini, skilmála og samninga, sem og upplýsinga- og auglýsingatexta.

Mikilvægi sjálfvirkrar málrýni fer eftir því hvernig ofantaldir þættir fara saman: því minni sem færni og tími er en kröfur um gæði texta meiri því mikilvægari er slík aðstoð.

Eins og áður var nefnt gegnir málrýni einnig mikilvægu hlutverki í öðrum máltæknihugbúnaði. Það geta verið forrit þar sem málrýni er kjarninn, eins og ritstoð eða kennsluforrit, en einnig annar hugbúnaður eins og talgreining, talgerving, vélþýðingar eða leitarvélar. Nánast allur flóknari máltæknihugbúnaður þarf að treysta á sjálfvirka málrýni.

Þriðja notkunarsvið málrýni eru leiðréttingar á ljóslesnum textum (e. *Optical Character Recognition*, *OCR*). Mikil vinna hefur verið lögð í það hér á landi að ljóslesa efni útgefið á prenti frá seinni hluta nítjándu aldar til dagsins í dag, nægir þar að nefna vefsíðuna http://timarit.is/. Til þess að þessir textar verði að fullu nothæfir í stafrænu umhverfi, þarf að leiðrétta villur sem óhjákvæmilega verða til við yfirfærslu textanna á stafrænt form.

Málrýni getur þurft að aðlaga fyrir hvern notendahóp og notkunarsvið. Villur sem börn gera geta verið öðruvísi en villur fullorðinna og villur sem talgreinar og ljóslestrarkerfi gera öðruvísi en þær sem fólk gerir svo dæmi séu tekin.



Mynd: Málrýni í fjölbreyttu samhengi

2.4.1 HVAÐ ERU RITVILLUR?

Villur í rituðum texta eru orð eða setningar sem ekki samræmast stafsetningar- eða málfræðireglum/málvenjum viðkomandi tungumáls. Fyrir lifandi tungumál eru slíkar reglur ekki alltaf ótvíræðar og með tímanum ryðja sér jafnvel málvenjur til rúms sem brjóta í bága við gildandi reglur. Sem dæmi má nefna hina alræmdu þágufallssýki (t.d. *mér langar* í stað *mig langar*) sem telst ekki enn viðurkennt (rit)mál en sem breiðist æ meir út og nýtur meira samþykkis í málsamfélaginu. Við þróun málrýna þarf því að hafa í huga að ekki er alltaf auðvelt að meta „rétt“ og „rangt“ í máli.

Við gerð málrýnis er gagnlegt að skipta villum upp í flokka eftir því hvaða greiningu þarf til þess að finna þær. Fyrst er villunum skipt í villur sem varða óþekkt orð, svokallaðar non-word-villur, og villur sem varða þekktar orðmyndir, svokallaðar real-word-villur. Í fyrri flokkinn falla villur sem mynda form sem ekki eru til í íslensku (stakorðavillur og í mörgum tilfellum orðskiptingarvillur/orðabilavillur). Real-word-villur eru hins vegar villur þar sem rangskrifuðu orðin eru til sem orðmyndir í íslensku en eru röng í því samhengi sem þau standa. Eftirfarandi listi er ekki tæmandi en gefur hugmynd um þau mismunandi atriði sem huga þarf að við gerð málrýnis:

- **Stakorðavillur** Stafsetningarvillur sem hægt er að greina án þess að líta á önnur orð. Það er hægt ef orðmyndin er ógild í íslensku, t.d. *pisla* í stað *pistla*, en ekki ef rangskrifaða orðið er til, t.d. *neita* í stað *neyta*.

- **Málfræðivillur** Villur þar sem nauðsynlegt er að greina orð eða jafnvel setningar eða setningarhluta málfræðilega til þess að finna villuna. Algengar villur varða t.d. fallbeygingu og tíðir/hætti sagna: *Hann er langbestur leikmaður í heimi* í stað *Hann er langbesti leikmaður í heimi*; *Guðni* segir *í samtali við Vísi að undirbúningur málþingsins hófst í byrjun*

árs í stað *Guðni segir í samtali við Vísi að undirbúningur málþingsins hafi hafist í byrjun árs.*

- **Samhengisháðar villur** Til þess að finna slíkar villur þarf að greina merkingarlegt og/eða málfræðilegt samhengi: *sem var við líði á þeim tíma* í stað *sem var við lýði á þeim tíma*; *græða sem mest á sem skemmtum tíma* í stað *græða sem mest á sem skemmstum tíma.*

- **Orðskiptingarvillur/orðabilavillur** Villur þar sem orðabil vantar eða það er á röngum stað: *sá hvorugur þeirrahjólreiðamanninn*; *sækjast eftir háskóla námi*; *réttis em innihalda kjöt* í stað *rétti sem innihalda kjöt.*

- **Orði sleppt eða því ofaukið** Oft tengjast slíkar villur orðasamböndum: *bíða með fram vormánuði* í stað *bíða með fram á vormánuði*; *þar biðu þeir bara á þangað til* í stað *þar biðu þeir bara þangað til.*

- **Orðræðu-/samræðuvillur** Greining á slíkum villum krefst þess að innihald textans sé túlkað. Til þess að greina villuna í seinni málsgreininni þarf að túlka fyrri setninguna: *Katrín Tanja hafði á endanum betur eftir hörkukeppni en það munaði aðeins þremur sekúndum á stelpunum. Katrín kláraði á 6:53 mínútum en Sara á 6:53 mínútum.*

- **Villur í orðanotkun** Merking og merkingarblæbrigði, en einnig hefð, ráða því hvaða orð ganga merkingarlega með öðrum orðum og hvaða orð eru venjulega notuð í ákveðnu samhengi: *ýsan er útrunnin* í stað *ýsan er gömul/úldin/óæt*; Hér er rangt orð, sem ekki er endilega líkt orði sem mætti nota í staðinn, notað í ákveðnu samhengi. Möguleg leiðrétting er heldur ekki endilega eitt ákveðið orð.

- **Greinarmerkjavillur** Greinarmerki vantar eða er ofaukið/þau eru á röngum stað: *Síðan kom annar þáttur til greina en það var snákurinn sem sagði þeim að borða ávöxtinn tel ég að ráða megi bót á öllum þessum ...*

Villur verða ýmist til vegna fljótfærni, fyrir mistök eða af því að viðkomandi veit ekki betur eða er óöruggur. Hér er oft ekki hægt að greina á milli (veit sá sem skrifar *noskra* í stað *norskra* ekki betur eða er þetta fljótfærni?) og það er heldur ekki nauðsynlegt í hverju tilfelli fyrir sig. Þekking á því hvernig villur verða til gagnast þó við þróun málrýna.

2.4.2 STAÐA TÆKNINNAR OG HELSTU AÐFERÐIR

Sjálfvirk málrýni er framkvæmd í tveimur skrefum: Fyrst er ákveðnum aðferðum beitt til þess að finna villur og í næsta skrefi eru leiðréttingar eða tillögur að leiðréttingum settar fram.

2.4.2.1 RITVILLULEIT

Aðferðir við sjálfvirka málrýni hafa verið í þróun að minnsta kosti síðan um 1960. Lengi vel var höfuðáherslan lögð á að finna stakorðavillur. Það var gert með því að bera orð í texta saman við gildar n-stæður (röð bókstafa sem geta staðið saman í tungumálinu) eða að nota orðalista. Í dag eru orðalistar langmest notaðir við hefðbundna villuleit en n-stæður frekar við leiðréttingu ljóslesinna texta. Grunnaðferðin við stakorðavilluleit er því einföld: Ef orð finnst ekki í orðalista hugbúnaðarins þá er það merkt sem villa. Þetta getur þó einungis verið fyrsta skrefið í villuleitinni. Orðalisti leiðréttingarforrits getur aldrei innihaldið öll möguleg orð og orðmyndir tungumálsins og til þess að trufla ekki notandann með of mörgum röngum villuskilaboðum (e. *false positives*) þarf því að greina orð nánar sem finnast ekki. Í tungumálum eins og íslensku, þýsku og eistnesku, sem innihalda mikinn fjölda samsettra orða og ný samsett orð verða stöðugt til, eru óþekkt gild orð líklegust til þess að vera samsett orð eða sérnöfn. Því beita málrýnar fyrir slík tungumál einhvers konar greiningu á óþekkt orð til þess að meta líkindin á því hvort um sé að ræða gilt samsett orð eða eiginnafn. Einnig er hægt að beita orðhlutagreiningu til þess að meta hvort hér sé á ferðinni gild orðmynd þekkts orðs sem vantar í orðalistann.

Orðalisti leiðréttingarforrits getur aldrei innihaldið öll möguleg orð og orðmyndir tungumálsins.

Til þess að meta hvort um villu sé að ræða þrátt fyrir að orð finnist í orðalista þarf að greina textann frekar. Það fer eftir gerð villunnar hvernig og hve nákvæm greiningin þarf að vera. Ein tiltölulega einföld aðferð er að skilgreina ruglingsmengi (e. *confusion sets*). Það eru pör eða listar af orðum sem er gjarnan ruglað saman, t.d. af því að þau eru borin eins fram. Í íslensku eru þetta orð eins og *leiti – leyti*; *neita – neyta*; *sína – sýna*; *list - lyst* o.fl. Í hvert sinn sem forritið finnur eitthvert þessara orða er reglum viðkomandi ruglingsmengis flett upp og þær bornar saman við samhengið sem orðið stendur í. Dæmi: *Verkinu var að miklu leiti lokið.* Orðið *leiti* finnst í ruglingsmengi og dæmi um samhengi fylgja. Reglan er sú að í samhenginu að miklu … á að standa *leyti*: {leiti, leyti: að litlu/miklu/öllu leyti; á næsta leiti}. Reglurnar geta verið á ýmsu formi, t.d. nákvæmt samhengi tilgreint, nálæg merkingarbær orð, orðflokka- eða setningamynstur.

Villur sem tengjast orðum sem finnast í orðalista er annars erfitt að finna nema að greina textann málfræðilega.

Villur sem tengjast orðum sem finnast í orðalista er annars erfitt að finna nema að greina textann málfræðilega. Orð eru mörkuð með viðeigandi málfræðiupplýsingum (nafnorð í íslensku, t.d.: nafnorð-kyn-tala-fall-greinir-sérnafn) og setningar jafnvel greindar eftir ákveðnum málfræðilegum þáttum. Niðurstöður mörkunar eru svo bornar saman við reglusafn forritsins og villur merktar þar sem það á við. Sem dæmi skilgreinir sænska málfræðileiðréttingarforritið GRANSKA reglur eins og: Laus greinir verður að hafa sama kyn og tölu og nafnorðið sem hann stendur með. Ef forritið finnur ósamræmi í mörkun greinis og nafnorðs er villa merkt og notanda bent á þessa reglu.

2.4.2.2 RITVILLULEIÐRÉTTINGAR

Þegar villa er fundin þarf að finna líklega leiðréttingu. Í kerfi sem byggist á reglum geta leiðréttingar eða leiðbeiningar verið hluti af reglunum.

Algengasta leiðin til þess að leiðrétta stakorðavillur er þó byggð á villurannsóknum frá sjöunda áratugnum. Rannsókn Dameraus (1964) notaði eftirfarandi skilgreiningar á stafafjarlægð (e. *edit distance*): einum staf er ofaukið (innsetning), einn staf vantar (eyðing), einum staf er skipt út fyrir annan (skipting) eða röð á tveimur stöfum víxlað (stafavíxl). Þessar skilgreiningar hafa síðan verið kallaðar Damerau-Levenshtein-fjarlægð. Fyrir ensku hefur verið sýnt fram á að u.þ.b. 80% af öllum rangstöfuðum orðum innihalda einungis eina villu, annaðhvort innsetningu, eyðingu eða skiptingu. Þessar niðurstöður hafa verið grunnurinn fyrir leiðréttingar samkvæmt Damerau-Levenshtein-fjarlægðinni. Hvort sem kerfi byggist á reglum eða tölfræðilegum aðferðum þá nýta flest kerfi útreikninga á stafafjarlægð til þess að finna líklegar leiðréttingar.

Þess ber þó að geta að þó að algengustu villurnar séu yfirleitt mjög nálægt rétta orðinu, þá hafa rannsakendur rekið sig á að munað getur allt að sjö bókstöfum á réttri og rangri útgáfu af orði í þýsku til dæmis.

Algengasta tölfræðilega aðferðin við að leiðrétta stafsetningarvillur byggist á líkani um truflanir í samskiptarásum (e. *noisy channel model*). Hugmyndin að baki því er að stafsetning rétta orðsins hafi riðlast með því að hafa orðið fyrir truflunum í samskiptarásinni og komi því rangt skrifað út. Leiðréttingarforritið þarf að reikna líkurnar á því að ákveðið orð hafi riðlast á þann hátt sem lýsir sér í rangskrifaða orðinu. Fyrst er kenningin um litla stafafjarlægð (Damerau-Levenshtein) nýtt til þess að finna líklegustu réttu orðin og síðan er með bayeskri ályktunarfræði metið hvert þeirra er líklegasta leiðréttingin. Finnist til að mynda villan *firir* eru líkurnar fyrir

nálægustu orðin *fyrir* (skipting i-y) og *firðir* (innsetning ð) reiknaðar út. Sá útreikningur tekur tillit til þess hve líklegt er að villuorðið sé í raun misritun fyrir orðið sem um ræðir og til þess hve líklegt er að orðið eigi að vera á þessum stað í textanum, óháð því hvernig villan lítur út.

Þegar málrýnir er notaður til þess að aðstoða við skrif er nægilegt að birta notanda líklegar leiðréttingar raðaðar eftir því hversu líklegar þær eru. Ef hins vegar á að leiðrétta texta innan hugbúnaðarkerfis þá þarf leiðréttingarforritið að taka ákvörðun um líklegustu leiðréttinguna og leiðrétta.

Með tilkomu gríðarlegs textamagns á vefnum og nýrra aðferða og tölvubúnaðar til þess að vinna með slíkt magn af gögnum hafa verið gerðar tilraunir með að nýta texta af vefnum til þess að þjálfa leiðréttingaforrit sem þurfa hvorki handgerðar reglur né orðalista.

2.4.2.3 DUDEN KORREKTOR

Skýrsluhöfundar hittu höfunda tveggja leiðréttingarforrita í fremstu röð fyrir þýsku og eistnesku. Bæði þessi forrit vinna eingöngu með reglur. EPC (http://www.epc.de/) er núverandi rétthafi Duden Korrektor-málrýnisins fyrir þýsku sem hefur verið í þróun frá árinu 2001. Það nýtir gríðarlega umfangsmikla gagnagrunna Duden-forlagsins sem er leiðandi í útgáfu orðabóka og málfræðibóka í Þýskalandi. Reglur Duden Korrektor eru þróaðar ofan á öflugan setningagreini sem annað fyrirtæki þróaði. Duden Korrektor er hægt að nota með MS Office-forritunum en einnig hefur verið þróaður umfangsmeiri hugbúnaður, Duden proof factory. Hann býður upp á orðskiptingu, prófun á málsniði, að tekið sé tillit til mállýsku, misstranga leiðréttingu o.s.frv., auk stafsetningar- og málfræðileiðréttinga úr Duden Korrektor. Samkvæmt reynslu EPC er mikilvægt að fylgja stöðlum frá Microsoft þegar hugbúnaður á að tengjast MS Office, að öðrum kosti skapast vandræði við hverja uppfærslu frá Microsoft. Duden Korrektor er í stöðugri þróun og daglegar breytingar eru keyrðar á sérstakri málheild til þess að meta áhrif þeirra á heildargæði forritsins.

Reglur Duden Korrektor eru þróaðar ofan á öflugan setningagreini.

2.4.3 MÁLRÝNAR FYRIR ÍSLENSKU

Nokkrir málrýnar hafa verið þróaðir fyrir íslensku. Enginn þeirra er þó opinn fyrir utanaðkomandi til þess að þróa áfram og aðlaga að mismunandi þörfum. Þeir málrýnar sem ætla má að helst séu í notkun í dag eru íslensku forritin Púki ritvilluvörn og Skrambi, auk íslenskrar útgáfu af Hunspell-forritinu. Þau finna og leiðrétta fyrst og fremst stakorðavillur, en íslensku

forritin framkvæma að einhverju leyti greiningu á samsettum orðum og Skrambi er með nokkur innbyggð ruglingsmengi. Hér á eftir verður farið stuttlega yfir hvern þessara málrýna.

Púki ritvilluvörn hefur verið í þróun síðan árið 1984, en fyrsta útgáfa hans kom út árið 1987. Lengst af var forritið sniðið að notkun með Word og Windows og hefur verið uppfært og endurbætt samhliða nýjum útgáfum af MS Office og Windows-stýrikerfinu. Árið 2014 kom Púki út fyrir Mac tölvur. Þar virkar forritið með MS Office-pakkanum og með öllum þeim forritum sem styðja kerfislægan yfirlestur. Púki tengist forritum í MS Office-pakkanum eins og innbyggðu leiðréttingarforritin og nýtir þannig marga þeirra möguleika sem MS Office býður upp á. Hægt er að bæta við orðum í orðabókina sem Púki notar til þess að meta hvort orð er gilt eða ekki, finna samheiti, nýta sjálfvirka leiðréttingu, hunsa villuskilaboð og virkja orðskiptiforrit. Púka fylgir einnig forrit til þess að fletta upp beygingarmyndum orða.

Púki greinir einungis stakorðavillur og framkvæmir enga greiningu á samhengi. Óþekkt samsett orð eru ekki merkt sem villur svo fremi sem Púki meti að þau fylgi íslenskum orðmyndunarreglum. Púki er þróaður og seldur af einkafyrirtækinu, Friðrik Skúlason ehf., og aðferðirnar sem beitt er til leiðréttingar eru ekki opnar. Í lýsingu á tilurð forritsins er þó nefnt að það geti borið kennsl á beygingarmyndir orða og metið þannig hvort orð er rangt stafsett eða erlent. Púki vinnur því með orðalista, orðhlutagreiningu og aðferðir til þess að greina samsett orð og er að öllum líkindum alveg byggður á reglum.

Skrambi er málrýnir sem var þróaður sem hluti af og í framhaldi af meistaraverkefni Jóns Friðriks Daðasonar „Post-Correction of Icelandic OCR Text“ (Leiðréttingar á ljóslesnum texta, 2012). Aðferðum sem Skrambi notar er lýst í ritgerðinni og forritið er hægt að nota fyrir stutta texta í gegnum vefviðmót. Skrambi byggist á kenningunni um truflaðar samskiptarásir sem lýst var í kafla 2.4.2.2 og inniheldur einnig nokkur ruglingsmengi. Rétt orð úr slíku mengi er valið með tilliti til samhengis með hjálp flokkunarreglna. Orðalistinn sem Skrambi notar samanstendur af öllum orðmyndum í Beygingarlýsingu íslensks nútímamáls (BÍN) ásamt lista af óbeygjanlegum orðum. Að auki eru sérstakir orðalistar notaðir við greiningu samsettra orða, þ.e. Skrambi athugar hvort óþekkt orð geti verið gilt samsett orð áður en hann merkir það sem villu.

Hunspell er opinn leiðréttingahugbúnaður sem notaður er í forritum eins og LibreOffice, Photoshop, InDesign og ýmsum vöfrum svo eitthvað sé

nefnt. Hunspell er ekki sniðinn að neinu ákveðnu tungumáli, einungis þurfa að vera til orða- og orðhlutalistar þess tungumáls sem á að leiðrétta. Slíkir listar eru til fyrir íslensku og er hægt að sækja þá á netinu til þess að nota með þeim forritum sem nota Hunspell.

Við prófun á íslenskum leiðréttingaforritum verður fljótt ljóst að hinn gríðarlegi fjöldi orðmynda í íslensku gerir stakorðavillugreiningu án samhengis mjög erfiða. Ekkert forritanna gefur sig út fyrir að finna villur út frá samhengi, að nokkrum sérvöldum orðum frádregnum í Skramba. Innsláttarvillur og aðrar stakorðavillur er oft ómögulegt að finna án samhengis, þar sem rangskrifaða orðið verður að beygingarmynd annars orðs. Dæmi: *Þeir vilja bara græða á sem skemmtum tíma*. Ekkert forritanna greinir villuna í skemmtum þar sem það er fullgild íslensk orðmynd. Forritin þrjú voru prófuð á nokkrum setningum af vefnum og textabrotum úr villumerktu safni setninga þar sem setningarnar innihalda a.m.k. eina villu hver. Safnið var sett saman í norsku verkefni: „Feilkorpus for å testa stavekontrollar for grønlandsk, islandsk, lulesamisk og nordsamisk“ árið 2013. Niðurstöður má sjá í töflunni hér fyrir neðan. Alls var 151 villa merkt handvirkt, bæði stakorðavillur og samhengisháðar villur. Farið var yfir hve margar af þessum villum forritin fundu en einnig hvað þau greindu mörg rétt orð sem villur (textabrotin innihéldu erlend nöfn sem yfirleitt voru merkt sem villur, þ.e. ekki öll rangmerkt orð voru fullgild íslensk orð).

Hinn gríðarlegi fjöldi orðmynda í íslensku gerir stakorðavillugreiningu án samhengis mjög erfiða.

	Fundnar villur	Rétt orð merkt sem villa	Nákvæmni	Heimt	F-gildi
Púki	96	11	90%	63,6%	0,745
Skrambi	71	16	81,60%	47%	0,597
Hunspell	88	41	68,20%	58,3%	0,629

Tafla: Niðurstöður prófana á málrýnum fyrir íslensku

Þó að leiðréttingaforritin, sem hér eru nefnd, komi að sjálfsögðu að miklu gagni veita þau einnig að einhverju leyti falskt öryggi. Notendur búast ekki við að málfræðivillur eða aðrar samhengisháðar villur séu greindar en forritunum sást einnig yfir fjölda einfaldra innsláttarvillna.

Í kafla 2.4.1 voru níu flokkar af ritvillum nefndar. Það er ekki tæmandi listi en gefur hugmynd um stærð verkefnisins. Forritin sem til eru fást

nær eingöngu við einn villuflokk, stakorðavillur óháð samhengi. Skrambi athugar að auki samhengi örfárra fyrirfram skilgreindra orða.

Það er því brýnt að þróa grunn að málrýni sem getur greint texta málfræðilega og merkingarlega og þannig fundið fleiri villutegundir.

Það er því brýnt að þróa grunn að málrýni sem getur greint texta málfræðilega og merkingarlega og þannig fundið fleiri villutegundir. Verði til opinn hugbúnaður af þessu tagi má þróa hann áfram og nýta í almennum notendahugbúnaði, kennsluforritum og öðrum máltæknihugbúnaði.

2.4.4 GÆÐAMAT

Gæðamat á málrýni getur verið með mismunandi hætti. Grunnmælikvarðinn ætti að vera nákvæmni, heimt og F-gildi miðað við staðlaða prófunarmálheild. Nákvæmni þarf að mæla fyrir villuleit annars vegar og villuleiðréttingar hins vegar. Nákvæmni villuleiðréttinga segir til um hve mikið af þeim orðum/orðasamböndum sem forritið merkir sem villur eru í raun villur og að auki hvort forritið sýni rétta leiðréttingu. Við prófun skal mæla gildin miðað við allar villur í prófunarmálheild en einnig einungis miðað við þær villur sem forritið samkvæmt markmiðum á að geta fundið og leiðrétt. Til samanburðar má fá óháðan sérfræðing, t.d. prófarkalesara til þess að leiðrétta prófunarmálheildina og reikna einnig út nákvæmni, heimt og F-gildi miðað við áður merktar villur í málheildinni.

