META-NET

White Paper Series Hvítbókaröð

THE ICELANDIC LANGUAGE IN THE DIGITAL AGE

ÍSLENSK TUNGA Á STAFRÆNNI ÖLD

Eiríkur Rögnvaldsson
Kristín M. Jóhannsdóttir
Sigrún Helgadóttir
Steinþór Steingrímsson

Springer

White Paper Series Hvítbókaröð

THE ICELANDIC LANGUAGE IN THE DIGITAL AGE

ÍSLENSK TUNGA Á STAFRÆNNI ÖLD

Eiríkur Rögnvaldsson Háskóla Íslands
Kristín M. Jóhannsdóttir Háskóla Íslands
Sigrún Helgadóttir Árnastofnun
Steinþór Steingrímsson Háskóla Íslands

Georg Rehm, Hans Uszkoreit
(ritstjórar, editors)

FORMÁLI PREFACE

Þessi hvítbók er hluti af ritröð til kynningar á máltækni og möguleikum hennar. Henni er einkum beint til fólks sem starfar í menntageiranum, á fjölmiðlum, í stjórnmálum – og í raun til málsamfélagsins í heild. Aðgengi að máltækni og notkun hennar er mjög mismunandi milli tungumála í Evrópu. Þar af leiðir að aðgerðir sem nauðsynlegar eru til að styðja rannsóknir og þróunarstarf í máltækni eru einnig ólíkar milli mála. Ýmsir þættir hafa áhrif á það hvaða aðgerða er þörf, svo sem stærð málsamfélagsins og hversu flókið tungumálið er.

Á vegum META-NET, sem er öndvegisnet fjármagnað af Evrópusambandinu, hefur verið lagt mat á núverandi stöðu í málföngum og máltækni (sjá bls. 73). Þessi greining tók til hinna 23 opinberu mála Evrópusambandsins auk annarra mikilvægra þjóðtungna og svæðisbundinna tungumála í álfunni. Niðurstöður þessarar greiningar benda til að í öllum málunum skorti rannsóknir á mikilvægum sviðum. Nákvæmari greining sérfræðinga og mat á núverandi stöðu mun hjálpa til við að hámarka árangur viðbótarrannsókna og lágmarka áhættu.

META-NET tengir saman 54 rannsóknarsetur í 33 löndum (í nóvember 2011, sjá bls. 69). Þau vinna með hagsmunaaðilum úr viðskiptalífinu (hugbúnaðarfyrirtækjum, tæknifyrirtækjum og notendum), frá opinberum stofnunum, rannsóknarstofnunum, sjálfstæðum félagasamtökum, fulltrúum málsamfélaga og evrópskum háskólum. Í samstarfi við þessa aðila vinnur META-NET að þróun heildstæðrar tæknisýnar og útfærðri rannsóknarstefnu handa margmála Evrópu árið 2020.

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others. The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differs. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies in this white paper series (p. 73). The analysis focused on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of additional research.

As of November 2011, META-NET consists of 54 research centres from 33 European countries (p. 69). META-NET is working with stakeholders from economy (software companies, technology providers, users), government agencies, research organisations, non-governmental organisations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

Höfundar þessa rits þakka höfundum hvítbókar um þýsku fyrir leyfi til að endurnýta almenna kafla úr verki þeirra [1].

Gerð þessarar hvítbókar var kostuð af Sjöundu rammaáætlun Evrópusambandsins og Stefnumótunaráætlun Evrópusambandsins í upplýsinga- og samskiptatækni samkvæmt samningum við T4ME (styrksamningur 249119), CESAR (styrksamningur 271022), METANET4U (styrksamningur 270893) og META-NORD (styrksamningur 270899).

The authors of this document are grateful to the authors of the White Paper on German for permission to re-use selected language-independent materials from their document [1].

The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under the contracts T4ME (Grant Agreement 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899).

EFNISYFIRLIT CONTENTS

ÍSLENSK TUNGA Á STAFRÆNNI ÖLD

THE ICELANDIC LANGUAGE IN THE DIGITAL AGE

1

YFIRLIT

Upplýsingatæknin hefur breytt hversdagslífi okkar. Við notum tölvur til að skrifa og vinna með texta, reikna, leita upplýsinga, og sífellt meira einnig til að lesa, hlusta á tónlist, skoða myndir og horfa á kvikmyndir. Við göngum með snjallsíma og spjaldtölvur á okkur og notum til að hringja, senda tölvupóst, afla okkur upplýsinga og stytta okkur stundir, hvar sem við erum stödd. Hvaða áhrif hefur þessi víðtæka stafræna bylting í upplýsingum, þekkingu og hversdagssamskiptum á tungumál okkar? Mun það breytast eða jafnvel deyja út? Hvaða möguleika hefur íslenska á að lifa af?

Mörg hinna 6.000 tungumála heimsins munu ekki lifa af í hinu hnattræna stafræna upplýsingasamfélagi. Talið er að a.m.k. 2.000 tungumál deyi út á næstu áratugum. Önnur munu lifa af inni á heimilum og í daglegum samskiptum, en ekki verða notuð í viðskiptalífinu eða vísindum og fræðum. Staða tungumálsins ræðst ekki bara af fjölda málnotenda, eða fjölda bóka, kvikmynda og sjónvarpsstöðva þar sem málið er notað, heldur einnig af hlutverki málsins í hinum stafræna upplýsingaheimi og innan hugbúnaðargeirans.

Á þessu sviði er íslenska ekki sérlega vel stödd. Í lok 20. aldar var íslensk máltækni nánast ekki til. Við áttum allgóðan stafrýni (*Púka*), ófullkominn talgervil, og þar með upp talið. Enginn íslenskur háskóli bauð upp á námsleiðir eða jafnvel einstök námskeið í máltækni eða tölvumálvísindum, engar rannsóknir voru stundaðar á þessu sviði, og engin íslensk hugbúnaðarfyrirtæki unnu að máltækniverkefnum [2].

Þetta fór að breytast eftir að sérstakur starfshópur skilaði skýrslu um máltækni til menntamálaráðherra árið 1999 [3]. Í þessari skýrslu voru settar fram tillögur um ýmsar aðgerðir til að koma íslenskri máltækni á laggirnar. Árið 2000 setti ríkisstjórnin af stað sérstaka máltækniáætlun með það að markmiði að styðja stofnanir og fyrirtæki til að koma upp undirstöðumálföngum – gagnasöfnum og hugbúnaði – fyrir íslenska máltækni. Þetta frumkvæði gat af sér ýmis verkefni sem hafa lagt grundvöll að íslenskri máltækni [2].

Eftir að máltækniáætluninni lauk árið 2004 ákváðu fræðimenn frá þremur stofnunum (Háskóla Íslands, Háskólanum í Reykjavík og Stofnun Árna Magnússonar í íslenskum fræðum) að taka höndum saman og mynda samstarfsvettvang sem nefnist Máltæknisetur (Icelandic Centre for Language Technology, ICLT) [4] til að fylgja viðfangsefnum áætlunarinnar eftir. Frá 2005 hafa fræðimenn Máltækniseturs ýtt úr vör ýmsum verkefnum sem hafa fengið styrki frá Rannsóknasjóði og Tækniþróunarsjóði.

Þrátt fyrir að talsvert hafi áunnist sýnir þessi skýrsla að það er einungis á sviði grundvallarbúnaðar og málfanga svo sem málfræðimörkunar, setningafræðilegrar þáttunar, málheilda og trjábanka sem staða íslenskunnar er viðunandi. Á flóknari sviðum eins og í merkingargreiningu setninga og texta, samræðukerfum, upplýsingaheimt, málmyndun, samantekt texta, merkingargreindum málheildum o.s.frv., er ekkert til fyrir íslensku. Því er ljóst að mikið starf er óunnið við að tryggja framtíð íslenskunnar sem fullgilds þátttakanda í evrópsku upplýsingasamfélagi nútímans – og framtíðarinnar.

Upplýsinga- og samskiptatæknin er nú á þröskuldi nýrrar byltingar. Í kjölfar einkatölva, netvæðingar, marg-

miðlunar, spjaldtölva, snjallsíma og tölvuskýja fylgir næsta kynslóð tækninnar sem mun ala af sér hugbúnað sem skilur ekki aðeins bókstafi og málhljóð heldur heil orð og setningar, og gagnast notendum margfalt betur vegna þess að hann talar, kann og skilur tungumál þeirra. Undanfarar þessarar þróunar eru t. d. Google Translate, ókeypis netþjónusta sem þýðir milli 57 tungumála, ofurtölvan Watson hjá IBM sem hefur sigrað Bandaríkjameistarann í spurningaleiknum „Jeopardy“, og Siri-hugbúnaðurinn fyrir iPhone frá Apple sem getur brugðist við talskipunum og svarað spurningum á ensku, þýsku, frönsku og japönsku.

Næsta kynslóð upplýsingatækninnar mun ráða svo vel við mannlegt mál að fólk mun geta notað sitt eigið tungumál til samskipta með þessari tækni. Tæki munu geta brugðist við raddskipunum sem eru einfaldar í notkun með því að afla sjálfkrafa mikilvægustu frétta og upplýsinga úr stafrænum upplýsingabrunni heimsins. Búnaður sem byggist á máltækni mun geta þýtt á sjálfvirkan hátt eða aðstoðað túlka; gert útdrætti úr samtölum og skjölum; og liðsinnt notendum við nám. Til dæmis gæti slíkur búnaður hjálpað nýbúum til að læra íslensku og falla þannig betur að menningu landsins og samfélagi.