Annað gæðamat felst í notendaprófunum, þar sem notendur meta hversu hjálplegt forritið er eða hvort það er jafnvel til trafala að einhverju leyti, til að mynda af því það merkir of mörg rétt orð sem villu. Að síðustu þarf að meta hvaða áhrif leiðréttingarforrit sem byggt er inn í annan máltæknihugbúnað hefur á heildargæði þess hugbúnaðar.

2.4.5 ÞRÓUN INNVIÐA FYRIR ÍSLENSKA MÁLRÝNI

Þróun innviða fyrir málrýni er tvíþætt:

- Söfnun og greining á gögnum til þess að kortleggja algengar ritvillur og til þess að þjálfa og prófa málrýnihugbúnað.
- Þróun aðferða við málrýni sem ráða við að leiðrétta fyrirfram skilgreindar villutegundir upp að settu marki.

2.4.5.1 VILLUMÁLHEILDIR

Eins og lýst var í kafla 2.4.1 eru ritvillur af ýmsum toga. Mismunandi tegundir af villum krefjast oft mismunandi aðferða við villuleit og leiðréttingu. Það er því mikilvægt að kortleggja villur sem gerðar eru til þess að geta geta skilgreint markmið málrýnis. Tvennt þarf að hafa í huga þegar markmiðin eru skilgreind: hvað ákveðin villutegund er algeng og hversu auðvelt er að finna hana og leiðrétta með sjálfvirkum hætti. Þannig ætti að leggja töluverða vinnu í að finna og leiðrétta algengustu villurnar en sjaldgæfar villur, sem erfitt er að eiga við, ættu að bíða.

Annar tilgangur gagnasöfnunar og villugreiningar er að útbúa vönduð prófunargögn sem mæla gæði málrýni.

Ein villumálheild hefur verið búin til fyrir íslensku sem hluti af öðru verkefni, eins og áður var sagt. Fjöldi orða í henni er um 167 þúsund. Textarnir eru úr stúdentaritgerðum framhaldsskólanema og frétta- og bloggtextum. Ekki er ljóst hvernig leyfismálum fyrir málheildina er háttað eða aðgengi að henni. Nauðsynlegt er að búa til stærri og ítarlegri villumálheild, sniðna að þörfum þróunar og prófana á íslenskum málrýni. Gera má ráð fyrir að setja þurfi saman fjölbreytt textasafn sem er allt að 500 þúsund orð að stærð. Þessir textar myndu einnig verða felldir inn í Risamálheildina (sjá 2.5.1.3) en þyrftu að vera sérmerktir þar sem líklegt er að þeir innihaldi töluvert af ritvillum. Mesta vinnan við gerð villumálheildar felst í því að merkja ritvillur inn í málheildina. Þetta verður gert með tóli til handvirkrar mörkunar, eins og lýst er í kafla 2.5.3.1, en í upphafi er nauðsynlegt að skilgreina tegundir af villum og hvernig á að merkja þær. Mjög reyndur málfræðingur þarf að hafa yfirumsjón með þeirri vinnu og æskilegt er að þessi skilgreiningarvinna verði unnin af fleiri en einum þar sem ýmis álitamál geta komið upp. Mörkunina sjálfa geta MA-nemendur í íslensku tekið að sér en mikilvægt er að samræma og yfirfara þá vinnu með reglulegu millibili. Samkvæmt reynslu við gerð villumálheildarinnar sem nefnd var hér að ofan má gera ráð fyrir um 1000 klukkustunda vinnu við að yfirfara 500 þúsund orð.

Til þjálfunar á leiðréttingaforriti með tauganetsaðferðum þurfa að verða til annars konar málheildir: annaðhvort mjög stór málheild af vönduðum textum sem innihalda ekki mikið af ritvillum eða gríðarstór almenn málheild sem er það stór að hægt er að læra réttar og rangar útgáfur af ákveðnum orðum og frösum. Með „gríðarstór“ er átt við málheild sem inniheldur marga milljarða orða. Hér verður einungis gerð grein fyrir þróun markaðra villumálheilda.

2. KJARNAVERKEFNI

M.1: Almenn villumálheild

Stór málheild þar sem ritvillur eru merktar inn eftir ákveðnu kerfi.

Verkþættir:

- Söfnun texta og frágangur fyrir villumálheild
- Skilgreining á villutegundum og -merkingum
- Uppsetning mörkunarsetta í mörkunartóli
- Mörkun

Mannauður:

- Gagnasérfræðingur: 1,5 mánuður
- Reyndur málfræðingur: 2 mánuðir
- Málfræðingur eða MA-nemandi í málfræði: 6 mánuðir

Alls: 9,5 mannmánuðir

Til þess að hægt verði að sníða málrýninn að mismunandi þörfum þarf að safna textum frá mismunandi hópum: frá börnum og unglingum, frá lesblindum og þeim sem ekki hafa íslensku að móðurmáli. Afla þarf tilskilinna leyfa fyrir meðferð gagna ólögráða barna og unglinga. Búa þarf til málheildir úr öllum gögnum sem safnast og merkja og greina villur samkvæmt ferli sem skilgreint verður fyrir almennu ritvillumálheildina. Í hvaða röð sérhæfðum málheildum er safnað skiptir ekki höfuðmáli. Að lokinni forvinnu ætti að meta hvaða gögn er auðveldast að útvega og haga forgangsröðun eftir því.

M.2: Sérhæfðar villumálheildir

Málheildir með textum frá ákveðnum hópi fólks. Ritvillur merktar inn samkvæmt M.1

Verkþættir:

- Villumálheild með textum barna og unglinga
- Villumálheild með textum frá lesblindum
- Villumálheild með textum frá fólki sem ekki hefur íslensku að móðurmáli

Mannauður:

- Gagnasérfræðingur: 3 mánuðir
- Málfræðingur eða MA-nemandi í málfræði: 6 mánuðir
- Leyfismál: 1 mánuður

Alls: 10 mannmánuðir

Þegar vinna við villumálheildir er komin vel af stað þarf að vinna tölfræði úr mörkuninni til þess að kortleggja villutegundir. Hægt er að setja upp ferli til þess áður en mörkun er lokið þar sem einfalt er að endurtaka útreikninga eftir því sem vinnu við mörkunina vindur fram. Einnig þarf að setja saman vandaðar prófunarmálheildir úr hverri villumálheild sem munu verða nýttar fyrir öll stig í þróun málrýnisins. Ákveðnir hlutar prófunarmálheilda ættu að vera lokaðir, þ.e. að prófun og þróun séu aðskilin þegar ný útgáfa málrýnis er prófuð.

M.3: Tölfræðileg úrvinnsla og prófunarmálheildir

Verkþættir:

- Villutölfræði
- Prófunarmálheildir

Mannauður:

- Sérfræðingur í máltækni eða tölfræði: 2 mánuðir

2.4.5.2 ORÐALISTAR OG MÁLLÍKAN

Málrýnir þarf að búa yfir orðalista með gildum orðmyndum í íslensku. Hann ætti að setja saman með hliðsjón af Risamálheild og orðfræðigögnum, fyrst og fremst BÍN (sjá 2.5.1.8). Einnig getur málrýnir þurft á öðrum orðalistum að halda, til dæmis til þess að greina samsett orð. Þegar almenn villumálheild verður tilbúin er hægt að safna algengum misritunum í eins konar villuorðabók (t.d. eitthver er alltaf misritun fyrir einhver). Til þess að geta beitt tölfræðiaðferðum við leiðréttingar þarf að búa til mállíkön með aðstoð Risamálheildarinnar.

M.4: Orðalistar og mállíkön

Verkþættir:

- Setja saman orðalista með aðstoð orðfræðigagna (BÍN), villumálheildar og Risamálheildar.
- Þjálfa mállíkön á Risamálheild.

Mannauður:

- Sérfræðingur í máltækni: 3 mánuðir

2.4.5.3 AÐFERÐIR OG HUGBÚNAÐARHÖGUN

Ólíkt öðrum kjarnaverkefnum er ekki til opinn hugbúnaður til þess að þróa málrýni og þarf því að leggja töluverða vinnu í þróun grunnhugbúnaðarins.

Velja þarf grunnaðferðir sem á að beita við gerð málrýnisins, en líklegast þykir að sambland reglna og tölfræðiaðferða muni gefa besta raun. Þá þarf að huga að vali umhverfis sem hentar (forritunarmála o.þ.h.), hönnun, tengingu við stoðtól og hvaða atriði eru mikilvæg fyrir tækniyfirfærslu. Mikilvægt er að reyndur hugbúnaðarsérfræðingur komi að þessum verkþætti.

M.5: Aðferðir og hugbúnaðarhögun

Val á aðferðum og val og uppsetning á þróunarumhverfi

Verkþættir:

- Skilgreiningar á meginaðferðum sem beita á við málrýni.
- Val á þróunarumhverfi, hönnun hugbúnaðar, tenging við stoðtól sem til eru, skilgreining á stoðtólum sem vantar. Huga þarf að atriðum tengdum tækniyfirfærslu eftir því sem mögulegt er á þessu stigi.

Mannauður:

- Hugbúnaðarsérfræðingur og sérfræðingur í máltækni: 2 mánuðir

2.4.5.4 ÞRÓUN MÁLRÝNIS FYRIR ÍSLENSKU

Kerfisbundin þróun sem miðar að því að leiðrétta algengustu villur getur hafist þegar greiningu villumálheildar er lokið. Strax og uppsetningu þróunarumhverfis er lokið (M.5) er þó hægt að hefjast handa við þróun stakorðavillugreinis. Fyrsta útgáfa málrýnis þarf að finna og leiðrétta villur sem lýsa sér sem óþekkt orð (e. *non-word errors*). Óþekkt orð þarf að greina áður en þau eru merkt sem villur, t.d. hvort um sé að ræða sérnafn, samsett orð eða jafnvel erlent orð. Ef villa er greind þarf að finna líklegustu leiðréttingarmöguleika. Að þessari vinnu lokinni ætti að gefa út fyrstu útgáfu málrýnis.

M.6: Málrýnir fyrir stakorðavillur

Verkþættir:

- Einfaldur stakorðavillugreinir sem merkir við óþekkt orð og gefur líklega leiðréttingarmöguleika.
- Nánari greining óþekktra orða: samsett orð, sérnöfn o.s.frv.
- Uppsetning vefviðmóts og opins aðgangs.

Mannauður:

- Sérfræðingur í máltækni, sérfræðingur í máltækni með viðeigandi forritunarkunnáttu eða sérhæfður forritari: 12 mánuðir

2. KJARNAVERKEFNI

Þegar niðurstöður villugreiningar liggja fyrir (M.2) þarf að skilgreina næstu markmið málrýnisins. Velja þarf ákveðnar tegundir af villum sem næsta útgáfa forritsins á að geta leiðrétt. Þegar gerð fyrsta vinnupakkans er lokið skal meta gæðin með aðstoð villuprófunarmálheildarinnar og halda þróun áfram á þennan hátt – skilgreina sértæk og mælanleg markmið, þróa viðbætur við hugbúnað og loks meta gæði. Gæta þarf þess við mat á hverri ítrun að viðbætur séu aðeins til bóta en geri ekki niðurstöður fyrri ítrana verri.

Hér á eftir fara dæmi um markmið sem hægt væri að skilgreina fyrir hverja ítrun. Þessi markmið ætti þó að skilgreina og forgangsraða endanlega eftir að greining villumálheildar liggur fyrir:

- Forritið þekkir algengar villur sem auðvelt er að leiðrétta með aðstoð ruglingsmengja (*leiti – leyti*; *list – lyst* o.s.frv.), einföld samhengisháð leiðréttingu.
- Forritið þekkir ákveðnar samsetningar með forsetningum og villur sem gerðar eru í tengslum við þær, t.d. *leita að/*af*, **víst/fyrst að.*
- Forritið greinir valdar málfræðireglur/greinir málfræðilegt ósamræmi milli orða sem standa nálægt hvort öðru: *telur að mergæxli séu *algengir*; **mörgum vantar skriffæri.*

Á seinni stigum þróunar þyrfti forritið að finna villur sem þarf meira samhengi til þess að greina: *Guðni segir í samtali við Vísi að undirbúningur málþingsins *hófst í byrjun árs.*

Mikilvægt er að gera sér grein fyrir að leiðréttingarforrit er þróað í mörgum skrefum yfir langan tíma, en getur þó strax frá fyrstu útgáfu komið að gagni. Slíkt forrit nær þó aldrei að leiðrétta allar villur, illa uppbyggðar eða órökréttar setningar, eins og t.d. *Eigum við í forræðishyggjunni sé að gefa okkur að þær sem mæta ekki séu að mæta út af trassaskap?*; *Ég gleymdi næstum að segja að atvinulausir og lágmarkslauninn eru svo há*. Gott greiningarforrit þyrfti í þessum tilvikum að geta greint villu þótt það geti ekki gert tillögu um leiðréttingu. Markmiðið ætti að vera að auka heimt í villuleit verulega miðað við það sem leiðréttingaforrit ná í dag án þess að það gerist um of á kostnað nákvæmni. Við lok hverrar ítrunar skal uppfæra opinn aðgang ásamt skjölun og leiðbeiningum.

M.7: Kerfisbundin þróun málrýnis

Verkþættir:

- Þróun almenns málrýnis fyrir íslensku eftir ítrunarferli. Hver ítrun tekur 6 mánuði og lögð er áhersla á að bæta málrýninn samkvæmt forgangsröðun sem gerð er með hliðsjón af villugreiningu (M.2).

Mannauður:

- Sérfræðingur í máltækni, sérfræðingur í máltækni með viðeigandi forritunarkunnáttu eða sérhæfður forritari: 84 mánuðir

2.4.5.5 AÐGENGI OG AÐLÖGUN

M.8 Málrýni í snjalltækjum

Lagt er til að samvinna verði höfð við talgreinisverkefni um þróun lyklaborða fyrir snjalltæki. Þau muni þá innihalda málrýni ásamt talgreiningarhnappi, og einnig munu lyklaborðin geta sagt fyrir um með ritspá (e. *autocompletion*) hvaða orð notandi er líklega að skrifa eða ætlar að skrifa næst.

Verkþættir:

- Íslenskt lyklaborð með málrýni og ritspá fyrir Android-stýrikerfið
- Íslenskt lyklaborð með málrýni og ritspá fyrir iOS-stýrikerfið
- Íslenskt lyklaborð með málrýni og ritspá fyrir Windows Phone-stýrikerfið

Mannauður

- Forritari: 12 mánuðir
- Sérfræðingur í máltækni: 4 mánuðir
- Alls: 16 mánuðir

2. KJARNAVERKEFNI

M.9 Málrýni í ritvinnslukerfum

Málrýni þarf að tengja við algengan ritvinnsluhugbúnað og stýrikerfi. Flest búa þau yfir stöðluðum aðferðum við að tengjast málrýni og er mikilvægt að fylgja þeim til þess að allar uppfærslur gangi snurðulaust fyrir sig.

Verkþættir:

- Tenging við MS Office
- Tenging við Mac OS
- Tenging við hugbúnað eins og InDesign, vafra, og annað skv. óskum notenda

Mannauður:

- Forritari: 12 mánuðir

Til þess að nýta málrýni sem hluta af máltæknihugbúnaði þarf t.d. að aðlaga forritunarskil og úttak. Málrýnihlutinn þarf að geta tekið við upplýsingum frá öðrum hugbúnaði og unnið með þær og skilað annaðhvort ótvíræðum niðurstöðum, þ.e. framkvæma leiðréttingu sjálfvirkt eða veita upplýsingum áfram til næsta hugbúnaðarhluta. Hver hugbúnaðarpakki hefur eigin þarfir hvað þetta varðar og því er aðalhlutverk málrýnisteymis að undirbúa málrýnina þannig að tengingar og breytingar verði sem auðveldastar. Hér geta einnig orðið til sérverkefni sem nauðsynlegt er að málrýniteymið taki þátt í en slík verkefni tilheyra tækniyfirfærslu.

M.10 Aðlögun að máltæknihugbúnaði

Verkþættir:

- Aðlögun og uppsetning útgáfu málrýnis sem nýtist í öðrum máltæknihugbúnaði

Mannauður:

- Sérfræðingur í máltækni, forritari: 6 mánuðir

M.11 Villumódel fyrir ljóslestur

Villur sem verða til við ljóslestur texta eru af öðrum toga en við innslátt. Að sjálfsögðu eru innsláttarvillur og aðrar ritvillur hluti af textum sem eru ljóslesnir en engu að síður þarf annað villumódel fyrir ljóslestur en fyrir innsleginn texta.

Verkþættir:

- Hönnun villumódels fyrir ljóslestur

Mannauður:

- Sérfræðingur í máltækni: 6 mánuðir

2.4.5.6 STOÐTÓL FYRIR MÁLRÝNI

Þróun málrýni þarf að vera í nánu samstarfi við þróun stoðtóla. Velja þarf þá aðferðafræði sem á að fara eftir og markara og þáttara í samræmi við það. Þáttari er grundvallartól við þróun málrýni. Annaðhvort mun þáttari eins og IceParser verða þróaður áfram eða nýr þáttari byggður á venslamálfræði (e. *dependency grammar*) þróaður. Náin samvinna við teymi sem vinna að stoðtólum eins og tilreiðara, markara, orðskipti og nafnaþekkjara er nauðsynleg.

M.12 Þáttari og önnur stoðtól fyrir málrýni

Verkþættir:

- Áframhaldandi þróun og aðlögun þáttara eða þróun nýs þáttara í samræmi við valda aðferðafræði málrýnis
- Samvinna við önnur stoðtólateymi um þarfir málrýnis

Mannauður:

- Sérfræðingur í máltækni: 12 mánuðir

Í fyrsta verkþætti þróunar á samhengisháðri ritvilluleit verða væntanlega notuð ruglingsmengi. Setja þarf saman lista af orðum sem gjarnan er ruglað saman (*leiti - leyti*) og þjálfa t.d. Winnow-flokkara til þess að greina það samhengi sem á við hvora ritmynd.* Málrýnir og ritstoðartól eru þróuð í ítrunarferli. Í hverri ítrun myndast þörf á flóknari merkingargreiningu. Greining á orðasamböndum, samheitagreining, innihaldsgreining, greining

* Tvö verkefni sem fengist hafa við slíka greiningu fyrir íslensku eru: Anton Karl Ingason o fl.., 2009 og Jón Friðrik Daðason, 2012. Sjá nánar í heimildaskrá.

á rökréttu samhengi og greining á viðeigandi orðanotkun eru dæmi um verkefni í merkingargreiningu sem nauðsynleg eru fyrir málrýni og ritstoð.

M.13 Merkingargreining fyrir málrýni

Verkþættir:

- Útbúa tól og gögn fyrir leiðréttingu með hjálp ruglingsmengja, t.d. með Winnow-flokkara
- Þróun og aðlögun annarra nauðsynlegra merkingargreiningartóla fyrir málrýni og ritstoð

Mannauður:

- Sérfræðingur í máltækni: 24 mánuðir

2.4.5.7 MÁLRÝNI MEÐ TAUGANETUM

M.14 Málrýni með djúpum tauganetum

Samhliða þróun hefðbundins málrýnis ætti þróun málrýni með tauganetsaðferðum að eiga sér stað a.m.k. fyrstu tvö árin. Að því tímabili loknu þarf að meta hvort og hvernig þeirri þróun skuli haldið áfram

Verkþættir:

- Þróun málrýni með tauganetsaðferðum

Mannauður:

- Sérfræðingar í djúpum tauganetum: 24 mánuðir

2.4.6 TÆKNIYFIRFÆRSLA

Frá fyrstu útgáfu verður málrýnirinn opinn og aðgengilegur til sérhæfðrar þróunar.

Frá fyrstu útgáfu verður málrýnirinn opinn og aðgengilegur til sérhæfðrar þróunar. Með þeirri þekkingu sem til verður, til að mynda á mismunandi þörfum hópa, verður hægt að útbúa sérhæfða málrýni og annan tengdan hugbúnað. Meta verður í hverju tilfelli fyrir sig hvort sérhæfð þróun á sviði málrýni á sér viðskiptalegan grundvöll (sérhæfð ritstoð fyrir einstök fyrirtæki og fagsvið) eða eigi að vera opin og öllum aðgengileg (sérhæfð ritstoð fyrir lesblinda). Gert er ráð fyrir að frá og með ári tvö í máltækniáætluninni verði málrýni komin á það stig að tækniyfirfærsla geti átt sér stað.

Tækniyfirfærsla á sér stað á tvennan hátt: annars vegar sérhæfð málrýni fyrir ákveðna hópa notenda og hins vegar tenging við máltæknihugbúnað.

2.4.6.1 SÉRHÆFÐ MÁLRÝNI

Þróun almenns málrýnis miðar að því að aðstoða meðalnotandann, fullorðinn einstakling sem hefur nokkuð gott vald á rituðu íslensku máli. Innan verkefnisins verður þó einnig til þekking á sérstökum vandamálum í ritun sem einstakir hópar glíma við: grunnskólabörn og unglingar, lesblindir og einstaklingar sem ekki hafa íslensku að móðurmáli. Grunnhugbúnað almenna málrýnisins verður hægt að aðlaga að þessum verkefnum, bæði með því að leggja áherslu á annars konar villur og með því að sýna reglur og leiðbeiningar.

2.4.6.2 FORRIT FYRIR ÍSLENSKUNÁM

Tækniyfirfærslan sem á sér stað innan kjarnaverkefnisins einskorðast við að tengja málrýninn við ritvinnslu. Góður málrýnir opnar einnig möguleika á þróun hugbúnaðar sem treystir á slíka greiningu. Kennsluforrit í stafsetningu og málfræði verða að geta greint villur og leiðbeint nemendum. Einnig geta forrit sem aðstoða fólk við að læra íslensku orðið mun betri og sveigjanlegri ef málrýnikjarni getur greint það sem nemandinn skrifar og gefið vísbendingar og ráð.

2.4.6.3 RITSTOÐ

Í hugbúnaði sem inniheldur ritstoð er einnig þörf á málrýnikjarna. Ritstoð getur greint fleiri atriði en þau sem almennt flokkast sem ritvillur. Hún fer yfir texta og athugar til að mynda hvort mikið er um endurtekningar og sýnir orð og orðasambönd sem hægt væri að nota í staðinn, hvort stíll er viðeigandi, nægilega formlegur til dæmis, hvort orðanotkun er viðeigandi og fylgir ákveðnum stöðlum ef þörf er á, o.s.frv. Það getur verið mikilvægt innan stórra stofnana og fyrirtækja að sömu orð séu notuð yfir sömu hluti og að ákveðnum ritstíl sé fylgt. Sérhæfð ritstoð getur í þessum tilfellum sparað vinnu við leit og tryggt það að mikilvægar upplýsingar týnist ekki. Þeir sem vinna við texta til birtingar, til dæmis hjá útgáfufyrirtækjum, auglýsingastofum og fréttaveitum eru annar stór markhópur fyrir ritstoðarhugbúnað sem getur aðstoðað við að tryggja gæði texta.

Það getur verið mikilvægt innan stórra stofnana og fyrirtækja að sömu orð séu notuð yfir sömu hluti og að ákveðnum ritstíl sé fylgt.

2.4.6.4 MÁLRÝNI Í MÁLTÆKNIHUGBÚNAÐI

Innan kjarnaverkefnis verður búin til málrýnieining sem hægt verður að aðlaga og tengja inn í flóknari máltæknihugbúnað. Málrýni er þá ekki endilega höfuðmarkmið hugbúnaðarins, en hún getur engu að síður bætt virkni hans til muna. Má þar nefna leitarvélar sem geta greint villur í fyrirspurnum og leit í ljóslesnum textum sem hægt verður að koma á betra form. Einnig mun málrýni geta greint úttak talgreina og leiðrétt, leiðrétt texta fyrir inntak talgervla og þýðingarvéla svo eitthvað sé nefnt.