Næsta kynslóð upplýsinga- og samskiptatækninnar mun gera iðnaðar- og þjónustuvélmennum (sem verið er að þróa á rannsóknastofum) kleift að skilja nákvæmlega hvað notendur þeirra vilja láta þau gera, og gera síðan skýra grein fyrir árangri sínum. Þarna er komið á allt annað og hærra svið en þegar unnið er með einfaldar stafatöflur og orðasöfn, stafrýna og framburðarreglur. Tæknin verður að hverfa frá einföldum nálgunum og snúa sér að gerð altækra mállíkana sem taka einnig til setningagerðar og merkingar til að skilja fjölbreyttar og flóknar spurningar og veita innihaldsrík og markviss svör.

Evrópsk tungumál eru misvel búin undir þessa framtíð. Í eftirfarandi skýrslu er sett fram stöðumat fyrir 30 Evrópumál, byggt á fjórum meginþáttum; vélþýðingum, talvinnslu, textagreiningu og grundvallarmálföngum sem þarf til smíði máltæknibúnaðar. Málunum var skipað í fimm klasa. Það þarf ekki að koma á óvart að íslenska er í lægsta klasanum á öllum þessum fjórum sviðum. Hún er þar á sömu slóðum og önnur tungumál sem fáir tala, svo sem írska, lettneska, litháíska og maltneska. Þessi tungumál eru langt að baki stórþjóðamálum eins og t. d. þýsku og frönsku. En jafnvel málföng og máltæknitól fyrir þau tungumál ná hvorki sömu gæðum né yfirgripi og hliðstæð föng og tól fyrir ensku, sem er í fararbroddi á nær öllum sviðum máltækninnar.

Hvað þarf til ef við viljum tryggja framtíð íslensku í upplýsingasamfélaginu? Árið 1999 áætlaði starfshópur um máltækni að það myndi kosta u.þ.b. einn milljarð króna á þágildandi verðlagi að gera íslenska máltækni sjálfbæra. Eftir það átti markaðurinn að geta tekið við, vegna þess að hann hefði þá aðgang að málföngum sem hefðu verið þróuð á vegum máltækniáætlunar ríkisstjórnarinnar, og yrðu tiltæk á jafnréttisgrundvelli fyrir alla sem hygðust nota þau við gerð markaðsvara [3].

Enda þótt máltækniáætlunin hafi verið árangursrík og haft mikil áhrif á þróun íslenskar máltækni verður að hafa í huga að ráðstöfunarfé hennar frá 2000-2004 var aðeins um 1/8 af því sem starfshópur um máltækni taldi þurfa [2]. Það þarf því ekki að koma á óvart að íslensk máltækni er enn á bernskuskeiði. 330 þúsund málnotendur eru einfaldlega of fáir til að standa undir kostnaðarsamri þróun nýrra framleiðsluvara. Um þessar mundir vinna nær engin íslensk fyrirtæki á sviði máltækni vegna þess að þau sjá sér engan hag í því. Áframhaldandi opinber stuðningur við íslenska máltækni er nauðsynlegur til að tryggja nýtingu þess búnaðar og málfanga sem þegar hefur verið komið upp, svo og þeirrar þekkingar og reynslu sem safnast hefur saman meðal fræðimanna og fyrirtækja.

Íslenska er ekki í bráðri hættu, þrátt fyrir yfirburði enskunnar í máltækni og tölvumálvísindum. Á hinn bóg-

inn gæti staðan gerbreyst á svipstundu þegar ný kynslóð tækninnar fer fyrir alvöru að ráða við mannlegt mál á skilvirkan hátt. Með framförum í vélþýðingum mun máltæknin hjálpa mönnum til að sigrast á tungumálaþröskuldum, en aðeins milli þeirra mála sem geta bjargað sér í hinum stafræna heimi. Tungumál sem jafnvel mjög fáir tala geta lifað af, verði fullnægjandi máltæknibúnaður tiltækur. Án slíks búnaðar munu jafnvel stórþjóðatungumál verða í mikilli hættu. Eigi íslenska að vera lífvænleg þjóðtunga í þróuðum heimi verður hún að geta staðið undir kröfum upplýsingatækninnar. Fjárfesting í máltækni verður því að vera grunnþáttur í framkvæmd íslenskrar málstefnu.

Langtímamarkmið META-NET er að innleiða hágæða máltækni fyrir öll tungumál þannig að menningarleg fjölbreytni stuðli að eflingu pólitískrar og efnahagslegrar einingar. Tæknin mun brjóta múra milli tungumála í Evrópu og smíða brýr milli þeirra í staðinn. Þetta krefst þess að allir hagsmunaaðilar – í stjórnmálum, rannsóknum, viðskiptum, og samfélaginu öllu – sameini krafta sína í þágu framtíðar.

Þessi hvítbókaröð tengist öðrum markvissum aðgerðum sem META-NET stendur að. Nýjustu upplýsingar eins og framtíðarsýn [5] META-NET og útfærða rannsóknarstefnu (Strategic Research Agenda, SRA) er að finna á vefsetri META-NET: http://www.meta-net.eu.

2

HÆTTUR SEM STEÐJA AÐ TUNGUMÁLINU: ÖGRUN FYRIR MÁLTÆKNI

Við verðum um þessar mundir vitni að stafrænni byltingu sem hefur gífurleg áhrif á samskipti og samfélag. Nýleg þróun í stafrænni upplýsinga- og samskiptatækni er stundum borin saman við það þegar Gutenberg fann upp prentverkið. Hvað getur sú samlíking sagt okkur um framtíð evrópsks upplýsingasamfélags og sérstaklega tungumála okkar?

> Við verðum um þessar mundir vitni að stafrænni byltingu sem hefur sambærileg áhrif og uppfinning prentverksins á sínum tíma.

Eftir uppfinningu Gutenbergs voru stigin tímamótaskref í samskiptum og deilingu þekkingar með verkum eins og t. d. þýðingu Lúthers á Biblíunni yfir á þjóðtungur. Á þeim öldum sem síðan eru liðnar hafa verið þróaðar menningarbundnar aðferðir til að sinna betur málvinnslu og deilingu þekkingar:

- Stöðlun stafsetningar og málfræðireglna helstu tungumála skapaði möguleika á hraðri útbreiðslu nýrra vísindalegra og vitsmunalegra hugmynda;
- þróun opinberra tungumála gerði fólki kleift að hafa samskipti innan ákveðinna (oft pólitískra) landamerkja;
- tungumálakennsla og þýðingar milli mála gerðu það mögulegt að eiga samskipti þvert á tungumál;
- ritstjórnarreglur og bókfræðileg viðmið tryggðu gæði prentaðs efnis og aðgengi að því;
- tilkoma margvíslegra fjölmiðla, svo sem dagblaða, útvarps, sjónvarps, bóka o.fl. fullnægði mismunandi samskiptaþörfum.

Á síðustu tuttugu árum hefur upplýsingatæknin átt sinn þátt í því að greiða fyrir mörgum ferlum og gera þau sjálfvirk:

- Ritvinnslu- og umbrotskerfi hafa komið í stað vélritunar og setningar;
- Microsoft PowerPoint hefur komið í staðinn fyrir glærur og myndvarpa;
- með tölvupósti eru skjöl send og tekið á móti þeim mun hraðar en með bréfasíma;
- Skype býður upp á ódýr netsímtöl og skapar vettvang fyrir fjarfundi;
- snið hljóð- og myndbandaskráa gerir auðvelt að skiptast á margmiðlunarefni;
- leitarvélar greiða notendum aðgang að vefsíðum með leit byggðri á lykilorðum;
- netþjónusta eins og Google Translate skilar sæmilega réttum þýðingum á svipstundu;
- félagsmiðlar eins og Facebook, Twitter og Google+ greiða fyrir samskiptum, samvinnu og deilingu upplýsinga.

Þrátt fyrir gagnsemi slíkra tóla og búnaðar dugir þetta ekki til að standa undir sjálfbæru margmála evrópsku samfélagi fyrir alla, með frjálsu flæði upplýsinga og varnings.

2.1 TUNGUMÁLAÞRÖSKULDAR STANDA Í VEGI FYRIR EVRÓPSKU UPPLÝSINGASAMFÉLAGI

Við getum ekki vitað nákvæmlega hvernig upplýsingasamfélag framtíðarinnar mun líta út. En miklar líkur eru á því að bylting í samskiptatækni muni skapa nýja tegund tengsla milli fólks sem talar mismunandi tungumál. Þetta setur aukinn þrýsting á fólk að læra ný tungumál og þó sérstaklega á hönnuði að búa til nýjan tæknibúnað sem tryggi gagnkvæman skilning og aðgang að deilanlegri þekkingu. Í alþjóðasamfélagi viðskipta og upplýsinga tengjast sífellt fleiri tungumál og málnotendur sífellt hraðar með hjálp nýrra miðla. Vinsældir félagsmiðla (Wikipedia, Facebook, Twitter, YouTube og nú nýlega Google+) eru einungis toppurinn á ísjakanum.