2.5 MÁLFÖNG

Málföng eru það sem á ensku kallast *language resources*. Innan hugtaksins málföng geta rúmast innviðir eins og þeir sem fjallað hefur verið um í undanförnum köflum en hér notum við það aðeins yfir gögn og stoðtól fyrir máltækni.

Gögn fyrir máltækni skiptast í textasöfn og málheildir, orðfræðigögn og hljóðgögn. Þau eru nauðsynleg til þjálfunar og prófunar á máltæknihugbúnaði og oft getur verið nauðsynlegt að tengja orðfræðigögn beint við hugbúnað.

2.5.1 TEXTAGÖGN

Textagögn sem nýtast í máltækni geta verið málheildir, samhliða textar, mállýsingar og orðasöfn. Mállýsingar og orðasöfn eru gögn á borð við orðabækur og merkingargögn, íðorðasöfn, beygingarlýsingar, framburðarlýsingar og önnur gögn sem gera grein fyrir merkingu orða eða notkun. Málheildir (e. *corpora*) eru söfn fjölbreyttra texta sem eru geymdir á stöðluðu sniði á rafrænu formi. Til þess að textarnir nýtist sem best við málrannsóknir eða smíði mállíkana eru þeir greindir á margvíslegan hátt.

Samhliða textar eru yfirleitt á tveimur tungumálum, þar sem annar textinn er þýðing á hinum. Setningum eða málsgreinum sem hafa sömu merkingu er þá raðað saman hlið við hlið. Safn samhliða texta er forsenda þess að hægt sé að þróa þýðingarvélar.

Það er talað um að málheild sé mörkuð þegar hver orðmynd hefur verið greind og við hana hengdur greiningarstrengur eða mark (e. *tag*) sem sýnir orðflokk og málfræðileg atriði eins og fall, tölu og kyn fallorða og persónu, tölu og tíð sagna. Auk þess fylgir nefnimynd (e. *lemma*) með hverri orðmynd, t.d. nefnifall í eintölu fyrir fallorð og nafnháttur sagna.

Málheildir í trjábanka hafa verið greindar setningafræðilega og þá fylgja mörk setningarhlutum sem gera grein fyrir setningafræðilegri stöðu orðs eða orðasambands.

Hverjum texta í málheildinni fylgja jafnframt lýsigögn (e. *metadata*) sem gera grein fyrir textanum, hvaðan hann kemur, hvers eðlis hann er, höfundi textans og fleiru sem komið gæti að gagni.

2. KJARNAVERKEFNI

Málheildir eru mikilvægar öllum máltækniverkefnum sem byggjast á tölfræðilegum aðferðum. Upp úr þeim er hægt að vinna mállíkön sem nauðsynleg eru til útreikninga á líkindum, en slíkir útreikningar eru kjarninn í tölfræðilegum aðferðum í máltækni. Almennt gildir það að því stærri sem málheildirnar eru því betri mállíkön er hægt að smíða upp úr þeim. Þó þarf að hafa aðra þætti í huga eins og gagnagæði, hvers konar textar eru í málheildinni og hvernig unnið hefur verið úr þeim.

2.5.1.1 TILTÆKAR MÁLHEILDIR

Þrjár tilbúnar íslenskar málheildir eru aðgengilegar með vel skilgreindum leyfum: Íslensk orðtíðnibók, Fornritamálheild og Mörkuð íslensk málheild (MÍM).

Íslensk orðtíðnibók

Íslensk orðtíðnibók var gefin út 1991. Þar eru birtar niðurstöður viðamikilla rannsókna á íslensku nútímamáli sem beindust að tíðni orða og málfræðiatriða í textum af ýmsu tagi.

Búið var til sérstakt textasafn fyrir gerð bókarinnar. Í textasafninu eru brot úr 100 textum sem voru gefnir út á tímabilinu 1980–1989, hvert með um 5.000 lesmálsorðum. Samtals er málheildin því um 500 þúsund orð. Textarnir voru markaðir og lemmaðir og hafa verið gerðir leitarbærir þannig á netinu. Unnt er að sækja flesta textana og nota þá við málrannsóknir og í máltækniverkefnum. Það á þó ekki við um alla textana í Orðtíðnibókinni því að þýddir textar eru ekki í þeim flokki. Veittur er aðgangur að þessum textum með sérsmíðuðu leyfi.

Fornritin

Í þessari málheild eru rafrænir textar Íslendingasagna, Sturlungu, Heimskringlu og Landnámabókar. Stafsetning allra textanna hefur verið umrituð til nútímastafsetningar. Einnig hefur nokkrum beygingarendingum verið breytt til nútímaíslensku. Leita má í textunum og einnig er unnt að sækja textana og nota þá við málrannsóknir og í máltækniverkefnum. Textarnir hafa verið markaðir og lemmaðir. Fornritamálheildin er gefin út með CC BY 3.0-leyfi.

MÍM

Vinnu við Markaða íslenska málheild (MÍM) lauk árið 2012 en hún hefur að geyma texta frá tímabilinu 2000–2010. Stefnt var að því að í málheildinni yrðu um 25 milljónir orða úr textum af ýmsu tagi sem gæfu sem raunsannasta mynd af ritaðri íslensku á tímabilinu. Í MÍM eru um 25 milljónir orða úr fjölbreyttum textum sem eru geymdir í stöðluðu sniði á rafrænu formi. Orð í textunum eru greind málfræðilega og hverjum texta fylgja lýsigögn, bókfræðilegar upplýsingar um verkið sem textinn er úr. Málheildin er ætluð fyrir málrannsóknir og til notkunar í máltækniverkefnum. Leyfilegt er að nota MÍM í hvers kyns máltækniverkefni og rannsóknir. Málheildin er aðgengileg með sérsmíðuðu leyfi, MÍM-leyfinu. Rétthafar efnis í málheildinni hafa allir samþykkt notkun með þeim notkunarskilmálum.

2.5.1.2 GULLSTAÐALLINN

Gullstaðallinn er málheild með um einni milljón orða. Orðin voru mörkuð með sjálfvirkum aðferðum og síðan leiðrétt handvirkt. Textar í málheildinni voru valdir úr textum Markaðrar íslenskrar málheildar (MÍM). Fyrir notkun Gullstaðalsins gildir því sama leyfi og fyrir málheildina. Gert er ráð fyrir að málheildin verði notuð sem gullstaðall fyrir þjálfun tölfræðilegra markara. Lokaútgáfa Gullstaðalsins á að verða tilbúin vorið 2017.

2.5.1.3 RISAMÁLHEILD

Hafist var handa við smíði nýrrar risastórrar málheildar árið 2015 með styrk úr Innviðasjóði Rannís. Markmiðið er að í málheildinni verði textar úr ýmsum áttum, samtals með a.m.k. einum milljarði orða. Gerðir hafa verið samningar við stóra rétthafa, svo sem útgefendur stærstu dagblaða og tímarita, og verður þorri málheildarinnar efni úr fjölmiðlum. Þar að auki verða í málheildinni opinberir textar, þingræður og fleira.

Textarnir í málheildinni verða markaðir og lemmaðir. Þeir verða leitarbærir á netinu í leitarvélum sem sniðnar eru að málvísindarannsóknum og hægt verður að sækja þá til nota í máltækni. Ekki voru allir rétthafar tilbúnir til þess að sett yrði alveg opið leyfi á gögnin þeirra. Því verða tvenns konar leyfi á gögnunum, annars vegar sérstakt leyfi, sambærilegt MÍM-leyfinu sem fyrr er getið, hins vegar opið CC-BY 4.0-leyfi.

Þessari málheild er ætlað að vera stöðugt uppfærð með nýjum gögnum. Eftir að fyrsta útgáfa kemur út þarf því að halda áfram að safna gögnum frá þeim rétthöfum sem veitt hafa leyfi til þess. Það hefur ýmsa kosti. Það

stækkar málheildina, sem er gagnlegt í öllum máltækniverkefnum sem nýta hana, það auðveldar rannsóknir á samtímamáli og ekki síst yrði sístækkandi málheild afar mikilvæg fyrir orðtöku, t.d. í BÍN-verkefninu (sjá 2.5.1.8).

Með upphaflegri fjármögnun verkefnisins úr Innviðasjóði hefur verið mögulegt að koma verkefninu á laggirnar og ljúka um 75% af þeirri vinnu sem þarf til að gefa út fyrstu útgáfu málheildarinnar eins og henni var lýst í upphafi. Áætlað er að til að ljúka fyrstu útgáfu málheildarinnar þurfi um átta mannmánuði til viðbótar en til að setja upp og aðlaga hugbúnað til að nýta málheildina fjóra mánuði að auki. Til að halda Risamálheildinni við eftir útgáfu þarf u.þ.b. hálft stöðugildi, en sá sem sér um málheildina gæti einnig tekið þátt í öðrum máltækniverkefnum, sérstaklega gagnaverkefnum.

G.1 Risamálheild

Verkþættir:

- Ljúka við að safna og vinna texta sem leyfi hafa fengist fyrir.
- Setja upp rannsóknarhugbúnað fyrir málheildir (Korp).
- Setja upp n-stæðuskoðara fyrir öll tímasett gögn.
- Varpa gögnum á staðlað snið til dreifingar.

Mannauður:

- Gagnaforritari: 6 mánuðir
- Sérfræðingur í mállegum gagnasöfnum: 6 mánuðir til að ljúka fyrstu útgáfu og svo 6 mánuðir á hverju ári áætlunar til að halda Risamálheildinni við.

Alls: 36 mannmánuðir

2.5.1.4 STÓR SÖGULEG MÁLHEILD

Skoða þarf hvort hægt sé að koma upp risastórri sögulegri málheild sem vinnur úr öllu efni sem aðgengilegt er á bækur.is, timarit.is og því sem hefur verið skrifað upp eftir eldri handritum. Með þessu móti væri hægt að búa til mállíkan úr eins miklu af textagögnum og mögulega eru til fyrir íslensku.

2.5.1.5 MÁLHEILD FYRIR NAFNAÞEKKJARA

Ekki er til nein málheild fyrir íslensku sem sniðin hefur verið að þjálfun nafnaþekkjara (sjá kafla 2.5.3.7) og rannsóknum á nákvæmni. Komið verður upp málheild fyrir nafnaþekkjara með því að fara yfir Gullstaðalinn með tilliti til slíkra nota og bæta við tilheyrandi mörkun.

G.2 Málheild fyrir nafnaþekkjara

Verkþættir:

- Forrit smíðað sem merkir við allar einingar í Gullmálheildinni sem mögulega geta verið nafnaeiningar.
- Farið handvirkt yfir merktar einingar og þær einingar markaðar sem á að marka.

Mannauður:

- Forritari: Hálfur mánuður
- Málfræðingur: 4 mánuðir
- Málfræðingur/nemi (mörkun): 3 mánuðir

Alls: 7,5 mannmánuðir

2.5.1.6 ÍSLENSKUR ORÐASJÓÐUR

Ómörkuð málheild með textum fengnum úr vefsöfnun Landsbókasafns Íslands – Háskólabókasafns. Orðasjóðurinn er settur saman og hýstur hjá Háskólanum í Leipzig.

2.5.1.7 SÖGULEGUR ÍSLENSKUR TRJÁBANKI

Sögulegur íslenskur trjábanki er safn setningafræðilega greindra texta frá 12. til 21. aldar. Í trjábankanum er alls ein milljón orða úr rúmlega 60 textum sem spanna um 800 ára tímabil. Frumgreining textanna var vélræn en farið var yfir greininguna handvirkt. Í trjábankanum eru setningafræðileg mörk, lemmur og föll fallorða.

Trjábankinn var annars vegar gerður til nota í máltækni en nákvæmar upplýsingar um setningagerð eru mikilvæg forsenda fyrir gerð ýmiss konar máltæknibúnaðar, svo sem leiðréttingarforrita, vélrænna þýðinga o.fl. Hins

vegar er bankinn ætlaður til málrannsókna, einkum á setningagerð og setningafræðilegum breytingum.

Trjábankinn er algerlega opinn og ókeypis, notkun hans er án allra takmarkana og ekki háður neinum leyfum.

Algengasta snið á trjábankagögnum síðustu ár er það snið sem notað er fyrir Universal Dependencies-trjábanka. Universal Dependencies trjábankar eru með samhæfðu sniði sem á að uppfylla þarfir allra tungumála. Með þeim er auðvelt að samnýta verkfæri á milli tungumála, sem annars getur verið erfitt eða ómögulegt. 1. mars 2017 voru 70 UD-trjábankar á 50 tungumálum aðgengilegir í gegnum CLARIN-málfanganetið. Með því að varpa íslenska trjábankanum á þetta snið myndu notkunarmöguleikar hans aukast til muna og þeir sem vinna með málleg gögn væru líklegri til að gera rannsóknir á íslensku sem væri styrkur fyrir þróun íslenskrar máltækni.

G.3 Sögulegur íslenskur trjábanki

Verkþáttur:

- Varpa íslenska trjábankanum í UD.

Mannauður: Sérfræðingur í trjábönkum: 8 mánuðir

2.5.1.8 BEYGINGARLÝSING ÍSLENSKS NÚTÍMAMÁLS

Beygingarkerfi íslenskunnar er flókið. Sértilvik og undantekningar eru margar. Algengt er að sama ending sé notuð fyrir margar beygingarformdeildir sama orðs og í sumum tilvikum geta sömu orðin haft fleiri en eina mögulega beygingarmynd innan sömu beygingarformdeildar. Vegna þess er ekki hægt að smíða almennar reglur sem hægt er að nota til að vinna með beygingarmyndir í máltæknihugbúnaði. Beygingarlýsing íslensks nútímamáls (BÍN) er því afar mikilvægt gagnasafn og nauðsynlegt að búa þannig um það að það nýtist sem flestum sem best. Þá þarf að hafa í huga að þannig sé búið um gögnin að þau nýtist rétt.

Mikilvægt er að leyfi til að nota beygingarmyndir til greiningar í máltæknitólum sé þannig úr garði gert að þegar gögnin eru notuð við smíði eða þjálfun hugbúnaðar sé leyfilegt að dreifa hugbúnaðinum með þeim hætti sem best hentar hverju sinni. Um leið þarf að tryggja að gögnin í beygingarlýsingunni verði ekki birt með þeim hætti að það sýni misvísandi upplýsingar, en það væri í andstöðu við lögbundnar skyldur Stofnunar Árna Magnússonar í íslenskum fræðum, þar sem BÍN hefur verið þróuð og unnin.

Því er nauðsynlegt að fara í endurbætur á gagnagrunni og umsjónarkerfi BÍN sem miðar að því að hægt sé að veita aðgang að birtingargögnum í gegnum aðgangsstýrð forritaskil og því að hægt sé að dreifa beygingarmyndum í sérstökum útgáfustýrðum pakka. Þeim pakka verður dreift á XML-sniði og hann fær annað heiti en BÍN, t.d. Íslenskur beygingargrunnur fyrir máltækni (ÍBM), þar sem hann er eingöngu ætlaður til máltæknivinnu. Aðgangur að BÍN-forritaskilum verður veittur til skilgreindra nota en ÍBM verður aðgengilegur til niðurhals með opnu leyfi.

Það er afar mikilvægt að gefa gögnin í þessum tveimur gagnasöfnum út með opnum leyfum. Til að það sé hægt þarf að búa um gögnin með réttum hætti. Það eru mismunandi vandamál í hvoru tilvikinu fyrir sig sem þarf að leysa.

Talsverða vinnu þarf að inna af hendi til að hægt sé að búa þannig um gögnin að þau séu nýtileg með þessum hætti og til að auðvelda vinnu í BÍN-gögnunum:

- BÍN verður tengd við textasafn, t.d. Risamálheildina (sjá kafla 2.5.1.5).
- Gagnalíkan BÍN verður endurskoðað með tilliti til þessara þarfa. Skilgreina þarf mögulegar birtingar og skilgreina möguleg notkunarsvið beygingarmynda (t.d. stílbundið, bundið merkingarflokki, minna viðurkenndar orðmyndir eða ritmyndir, t.d. úr talmáli o.s.frv.). Það er grundvöllur fyrir hvers kyns úrtaki, sem BÍN-forritaskil þurfa að geta boðið upp á.
- Farið verður yfir þau orð sem hafa fleiri en eina mögulega beygingarmynd í einni eða fleiri beygingarformdeildum og skilgreina notkunarsvið hverrar beygingarmyndar.
- BÍN hefur ekki haft neinn ritstjórnarham. Smíðað verður öflugt ritstjórnarviðmót til að auðvelda allt viðhald og viðbætur við gagnasafnið.
- Sett verður upp öflugt orðtökutól sem auðveldar ákvarðanir um viðbætur og stækkun gagnasafnsins.

G.4 Beygingarlýsing íslensks nútímamáls fyrir máltækni

Verkþættir:

- Aðlögun orðtökutóls að BÍN.
- Endurskoðun gagnalíkans.
- Farið yfir margræðar beygingarmyndir.
- Smíði ritstjórnarviðmóts.
- Smíði API.
- Skilgreiningar á XML-sniði fyrir máltæknigögn og vörpun gagna á sniðið.

Mannauður:

- Forritari: Eftir verkþáttum: 3+1+1+4+2+1 mánuðir
- Málfræðingur: Eftir verkþáttum 2+3+12+1+1+1 mánuðir

Alls: 32 mannmánuðir

2.5.1.9 ORÐSKIPTINGAR

Orðalisti með 203.964 orðum sem byggist á uppflettiorðum úr Íslensk-danskri orðabók Sigfúsar Blöndals er tiltækur með orðskiptingum. Orðin voru tölvuskráð á Íslenskri málstöð á miðjum 9. áratug síðustu aldar. Orðskiptingar voru búnar til vélrænt en svo leiðréttar handvirkt.

Þennan orðalista þarf að stækka með gögnum úr BÍN. Listi með orðskiptingum allra orða í BÍN væri afar gagnlegur í stafsetningar- og málfarsleiðréttingartólum auk þess sem hann gagnast auðvitað í orðskiptingartólum í umbrotshugbúnaði. Leiðin til þess að stækka listann er að þjálfa orðskiptingarforrit með gögnunum sem til eru og keyra það á orðalista úr BÍN. Málfræðingur færi svo handvirkt yfir tillögur tölvuforritsins.

G.5 Orðskiptingar

Verkþættir:

- Þjálfa forrit sem giskar á orðskiptingar.
- Stækka lista með orðum úr BÍN.
- Fara yfir nýjar orðskiptingar og leiðrétta.

Mannauður:

- Forritari: 1 mánuður
- Málfræðingur: 5 mánuðir

Alls: 6 mannmánuðir

2.5.1.10 FRAMBURÐARORÐABÓKIN

Framburðarorðabókin varð til í Hjal-talgreiniverkefninu árið 2003. Hún inniheldur milli 50 og 60 þúsund hljóðritaðar orðmyndir og er aðgengileg með CC-BY 3.0-leyfi. Orðin eru rituð í SAMPA-hljóðritunarstaðlinum og hljóðrituninni hefur einnig verið varpað yfir í IPA, alþjóðlega hljóðritunarstafrófið (e. *International Phonetic Alphabet*). Orðin í Framburðarorðabókinni eru oft hljóðrituð á fleiri en einni framburðarmállýsku. Upplýsingar um hvaða framburðarmállýsku er að ræða eru þó ekki fyrir hendi. Til að gagnasafnið geti nýst við talgervingu þarf að skrá þær. Til að hægt sé að búa til talgervla með mismunandi framburðarmállýskum þarf að ganga úr skugga um að dæmi um mismunandi framburð dekki hverja framburðarmállýsku fyrir sig, dæmi séu um allar gerðir tilbrigða sem greina hverja framburðarmállýsku frá öðrum. Þá þarf einnig að merkja inn upplýsingar um orðflokka, sérstaklega í þeim tilvikum þegar framburður getur verið tvenns konar. Dæmi um slíkt tengjast langoftast sérnöfnum og -ll- inni í orðum: gellur (so.), gellur (no.); Valla (sérnafn), valla (no.) o.s.frv. Það er því þörf á endurbótum og stækkun á Framburðarorðabókinni. Markmiðið ætti að vera að ná utan um undantekningar og nógu mörg dæmi um framburð til að hægt sé að þjálfa hljóðritunartól sem hljóðritar því sem næst skammlaust allar helstu framburðarmállýskur í íslensku.

G.6 Framburðarorðabók

Verkþættir:

- Smíða tól til að halda utan um endurbætur á Framburðarorðabókinni. Tólið passar upp á að aðeins rétt hljóðritunartákn séu notuð og býður upp á að hljóðritaður strengur sé spilaður með eldri talgervli.
- Framburðarmyndir merktar eftir framburðarmállýskum og orðflokkum.
- Dekkun könnuð og bætt inn í gagnasafnið eftir þörfum til að bæta dekkun.

Mannauður:

- Málfræðingur: Eftir verkþáttum 0+4+8 mánuðir
- Forritari: Eftir verkþáttum 3+1+2 mánuðir

Alls: 18 mannmánuðir

2.5.1.11 ÍSLENSKT ORÐANET

Íslenskt orðanet Jóns Hilmars Jónssonar lýsir merkingar- og setningarlegum venslum íslenskra orða og orðasambanda. Til grundvallar liggur safn orðasambanda og samsetninga sem hefur að geyma rösklega 200 þúsund orðasambönd af ýmsu tagi og um 100 þúsund samsetningar. Öll gögnin eru unnin úr samfelldum textum og eru því textalegur vitnisburður um merkingar- og setningarleg vensl.

Þessi gögn nýtast á ýmsan hátt. T.d. er hægt að fletta upp í þeim og fá sambærilegar upplýsingar við það sem er að finna í samheitaorðabókum og hugtakaorðabókum. Möguleikar til nýtingar í máltækni eru hins vegar vannýttir. Merkingarlegar upplýsingar á borð við þær sem orðanetið býr yfir geta nýst til að bæta árangur í leitarvélum, við upplýsingaheimt og efnisgreiningu og hafa verið nýttar til að bæta árangur þýðingarvéla. Þá nýtast mörkuð orðasambönd við samhengisháða leiðréttingu í ritstoðarkerfum. Í tilviki íslenska orðanetsins er vandinn hins vegar sá að gögnin eru ekki aðgengileg með öðrum hætti en til leitar. Til að gera megi þau aðgengileg þarf að skilgreina gagnasnið sem fellur að gögnunum og varpa þeim yfir á það snið. Það krefst nokkurrar rannsóknarvinnu ef ganga á úr skugga um að valin leið nýtist rétt. Þegar snið hefur verið valið fyrir gögnin þarf að varpa þeim yfir á það snið. Þá væri afar gagnlegt að yfirfara hvort ákveðnir flokkar orða séu gisnir og gera grein fyrir því, helst að leggja einhverja vinnu í að

fylla upp í slíkar gloppur. Í þá vinnu myndu stórar málheildir nýtast vel og orðtökutól sem einnig er fjallað um í kaflanum um BÍN. Að lokum þarf að búa til lýsigögn og setja opið leyfi á allan gagnapakkann, t.d. CC BY-SA 4.0. Gæta þarf þess að grunninum sé stöðugt haldið við, eðlis gagnanna vegna. Merking flyst til og ný orð og orðasambönd líta dagsins ljós. Gera má ráð fyrir að með 25% vinnuhlutfalli, eftir að gengið hefur verið frá fyrstu útgáfu gagnapakkans, væri hægt að halda orðanetinu við með góðu móti.