> Sífellt fleiri tungumál og málnotendur tengjast sífellt hraðar með hjálp nýrra miðla.

Nú á dögum getum við flutt margra gígabæta texta um heiminn þveran og endilangan á örfáum sekúndum áður en við áttum okkur á því að hann er á máli sem við skiljum ekki. Samkvæmt nýrri skýrslu frá framkvæmdastjórn Evrópusambandsins kaupa 57% evrópskra netnotenda vörur og þjónustu með því að nota tungumál önnur en móðurmál sitt. (Enska er algengasta erlenda tungumálið á þessu sviði en þar á eftir koma franska, þýska og spænska.) 55% notenda lesa erlent mál sér til gagns en aðeins 35% nota annað tungumál til þess að skrifa tölvupóst eða gera athugasemdir á vefnum [6]. Fyrir nokkrum árum var enska tungumál netsins – megnið af því efni sem þar var að finna var skrifað á ensku – en þetta hefur nú gerbreyst. Algjör sprenging hefur orðið í textamagni á öðrum Evrópumálum á netinu (og sama gildir um tungumál Asíu og Mið-Austurlanda).

Það sætir furðu að hin altæka stafræna gjá sem munur tungumála skapar skuli ekki hafa fengið mikla athygli í opinberri umfjöllun; samt sem áður vekur hún mjög brýna spurningu: Hvaða Evrópumál munu dafna í netvæddu upplýsinga- og þekkingarsamfélagi og hver eru dæmd til að hverfa?

2.2 TUNGUMÁL OKKAR Í HÆTTU

Þótt prentverkið hraðaði deilingu upplýsinga í Evrópu olli það því einnig að mörg evrópsk tungumál liðu undir lok. Textar á svæðisbundnum málum og minnihlutamálum komust sjaldan á prent og því voru tungumál eins og korníska og dalmatíska eingöngu notuð sem talmál og notkunarsvið þeirra þar með takmarkað. Mun netið hafa sambærileg áhrif á tungumál okkar?

Hin u.þ.b. 80 tungumál Evrópu eru ein ríkulegustu og mikilvægustu menningarverðmæti álfunnar og grundvallarþáttur í hinni einstöku samfélagsgerð hennar [7]. Þótt tungumál eins og enska og spænska muni að öllum líkindum halda stöðu sinni á hinu stafræna markaðstorgi sem er að verða til gætu mörg evrópsk tungumál orðið gagnslaus í netvæddu samfélagi. Slík þróun myndi veikja alþjóðlega stöðu Evrópu og stangast á við markmið um jafna samfélagsþátttöku allra Evrópuþegna óháð tungumáli.

> Hin fjölbreyttu tungumál Evrópu eru ein ríkulegustu og mikilvægustu menningarverðmæti álfunnar.

Í skýrslu UNESCO um fjöltyngi er lögð áhersla á að tungumál séu ómissandi tæki til þess að gera mönnum kleift að njóta grundvallarmannréttinda, svo sem tjáningarfrelsis, menntunar og þátttöku í samfélaginu [8].

2.3 MÁLTÆKNI ER GRUNDVALLARSTUÐNINGSTÆKNI

Áður fyrr beindust aðgerðir til að vernda og varðveita tungumál einkum að tungumálakennslu og þýðingum. Giskað hefur verið á að evrópski markaðurinn á sviði þýðinga, túlkunar, staðfærslu hugbúnaðar og alþjóðavæðingar vefsetra hafi velt 8,4 milljörðum evra árið 2008 og er talinn munu vaxa um tíu prósent á ári [9]. Samt sem áður fullnægir þessi upphæð einungis litlum hluta núverandi þarfar og framtíðarþarfa fyrir samskipti milli tungumála. Augljósasta aðferðin til að tryggja breidd og dýpt málnotkunar í Evrópu framtíðarinnar er að nota viðeigandi tækni, rétt eins og við notum tæknina til að leysa þarfir okkar í samgöngum, orku og stuðningi við fatlaða, svo að eitthvað sé nefnt.

Stafræn máltækni sem beinist að öllum myndum ritaðs máls og talsamskipta gerir fólki kleift að vinna saman, stunda viðskipti, deila þekkingu og taka þátt í félagslegum og pólitískum rökræðum óháð tungumáli og tölvufærni. Hún er oft hulinn hluti af flóknum hugbúnaði sem við nýtum okkur þegar við:

- öflum upplýsinga með notkun leitarvéla á netinu;
- rýnum stafsetningu og málfræði í ritvinnslukerfi;
- skoðum umsagnir um vörur í netverslun;
- hlustum á talaðar leiðbeiningar leiðsagnarkerfis í bíl;
- þýðum vefsíður með hjálp netþjónustu.

Máltækni felst í ýmsum grundvallarbúnaði sem margvísleg ferli innan stærri hugbúnaðarkerfa byggjast á. Tilgangur hvítbókaraðar META-NET er að skerpa sýn okkar á það hversu þroskuð þessi grunntækni sé fyrir hin ýmsu Evrópumál.

Evrópa þarfnast traustrar og ódýrrar máltækni fyrir öll tungumál álfunnar.

Til að viðhalda stöðu sinni í fararbroddi nýsköpunar á heimsvísu þarfnast Evrópa máltækni sem er löguð að öllum evrópskum tungumálum og er traust, ódýr og vel samþættuð helstu hugbúnaðarumhverfum. Án máltækni munum við ekki öðlast gjöfulan margmála reynsluheim, byggðan á gagnvirkni og margmiðlun, í náinni framtíð.

2.4 TÆKIFÆRI MÁLTÆKNINNAR

Í prentheiminum varð stærsta tæknibyltingin þegar farið var að fjölfalda ímynd texta með notkun prentvéla. Menn þurftu áfram að fletta upp þekkingaratriðum, lesa, þýða, og taka saman þekkingu. Það þurfti að bíða eftir Edison með upptökur á talmáli – en sú tækni bjó þó einnig aðeins til afrit.

Stafræn máltækni getur nú gert sjálfvirkt allt ferlið við þýðingu, samningu efnis og þekkingarstjórnun fyrir öll evrópsk tungumál. Hún getur einnig raungert þróun eðlilegs stýriviðmóts sem byggt er á máli og tali fyrir heimilisraftæki, vélar, bifreiðar, tölvur og vélmenni. Þróun viðskipta- og iðnaðarverkbúnaðar er enn á frumstigi, en áfangar í rannsóknum og þróun á þessu sviði eru þó farnir að opna mikla möguleika. Til dæmis eru vélþýðingar nú þegar sæmilega nákvæmar á afmörkuðum sviðum og tilraunabúnaður skilar margmála upplýsingum og sinnir þekkingarstjórnun og samningu efnis á mörgum Evrópumálum.

Eins og oftast er með tækni var fyrsti máltæknibúnaðurinn, svo sem raddstýrð notendaviðmót og samræðukerfi, þróaður með mjög sérhæfða notkun í huga og sýnir því oft takmarkaða hæfni. En geysimikil markaðstækifæri er að finna í menntageiranum og skemmtanaiðnaðinum þar sem hægt væri að nýta máltækni í leikjum, menningarminjasetrum, menntandi skemmtun, bókasöfnum, hermun og æfingaáætlunum. Upplýsingaþjónusta í farsíma, hugbúnaður fyrir tölvustutt tungumálanám, fjarnámsumhverfi, sjálfsmatstól og forrit til að uppgötva ritstuld eru fáein dæmi þar sem máltækni getur

leikið mikilvægt hlutverk. Vinsældir félagsmiðla eins og Twitter og Facebook benda til þess að þörf sé á háþróaðri máltækni sem getur haldið utan um póst, gert útdrætti úr umræðum, bent á hneigð í skoðunum, greint tilfinningar í svörum, bent á brot á höfundarétti eða haft uppi á misnotkun.

Máltækni hjálpar fólki að sigrast á þeirri „fötlun" sem felst í málfræðilegum fjölbreytileik.

Í máltækni felast gífurleg tækifæri fyrir evrópskt samstarf. Hún getur hjálpað okkur að takast á við hið flókna málumhverfi í Evrópu – þá staðreynd að mismunandi tungumál lifa eðlilegu samlífi í evrópskum viðskiptum, samtökum og skólum. En þegnarnir þurfa að geta haft samskipti yfir þessi tungumálamörk sem skera hinn sameiginlega evrópska markað þvert og endilangt og með aðstoð máltækni má sigrast á þessari hindrun en styðja um leið við óhefta notkun einstakra tungumála.

Ef við horfum enn lengra fram í tímann mun nýskapandi margmála evrópsk máltækni verða viðmiðun fyrir aðra í alþjóðasamfélaginu þegar þeir fara að virkja sín eigin margmála samfélög. Líta má á máltækni sem eins konar „stuðningstækni" sem aðstoðar okkur við að yfirstíga „fötlunina" sem fylgir fjölbreytilegu tungumálaumhverfi og gerir málsamfélögin aðgengilegri hvert öðru. Að lokum má nefna virkt rannsóknarsvið innan máltækninnar sem er notkun máltækni við björgunaraðgerðir á hamfarasvæðum, þar sem rétt framkvæmd getur skipt sköpum. Í framtíðinni gætu greind vélmenni búin hæfileikum til margmála málnotkunar bjargað mannslífum.