G.7 Íslenskt orðanet

Verkþættir:

- Gagnasnið valið eða skilgreint.
- Gögn yfirfarin og þeim varpað á valið snið.
- Farið yfir gloppur.
- Orðtökutól aðlagað að orðaneti.
- Fyllt upp í gloppur.
- Lýsigögn búin til og gengið frá gögnum til útgáfu.

Mannauður:

- Forritari: Eftir verkþáttum 2+1+1+3+1+1 mánuðir
- Málfræðingur: Eftir verkþáttum 2+0+2+1+6+1 mánuðir

Alls: 21 mannmánuður

Athugasemd: Vinnu við endurbætur og frágang þarf að vinna í samráði við eiganda gagnanna. Hann hefur lýst áhuga á því að búið sé um gögnin með þeim hætti sem hér er lýst og þau opnuð.

2.5.1.12 MERKOR

MerkOr er svokallaður merkingarbrunnur fyrir íslensku, til aðgreiningar frá hefðbundnari orðanetum. MerkOr var unninn á árunum 2010–2012 og eru niðurstöður verkefnisins tvíþættar: annars vegar algrím og aðferðir til þess að vinna merkingarupplýsingar sjálfvirkt úr textum og hins vegar merkingarbrunnurinn sjálfur sem inniheldur um 110 þúsund orð. Hann er hannaður til nota í máltækni og er gagnagrunnurinn ásamt forritunarviðmóti (API) aðgengilegur á https://github.com/bnika/MerkOrCore. Merkingarbrunnurinn inniheldur vensl byggð á setningamynstrum og skyldleikavensl

og orðaþyrpingar byggðar á tölfræðiaðferðum. Með því að skoða vensl og orðaþyrpingar má oft fá gleggri heildarmynd af merkingarumhverfi orða en mögulegt er úr hefðbundnum orðabókum og orðanetum. Eins má finna þau orð sem eru dæmigerð fyrir ákveðið efni (íþróttir, stjórnmál o.s.frv.). Slíkar upplýsingar geta nýst m.a. við að ákvarða umfjöllunarefni texta eða merkingarlegar tengingar milli texta.

MerkOr-merkingarbrunnurinn er tilbúinn til notkunar en til þess að hann komi að raunverulegu gagni þyrfti að uppfæra hann með því að vinna merkingarvensl úr mun meira textamagni en mögulegt var á sínum tíma. Um leið þyrfti að fara yfir greiningarhugbúnaðinn sjálfan og uppfæra hann. Einnig myndi tenging við Íslenskt orðanet skapa möguleika á enn öflugri merkingargagnagrunni.

G.8 Uppfærsla á MerkOr

Verkþættir:

- Uppfæra hugbúnað.
- Keyra hugbúnað á stórt textasafn, t.d. Risamálheild.
- Útgáfa hugbúnaðar og merkingarbrunns.

Mannauður:

- Sérfræðingur í máltækni: 6 mánuðir.

2.5.1.13 ICEWORDNET

IceWordNet er íslensk útgáfa af Princeton Core WordNet, aðgengileg með CC-BY 3.0-leyfi. Það samanstendur af tæplega fimm þúsund íslenskum þýðingum á orðum úr kjarnalista Princeton ásamt íslenskum samheitum orðanna. Princeton WordNet hefur verið notað í leitarvélum og hugbúnaði fyrir upplýsingaheimt. Stór íslensk útgáfa af WordNet gæti gagnast vel í slíkum tólum. Líklega gæti þó Íslenskt orðanet, sem nú þegar er stórt og viðamikið, gagnast eins vel í því skyni.

2.5.1.14 ÍSLENSK NÚTÍMAMÁLSORÐABÓK

Íslensk nútímamálsorðabók er unnin á Stofnun Árna Magnússonar í íslenskum fræðum og nýtir sama grunnorðalista og ISLEX. Gert er ráð fyrir að um 50 þúsund flettur verði tilbúnar árið 2018. Leyfismál hafa ekki verið ákveðin fyrir orðabókargögnin.

2.5.1.15 RITMÁLSSAFN ORÐABÓKAR HÁSKÓLANS

Í Ritmálssafni Orðabókar Háskólans eru 2,6 milljón dæmi um orðnotkun, allt frá 1540 til loka 20. aldar. Ritmyndirnar eru stafréttar, þ.e. óbreyttar frá því að þær voru skrifaðar en eru tengdar við staðlaðar myndir orðanna með nútímastafsetningu. Þau gögn má nota við þróun á hugbúnaði til leitar og upplýsingaheimtar úr eldri textum.

2.5.1.16 ÍÐORÐABANKINN

Íðorðabankinn er safn íðorðasafna sem unnin hafa verið af orðanefndum á ákveðnum fræða- eða sérsviðum. Íðorðasöfnin eru eign orðanefndanna en ríflega helmingur þeirra hefur gefið leyfi fyrir því að sínum íðorðasöfnum sé dreift með CC-BY-SA 3.0-leyfi. Orðasöfnin geta nýst við vélþýðingar og við efnisflokkun texta.

2.5.1.17 ISLEX

ISLEX eru tvímálaorðabækur fyrir íslensku annars vegar og Norðurlandamálin hins vegar (íslensk-sænsk, íslensk-dönsk, íslensk-norsk (bókmál og nýnorska), íslensk-færeysk og íslensk-finnsk). Í orðabókinni eru um 50 þúsund flettur fyrir hvert tungumálapar. Orðabókin getur nýst við vélþýðingar. Leyfismál eru þó ekki einföld þar sem stofnanir í hverju landi fyrir sig fara með leyfismál fyrir sitt tungumál. Gögnunum hefur verið dreift með CC-BY-NC-ND 3.0-leyfi.

2.5.1.18 HUGTAKASAFN ÞÝÐINGAMIÐSTÖÐVAR UTANRÍKISRÁÐUNEYTISINS

Þýðingamiðstöð utanríkisráðuneytisins hefur verið rekin frá árinu 1990. Hlutverk hennar er að þýða lagatexta og reglugerðir sem falla undir samninginn um Evrópska efnahagssvæðið. Í þeirri vinnu fer fram mikið íðorðastarf og hjá Þýðingamiðstöðinni hefur orðið til hugtakasafn sem telur yfir 80 þúsund flettur. Orðasafnið getur gagnast við efnisgreiningu og vélþýðingar en gögnin eru ekki aðgengileg og hafa ekki verið gefin út, hvorki með opnu leyfi né takmarkandi leyfi.

2.5.1.19 ÞÝÐINGARMINNI ÞÝÐINGAMIÐSTÖÐVAR

Í þýðingarvinnu Þýðingamiðstöðvar utanríkisráðuneytisins hefur orðið til töluvert stórt þýðingarminni sem þýðendurnir nota til að flýta fyrir vinnu sinni og til að samræmis sé gætt í þýðingum. Þýðingarminnið telur nú um

1,2 milljónir setningapara. Þýðingarminnið er háð notkunartakmörkunum þar sem það er ekki gefið út með opnu leyfi. Háskólarnir og Stofnun Árna Magnússonar í íslenskum fræðum hafa þó fengið vilyrði fyrir því að nota megi gögnin til rannsókna í vélþýðingum.

2.5.1.20 OPEN SUBTITLES

Á vefnum Open subtitles er safn skjátexta sem eru öllum aðgengilegir. Þar er að finna 1,4 milljónir setningapara á milli ensku og íslensku. Setningapörin má finna á vef OPUS-verkefnisins sem hefur að geyma opin samhliða textasöfn á milli Evrópumála. Þessa texta má nota í tilraunum á þjálfun þýðingarvéla.

2.5.2 HLJÓÐGÖGN

2.5.2.1 MÁLRÓMUR

Málrómur er opið upptökusafn. Gögnunum var safnað í samstarfi við Google á árunum 2011–2012. 563 þátttakendur tóku þátt. Tekin voru upp 127 þúsund raddsýni, alls 152 klukkustundir. Farið var yfir upptökurnar og merkt við þær skrár þar sem textinn sem átti að lesa er ekki í samræmi við það sem er í upptökunni. Yfir 108 þúsund skrár voru réttar, samtals um 135 klukkustundir af upptökum. Málrómur er gefin út með CC BY 4.0-leyfi.

2.5.2.2 HLJÓÐUPPTÖKUR GOOGLE VEGNA TALGERVILS

Háskólinn í Reykjavík safnaði röddum tuttugu þátttakenda fyrir talgervilsverkefni í samstarfi við Google-verkefnið Unison. Stuðst var við forritið ChitChat til þess að safna gögnunum en um 250 setningar voru teknar upp fyrir hvern þátttakanda (45–60 mínútur af uppteknu efni fyrir hvern). Markmiðið með gagnasöfnuninni var að búa til stikað talgervilskerfi þar sem myndaða röddin líkist ekki endilega neinum þeirra sem lesa upp. Stefnt er að því að gefa gagnasafnið út með opnu CC BY 4.0-leyfi.

2.5.2.3 ISLEX-UPPTÖKURNAR

Við hvert uppflettiorð í ISLEX-orðabókinni er gefinn framburður í formi hljóðskrár. Um 49 þúsund stök orð voru tekin upp og auk þess rúmlega 700 orðasambönd. Upptökurnar voru gefnar út í fullum upptökugæðum með CC-BY-NC-ND 3.0-leyfi.

2.5.2.4 ALÞINGISUMRÆÐUR

Í þessu upptökusafni eru upptökur frá umræðum á Alþingi veturinn 2004–2005. Alls eru upptökurnar tæplega 21 klukkustund. Þær hafa verið umritaðar nákvæmlega, með tímastimplum. Textaskrár fylgja ásamt upplýsingum um þá sem taka til máls, s.s. aldur og kyn. Gögnin eru aðgengileg með CC BY 3.0-leyfi.

2.5.2.5 HJAL

Hjal-upptökurnar voru gerðar við þróun á stakorðagreini árið 2003. Upptökurnar fóru fram í gegnum síma, málhafar voru samtals 883. Frá hverjum málhafa eru um 47 upptökuskrár sem hver um sig er allt frá einu orði og upp í heila setningu. Samtals eru hljóðskrárnar því ríflega 40 þúsund. Skrárnar eru teknar upp á 8 kHz og heildarlengd þeirra er um 52 klst. Ekki er óalgengt að skrár innihaldi þagnir svo gera má ráð fyrir að nokkuð minna sé af tali, líklega u.þ.b. 40 klst. Hjal-upptökurnar eru aðgengilegar með CC-BY 3.0-leyfi.

2.5.2.6 ÍSTAL

ÍsTal er talmálsbanki sem hefur að geyma um 20 klst. af upptökum á sjálfsprottnum samtölum. Gögnin voru greind og merkt. Samtölin voru skrifuð upp með mörkum sem segja til um hver talar hverju sinni, innskot, framígrip, hik o.s.frv. ÍsTal-verkefnið fór fram á árunum 1999–2002. Gögnin hafa ekki verið gerð aðgengileg og fyrir því liggja ýmsar ástæður. Til að mynda var ekki fengið leyfi til þess í upphafi hjá þátttakendum og þar að auki er viðkvæmt efni í upptökunum sem þyrfti að eyða úr þeim. Ef gera ætti ÍsTal aðgengilegt þyrfti því að fara fram talsverð vinna við að afla leyfa hjá þátttakendum og yfirfara efnið.

G.9 ÍsTal

Verkþættir:

- Hafa samband við þátttakendur og afla skriflegra leyfa.
- Yfirfara gögnin og gera þau tilbúin til útgáfu.

Mannauður:

- Málfræðingur: Eftir verkþáttum 2+4 mánuðir

2.5.2.7 ÖNNUR MINNI SÖFN

Nokkur önnur minni hljóðupptökusöfn eru til, t.a.m. Jensson-málheildin, Þór-málheildin og Rúv-málheildin á Málföng.is. Þessir minni pakkar eiga það sameiginlegt að innihalda lítið af gögnum, örfáar klukkustundir og ekki hefur verið gengið frá leyfismálum. Því þykir okkur ekki ástæða til að fjalla ítarlega um þá hér.

2.5.3 STOÐTÓL

Stoðtól eru tól sem aðstoða við að vinna skipulögð gagnasöfn úr hráum gögnum eða framkvæma grunnmálgreiningu. Þau fást yfirleitt við eitt einangrað verkefni, t.d. að greina orðflokka og beygingar orða og mynda oft greiningarkeðju sem skilar niðurstöðum fyrir flóknari hugbúnað. Hágæða-stoðtól skipta sköpum fyrir máltækni. Allar villur, sem slíkur hugbúnaður gerir, smitast áfram til seinni stiga hugbúnaðarþróunar og stoðtólin hafa þannig mikil áhrif á gæði tilbúinna máltæknilausna.

Til eru tveir hugbúnaðarpakkar sem innihalda stoðtól fyrir íslensku: IceNLP og Greynir. IceNLP er máltæknitól sem inniheldur einingar til þess að greina íslenskan texta: tilreiðara (e. *tokenizer*), markara (e. *part-of-speech tagger*), lemmald (e. *lemmatizer*), hlutaþáttara (e. *shallow parser*) og nafnaþekkjara (e. *named entity recognizer*). Einingar úr IceNLP hafa sennilega verið notaðar í flestum þeim máltækniverkefnum sem unnin hafa verið á Íslandi síðan hann kom út. IceNLP er skrifaður í Java og JFlex og er kóðinn aðgengilegur á github undir GNU LGPL v3-leyfinu (sjá kafla 3.1.1): https://github.com/hrafnl/icenlp. Á Sourceforge er hægt að sækja pakkann á binary formi undir nafninu IceNLPCore (https://sourceforge.net/projects/icenlp/).

Greynir, málgreinir fyrir íslensku, er nýlegt tól sem greinir íslenskar setningar og býður upp á ýmsa möguleika í greiningu texta (https://greynir.is/about). Greynir safnar textum af fréttamiðlum, setningagreinir þá, finnur upplýsingar um fólk, skylda texta og fleira. Í undirköflum um tilreiðara og þáttara verður fjallað um innviði Greynis.

Greyni er hægt að nálgast á github: https://github.com/vthorsteinsson/Reynir. Til þess að keyra Greyni á eigin tölvu er nauðsynlegt að sækja BÍN samkvæmt þeim skilmálum sem um hana gilda (sjá kafla 2.5.1.8). Greynir er að mestu leyti skrifaður í Python og er undir GNU GPL v3-leyfinu.

Að hvaða leyti ákveðnir hlutar IceNLP og Greynis nýtast við gerð opinna stoðtóla innan máltækniáætlunar verður að meta sérstaklega út frá gæðum einstakra eininga og þörfum annarra máltækniverkefna.

Hér á eftir verður fjallað um þau stoðtól sem þarf að þróa eða þróa áfram innan áætlunarinnar.

2.5.3.1 TÓL FYRIR HANDVIRKA MÖRKUN (E. ANNOTATION TOOL)

Tól til þess að nota við handvirka mörkun texta eru mikilvæg hjálpartól í ýmissi gagnavinnu. Texta þarf oft að merkja handvirkt að öllu eða einhverju leyti þegar verið er að útbúa gögn fyrir máltæknitól. Textar geta verið markaðir með málfræðiupplýsingum, merkingarlegum upplýsingum, einstökum þáttum eins og mannanöfnum, örnefnum eða dagsetningum, ritvillum o.s.frv. Mörkun, hvort sem hún er gerð handvirkt eða sjálfvirkt, er framkvæmd til þess að annar hugbúnaður geti lesið upplýsingar úr stöðluðum merkingum.

Markmiðið með handvirkri mörkun er yfirleitt að útbúa nægilegt magn af gögnum sem hugbúnaður getur lært af til þess að mörkunin geti að lokum verið framkvæmd sjálfvirkt. Einnig getur verið mikilvægt að nota handvirkt mörkuð gögn til þess að prófa gæði hugbúnaðar sem á að framkvæma sömu greiningu sjálfvirkt.

Dæmi um mörkun á sérnöfnum:

> Guðni Th. Jóhannesson [MANNANAFN], forseti Íslands [STAÐARHEITI], skaut léttum skotum að Vladimir Pútin [MANNANAFN], forseta Rússlands [STAÐARHEITI], á Arctic Forum [VIÐBURÐUR] ráðstefnunni í Arkhangelsk [STAÐARHEITI] í Rússlandi [STAÐARHEITI] í gær

Ef tól er fyrir hendi, sem getur markað texta með þeim upplýsingum sem á að marka, flýtir fyrir að keyra fyrst sjálfvirka mörkun sem er svo leiðrétt handvirkt. Það er því mikilvægt að auðvelt sé að varpa úttaki sjálfvirkra markara á það form sem mörkunartólið vinnur með. Sömuleiðis að úttak handvirkrar mörkunar sé auðvelt að nota sem inntak fyrir þau máltæknitól sem þurfa að nota það. Annar mikilvægur eiginleiki mörkunartóls er að auðvelt sé að vinna með það og að það sé öflugt og áreiðanlegt. Almenn tölvukunnátta á að nægja til þess að geta markað texta handvirkt.

2. KJARNAVERKEFNI

Ekkert ákveðið mörkunartól hefur verið nýtt við handvirka mörkun hingað til við vinnslu íslenskra málheilda. Til eru opin tól sem hægt er að nýta en leggja þyrfti vinnu í að ígrunda hvaða tól væri hentugast m.t.t. ofangreindra atriða. Best væri að koma upp alhliða umhverfi sem byði upp á að marka texta með öllum þeim mismunandi upplýsingum sem mismunandi máltæknitól vinna með. Það þarf að vera hægt að skilgreina ný mörkunarsett fyrir ný verkefni, sem m.a. geta einnig falið í sér að marka vensl á milli markaðra eininga. Þetta kostar mögulega meiri vinnu í upphafi en gerir alla þróun einfaldari til lengri tíma litið.

Dæmi um opin tól til handvirkrar mörkunar:

- GATE https://gate.ac.uk/family/developer.html
- brat http://brat.nlplab.org/index.html
- Flat https://github.com/proycon/flat. Vefumhverfi sem byggist á FoLiA annotation-forminu sem aftur byggist á XML: http://proycon.github.io/folia/

I.1 Val og aðlögun á tóli fyrir handvirka mörkun

Verkþættir:

- Þarfagreining fyrir mörkunartól, val á opnum hugbúnaði
- Uppsetning og aðlögun, leiðbeiningar

Mannauður:

- Gagnasérfræðingur og/eða sérfræðingur í máltækni, forritari: 3 mánuðir

2.5.3.2 ORÐTÖKUTÓL

Við uppbyggingu og viðhald gagnasafna á borð við BÍN og Íslenskt orðanet er nauðsynlegt að hafa yfirsýn yfir gagnasafnið, hvort sem það snýr að eyðum í gagnasafninu eða upplýsingum sem veittar eru í dæmum. Með því að orðtaka skipulega stór gagnasöfn, þar sem óþekktum orðmyndum er safnað ásamt tölfræði um þekktar og óþekktar orðmyndir, geta ritstjórar gagnasafnanna ekki aðeins áttað sig á því hvar eyður eru í þeirra gagnasöfnum og því þegar ný orð ná fótfestu, heldur auðveldar það þeim líka að sjá í hvaða samhengi orðin eru notuð. Orðtökutól getur líka gagnast við orðabókargerð og við rannsóknir á nútímamáli.

I.2 Orðtökutól

Verkþættir:

- Smíða sveigjanlegt orðtökutól sem nýtt getur Risamálheildina og aðrar málheildir sem settar eru upp með sama sniði.

Mannauður:

- Forritari: 4 mánuðir

2.5.3.3 TILREIÐARI (E. SENTENCE DETECTOR, TOKENIZER)

Eitt af grunnskrefunum í allri málvinnslu með texta er að skipta textanum upp í einingar, yfirleitt setningar og tóka (e. *tokens*). Villur sem gerðar eru á þessu frumstigi gagnaundirbúnings halda sér áfram í gegnum allt vinnsluferlið. Það er því mikilvægt að gæðin séu sem mest.

Helsta áskorunin fyrir setningaþekkjara (e. *sentence detector*, *sentence segmentizer*) er að greina hvenær punktur táknar lok setningar og hvenær ekki og/eða hvenær stór stafur táknar upphaf setningar og hvenær ekki. Tókaþekkjari (e. *tokenizer*) í sinni einföldustu mynd skiptir texta niður í tóka sem afmarkast af bilum. Þetta er þó í fæstum tilfellum fullnægjandi, þar sem ýmis merki, tákn og tölur geta myndað eigin tóka án þess að afmarkast af bili. Einfalt dæmi eru greinarmerki sem þarf yfirleitt að skilja frá orðinu sem þau standa á eftir.

Áskoranir í þessum verkþætti eru áþekkar fyrir íslensku og önnur tungumál. Engu að síður þarf séríslenska útgáfu af tilreiðara sem þekkir t.d. íslenskar skammstafanir, tímaeiningar, dagsetningar o.s.frv. Tilreiðari þarf einnig að vera stillanlegur því það þarf að vera hægt að aðlaga úttakið að mismunandi verkefnum. Sem dæmi má nefna að táknin '#' og '@' hafa ákveðna merkingu á Twitter og sveigjanlegur tilreiðari gæti ýmist litið á þessi tákn sem hluta af tóka eða ekki ('#stefnuræða' vs. '#', 'stefnuræða').

Hér á eftir verður fjallað nánar um tilreiðara í IceNLP og Greyni.

Setningaþekkjarinn í IceNLP skiptir texta upp í setningar samkvæmt svokölluðum SRX-reglum (e. *Segmentation Rules eXchange*). Almenna reglan er sú að punktar (einnig spurningarmerki og upphrópunarmerki) segja til um lok setningar. SRX-reglurnar fjalla um undantekningar frá þeirri reglu. Flestar reglurnar innihalda skammstafanir og segja til um að punktur í skammstöfun eigi ekki að túlka eins og punkt í lok setningar. Einnig eru

reglur um dagsetningar, tíma og aðrar raðtölur. Setningaþekkjaranum tekst þó ekki alltaf að greina setningaskil á réttum stöðum: *Prátt fyrir botngjöf á ca. [setningaskil] 100 km. hraða …*

Tókaþekkjarinn í IceNLP inniheldur setningaþekkjarann og skiptir setningunum áfram upp í tóka. Hægt er að velja um nokkrar stillingar, t.d. hvort skammstöfunum er skipt upp og hvort tókaþekkjarinn á að skipta strangt eða ekki strangt, t.d.: *delta§(4)* eða *delta § (4)*. Þrátt fyrir þessar stillingar er tókaskiptingin ekki nægilega fyrirsjáanleg, þ.e. ekki er hægt að treysta á að niðurstöðurnar fylgi þeirri stillingu sem valin er.

Greynir er fyrst og fremst hugsaður sem heildstætt máltæknitól sem sett er upp sem vefviðmót. Það er því engin eining eða forritunarskil innan Greynis sem bjóða upp á hreina textatilreiðingu án málgreiningar. Setningaþekkjarinn í Greyni vinnur með niðurstöður tókaþekkjara sem merkir tóka m.a. sem greinarmerki og mögulega byrjun eða lok setningar ef við á. Hér er því farin önnur leið en sú hefðbundna að skipta texta fyrst upp í setningar og láta tókaþekkjara svo vinna setningu fyrir setningu. Greynir skilar svipuðum niðurstöðum í setningaskiptingu og IceNLP, þ.e. hann túlkar of marga punkta sem lok setningar og setur þannig setningaskil inn í miðjar setningar.

Tókaþekkjarinn í Greyni býður ekki upp á mismunandi stillingar. Hann skiptir strengjum strangt upp í tóka þar sem tákn koma fyrir. Hann þekkir þó raðtölupunkta *4. september* og punkta í tölum *40.000* og skilur þá ekki frá. Úttakið er því nokkuð fyrirsjáanlegt en um leið ósveigjanlegt.