2.5 ÖGRANIR SEM MÁLTÆKNI STENDUR FRAMMI FYRIR

Þótt töluverðar framfarir hafi orðið í máltækni á síðustu árum er hraði tækniframfara og nýsköpunar í framleiðsluvörum enn of lítill. Sá máltæknibúnaður sem mest er notaður, svo sem málfræði- og stafrýnar ritvinnslukerfa, er venjulega einmála og þar að auki einungis til fyrir fáein tungumál.

Núverandi hraði tæknilegra framfara er of lítill.

Þótt vélþýðingar á netinu séu gagnlegar til að fá þokkalega hugmynd um efni skjala glíma þær við alls kyns vandamál þegar þörf er á mjög nákvæmum og fullkomnum þýðingum. Vegna þess hve mannlegt mál er flókið er það bæði langt og dýrt ferli sem krefst langtíma fjármögnunar að skrifa hugbúnað sem líkir eftir mannlegu máli og prófa hann við eðlilegar kringumstæður. Til að halda brautryðjendahlutverki sínu í því að takast á við þær tæknilegu ögranir sem fylgja margmála samfélagi verður Evrópa því að beita nýjum aðferðum til að hraða þróuninni. Hér gæti bæði verið um að ræða framfarir í tölvutækni og aðferðir eins og lýðvirkjun.

2.6 MÁLTAKA MANNA OG VÉLA

Til að útskýra hvernig tölvur fást við tungumál og hvers vegna það er svo erfitt að forrita þær til þess skulum við líta sem snöggvast á það hvernig við tileinkum okkur móðurmálið og önnur mál, og skoða síðan hvernig máltæknikerfin virka.

Mannfólkið öðlast málkunnáttu á tvo mismunandi vegu: Lærir af dæmum og lærir reglurnar sem liggja þar að baki.

Mannfólkið lærir tungumál á tvo mismunandi vegu. Ungbörn læra móðurmál sitt með því að hlusta á samskipti foreldra sinna, systkina og annarra fjölskyldumeðlima. Um það bil tveggja ára gömul fara þau að mynda fyrstu orðin og stuttar setningar. Þetta er því aðeins

mögulegt að börn hafa meðfæddan hæfileika til máls, og til að herma eftir því sem þau heyra og binda það í kerfi. Nám annars máls síðar á ævinni krefst meiri áreynslu, einkum vegna þess að nemandinn er ekki umlukinn málsamfélagi sem hefur málið að móðurmáli. Í skólum eru erlend mál venjulega numin með því að læra málfræðilega formgerð, orðaforða og stafsetningu með mynsturæfingum sem lýsa málfræðilegri kunnáttu í formi óhlutstæðra reglna, tafla og dæma. Nám erlends tungumáls verður erfiðara með aldrinum.

Hinar tvær megingerðir máltæknikerfa „nema“ tungumál á svipaðan hátt og mennirnir. Tölfræðilegar (eða gagnaknúnar) aðferðir afla málþekkingar úr gífurlega umfangsmiklum textasöfnum. En þótt nægjanlegt sé að nota texta á einu máli til að þjálfa t. d. stafrýna eru samhliða textar á tveim eða fleiri málum nauðsynlegir þegar kemur að þjálfun vélrænna þýðingarkerfa. Algrím vélræns náms „lærir“ þá mynstur sem sýna hvernig orð, orðasambönd og heilar setningar eru þýdd.

Þessi tölfræðilega nálgun getur krafist milljóna setninga og gæði útkomunnar aukast í réttu hlutfalli við magn greinds texta. Þetta er ein ástæða þess að þeir sem reka leitarvélar eru áfjáðir í að safna eins miklu af rituðu efni og hægt er. Stafrýnar í ritvinnslukerfum og netþjónustur eins og Google Search og Google Translate byggjast á tölfræðilegum aðferðum. Meginkostur tölfræðinálgunarinnar er sá að vélin lærir fljótt í samfelldri röð þjálfunarferla, jafnvel þótt gæðin geti verið með ýmsu móti.

Hin meginaðferðin í máltækni og vélþýðingum er að smíða reglukerfi. Þá þurfa sérfræðingar á sviði málvísinda, tölvumálvísinda og tölvunarfræði fyrst að skrá málfræðigreiningu (þýðingarreglur) og búa til orðalista (orðasöfn). Þetta tekur langan tíma og kostar mikla vinnu. Reglukerfin krefjast einnig sérfræðiþekkingar. Sum helstu reglubyggðu vélþýðingarkerfin hafa verið í stöðugri þróun í meira en tuttugu ár. Meginkosturinn við reglukerfin er að sérfræðingarnir hafa meiri stjórn á málvinnslunni. Þetta gerir það mögulegt að laga kerfisbundið villur í hugbúnaðinum og veita notendum nákvæma endurgjöf, sérstaklega þegar reglukerfin eru notuð í tungumálanámi. En vegna þess hversu kostnaðarsöm þessi vinna er hefur reglubyggð máltækni til þessa einungis verið þróuð fyrir stærstu tungumálin.