Textatilreiðarinn í IceNLP er tilbúið tól sem hefur verið töluvert notað í íslenskum máltækniverkefnum. Greynir inniheldur tilreiðara, en hann er ekki til sem sjálfstætt tól eða viðmót. Vilyrði hefur fengist hjá rétthafa Greynis til þess að útbúa sjálfstæða einingu byggða á tilreiðara Greynis. Sú eining mætti vera gefin út undir Apache 2.0-leyfinu sem samræmist máltækniáætluninni. Við leggjum til að það verði gert.

Báðir tilreiðararnir þarfnast frekari þróunar. Nauðsynlegt er að til verði a.m.k. einn verulega góður tilreiðari fyrir íslenska texta sem býður upp á sveigjanlegar stillingar. Útbúa þarf prófunarsett sem tekur til mismunandi inntaks og þarfa í úttaki til þess að áframhaldandi þróun verði markviss.

I.3 Textatilreiðari

Verkþættir:

- Útbúa sjálfstæðan textatilreiðara sem hægt er að keyra með mismunandi stillingum. Mögulega best að nýta tilreiðarann úr Greyni, einangra hann og þróa áfram
- Útbúa prófunarsett fyrir tilreiðara sem tekur tillit til mismunandi stillinga

Mannauður:

- Sérfræðingur í máltækni (forritari): 6 mánuðir

2.5.3.4 MÁLFRÆÐILEGUR MARKARI (E. PART-OF-SPEECH TAGGER)

Fyrsti málfræðimarkarinn fyrir íslensku var smíðaður af Stefáni Briem í undirbúningsvinnu fyrir Íslenska orðtíðnibók sem kom út árið 1991. Markarinn var aldrei gefinn út en hann var notaður til að marka textana sem fóru í bókina áður en öll mörk voru handyfirfarin. Á árunum 2001–2003 var fjöldi markara þjálfaður á mörkuðu textunum úr Orðtíðnibókinni. Bestu niðurstöðurnar fengust með TnT-markara. Búinn var til pakki, CombiTagger sem notaði 5–6 mismunandi markara og valið var það mark sem hlaut atkvæði flestra markaranna. Með þeim hætti náðist allt að 93,41% mörkunarnákvæmni. IceTagger er hluti af IceNLP-pakkanum og í prófunum árið 2009 sýndi hann meðalnákvæmni upp á 91,59%. Mesta mörkunarnákvæmni sem náðst hefur var með IceStagger sem komst upp í 93,84% nákvæmni.

Ekki hafa verið gerðar miklar rannsóknir á mörkun íslenskra texta síðan IceStagger var smíðaður árið 2012. Það þarf að setja upp verkefni sem kanna hversu mikla nákvæmni hægt er að ná með nýjustu aðferðum, t.d. tauganetum á borð við LSTM. Eins væri gagnlegt að gera rannsókn á því hversu mikill hluti af því sem vélmarkari nær ekki er í rauninni ómögulegt að ná, t.d. vegna tvíræðni í textum.

Nákvæm mörkun texta er mikilvæg í flestum máltækniverkefnum vegna þess að mörkunin er helsta stoð tölvunnar til að skilja undirliggjandi kerfi málsins. Betri mörkunarnákvæmni getur því í mjög mörgum tilvikum bætt gæði máltæknibúnaðar. Það er því mikilvægt að haldið sé áfram að reyna að bæta mörkunaraðferðir fyrir íslensku og greina hversu langt er mögulegt að

komast, bæði fræðilega (þ.e. hversu mikil væru líkindin ef tveir mjög færir málfræðingar mörkuðu sama textann og afraksturinn væri borinn saman) og tæknilega.

I.4 Málfræðilegur markari

Verkþættir:

- Rannsaka hversu nákvæm mörkun getur orðið í íslensku með það markamengi sem notað hefur verið
- Tilraunir með nýjustu gerðir markara með það að markmiði að komast yfir 94% nákvæmni

Mannauður:

- Málfræðingur, sérfræðingur í máltækni: 18 mánuðir

2.5.3.5 ÞÁTTARI (E. PARSER)

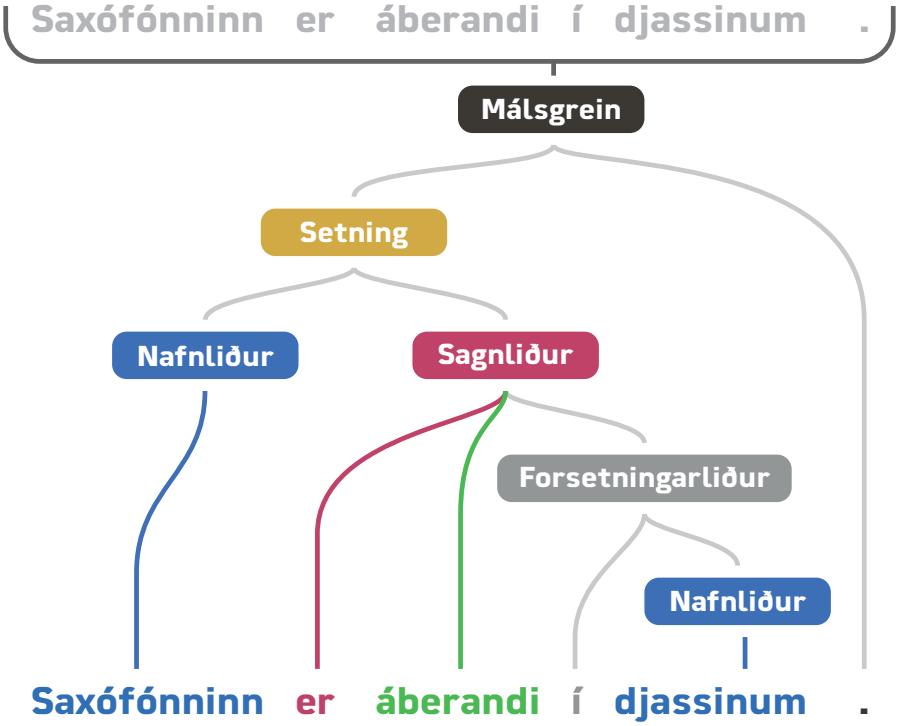
Setningagreinir eða þáttari nýtir úttak markara og greinir setningabyggingu eftir einhverri fyrirfram skilgreindri setningafræði. Bæði Greynir og IceNLP geta framkvæmt setningagreiningu. Beitt er ólíkum aðferðum sem skila ólíku úttaki og beinn samanburður er því ekki mögulegur. Ekki er heldur til neinn gullstaðall til þess að prófa gæði setningagreiningar.

IceParser í IceNLP greinir markaðan texta (frá IceTagger) eftir liðgerðarreglum. Setningarnar eru ekki fullgreindar heldur hlutaþáttaðar (e. *shallow parsing*, *chunking*), hver setningarliður er greindur út af fyrir sig en engin heildargreining fyrir málsgreinina framkvæmd. Hægt er að láta IceParser merkja þá liði sem við á með setningarhlutverki þeirra, frumlag, andlag o.s.frv. og er þá einnig merkt með hvaða sagnlið frumlag og andlag stendur.

Hrafn Loftsson og Eiríkur Rögnvaldsson hafa gert prófanir á IceParser. Hlutaþáttunin nær 96,7% F-gildi ef mörkunin á inntakinu er úr Gullstaðlinum (sjá kafla 2.5.1.2) en 91,9% ef IceTagger er notaður til þess að marka inntakið. Þáttun með setningarhlutverkum nær 84,3% F-gildi á inntaki með Gullstaðalsmörkun en 75,3% með inntaki frá IceTagger. Þess ber að geta að þessar prófanir eru frá árinu 2007.

Þáttari Greynis fullþáttar setningar samkvæmt context free-málfræðireglum sem samdar voru sérstaklega fyrir Greyni. Greynir býr til allar mögulegar

trjágreiningar samkvæmt málfræðinni og velur þá líklegustu. Fyrir setninguna *Saxófónnin er áberandi í djassinum* finnur Greynir 167 trjágreiningar. Fyrir flóknari setningar skipta möguleikarnir þúsundum. Meðfylgjandi mynd sýnir hvernig líklegasta greiningin er sýnd í vefviðmóti Greynis. Engin prófunargögn eru til fyrir Greyni og því ekki hægt að leggja mat á gæði þáttarans.



Mynd: Dæmi um setningagreiningu Greynis.

IceParser og Greynir eru tveir mjög ólíkir þáttarar fyrir íslensku. Oft nægir hlutaþáttun eins og IceParser framkvæmir, en fyrir önnur verkefni þarf fullþáttun eins og Greynir býður upp á. IceParser þyrfti að gæðaprófa í því ástandi sem hann er í dag þar sem síðustu skráðu próf voru gerð árið 2007. Greynir er tiltölulega nýr og gæði þáttarans hafa ekki verið metin. Til þess að geta notað þáttara Greynis sem sjálfstætt tól þarf að leita til rétthafa með að gefa hann út undir alveg opnu leyfi, t.d. Apache 2.0.

Mögulega henta þó enn aðrar þáttunaraðferðir fyrir einstök verkefni. Ber þar helst að nefna mörkun og þáttun samkvæmt *Universal Dependencies* sem hefur verið að ryðja sér til rúms á undanförnum árum (http://universaldependencies.org/). UD-trjábankar eru til fyrir fjölmörg tungumál, m.a. ensku, dönsku, norsku, sænsku, finnsku, eistnesku og þýsku. Eitt markmiða UD er að auðvelda það að samnýta máltæknitól milli tungumála þar sem mörkun og setningargreining er á samræmdu formi.

I.5 Þáttarar

Verkþættir:

- Gæðaprófun á þátturum IceNLP og Greynis, mat á áframhaldandi þróun.
- Aðlögun *Universal Dependencies*-þáttara fyrir íslenska UD-málfræði.
- Samstarf við önnur teymi um þarfir í sjálfvirkri setningagreiningu.

Mannauður:

- Sérfræðingur í máltækni: 18 mánuðir

2.5.3.6 LEMMALD

Lemmald leitar að uppflettimynd orðmynda í texta og notar til þess upplýsingar á borð við málfræðimörk, beygingarlýsingu og samhengi sem orðin standa í. Smíðuð hafa verið tvö forrit til að lemma íslenska texta. Það fyrra gaf góða raun í prófunum þegar það var þróað en við notkun í Markaðri íslenskri málheild eru áberandi villur sem gerðu það að verkum að hafist var handa við að smíða lemmald með öðrum aðferðum. Sú vinna hefur lofað góðu en er talsvert frá því að vera fulllokið. Nýja lemmaldið nýtir BÍN til að finna þekktar beygingarmyndir en notar reglur til að lemma óþekkt orð og til að greina á milli margræðra orðmynda. Nýja lemmaldið er gefið út með Apache-leyfi. Nokkuð vantar þó upp á að ljúka vinnu við það og þá á eftir að gera rannsókn á nákvæmni þess svo hægt sé að bera það saman við gamla lemmaldið.

I.6 Lemmald

Verkþættir:

- Ljúka vinnu við Lemmald sem hefur verið í þróun
- Rannsaka gæði og bera saman við eldra lemmald

Mannauður:

- Sérfræðingur í máltækni: 6 mánuðir

2.5.3.7 NAFNAÞEKKJARI

Hlutverk nafnaþekkjara er að finna og flokka sérnöfn, svo sem nöfn fyrirtækja og stofnana, mannanöfn og staðarnöfn, auk tölulegra eininga á borð við tíma, magntölur, verð og hlutfallstölur. Einhver vinna hefur verið unnin við nafnaþekkjara fyrir íslensku. Í IceNLP er nafnaþekkjaraeining sem náð hefur 71%–79% nákvæmni. Einnig er vísir að nafnaþekkjara í Greyni en nákvæmni hans hefur ekki verið mæld. Fyrir önnur tungumál hafa bestu kerfin náð 93–94% nákvæmni. Þegar málheild fyrir nafnaþekkjara verður tilbúin (sjá 2.5.1.5) þarf að gera tilraunir með mismunandi aðferðir, þar á meðal tauganet sem skilað hafa góðum niðurstöðum á þessu sviði á allra síðustu árum.

I.7 Nafnaþekkjari

Verkþættir:

- Tilraunir með mismunandi aðferðir þjálfunar.
- Val á aðferð og útgáfa á nafnaþekkjara.

Mannauður:

- Sérfræðingur í máltækni: 6 mánuðir

2.5.3.8 MERKINGARGREINING

Verkefni í merkingargreiningu eru margvísleg: að tengja orð í texta við ákveðna merkingu og einræða ef þörf er á, að leysa úr endurvísunum (e. *anaphora resolution*), og finna staðgengla (e. *co-reference*), að meta merkingarvensl milli orða eða hversu lík merking orða eða texta er, að greina merkingarfræðileg hlutverk (þolandi, gerandi o.s.frv.) svo eitthvað sé nefnt. Einnig er hægt að nýta aðferðir merkingargreiningar til þess að vinna merkingarupplýsingar úr miklu magni texta til þess að búa til eða bæta við merkingarorðasöfn eins og orðanet.

Verkefni í merkingargreiningu eru skilgreind í kafla 2.4.5.6 um málrýni. Önnur áríðandi verkefni í merkingargreiningu lúta að einræðingu, greiningu orðasambanda, endurvísunum og staðgenglum. Í textanum á myndinni á bls. 134 má sjá dæmi um öll þessi atriði: *léttari hluti* þarf að einræða, merkir hér *skemmtilegar/léttvægar hliðar á einhverju* en ekki til dæmis *áþreifanlegan hlut sem ekki er þungur*; sagnasambandið *unnið fyrir því* þarf að greina til þess að *unnið* sé ekki greint í merkingunni *að vinna fótboltaleik*; mikið er um

endurvísanir þar sem *hann* eða *honum* vísar í Gylfa Þór Sigurðsson og hann er ýmist nefndur *Gylfi*, *Gylfi Þór*, *Gylfi Sigurðsson* eða *Gylfi Þór Sigurðsson*; *úrvalsdeildinni* og *deildinni* eru staðgenglar fyrir *ensku úrvalsdeildinni* og með *lið sem berst við falldrauginn* er átt við Swansea:

Mynd: Endurvísanir og staðgenglar: það sem er í sama lit á við það sama.

„**Gylfi Sigurðsson** á skilið sæti í liði ársins."

Þetta er fyrsta setningin í myndbandi sem Facebook-síðan Dream Team gerði um **Gylfa Þór Sigurðsson**, leikmann **Swansea** í **ensku úrvalsdeildinni**. Dream Team er vinsæll hluti enska götublaðsins The Sun þar sem farið er aðeins yfir léttari hluti fótboltans.

Í myndbandinu er bent réttilega á það að **Gylfi** er búinn að spila frábærlega fyrir **Swansea** og er stoðsendingahæstur í **úrvalseildinni**. Kevin de Bruyne hjá Manchester City er reyndar búinn að ná **honum**.

Tekið er fram að **Gylfi Þór** er búinn að skora fleiri mörk í **deildinni** en Philippe Coutinho hjá Liverpool og Pedro hjá Chelsea. Þá hefur **hann** komið að fleiri mörkum en Dele Alli hjá Tottenham.

Ólíkt öðrum sem verða vafalítið í liði ársins er **Gylfi Þór** að spila fyrir **lið sem berst við falldrauginn** en samt sem áður er **hann** búinn að skora eða leggja upp ríflega helming allra marka **liðsins**.

„**Gylfi Þór Sigurðsson** á skilið sæti í liði ársins því **hann** hefur unnið fyrir því"

(http://www.visir.is/g/2017170409191/-gylfi-sigurdsson-a-skilid-saeti-i-lidi-arsins-)

I.8 Merkingargreining: Einræðing, orðasambönd, endurvísanir og staðgenglar

Verkþættir:

- Einræðing stakra orða.
- Greining samsettra sagna og orðasambanda.
- Þróun aðferða til greiningar á endurvísunum.
- Þróun aðferða til greiningar á staðgenglum.

Mannauður:

- Sérfræðingur í máltækni: 36 mánuðir.

3 LEYFISMÁL OG AÐGENGI MÁL-FANGA

3. LEYFISMÁL OG AÐGENGI MÁLFANGA

Markmiðið er að þau verkefni, sem unnin verða innan áætlunarinnar, megi nýta til áframhaldandi þróunar, hvort sem er í rannsóknar- eða viðskiptaskyni.

Til þess að tryggja aðgengi að tólum og gögnum sem unnin verða innan áætlunarinnar þarf að setja ákveðin rammaskilyrði fyrir hvert verkefni. Markmiðið er að öll verkefnin megi nýta til áframhaldandi þróunar, hvort sem er í rannsóknar- eða viðskiptaskyni.

3.1 LEYFI

Ef hugbúnaður eða önnur gögn eru gefin út án sérstaks leyfis (e. *license*) er hann ekki frjáls og opinn í þeim skilningi að aðrir megi breyta, nota, dreifa eða selja. Það er því mikilvægt að tekin sé vel ígrunduð ákvörðun um hvaða leyfi eigi við í hvert sinn. Við leggjum til að allur hugbúnaður á vegum áætlunarinnar verði með opnu leyfi á borð við Apache 2.0 og gögn verði með eins opnum leyfum og unnt er með tilliti til höfundarréttar- og persónuverndarsjónarmiða.

3.1.1 ÚTBREIDD LEYFI FYRIR HUGBÚNAÐ

Apache License 2.0

Apache 2.0-leyfið er mjög opið leyfi sem leyfir ótakmarkaða notkun hugbúnaðar og hugbúnaðarkóða. Þar með talið notkun í viðskiptalegum tilgangi og endurútgáfu á kóðanum með eigin breytingum. Upprunalegum eða breyttum kóða verður að dreifa eða endurútgefa undir Apache 2.0-leyfinu. Ekki þarf hinsvegar að gefa út kóða hugbúnaðar sem þróaður er með því að nýta Apache 2.0-kóða.

GNU General Public License (GNU GPL v3)

Þetta er leyfi sem leggur áherslu á „fjórfrelsið“: 1) frelsi til þess að nýta hugbúnað í hvaða tilgangi sem er, 2) frelsi til þess að aðlaga hugbúnað að eigin þörfum, 3) frelsi til þess að deila hugbúnaði og 4) frelsi til þess að deila breytingum sem gerðar eru.

Um leið eru ákveðnar kvaðir sem fylgja því að nýta hugbúnað með GPLv3-leyfi: Allur kóði hugbúnaðar sem nýtir kóða með GPLv3-leyfi (óbreyttan eða breyttan) verður líka að vera opinn kóði. Það er því ekki alltaf fýsilegt fyrir fyrirtæki í samkeppni að nýta kóða með þessu leyfi.

GNU Lesser General Public License (GNU LGPL v3)

Helsti munur á GNU GPL og GNU LGPL er að LGPL leyfir notkun í hugbúnaði án þess að kóðinn í heild sinni sé opinn.

3.1.2 LEYFI FYRIR GÖGN

Ófrávíkjanleg regla er að gögn sem unnið er með innan áætlunarinnar séu með skilgreindum notkunarleyfum. Því skulu gögn sem verða til eða eru þróuð innan áætlunarinnar gefin út með opnum alþjóðlegum leyfum. Markmið áætlunarinnar er að hámarka notkunarmöguleika og nýtingu gagnanna. Leitast skal við eins og mögulegt er að gögn sem aflað er frá þriðja aðila til að nota í máltækniverkefnum innan áætlunarinnar verði gefin út með opnu leyfi. Þegar því verður ekki komið við skal leitast við að gefa gögnin út með eins litlum takmörkunum og mögulega er unnt. Þegar leyfi eru valin þarf að taka afstöðu til eftirfarandi þátta: Má endurútgefa gögnin og þá með hvaða skilmálum? Má breyta gögnunum? Má nýta gögnin í verkefni í ágóðaskyni? Almennt er þess krafist í flestum leyfum að uppruna eða höfundar gagnanna sé getið.

Creative Commons (CC) eru dæmi um vel þekkt leyfi fyrir gögn. Kosturinn við að nota vel þekkt leyfi er að þeir sem hyggjast nýta gögnin eiga auðvelt með að átta sig á því hvað má og hvað ekki og hvaða skilmálum ber að fylgja. Sérsniðin leyfi geta fælt frá, sérstaklega ef þau eru ítarleg og flókin. Til eru nokkrar gerðir CC-leyfa. Opnasta leyfið leyfir alla notkun með því skilyrði að uppruna sé getið. Þetta þýðir t.d. að hver sem er má fá eintak af gögnunum, afrita þau og selja ef honum sýnist svo. Hann gæti einnig breytt þeim og birt breytta útgáfu, notað þau til rannsókna eða gert hvað annað sem honum dytti í hug.

Önnur CC-leyfi takmarka notkunarmöguleika með einhverjum hætti. T.d. banna sum leyfin alla notkun í ágóðaskyni, sum banna að gögnunum sé breytt og sum krefjast þess að allar afurðir sem út úr gögnunum koma verði dreift með sams konar leyfi og upprunalegum gögnum. Stefna ætti að því að nota leyfi sambærileg við CC-leyfin.

3.1.3 STAÐLAR

Í upphafi hvers verkefnis þarf að ákveða hvaða stöðlum á að fylgja. Mismunandi staðlar hafa verið skilgreindir fyrir ólíkar tegundir gagna, ákveðnir staðlar eru til fyrir orðabókargögn og aðrir fyrir hljóðrituð gögn svo dæmi

séu tekin. Mikilvægt er að staðlarnir séu opnir og vel þekktir á því sviði sem um ræðir, sérstaklega í nágrannalöndum okkar og öðrum Evrópulöndum þar sem líklegt er að áhugi sé fyrir samstarfi. Hugbúnaður þarf að sjálfsögðu að fylgja stöðlum þess umhverfis sem unnið er í og einnig þarf að gæta að því að vanda val á hugbúnaði þriðja aðila þar sem það á við, að þess sé gætt að um sé að ræða virkt viðhald og útbreidda notkun. Allur hugbúnaður, sem skilað er innan verkefnisins sem opnum hugbúnaði, skal vera vandlega prófaður og skjalaður.

3.2 AÐGENGI OG YFIRFÆRSLA

Við undirbúning verkefna og í áætlanagerð er æskilegt að koma snemma á samskiptum við hugsanlega notendur til þess að tækniyfirfærsla gangi sem best fyrir sig. Hér ber að nefna hluti eins og forritaskil, forritunarmál og stýrikerfi. Höfundar hugbúnaðar skulu kappkosta að hann nýtist sem víðast og gera grein fyrir því sem þarf að breyta/aðlaga fyrir mismunandi vinnuumhverfi.

Öllum verkefnum skal skilað vel skjöluðum í miðlæga varðveislustöð fyrir íslenska máltækni. Þar þarf að vera virkt utanumhald þannig að þeir sem hyggjast nýta gögn og/eða hugbúnað hafi ætíð greiðan aðgang að miðstöðinni, t.d. í gegnum vefviðmót. Tillögur um skipan varðveislu-, viðhalds og aðgengismála eru í kafla 6.3.

4 ÖNNUR MÁLTÆKNI-VERKEFNI

4. ÖNNUR MÁLTÆKNIVERKEFNI

Kjarnaverkefnin sem hér hafa verið skilgreind eru grunnur fyrir mörg flókn-ari máltæknitól. Að geta greint innihald tals og texta (fyrirspurna, athuga-semda, lengri texta) og tengt við aðra texta, gagnagrunna, þekkingargrunna eða aðrar upplýsingar og unnið úr upplýsingum og venslum, jafnvel með aðstoð gervigreindar, er grundvöllur allra „snjallra“ samskipta við og í gegnum tæki. Upplýsingaútdráttur, viðhorfsgreining, upplýsingaheimt, samræðugreining og margmiðlunargreining eru meðal þeirra þátta sem gera þróun öflugra og snjallra samskipta- og upplýsingakerfa fyrst mögulega.

Hér á eftir verður fjallað stuttlega um nokkur tól sem eru sem stendur utan kjarnaverkefna. Fleiri tól má nefna eins og textasamantekt og efnisflokkun skjala. Ráðast ætti í þróun slíkra tóla um leið og ráðrúm gefst hvað varðar mannafla og nauðsynleg grunntól.

4.1 UPPLÝSINGAÚTDRÁTTUR

Upplýsingaútdráttur (e. *information extraction*) vinnur staðreyndir úr textum eftir ákveðnum mynstrum og nýtir við áframhaldandi vinnslu og/eða vistar í þekkingargagnagrunni. Fyrsta skrefið í upplýsingaútdrætti er að nota nafnaþekkjara til þess að finna sérnöfn og fleiri einingar eins og lýst er í kafla 2.5.3.7 um nafnaþekkjara. Ýmsar aðferðir er svo hægt að nota til þess að greina sambönd og vensl í textum. Úr fréttatextanum *Guðni Th. Jóhannesson, forseti Íslands, skaut léttum skotum að Vladimír Pútín, forseta Rússlands, á Arctic Forum-ráðstefnunni í Arkhangelsk í Rússlandi í gær* mætti til að mynda vinna eftirfarandi staðreyndir:

Guðni Th. Jóhannesson er forseti Íslands

Vladimír Pútín er forseti Rússlands

Arctic Forum er ráðstefna

Arctic Forum var haldin í Arkhangelsk

Arkhangelsk er í Rússlandi

Guðni Th. Jóhannesson og Vladimír Pútín voru á Arctic Forum

Að auki getur verið mikilvægt að tímasetja atburði. Með því að tengja *í gær* við útgáfudag fréttarinnar má tímasetja Arctic Forum og um leið greina að á þessum tímapunkti voru Guðni Th. og Pútín forsetar landa sinna.

Greynir inniheldur nú þegar grunn að nafnaþekkjara og upplýsingaútdrætti. Úr setningu eins og *Ellen Calmon, formaður Öryrkjabandalagsins* getur Greynir áttað sig á starfi Ellen Calmon og vistað í gagnagrunni. Enn sem komið er virðast þær upplýsingar sem fundnar eru um persónur ekki vera greindar nánar heldur vistaðar í heild sinni, þannig að ef hlutverk er orðað með öðrum hætti (*formaður ÖBÍ*, *formaður félagsins*) verða til ný hlutverk í gagnagrunninum. Ekki er ljóst hvort og þá með hvaða hætti sjálfstæð áframhaldandi þróun nafnaþekkjara og upplýsingaútdráttar úr Greyni er möguleg.

4.2 ÁLITSGREINING/ VIÐHORFSGREINING

Álits- eða viðhorfsgreining (e. *sentiment analysis*) greinir hvort ákveðin segð eða texti er jákvæður eða neikvæður í garð t.d. fyrirtækis eða vöru. Setningar geta verið metnar jákvæðar (*Frábær veitingastaður, besti matur sem ég hef fengið!*); neikvæðar (*Ég ætla aldrei að skipta við þetta fyrirtæki aftur, ömurleg þjónusta*); eða hlutlausar (*Ég fór út að borða í gærkvöldi*). Til þess að flokka slíkar setningar sjálfvirkt er ýmist hægt að nýta sérstaka viðhorfsorðalista, þar sem *aldrei aftur*, *ömurleg* væru merkt sem neikvæð og *frábær*, *besti* sem jákvæð, eða markaða texta þar sem kerfi getur lært hvernig setningar og textar eru jákvæðir/neikvæðir/hlutlausir. Mikilvægur þáttur í álitsgreiningu er einnig greining á „tjáknum“ (e. *emojis*), þar sem gildishlaðnir textar innihalda þau oft í miklum mæli.

Álit fjöldans hefur núorðið mun meiri áhrif á (mögulega) viðskiptavini en auglýsingar og því er augljóst kappsmál fyrir fyrirtæki að greina umræðuna. Sjálfvirkni á þessu sviði gerir þessa greiningu mun hagkvæmari og gerir mögulegt að komast yfir meira efni en ella.

4.3 UPPLÝSINGAHEIMT

Markmið upplýsingaheimtar (e. *information retrieval*) er að finna skjöl eða vefsíður sem eru líklegust til þess að innihalda þær upplýsingar sem leitað er að með ákveðinni fyrirspurn. Máltæknihliðin á upplýsingaheimt snýr að því að greina texta- og jafnvel hljóðgögn en upplýsingaheimt getur einnig falið í sér greiningu á myndum og kvikmyndum. Hún er umfangsmikið fagsvið sem snýr m.a. að skipulagningu og greiningu gagna, greiningu á fyrirspurnum og hvernig má á sem skemmstum tíma finna það sem best á við. Augljósasta dæmið um upplýsingaheimt í daglegri notkun eru leitarvélar eins og Google, sem skila leitarniðurstöðum í lista þar sem það sem best á við er efst.*

Þróunin á þessu sviði færist nú sífellt frá almennri leit yfir í sérhæfðar lausnir. Áherslan er á að greina og tengja saman ákveðnar upplýsingar úr miklu magni gagna á mismunandi formi og er oft hluti af flóknari viðskiptahugbúnaði. Fyrirtæki vista mikið magn gagna á fjölbreyttu formi, á mismunandi tungumálum og þar geta sérhæfð kerfi til upplýsingaheimtar hjálpað til m.a. við að flýta fyrir ferlum og auka virði gagna.

4.4 SPURNINGASVÖRUN

Upplýsingaheimt skilar gögnum sem eru líkleg til þess að innihalda svör við spurningum notandans en hann þarf sjálfur að lesa í gegnum skjölin. Til eru einnig kerfi til spurningasvörunar (e. *question answering*, *QA-systems*) sem skila einföldu svari við spurningu í stað þess að skila einungis tengdum gögnum. Slík kerfi þurfa að greina spurninguna sem borin er upp og leita svo að mögulegum svörum. Öflugustu kerfin nýta bæði þekkingar-gagnagrunna og verufræðinet, sem hægt er að fletta beint upp í, og aðferðir upplýsingaheimtar þar sem bestu niðurstöður eru greindar frekar og svör mynduð út frá þeim upplýsingum. Hér má nefna Watson-kerfið frá IBM sem vann sér það til frægðar að sigra stóra spurningakeppni árið 2011.

Google hefur samþætt upplýsingaheimt og spurningasvörun í leitarvél sinni. Spyrji maður á ensku *Who is the president of Finland?* kemur svarið um hæl, með mynd: Sauli Niinistö. Leitarvélin skilur spurninguna og kemur með svarið og vísar ekki einungis í vefsíður þar sem svarið gæti verið að

* Google notar þó ekki einungis aðferðir hefðbundinnar upplýsingaheimtar við að velja og raða leitarniðurstöðum.

finna. Spyrji maður hinsvegar á íslensku *Hver er forseti Finnlands?* er efsta niðurstaðan tengill á Wikipediasíðuna um Finnland. Einungis er leitað að orðunum í leitarstrengnum.

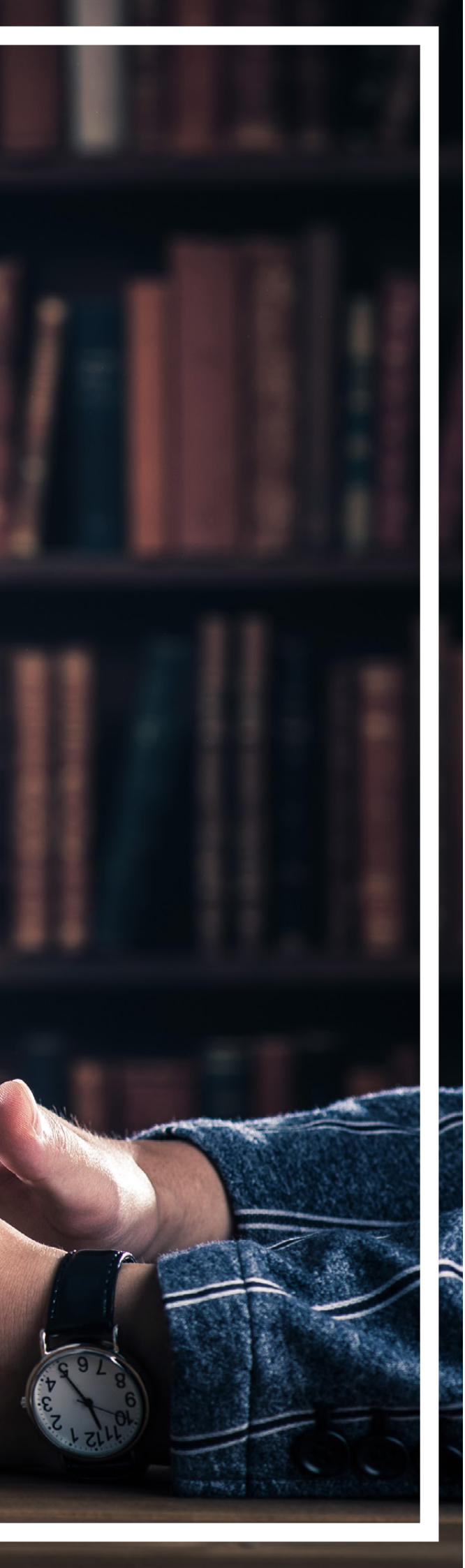
4.5 SAMRÆÐUKERFI

Samræðukerfi (e. *spoken dialogue systems*) gera samskipti í gegnum talað mál milli manns og tölvu möguleg. Hér eru símsvörunarkerfi algengt notkunartilvik, þar sem tölvukerfi geta leyst úr einföldum fyrirspurnum og verkefnum. Samræðukerfi má einnig finna í bílum og víðar. Algengast og einfaldast er að kerfið sé hannað þannig að tölvan stýri samtali um afmarkað efni en opnari kerfi eru flóknari, bæði hvað varðar innihald og stjórn á samtali. Nýjasta þróunin á þessu sviði eru gerviþjónarnir Alexa (Amazon), Google Assistant, Cortana (Microsoft) og Siri (Apple), kerfi sem finna upplýsingar og framkvæma verkefni eins og að hringja, panta, velja tónlist o.fl. Þessi kerfi tengjast einnig í auknum mæli „snjallheimilinu“, þar sem ljósum, gluggatjöldum, ofnum og öðrum húsbúnaði er stjórnað með tölvukerfi.

Hefðbundin samræðukerfi eru byggð upp eins og mynd á bls. 31 sýnir. Gervigreindarhlutinn er þá samræðukjarni: fyrst þarf að framkvæma talgreiningu, þá framkvæmir málgreiningarhluti sína vinnu og sendir áfram til samræðukjarnans (samræðugreiningar og -stýringarhluta), sem aftur sendir upplýsingar áfram til málmyndunarhluta, og að lokum les talgervill það sem kerfið vill segja. Til eru nokkrar aðferðir til þess að hanna samræðukjarna og tengist gerð málgreiningar- og málmyndunarhluta því hvaða aðferð er valin: gröf, rammar, gervigreindaraðferðir eða tölfræðiaðferðir. Hönnun samræðukerfa fylgir einnig núverandi þróun í notkun (djúp) tauganeta.

4.6 MARGMIÐLUNAR-GREINING/HLJÓÐ OG MYND

Í margmiðlunargreiningu (e. *multimedia content analysis*) er innihald (kvik) mynda, hljóðs (talmáls) og texta greint. Til að mynda er hægt að greina talgögn sjónvarpsefnis og flokka og tengja saman efni með svipuðu innihaldi, finna tengt efni af frétta- og samfélagsmiðlum, greina vörumerki og texta í myndum og tengja við aðra umfjöllun, og almennt að tengja saman innihald ólíkra miðla, jafnvel óháð tungumáli. Sífellt meira efni er á kvikmynda/vídeóformi og því er mikilvægt að finna leiðir til þess að vinna innihaldupplýsingar úr því efni ekki síður en textum. Myndin að neðan sýnir dæmi um „tíst“ með mynd: Nauðsynlegt er að beita myndgreiningu til þess að greina texta og jafnvel merki á myndinni til þess að skilja textann í tístinu.



5

NÝSKÖPUN Í MÁLTÆKNI

5. NÝSKÖPUN Í MÁLTÆKNI

Markmið Máltækniáætlunar fyrir íslensku 2018–2022 er að gera fólki kleift að nota íslensku í samskiptum sínum við hugbúnað og upplýsingakerfi. Því er ekki nóg að byggja upp þá nauðsynlegu innviði, sem gerð hefur verið grein fyrir, heldur er einnig mikilvægt að hvetja iðnaðinn til nýsköpunar í máltækni. Þannig eru innviðirnir nýttir í að leysa þá þörf sem er á að búa til stað fyrir íslenskuna í stafrænum heimi. Með góðum og víðtækum máltækniinnviðum skapast fjölmörg viðskiptatækifæri og möguleikar á spennandi tæknilausnum margfaldast. Það er mjög mikilvægt fyrir tungumálið að þessi tækifæri verði nýtt til hins ítrasta og því er jafnframt lagt til að tækniþróun og nýsköpun í máltækni verði studd í kjölfarið á þeirri innviðauppbyggingu sem unnin er í áætluninni.

Lagt er til að komið verði á hvatakerfi til nýsköpunar í máltækni með sérstökum tækniþróunarstyrkjum til fyrirtækja og stofnana sem útfæra og smíða máltæknilausnir. Hvatakerfið verður byggt á styrkjum sem koma til móts við fjárfestingu í nýsköpun í máltækni á svipaðan hátt og styrkir Tækniþróunarsjóðs virka almennt fyrir nýsköpun. Þar sem fyrirtæki eða stofnanir sem hefja svona þróun fjárfesta sjálf í þessum verkefnum eignast þau hugverkin sem verða til og ráða því hvort notkun á þeim verður opin eða ekki. Hér er gerð grein fyrir því hvernig þessi þáttur áætlunarinnar verður skipulagður og gefin dæmi um möguleg verkefni. Enn fremur verður fjallað um þá þekkingaryfirfærslu sem máltækni hefur yfirleitt í för með sér og tækifæri til útflutnings á þekkingu og þjónustu reifuð.

5.1 TÆKNIÞRÓUN Í MÁLTÆKNI

Því verður lögð áhersla á að mynda gott tengslanet milli þeirra fyrirtækja og stofnana sem stunda þróun og viðskipti með máltækni og þeirra sem vinna við að smíða og viðhalda innviðum.

Heildarmarkmið áætlunarinnar er að koma íslenskri máltækni í notkun hjá almenningi á sem flestum sviðum samfélagsins. Það er því mikilvægt að skapa og rækta nýsköpunarumhverfi í kringum máltækni þannig að fyrirtæki og stofnanir geti hafist handa við að skapa lausnir og veita þjónustu sem byggð er á máltækni. Því verður lögð áhersla á að mynda gott tengslanet milli þeirra fyrirtækja og stofnana sem stunda þróun og viðskipti með máltækni og þeirra sem vinna við að smíða og viðhalda innviðum.

Til þess að hvetja til þátttöku í þessu starfi verður settur á laggirnar samkeppnissjóður um tækniþróun í máltækni. Faghópur verður skipaður sem sér um að meta umsóknir í sjóðinn og taka ákvarðanir um úthlutun.

Innan faghópsins þarf að vera nógu góð þekking á máltækni og greinum tengdum henni til þess að meta megi umsóknirnar af þekkingu en varast ber hagsmunaárekstra. Þar sem áætlunin stendur yfir í takmarkaðan tíma er mælst til að úthlutað verði úr sjóðnum tvisvar til fjórum sinnum á ári og að ekki verði settur umsóknarfrestur á umsóknir heldur verði þær teknar fyrir jafnóðum og þær berast.

Unnið verður að því að að tengja saman fyrirtæki og stofnanir sem hyggjast stunda tækniþróun í máltækni og þá sem vinna við innviðauppbyggingu þannig að tryggt sé að margar góðar umsóknir komi í sjóðinn. Skipulagi og utanumhaldi á þessu starfi er lýst nánar í 6. kafla.

5.2 DÆMI UM TÆKNIÞRÓUNARVERKEFNI

Í lýsingu á innviðaverkefnum í 2. kafla var nokkrum tækniþróunarverkefnum lýst sem væru rökrétt framhald af þróun innviða sem fram fer í áætluninni. Hér er fleiri dæmum um slík verkefni lýst enda þurfa þau ekki endilega að vera í samhengi við einstaka innviði sem verið er að þróa heldur geta þau nýtt sér málföng og verkfæri frá fleiri en einu sviði máltækninnar. Tilgangurinn með að telja upp þessar hugmyndir er fyrst og fremst að sýna fram á þá möguleika sem opnast með góðum máltækniinnviðum. Þessar tillögur eru þó settar fram með tveimur fyrirvörum. Í fyrsta lagi getur svona listi úrelst fljótt. Tækniheimurinn breytist hratt og því getur verið að sumar hugmyndir á listanum eigi ekki við eftir nokkur ár og að aðrar komi í staðinn. Í öðru lagi er gert ráð fyrir því í verkefninu að frumkvæði að því að vinna þessi eða álíka verkefni komi frá nýsköpun í tækniiðnaði. Þannig eru það notendurnir sjálfir í gegnum viðskipti sín við tækni- og þjónustufyrirtæki sem hafa áhrif á þróunina. Frumkvæði atvinnulífsins veitir aðhald til að í máltækniáætluninni sé hraðri þróun tækninnar fylgt og brugðist sé við breytingum. Það er því mikilvægt að hafa þessa fyrirvara í huga þegar eftirfarandi listi er skoðaður.

Frumkvæði atvinnulífsins veitir aðhald til að í máltækniáætluninni sé hraðri þróun tækninnar fylgt og brugðist sé við breytingum.

5.2.1 SJÁLFVIRK LESTRARKENNSLA FYRIR BÖRN

Talgreining og sú tækni sem henni fylgir býður upp á marga möguleika í sjálfvirkri aðstoð við nám. Til dæmis er hægt að búa til þjálfunarforrit fyrir

börn þannig að þau geti lesið texta upp fyrir tölvu sem metur hversu vel þeim tekst til. Hægt er að útfæra þetta beint með talgreini en ef textinn sem lesa á upp er þekktur fyrirfram má fá nákvæmari niðurstöður með samröðun. Ef barninu tekst að lesa upp setninguna fær það mörg stig og þannig er hægt að búa til leik úr því að læra að lesa.

5.2.2 TÖLVUSTUDD TUNGUMÁLAKENNSLA

Hægt er að nýta máltækni til að styðja við kennslu í íslensku fyrir fólk sem ekki hefur hana að móðurmáli. Hægt er að búa til orðalista þannig að fólk geti aukið orðaforða sinn og lært orðmyndir eins og fallbeygingu fallorða og hætti og myndir sagna. Upplestur og framburð má læra á svipaðan hátt og lagt er til í verkefni sem lýst er í Sjálfvirk lestrarkennsla fyrir börn.

5.2.3 SJÁLFVIRK SÍMAVER

Fyrirtæki og stofnanir geta notað talgervil og talgreiningu til að koma á sjálfvirkni í símaverum við miðlun upplýsinga til viðskiptavina og skjólstæðinga.

Fyrirtæki og stofnanir geta notað talgervil og talgreiningu til að koma á sjálfvirkni í símaverum við miðlun upplýsinga til viðskiptavina og skjólstæðinga. Einföldum fyrirspurnum er þá svarað strax og örugglega og tími gefst fyrir starfsfólk sem annast upplýsingamiðlun að einbeita sér að flóknari fyrirspurnum. Um leið minnkar biðtíminn. Hér er augljóslega um mikið hagræði að ræða en til þess að ná þessu markmiði þurfa innviðir á borð við talgreiningu og talgervingu að vera tilbúnir.

5.2.4 RADDSTÝRING TÆKJA OG VEFJA

Mörg tæki eru framleidd með forritaskilum (e. *application programming interface*, *API*) sem leyfa raddstýringu. Auðvelt er að styðja slík tæki með íslenskri talgreiningu og talgervlum en til þess þurfa innviðirnir að vera komnir og möguleikar á stuðningi til tækniþróunar. Á svipaðan hátt má þróa raddstýringu fyrir vefi þannig að hægt sé að nálgast efni með raddskipunum og hlusta á það í stað þess að horfa á það á vefsíðu. Þannig væri hægt biðja um fréttir frá vefmiðlum munnlega og hlusta á upplestur þeirra. Þannig verður fréttaneyslan handfrjáls.

5.2.5 MERKINGARGREINING OG SNJÖLL LEITAR- OG UPPLÝSINGAKERFI

Fyrirtæki og stofnanir geyma ógrynni af gögnum á fjölbreyttu formi, jafnt í skipulögðum gagnagrunnum sem og alls kyns óskipulegum gögnum. Þar

liggja verðmætar upplýsingar sem þó er oft erfitt og jafnvel ómögulegt að finna: Notandi verður að vita hvar hann á að leita og nákvæmlega hvernig hann þarf að orða fyrirspurn. Og jafnvel þó að fyrirspurn skili niðurstöðum þá hefur leitin að öllum líkindum farið fram hjá tengdum gögnum þar sem svipað efni er orðað á annan hátt. Sjálfvirk merkingargreining greinir fyrst fyrirspurn til þess að leita eftir merkingu frekar en eingöngu eftir leitarstreng. Snjöll leitarkerfi geta fundið tengdar upplýsingar hvort sem er í frjálsum texta eða í gagnagrunnum og þau búa einnig yfir einhvers konar þekkingargrunni (e. *knowledge base*) og/eða verufræðineti (e. *ontology*) þar sem þekking sem er mikilvæg fyrir fyrirtæki eða stofnun er vistuð á skipulegan hátt. Þannig geta snjöll leitarkerfi ekki einungis fundið einangraðar upplýsingar heldur einnig greint þær nánar, tengt saman og veitt verðmæta innsýn í gögn, oft sem hluti af viðskiptagreindarkerfi. Auk þess að vera augljóslega hagkvæm fyrir öll stærri fyrirtæki geta snjöll upplýsingakerfi til dæmis bætt þjónustu og ákvarðanatöku innan heilbrigðis- og dómskerfis, ásamt því að flýta fyrir ferlum þar sem minni tími fer í að tína saman nauðsynlegar upplýsingar.

Sjálfvirk merkingargreining greinir fyrst fyrirspurn til þess að leita eftir merkingu frekar en eingöngu eftir leitarstreng.

5.2.6 AUGNSTÝRÐ RITUN FYRIR ÍSLENSKU

Vorið 2017 gaf Microsoft út GazeSpeak, forrit sem gerir fólki með hreyfitaugungahrörnun (e. *ALS*) kleift að tala með augunum. Fólk með hreyfitaugungahrörnun er sumt í þeirri stöðu að hafa misst alla hreyfigetu og getur ekkert hreyft annað en augun. Hugbúnaðurinn gerir því kleift að velja orð úr listum með því að horfa upp, niður eða til hliðanna, þannig velur það upphafsstaf orðs eða flokk orða og fikrar sig svo nær orðinu sem það vill segja með frekari augnbendingum. Myndavél fylgist með augum þeirra og þannig getur tölvan fljótt fundið rétt orð úr orðalistum og með mállíkönum. Tölvusjón fylgist með notandanum og máltæknin nýtir tölvusjónina til að spá fyrir um rétt orð. Samkvæmt prófunum geta notendur náð að segja allt að 15 orð á mínútu með þessari tækni. Það yrðu líklega aldrei mjög margir sem myndu nota svona kerfi fyrir íslensku en þörfin væri engu að síður mjög brýn fyrir þá fáu sem þyrftu á því að halda.

5.3 ÞEKKINGARYFIRFÆRSLA

Uppbygging innviða og þróun á máltækni fyrir íslensku hefur marga kosti fyrir íslenskuna, samskipti fólks í landinu, viðskipti og aðgang að opinberri þjónustu eins og tíundað hefur verið í skýrslunni. Með framkvæmd

áætlunarinnar mun skapast mikil þekking og færni á sviðinu hérlendis sem nýtist í máltækni en erlendis sýnir reynslan að sérfræðimenntun og kunnátta í máltækni getur haft mjög góð áhrif á önnur svið í gegnum þekkingaryfirfærslu. Sérfræðingur í máltækni getur stundað tölfræðilega líkanagerð, merkjafræði, gagnagreiningu, þjálfun og beitingu djúptauganeta, hönnun á vitvélum og greiningu á kvikum kerfum svo eitthvað sé nefnt. Á sama tíma krefjast verkefni í máltækni oft fjölbreyttrar sérfræðiþekkingar. Þess vegna er þörf fyrir fólk með mismunandi menntun og þekkingu til þess að byggja upp öflugan máltækniiðnað. Langoftast er einnig hægt að beita þeirri tækni sem máltæknin nýtir sér á öðrum sviðum eins og til dæmis í lífvísindum, fjármálafræði og rekstrarverkfræði. Það er því ljóst að með fleiri starfstækifærum í máltækni skapast einnig tækifæri til að hafa jákvæð áhrif á aðrar greinar. Íslenskt atvinnulíf mun hagnast mjög á slíkri þekkingu.

Með fleiri starfstækifærum í máltækni skapast einnig tækifæri til að hafa jákvæð áhrif á aðrar greinar.

5.4 MÁLTÆKNI SEM ÚTFLUTNINGSVARA

Enn er mikil þörf á máltækniþekkingu í heiminum þrátt fyrir að mikið hafi áunnist og tækniframfarirnar hafi verið miklar undanfarið fyrir stærri málsvæði. Þessi þörf er af mörgum toga. Fyrir stóru málsvæðin á enn eftir að útfæra margar máltæknilausnir og tæknin nær ekki til allra, þar eru enn mörg tækifæri ónýtt. Þá á ennþá eftir að þróa máltæknilausnir fyrir mörg málsvæði með vaxandi tækniþarfir. Það er því af og frá að vettvangur þeirrar sérfræðiþekkingar sem skapast í áætluninni einskorðist við Ísland eða íslenskt tungumál. Sérfræðingar áætlunarinnar geta með ýmsum hætti látið til sín taka á erlendum vettvangi.

Beinn útflutningur á vörum eða þjónustu, sem byggð er á máltækni og tengdri tækni, er augljós leið fyrir not á íslenskri sérfræðiþekkingu sem myndast í máltækniáætlun fyrir íslensku 2018–2022. Nýsköpunarfyrirtæki gæti notað íslenska markaðinn sem tilraun áður en markaðsstarf hefst erlendis eins og algengt er með önnur íslensk nýsköpunarfyrirtæki. Þá er hægt að byggja á opnum og ókeypis innviðum fyrir íslenskuna til þess að sjá tæknilausnina virka áður en hafist er handa við að þróa málföng fyrir tungumál þeirra markaðssvæða sem stefnt er á.

Þátttaka í erlendum samstarfsverkefnum er einnig góð leið til þess að nýta þá sérfræðiþekkingu sem verður til í máltækniáætluninni. Verkefni sem til

dæmis eru skipulögð á vegum Evrópusambandsins eða Norðurlandaráðs eru kjörin tækifæri til þess að flytja út íslenska þekkingu í máltækni. Þá er alþjóðlegt samstarf um þróun á máltækni fyrir tungumál með rýr málföng* kjörinn vettvangur fyrir íslenska sérfræðinga til þess að leita nýrra leiða til að bæta íslenska máltækni, bera saman þróun á Íslandi og annars staðar og styðja við málsamfélög sem hafa minni möguleika á þróun máltækniinnviða.

* Ráðstefnan *Spoken Language Technologies for Under-resourced Languages* er haldin annað hvert ár og gengur út á að gera máltækni aðgengilega fyrir sem flest tungumál heimsins.

6

SKIPULAG ÁÆTLUNAR

6. SKIPULAG ÁÆTLUNAR

Mikilvægt er að samstarf stofnana, háskóla og atvinnulífs sé vel skilgreint. Við leggjum til að settur verði upp klasi með áhuga- og hagsmunaaðilum til að auðvelda samvinnu og samskipti á öllum stigum áætlunarinnar. Klasinn verður hýstur í miðstöð máltækniáætlunarinnar sem við leggjum til að verði hjá sjálfseignarstofnuninni Almannarómi og hún endurskipulögð og efld til að vera betur í stakk búin til að takast á við verkefnið. Sett verður á fót fagráð sem hefur það hlutverk að skipuleggja vinnu við innviði og hafa eftirlit með teymum sem vinna að þeim innviðaverkefnum sem skilgreind eru í þessari skýrslu. Gert er ráð fyrir að stór hluti innviðaverkefnanna fari fram á þeim stofnunum sem hafa áður unnið að máltækni og þar sem sérþekking er fyrir hendi en einnig verði kannaðir möguleikar á því að fá fleiri að borðinu þar sem því verður komið við. Þá verði stofnaður mótframlagssjóður með ákveðinni fjárveitingu á hverju ári áætlunarinnar til að gera fyrirtækjum og stofnunum mögulegt að hagnýta máltækni svo að tæknin sem þróuð er í áætluninni komist í notkun hratt og örugglega.

Þá verði stofnaður mótframlagssjóður með ákveðinni fjárveitingu á hverju ári áætlunarinnar til að gera fyrirtækjum og stofnunum mögulegt að hagnýta máltækni svo að tæknin sem þróuð er í áætluninni komist í notkun hratt og örugglega.

Tillögurnar eru byggðar á greiningu hópsins á sambærilegum áætlunum erlendis, viðtölum við sérfræðinga í máltækni á Íslandi, sérfræðinga eistnesku máltækniáætlunarinnar, sem staðið hefur yfir frá 2011 og lýkur í ár, viðtölum og bréfaskriftum við máltæknisérfræðinga í einkageiranum í Þýskalandi og Bretlandi og við sérfræðinga í vélrænum þýðingum hjá MT@EU. Enn fremur var rætt við fulltrúa CLARIN um mögulega þátttöku Íslands í því samstarfi og ávinning af því fyrir uppbyggingu máltækni hér á landi. Lýsing á þessum viðtölum og samskiptum má finna aftast í þessum kafla ásamt umfjöllun um máltækniáætlanir í öðrum löndum.

6.1 YFIRLIT

Tillaga hópsins um skipulag máltækniáætlunar fyrir íslensku 2018–2022 er að sett verði á laggirnar miðstöð fyrir áætlunina sem hefur það hlutverk að útfæra markmið hennar með þjónustusamningi við menntamálaráðuneytið. Sjálfseignarstofnunin Almannarómur var stofnuð árið 2013 til að standa að smíði máltæknilausna fyrir íslensku. Stofnaðilar voru á þriðja tug fyrirtækja, stofnana og félagasamtaka. Við leggjum til að hlutverk Almannaróms verði lagað að því að verða þessi miðstöð máltækniáætlunar. Aðalmarkmið miðstöðvarinnar er að sjá til þess að verkefni áætlunarinnar verði framkvæmd af þeim sérfræðingum, stofnunum og fyrirtækjum sem best þykja til þess fallin að vinna þau, sjá um samhæfingu milli verkefna og draga atvinnulífið

og aðra hagsmunaaðila að verkefnum þar sem því verður við komið. Þá skal miðstöðin leggja áherslu á að tryggja gott samstarf allra aðila innanlands og vinna að samstarfi við erlend fyrirtæki og stofnanir á sviðinu þannig að þeir innviðir og tækni sem þróuð verða í verkefninu komist í notkun. Almannarómur mun sjá um að skipa fagráð sem fer yfir og skipuleggur þau verkefni sem liggja fyrir á hverju ári og sér um faglegt eftirlit með starfinu.

6.1.1 FRAMKVÆMD KJARNAVERKEFNA

Innviðaverkefni sem tilgreind eru í 2. kafla skýrslunnar verða kostnaðargreind nánar og vörður og markmið skilgreind í ítarlegri verkefnalýsingum. Almannarómur auglýsir verkefnin og fagráð, skipað af Almannarómi, velur þann umsækjanda sem talinn er geta best valdið hverju verkefni fyrir sig. Almannarómur getur líka í einhverjum tilvikum leitað til ákveðinna aðila með verkefni ef fagráðið telur þá henta best. Verkefnastjóri, sem ber ábyrgð á verkefninu, verður skipaður og hann mun vera í samskiptum við Almannaróm um framgang þess. Mælt er með því að fyrir hvert kjarnaverkefni verði myndað sérstakt kjarnateymi sem tekur að sér að þróa ákveðna innviði út þann tíma sem áætlunin nær yfir. Þannig byggist upp þekking og reynsla í faginu og hæft fólk fæst til starfa til lengri tíma. Eitt öflugt teymi ætti að hafa yfirumsjón með þróun hvers innviðaverkefnis (talgreiningar, talgervingar, vélþýðinga og málrýni) en sökum fjölbreytni og umfangs verkefna á sviði málfanga þarf fleiri teymi í þann flokk. Gerð er grein fyrir tillögu um teymi í „Verkáætlunin í hnotskurn“, á bls. 22 í upphafi skýrslunnar.

Mælt er með því að fyrir hvert kjarnaverkefni verði myndað sérstakt kjarnateymi sem tekur að sér að þróa ákveðna innviði út þann tíma sem áætlunin nær yfir.

Þegar kemur að framkvæmdaaðilum sem sækja um að gera ákveðin verkefni er helst litið til Stofnunar Árna Magnússonar í íslenskum fræðum, Háskólans í Reykjavík og Háskóla Íslands. Þetta eru þær stofnanir sem leitt hafa þróun í máltækni og söfnun málfanga hingað til og því eðlilegt að fá þær inn í samstarfið og hafa með í ráðum frá upphafi áætlunarinnar. Sú þekking og reynsla sem þessar stofnanir búa yfir er mjög mikilvæg og lykill að því að góður árangur náist. Almannarómur skal jafnframt leitast við að fá aðra þátttakendur inn í verkefnið. Við Háskólann á Akureyri og í Nýsköpunarmiðstöð Íslands er að finna þekkingu sem gæti nýst við að safna málföngum og þróa máltækni en einnig krefjast mörg verkefnanna samstarfs við fyrirtæki og stofnanir sem búa yfir gögnum eða umsvifum sem nota má til að safna og útbúa málföng og prófa nýja tækni. Þar ber til dæmis að nefna Hljóðbókasafnið, Ríkisútvarpið, Dómstólaráð, 365 miðla og Creditinfo. Almannarómur getur liðkað fyrir samskiptum og auðveldað

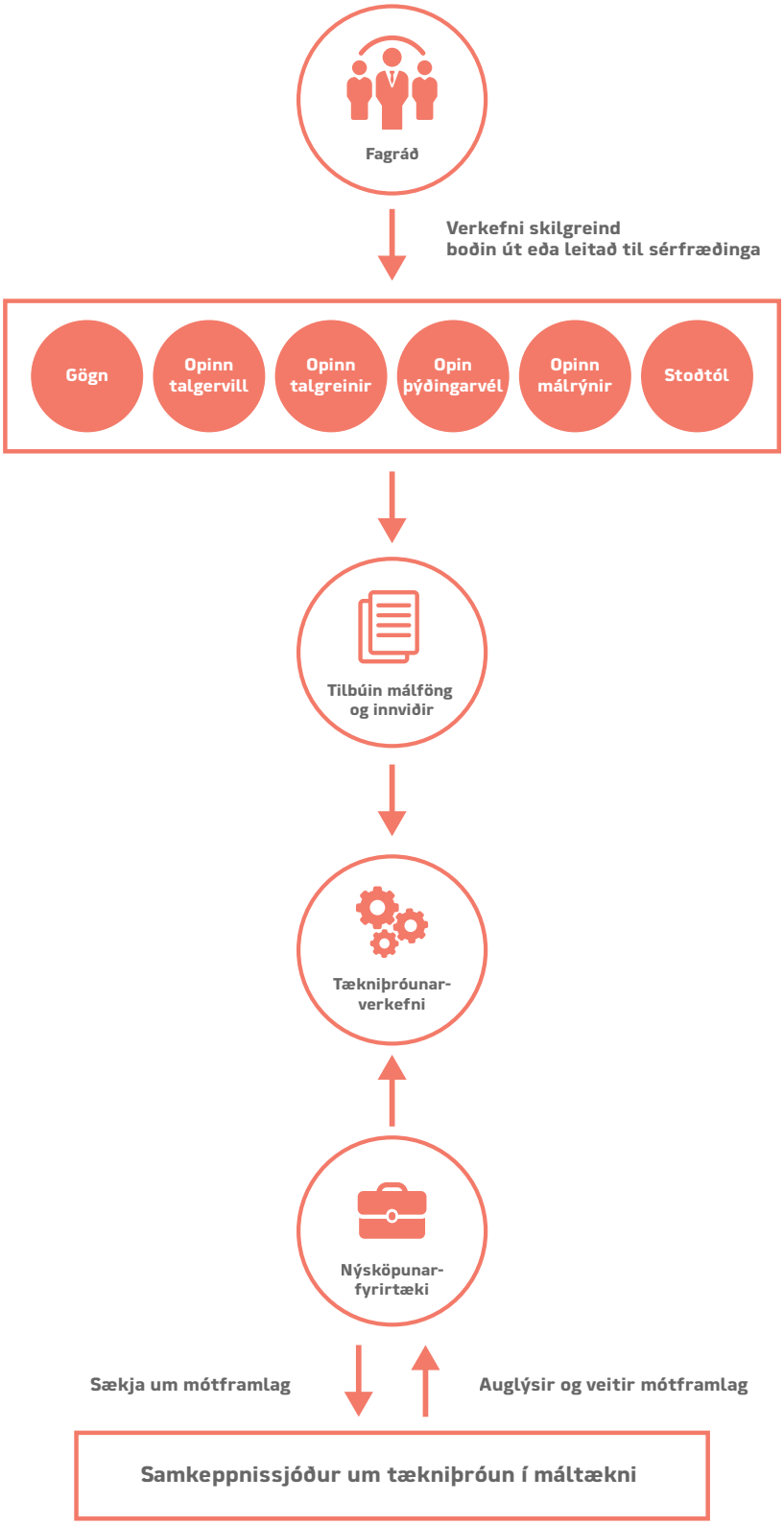
þátttöku með því að áætla og útvega það sem þarf í slíkt samstarf, til dæmis með sérfræðiþekkingu í leyfismálum gagna og tækniþekkingu.

6.1.2 FRAMKVÆMD HAGNÝTRA TÆKNIÞRÓUNARVERKEFNA

Lagt er til að fjármunum verði veitt í að styrkja nýsköpun og sprotafyrirtæki í máltækni í gegnum samkeppnissjóð um tækniþróun í máltækni. Farið væri fram á 50% mótframlag en verkefnin verða metin út frá viðskiptasjónarmiðum eins og tíðkast hjá Tækniþróunarsjóði og með það fyrir augum að koma máltækni í notkun fyrir íslensku. Þannig verður hægt að fá atvinnulífið með í að skapa máltækniiðnað í landinu. Frekari greiningar er þörf til að áætla hversu stór slíkur mótframlagssjóður fyrir máltækni þarf að vera en við gerum ráð fyrir að 50–200 m.kr. þurfi árlega, minnst fyrsta árið en meira eftir því sem líður á áætlunina.

Almannarómur sér til þess að aðilar í atvinnulífinu séu vel upplýstir um þá möguleika sem þeir innviðir sem verið er að þróa bjóða upp á.

Til þess að hvetja fyrirtæki og aðila í nýsköpun til þess að sækja um í sjóðinn mun Almannarómur halda reglulegar kynningar á þeim innviðaverkefnum sem eru í gangi hverju sinni og leiða saman fólk úr háskólum og stofnunum annars vegar og fólk úr atvinnulífinu hins vegar með það að markmiði að nýta þá innviði sem verið er að búa til í að útfæra hagnýt verkefni í máltækni. Þannig gætu nýsköpunarfyrirtæki til dæmis sótt um tækniþróunarstyrk með aðstoð eða í samvinnu við þá sem eru að þróa grunnhugbúnað og tækni. Almannarómur sæi þá um að mynda það mikilvæga tengslanet sem þarf til þess að koma slíku samstarfi af stað og sjá til þess að aðilar í atvinnulífinu séu vel upplýstir um þá möguleika sem þeir innviðir sem verið er að þróa bjóða upp á.



Framkvæmd innviðaverkefna er ákveðin og boðin út. Tækniþróunarverkefni verða til í grasrótinni hjá nýsköpunarfyrirtækjum

6.1.3 SAMVINNA VIÐ ÚTLÖND

Áríðandi er að þau opnu málföng sem verða til í áætluninni séu gerð aðgengileg þeim sem hanna og þróa máltæknilausnir erlendis.

Samskipti og samvinna í erlendum máltækniverkefnum er mjög mikilvæg fyrir áætlunina. Áríðandi er að þau opnu málföng sem verða til í áætluninni séu gerð aðgengileg þeim sem hanna og þróa máltæknilausnir erlendis. Hér er til dæmis átt við stórfyrirtæki á borð við Google, Apple, Microsoft, Amazon og Samsung en þessi fyrirtæki útbúa sína tækni með gerviþjónum (e. *virtual assistants*), t.d. Google Assistant, Siri, Cortana, Alexa og Bixby. Þó er um mun fleiri og fjölbreyttari hugbúnaðarlausnir að ræða á alþjóðavettvangi og mikilvægt er að grípa öll þau tækifæri sem bjóðast til að koma íslenskri máltækni að.

Ekki er augljóst hvernig við komum íslensku inn í tækni sem er í eigu erlendra fyrirtækja en helsti möguleikinn er að gera málföng og önnur verkfæri það opin og aðgengileg að hægt sé að fella íslenskuna á sem auðveldastan hátt inn í lausnirnar. Þá væri gagnlegt að nýta sömu staðla eða bjóða upp á gögnin og verkfærin með sömu stöðlum og notuð eru hjá þessum fyrirtækjum.

Þá er þátttaka í alþjóðlegum verkefnum einnig mikilvægur þáttur í því að íslenska verði hluti af stafrænum heimi framtíðarinnar. Evrópska CLARIN-verkefnið er samstarfsverkefni um málgögn. Þátttaka í því myndi veita okkur betri grundvöll til að koma okkar gögnum og verkfærum á framfæri með auðveldum hætti. Það myndi einnig tryggja varðveislu gagnanna og í CLARIN-samstarfið er hægt að sækja sérþekkingu sem ekki er fyrir hendi hér á landi. Tækifæri til að taka þátt í slíkum verkefnum ætti að nýta til hins ítrasta. Þá getur þekking, reynsla og tækni sem verður til í áætluninni einnig nýst öðrum málsvæðum með rýr málföng (e. *under-resourced languages*).

6.2 MIÐSTÖÐ MÁLTÆKNIÁÆTLUNAR

Við leggjum til eins og fyrr segir að Almannarómur verði gerður að miðstöð máltækniáætlunar 2018–2022 og skipulag Almannaróms verði lagað að því hlutverki. Helstu hlutverk miðstöðvar máltækniáætlunar eru eftirfarandi:

- Að tryggja virkt samstarf og samskipti þeirra sem vinna að máltækni á Íslandi og þeirra sem hug hafa á því.
- Að forgangsraða og skipuleggja vinnu við innviðaverkefni innan áætlunarinnar með aðstoð fagráðs.
- Að fylgjast með vinnu við verkefni sem eru í gangi og skipuleggja kynningar á verkefnum.
- Að veita þeim aðstoð og ráðgjöf sem hafa hug á að fara í tækniþróunarverkefni sem nota máltækni.
- Að kynna möguleika máltækni fyrir fyrirtækjum og stofnunum sem gætu hagnýtt sér máltæknilausnir í sínum rekstri.
- Að koma á samstarfi við erlend fyrirtæki sem þróa máltæknilausnir um að nota íslensku í þeim lausnum.
- Að fylgjast með möguleikum á fjölþjóðlegu þróunarsamstarfi í máltækni og kanna möguleika á því að íslenska verði með í slíkum verkefnum.
- Að kynna áætlunina og afurðir hennar á alþjóðlegum vettvangi.

Ljóst er að miðstöð eins og þessi þarf fastan starfsmann og skrifstofu. Því þarf að gera ráð fyrir fjárveitingum til rekstrar Almannaróms í fjárhagsáætlunum verkefnisins.

6.2.1 FAGRÁÐ

Fyrir innviða- og nýsköpunarverkefni þarf að stofna fagráð skipað sérfræðingum og öðrum fulltrúum þeirra sem koma að áætluninni. Fagráðin eru ráðgefandi fyrir úthlutanir og stýra úthlutunum verkefna. Mikilvægt er að þau komi saman þegar þörf er á og taki ákvarðanir á sem skemmstum tíma til þess að ekki verði óþarfa hlé á þróunarvinnu.

6.2.2 KLASI INNLENDRA ÞÁTTTAKENDA

Fyrir utan miðstöð máltækniáætlunar þyrfti að koma á breiðari samstarfsvettvangi allra þeirra sem vinna að máltækni, íslenskum máltækniklasa. Hann yrði vettvangur fyrir þróunarsamstarf, kynningarstarf og hugmyndavinnu. Slíkur klasi gæti einnig komið að menntun í máltækni í samstarfi við háskólana.

6.2.3 ÍSLENSKA Í ALLAN TÆKJABÚNAÐ

Hugbúnaður erlendra stórfyrirtækja á borð við Google, Apple og Microsoft er mikið notaður hér á landi. Sökum smæðar íslenska markaðarins sjá þessi fyrirtæki sér yfirleitt ekki hag í að útfæra máltæknilausnir fyrir íslensku. Það er því mikilvægt að leggja áherslu á að vera í góðum samskiptum við erlend hugbúnaðarfyrirtæki og leggja kapp á að útvega þau íslensku málföng sem þarf til þess að koma íslenskunni inn sem víðast. Miðstöð máltækniáætlunar þarf þá að miðla kröfum um frágang og uppsetningu málfanga til þeirra sem vinna að þróun þeirra svo að hægt sé að búa þannig um gögnin að þau nýtist öllum.

Það er því mikilvægt að leggja áherslu á að vera í góðum samskiptum við erlend hugbúnaðarfyrirtæki og leggja kapp á að útvega þau íslensku málföng sem þarf til þess að koma íslenskunni inn sem víðast.

6.2.4 ÞÁTTTAKA Í ERLENDUM VERKEFNUM

Miðstöð máltækniáætlunar kynnir sér og fylgist með erlendum máltækniverkefnum sem íslensku gæti verið hagur að. Leita skal eftir að taka þátt í slíkum verkefnum eins og mögulegt er. Þá skal miðstöðin kynna áætlunina og afurðir hennar á alþjóðlegum vettvangi og fylgjast með þeirri þróun sem á sér stað svo ákvörðunartaka í og í kringum áætlunina sé sem best.

6.3 VIÐHALD OG VARÐVEISLA MÁLFANGA

Í öllum gagna- og kjarnaverkefnum máltækniáætlunarinnar verða til gagnasöfn eða tæki með opnum leyfum og ætluð öllum til afnota. Tryggja þarf nákvæma skráningu, viðhald, varðveislu og aðgengi að þessum afurðum til framtíðar. Til að það sé unnt þurfa innviðir að vera fyrir hendi í stofnunum sem líklegar eru til að standa af sér mögulegt umrót sem framtíðin ber í skauti sér. Í nágrannalöndum okkar, á öllum Norðurlöndum og í 15 öðrum Evrópulöndum hafa slíkar stofnanir gengið inn í CLARIN-samstarfið. CLARIN stendur fyrir *Common Language Resources and Technology Infrastructure*. Meginmarkmið þessa verkefnis er að öll stafræn málföng, það er gagnasöfn um tungumál og önnur málsöfn og míllegar heimildir alls staðar að úr Evrópu, verði aðgengileg í gegnum einn sameiginlegan netaðgang til rannsókna í hug- og félagsvísindum og til tækniþróunar. Við leggjum til að Ísland sæki um aðild að CLARIN-samstarfinu og nýti sér þá innviðauppbyggingu sem þar hefur átt sér stað.

Tryggja þarf nákvæma skráningu, viðhald, varðveislu og aðgengi að þessum afurðum til framtíðar.

6.3.1 CLARIN

Síðan CLARIN var sett á laggirnar árið 2012 hefur verið unnið að því innan þess að byggja upp innviði til að styðja við uppbyggingu, skráningu, viðhald, varðveislu, notkun og samnýtingu míllegra gagna og búnaðar til rannsókna á þess konar gögnum í hug- og félagsvísindum. Lögð er áhersla á að skráning lýsigagna sé ítarleg og nákvæm svo að öll leit sé eins auðveld og hægt er og til að öll rétt gögn finnist þegar leitað er að gögnum í einhverjum ákveðnum tilgangi. Það flýtir fyrir allri vinnu og gerir það mögulegt að finna viðamikil endurnýtanleg gögn fyrir rannsóknir og þróun.

Með þessum innviðum veitir CLARIN aðgang með einföldum og varanlegum hætti að stafrænum málgögnum (textum, hljóði og mynd) sem vísindamenn geta nýtt sér á hvaða sviði sem þeir starfa. CLARIN býður upp á þróuð verkfæri til að skoða, greina og vinna með slík gagnasöfn, hvar sem þau eru staðsett. Það er gert í gegnum net CLARIN-miðstöðva, með einum netaðgangi fyrir allt vísindasamfélagið í þátttökulöndunum. Hugbúnaður og gögn frá mismunandi stöðum eru aðgengileg öllum CLARIN-miðstöðvum og eru samræmd.

6. SKIPULAG ÁÆTLUNAR

CLARIN-miðstöðvarnar eru reknar af samtökum (e. *consortium*) stofnana sem hag hafa af samstarfinu. Yfirleitt er einn slíkur hópur í hverju landi. Stofnanir sem taka þátt eru háskólar, bókasöfn, rannsóknarstofnanir og aðrar stofnanir sem vinna með málleg gögn. CLARIN-miðstöð yrði sett upp hjá einni af stofnunum í samtökunum.

Innan CLARIN er hægt að hýsa gögn og veita aðgang að þeim svo að þau nýtist sem best. CLARIN mælir með því að notkunarleyfi á gögnum og hugbúnaði séu sem opnust en þó er heimilt að setja strangari skorður á notkun ef ástæða þykir til þess.

Með þátttöku í CLARIN fengju þeir sem vinna að máltækni á Íslandi aðgang að ótal verkfærum og gagnasöfnum til að nýta í rannsóknum og tækniþróun.

Með þátttöku í CLARIN fengju þeir sem vinna að máltækni á Íslandi aðgang að ótal verkfærum og gagnasöfnum til að nýta í rannsóknum og tækniþróun. Þar að auki býður CLARIN upp á að ný gögn og tól séu hýst til frambúðar. Öll gagnasöfn fá einkvæmt númer, svokallað ISLRN-númer. Það gagnast við vísanir í gagnasöfn þegar þau eru notuð í rannsóknum eða þróun og með þeim er hægt að tryggja að verið sé að vinna með sömu gögn þegar þarf að endurtaka tilraunir eða þróa nýjar og betri aðferðir til að vinna með málleg gögn. Mikilvægur ávinningur af CLARIN fyrir íslenska máltækni tengist einmitt þessu.

Máltæknigögn og máltækniverkfæri þarf að vera hægt að geyma á stað þar sem auðvelt er að finna þau og veita sem flestum tækifæri á að nýta þau, innanlands sem utan. Það þarf að vera hægt að treysta því að gögnin séu hýst og gerð aðgengileg til frambúðar, óháð hugsanlegum breytingum á tæknilegu umhverfi. Auk þess getur sú þekkingarmiðlun sem lögð er áhersla á innan CLARIN verið afar mikilvæg fyrir íslenska máltækni þar sem sérfræðingar á sviði máltækni eru fremur fáir hér á landi. Þátttaka í CLARIN gæti því verið afar mikilvæg stoð fyrir máltækniáætlunina.

Þátttaka í CLARIN ætti þó ekki að vera háð máltækniáætluninni sjálfri. Máltækniáætlunin er til ákveðins tíma, líklega 5 ára, en þátttaka í CLARIN þyrfti að halda áfram að þeim tíma liðnum, einmitt til að tryggja að afurðir áætlunarinnar verði áfram aðgengilegar sem og önnur gögn sem tengjast ekki endilega máltækni. CLARIN hefur víðari skírskotun en aðeins til máltækni, til dæmis gæti samstarfsnetið gagnast vel áætlun um varðveislu menningararfleifðar á stafrænu formi, sem menntamálaráðuneytið vinnur nú að, og afrakstur þeirrar vinnu nýst betur, innan og utan landsteinanna.

Kostnaður við þátttöku í CLARIN er fyrst og fremst bundinn við þátttökugjöld og við rekstur CLARIN-miðstöðvar. Þátttökugjaldið fyrir Ísland ræðst af ákvörðun um kostnað fyrir minni ríki og er tengt hlutfalli af lands-

framleiðslu ríkja Evrópusambandsins. Gjald fyrir árið 2017 hefði verið 13.028 evrur en þátttökugjaldið hækkar um 2% á ári og er því þekkt langt fram í tímann.

Í flestum tilvikum eru CLARIN-miðstöðvar reknar innan einhverrar af þeim stofnunum sem mynda CLARIN-samtökin í hverju landi. Þetta geta t.d. verið háskólar eða rannsóknastofnanir. Sjá lista hér: https://www.clarin.eu/content/participating-consortia.

Gera má ráð fyrir að bæta þyrfti við a.m.k. einu stöðugildi hjá þeirri stofnun sem yrði CLARIN-miðstöð hér á landi. Það væri um hálft stöðugildi í tækniumsjón og hálft stöðugildi við umsýslu og upplýsingagjöf. Fyrstu 1–2 árin, þegar verið væri að setja allt upp og koma verkefninu af stað hér á landi, þyrfti þó að gera ráð fyrir fleirum, líklega tveimur stöðugildum í 18 mánuði.

Auk launakostnaðar og kostnaðar við aðstöðu starfsmanna þarf að gera ráð fyrir rekstrarkostnaði tölvubúnaðar og ferðakostnaði fyrir einn starfsmann einu sinni á ári á aðalfund CLARIN.

Kostnað á ári má því gróflega áætla svona:

Þátttökugjöld:	1.500.000 kr.
Rekstur tölvubúnaðar:	um 500.000 kr.
Ferðakostnaður:	um 500.000 kr.

Þá þarf að gera ráð fyrir tveimur stöðugildum í 18 mánuði og einu stöðugildi eftir það. Sú stofnun sem starfrækir CLARIN-miðstöð þarf að fá fjárveitingar til að standa undir því.

Heildarkostnaður:	2,5 milljónir króna + launakostnaður

6.4 AÐRAR MÁLTÆKNIÁÆTLANIR

Á alþjóðavettvangi hefur undanfarin ár mikið verið rætt um að framtíð tungumála velti á því að hægt sé að nýta þau í stafrænum heimi. Því hefur máltækniáætlunum verið hleypt af stokkunum fyrir allmörg tungumál. Vinnuhópurinn kynnti sér sérstaklega hollensk-flæmsku áætlunina STEVIN (*Essential Speech and Language Technology Resources*) sem var í gangi á árunum 2004–2011, spænska máltækniáætlun, sem nú er í gangi, og tvær áætlanir sem gerðar hafa verið fyrir eistnesku.

Holland

STEVIN-áætlunin var gerð til þess að festa hollensku og flæmsku í sessi á sviði máltækni. Markmið áætlunarinnar voru að auka almenna vitund um máltækni, sérstaklega innan atvinnulífsins, að standa fyrir rannsóknum og þróun málfanga til þess að fylla upp í göt í innviðum fyrir máltækni og að skipuleggja stjórnun, viðhald og dreifingu á málföngum. Verkefni innan áætlunarinnar voru ýmist rannsóknar- og þróunarverkefni, kynningarverkefni (e. *demos*) eða verkefni tengd menntun. Má hér nefna söfnun á talgögnum frá börnum, fólki með hollensku sem annað tungumál og eldra fólki, verkefni tengd merkingargreiningu, samhliða textum og málheildum, sérstök ritstoð fyrir lesblind börn, leitarvél fyrir dómstóla með talgreiningu réttarhalda og margt fleira. Stjórn var mynduð með fulltrúum þeirra stofnana sem stóðu að áætluninni ásamt sérfræðingum í máltækni. Hún hafði yfirumsjón með áætluninni og tók ákvarðanir um styrki. Hollenska tungumálastofnunin (Nederlandse Taalunie) sá um samhæfingu og fjármálastjórn en sérstök nefnd sérfræðinga, sem settu áætlunina fram, sáu um að stefnunni væri framfylgt, auglýstu eftir styrkumsóknum og ráðlögðu stjórninni við styrkjaákvarðanir. Alþjóðlegur ráðgjafahópur fór þó fyrst yfir allar styrkumsóknir. Sérstök skrifstofa máltækniáætlunarinnar var rekin af tveimur stofnunum og sá hún um verkefnastjórnun.

Til þess að auka vitund um máltækni í atvinnulífinu var stór ráðstefna skipulögð þar sem máltæknisérfræðingum og -fyrirtækjum var boðið að kynna máltæknilausnir á mörgum sviðum.

Til þess að auka vitund um máltækni í atvinnulífinu var stór ráðstefna skipulögð þar sem máltæknisérfræðingum og -fyrirtækjum var boðið að kynna máltæknilausnir á mörgum sviðum. Lögð var áhersla á að hafa sýninguna á sem breiðustum grunni og notkunarmöguleikar fyrir fjölmiðla, menntakerfið, heilbrigðisstofnanir, samgöngur, ferðaþjónustu, stjórnsýslu, fjarskipti og fjármálakerfi kynnt.

Spánn

Spænska máltækniáætlunin *Plan de Impulso de las Tecnologías del Lenguaje* stendur nú yfir. Hún hófst árið 2016 og lýkur 2020. Hvatar hennar eru m.a. þeir að tryggja samkeppnishæfni Spánar og Suður-Ameríku og að forða öðrum opinberum tungumálum á Spáni, fyrir utan spænsku, frá stafrænum dauða. Tungumálin í áætluninni eru spænska, katalónska, galisíska, baskneska og oksítanska. Markmiðin eru að fjölga málföngum fyrir þessi tungumál og auka gæði og aðgengi að þeim, að leggja áherslu á tækniyfirfærslu frá rannsóknum út í atvinnulífið og að auka gæði og skilvirkni opinberra stofnana með því að innleiða vélþýðingar og aðra máltækni. Spænska áætlunin leggur einnig áherslu á sýnileika máltækni, bæði hjá fyrirtækjum og nemendum, og hvetur þannig til þess að yfirfærsla tækni til atvinnulífsins fari fram og að hæfileikaríkt fólk sérhæfi sig í máltækni og tengdum fögum.

Stjórnsýslan á að vera leiðandi aðili í máltækni en innan heilbrigðiskerfis, dómskerfis, menntakerfis og ferðaþjónustu á að þróa sýnileg verkefni. Áætlaður heildarkostnaður áætlunarinnar eru 90 milljónir evra, jafngildi um 10 milljarða íslenskra króna. Það gerir að jafnaði um 2 milljarða króna á hvert tungumál.

Stjórnsýslan á að vera leiðandi aðili í máltækni en innan heilbrigðiskerfis, dómskerfis, menntakerfis og ferðaþjónustu á að þróa sýnileg verkefni.

Eistland

Vinnuhópurinn sat fundi með fulltrúum eistnesku máltækniáætlunarinnar sem staðið hefur yfir síðan 2011 og lýkur á þessu ári. Eistarnir miðluðu af reynslu sinni og svöruðu spurningum um tilhögun áætlunarinnar. Tvær máltækniáætlanir hafa verið gerðar í Eistlandi. Sú fyrri *National Programme for Estonian Language Technology 2006–2010* og sú seinni *National Programme for Estonian Language Technology 2011–2017*. Það er mennta- og rannsóknamálaráðuneyti Eistlands sem er ábyrgt fyrir áætlununum, en háskólarnir í Tallinn og Tartu ásamt Stofnun eistneskrar tungu (Eesti Keele Instituut) eru leiðandi við framkvæmd þeirra. Eistneska og umhverfi hennar er um margt líkt íslensku: Of fáir tala tungumálið til þess að fyrirtæki sjái sér hag í að leggja í kostnaðarsama þróunarvinnu í máltækni og samfélagið er tæknivætt, þ.e. fólk notar eða vill geta notað hugbúnað sem styðst við máltækni. Vilji er fyrir því að eistneskan haldi stöðu sinni og síðast en ekki síst er tungumálið, líkt og íslenska, með flókið beygingarkerfi og mjög virka orðmyndun, svo að vandamálin sem Eistar standa frammi fyrir í textagreiningu eru ekki ólík þeim sem við Íslendingar stöndum frammi fyrir.

6. SKIPULAG ÁÆTLUNAR

Innan eistnesku máltækniáætlananna tveggja hafa verið unnin verkefni á sviði talgervingar, talgreiningar, orðasafna og málheilda, stafsetningarleiðréttingar og vélþýðinga. Ágætur árangur hefur náðst við þróun innviða og hugbúnaður er nú þegar í notkun sem nýtir innviðatól. Má þar nefna sjálfvirkan upplestur á hljóðbókum og textuðu sjónvarpsefni, vefviðmót sem býður upp á að senda inn hljóðskrár og fá texta frá talgreini með tölvupósti, og umfangsmikið opið safn orðabóka. Stafsetningarleiðréttingaforrit fyrir eistnesku er hluti af MS Office og góður árangur hefur einnig náðst í vélþýðingum. Hægt er að prófa þýðingarvél í gegnum vefviðmót og bera saman við niðurstöður Google Translate. Markmið vélþýðingahópsins er þó fyrst og fremst að þróa sérhæfðar þýðingarvélar.

Kjarnaverkefni eistnesku áætlunarinnar ganga vel og eru að hluta komin í notkun í almennum hugbúnaði.

Kjarnaverkefni eistnesku áætlunarinnar ganga vel og eru að hluta komin í notkun í almennum hugbúnaði. Mikilvægur lærdómur sem draga má af áætlunum í Eistlandi er þó að of seint og of ómarkvisst hafi verið unnið að tengingu við atvinnulífið. Gert var ráð fyrir að áhugi væri ekki mikill og reglugerðir sem hindra beinan stuðning við fyrirtæki standa líka í veginum. Þó eru að minnsta kosti tvö öflug máltæknifyrirtæki í Eistlandi, Tilde-Eestim og Filosoft. Einnig töluðu forsvarsmenn áætlunarinnar um að líklega væri betra að hafa ákveðinn aðila, t.d. félag með eigin skrifstofu og „kennitölu“ til þess að halda utan um framkvæmd áætlunarinnar í stað þess að fela starfsmanni annarrar stofnunar að sjá um það, en í eistnesku áætluninni var starfsmaður áætlunarinnar einnig starfsmaður Háskólans í Tartu. Á sama stað er CLARIN-miðstöðin í Eistlandi, og sögðu þau hana gegna mikilvægu hlutverki við að halda utan um málföng.

LOKAORÐ

0101011101 arðmiði geisp griðungur 1010111 Skrúður 0101110
Róði gauskur 0101011101 glundroði 0101
101 arðmiði geisp griðungur 1010111 Skrúður

Í skýrslunni er fjallað um þau tækifæri sem felast í þróun máltækni fyrir íslensku. Nauðsynlegt er að koma af stað skipulegri uppbyggingu innviða, mynda þekkingarkjarna og skapa öflugt nýsköpunarumhverfi fyrir máltæknilausnir. Í ljósi þeirrar byltingar sem á sér stað um þessar mundir á sviði gervigreindar og máltækni er mikilvægt að máltækniáætlun fyrir íslensku komi til framkvæmda svo fljótt sem auðið er. Þannig getum við verið þátttakendur í þróuninni og fengið tækifæri til að nota okkar tungumál, íslenskuna, í tækni framtíðarinnar.

HEIMILDASKRÁ

UM ÍSLENSKU Á TÖLVUÖLD

Eiríkur Rögnvaldsson, Sigrún Helgadóttir og Hrafn Loftsson. 2015. Nefnd um notkun íslensku í stafrænni upplýsingatækni. Skýrsla unnin fyrir mennta- og menningarmálaráðherra.

Eiríkur Rögnvaldsson, Kristín M. Jóhannsdóttir, Sigrún Helgadóttir og Steinþór Steingrímsson. 2012. Íslensk tunga á stafrænni öld. Meta-Net hvítbókaröð. Springer.

Eiríkur Rögnvaldsson, Haraldur Bernharðsson, Sigrún Helgadóttir, Björgvin Ívar Guðbrandsson, Jóna Pálsdóttir og Sigurbjörg Jóhannesdóttir. 2012. Íslenska í tölvuheiminum. Mennta- og menningarmálaráðuneytið.

Íslenska til alls. Tillögur Íslenskrar málnefndar að íslenskri málstefnu. 2008. Menntamálaráðuneytið.

Tungutækni. Skýrsla starfshóps. 1999. Menntamálaráðuneytið.

AÐRAR MÁLTÆKNIÁÆTLANIR OG MÁLTÆKNISKÝRSLUR

National programme Estonian Language Technology 2006-2010. 2007. Mennta- og rannsóknamálaráðuneyti Eistlands.

National programme Estonian Language Technology 2011-2017. 2011. Mennta- og rannsóknamálaráðuneyti Eistlands.

Liin, K., Muischnek, K., Müürisep, K., og Vider, K. 2012. Eesti keel digiajastul – The Estonian Language in the Digital Age. Meta-Net hvítbókaröð. Springer.

Spyns, Peter og Elisabeth D'Halleweyn. 2012. The STEVIN Programme: Result of 5 years cross-border HLT for Dutch Policy Preparation. Í Peter Spyns og Jan Odijk (Hrsg.): Essential Speech and Language Technology for Dutch. Theory and Applications of Natural Language Processing. bls. 21-39.

Plan for the Advancement of Language Technology (Spænska máltækniáætlunin). 2015. Agenda Digital para España. Madrid, Spánn.

Rafel Rivera Pastor o.fl. 2017. Language equality in the digital age - Towards a Human Language Project. Scientific Foresight Unit (STOA).

Language Technologies. 2013. LT2013: Status and potential of the European language technology markets. LT-Innovate.

FRÆÐILEGT EFNI

Anton Karl Ingason, Skúli B. Jóhannsson, Eiríkur Rögnvaldsson, Hrafn Loftsson og Sigrún Helgadóttir. 2009. Context-Sensitive Spelling Correction and Rich Morphology. Proceedings of the 17th Nordic Conference of Computational Linguistics, NODALIDA. s. 231-234.

Carlberger, J., R. Domeij, V. Kann og O. Knutsson. 2004. The Development and Performance of a Grammar Checker for Swedish: A Language Engineering Perspective.

DePalma, Donald A., Vijayalaxmi Hegde, Hélène Pielmeier, Robert G. Stewart og Stephen Henderson. 2016. The Language Services Market: 2016. Common Sense Advisory, Boston, USA.

Edlund, Jens, C. Tånnander og J. Gustafson. 2015. Audience response system-based assessment for analysis-by-synthesis, Proceedings of ICPhS.

Etchegoyhen, Thierry, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard Van Loenhout, Arantza Del Pozo, Mirjam Sepesy Maucec, Anja Turner, og Martin Volk. 2014. Machine Translation for Subtitling: A Large-Scale Evaluation. Í Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.

Helfrich, Antje og Bradley Music. 2000. Design and Evaluation of Grammar Checkers in Multiple Languages. Proceedings of the 18th conference on Computational linguistics, Vol. 2, bls. 1036-1040.

Ingibjörg Elsa Björnsdóttir. 2016. Vélþýðingar á íslensku og Apertium-þýðingarkerfið. Orð og tunga 18, Reykjavík.

Jón Friðrik Daðason. 2012. Post-Correction of Icelandic OCR Text. Háskóli Íslands, Meistararitgerð.

Jurafsky, Dan og James H. Martin: Speech and Language Processing. Uppkast að þriðju útgáfu, 2017. https://web.stanford.edu/~jurafsky/slp3/ sótt 25.04.2017

Klein, G., Y. Kim, Y. Deng, J. Senellart og A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. ArXiv e-prints.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin og Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), Prag, Tékkland.

Kukich, Karen. 1992. Techniques for automatically correcting words in text. ACM 24(4):377-439.

Loftsson, Hrafn, Ida Kramarczyk, Sigrún Helgadóttir, og Eiríkur Rögnvaldsson. 2009. „Improving the PoS Accuracy of Icelandic Text.“ Í: Jokinen, Kristiina og Eckhard Bick (eds.): Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009, pp. 103-110. NEALT Proceeding Series 4. Northern European Association for Language Technology (NEALT), Tartu University Library.

Sennrich, Rico, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel L”aubli, Antonio Valerio Miceli Barone, Jozef Mokry og Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine Translation. In Proceedings of the Demonstrations at the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spánn.

Strömbergsson, Sofia, C. Tånnander og J. Edlund. 2014. Ranking severity of speech errors by their phonological impact in context. Interspeech, 1568-1572.

Taylor, P., Black, A., og Caley, R. 1998. The architecture of the Festival Speech Synthesis System. Proc. 3rd ESCA Workshop on Speech Synthesis, bls. 147-151, Jenolan Caves, Australia.

Tihanyi, László, Csaba Oravecz. 2017. First Experiments and Results in English–Hungarian Neural Machine Translation. XIII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Ungverjaland.

Tånnander, Christina. 2012. An audience response system-based approach to speech synthesis evaluation. Í The Fourth Swedish Language Technology Conference (SLTC 2012), bls. 74-75. Lund, Svíþjóð.

van den Oord, Aaron , Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, og Kavukcuoglu, Koray. 2016. Wavenet: A generative model for raw audio. ArXiv.org:1609.03499.

Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy (2005). Parallel corpora for medium density languages. Í Proceedings of the RANLP 2005, bls. 590-596.

Whitelaw, C., B. Hutchinson, G. Chung o.fl. 2009. Using the Web for Language Independent Spellchecking and Autocorrection. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, s. 890-899.

Wu, O. Watts, og S. King. 2016. Merlin: An open source neural network speech synthesis system," í 9th ISCA Speech Synthesis Workshop (SSW9), Sunnyvale, CA, USA.

Zhang, X., H. Kulkarni og M. Morris. 2017. Smartphone-Based Gaze Gesture Communication for People with Motor Disabilities. CHI 2017.

8.1 TENGLAR

http://malfong.is/

http://malid.is/

http://bin.arnastofnun.is/

https://greynir.is/

http://nlp.cs.ru.is/icenlp/

http://puki.is/

http://skrambi.arnastofnun.is/

http://www.epc.de/

http://hunspell.github.io/

https://www.clarin.eu