Þar sem styrkleikar og veikleikar tölfræðilegu kerfanna og reglubyggðu kerfanna eru á mismunandi sviðum beinast rannsóknir um þessar mundir að blönduðum aðferðum sem tengja þessar tvær gerðir saman. Enn sem komið er hafa slíkar aðferðir þó ekki reynst eins vel í markaðshugbúnaði og á rannsóknarstofunum.

Eins og fram hefur komið í þessum kafla byggist alls kyns búnaður sem notaður er í upplýsingasamfélagi nútímans á máltækni. Í Evrópu á þetta sérstaklega við á sviði viðskipta og upplýsinga vegna þess hversu margmála málumhverfið þar er. En þrátt fyrir að máltækni hafi tekið miklum framförum á síðustu árum eru enn miklir möguleikar á því að auka gæði máltæknikerfa. Hér á eftir verður hlutverki íslenskunnar í evrópsku upplýsingasamfélagi lýst og mat lagt á stöðu máltækni fyrir íslensku.

3

ÍSLENSKA Í EVRÓPSKU UPPLÝSINGASAMFÉLAGI

3.1 ALMENN ATRIÐI

Um það bil 330 þúsund manns eiga íslensku að móðurmáli. Flestir búa á Íslandi [10] en fjölmargir Íslendingar eru þó búsettir erlendis [11], svo sem annars staðar á Norðurlöndunum, á meginlandi Evrópu og í Norður-Ameríku. Þá er íslenska móðurmál fáeinna Vestur-Íslendinga af annarri og þriðju kynslóð [12] en þeir eru flestir komnir um og yfir sjötugt. Á síðustu árum hefur innflutningur til landsins aukist til muna og þar með hefur þeim fjölgað sem tala íslensku sem erlent mál þótt sá hópur sé enn tiltölulega lítill.

Íslenska er notuð á öllum stigum stjórnsýslu, í skólakerfinu, viðskiptum og í öllum almennum samskiptum í landinu.

Þótt ekki sé ákvæði um íslenska tungu í stjórnarskrá lýðveldisins hefur nýlega verið fest í lög að íslenska sé opinbert tungumál landsins [13]. Hún er notuð á öllum stigum stjórnsýslu, í skólakerfinu, í viðskiptum og öllum almennum samskiptum í landinu.

Lítið er um mállýskur í íslensku og vanalega er talað um smávægileg mállýskutilbrigði í framburði fremur en eiginlegar mállýskur. Lífseigast þessara mállýskutilbrigða er harðmælið þar sem lokhljóð eru fráblásin á milli sérhljóða á norðanverðu landinu en ófráblásin annars staðar, í orðum eins og *æpa*, *vita* og *taka*. Önnur mállýskuafbrigði eru smám saman að láta undan síga, svo sem raddaður framburður *l*, *m*, *n* á undan *p*, *t*, *k* í orðum eins og *úlpa*, *svampur*, *vanta*; vestfirskur einhljóðaframburður á undan *ng* og *nk* í orðum eins og *söngur*, *banki*, en í máli flestra er þar tvíhljóð; og hinn svokallaði *hv*-framburður þar sem borið er fram önghljóð í upphafi orða eins og *hver* þar sem flestir hafa lokhljóðið *k* [14]. Á hinn bóginn virðist sem ný mállýskutilbrigði séu að myndast, svo sem tvinnhljóðun á *tj* þar sem *tjald* fer að hljóma eins og það væri *tsjald* [15]. Einungis er um minniháttar mállýskuafbrigði að ræða í setningagerð og fæst þeirra eru landshlutabundin. Þó virðast einstaka breytingar vera að gerast, sérstaklega í máli yngra fólks, og má þar nefna hina svokölluðu nýju þolmynd, eins og í *það var barið mig* í stað *ég var barin(n)*, svo og útvíkkaða notkun framvinduhorfs, *vera að*, eins og í *ég er ekki að skilja þetta* og *þeir voru að spila mjög vel*. Slík notkun heyrist varla hjá eldra fólki. Íslenskuna sem töluð er í Vesturheimi má telja sérstaka mállýsku (eða mállýskur) enda hefur orðaforði þar þróast öðruvísi en á Íslandi. Þar má meðal annars nefna vestur-íslensku orðin *telefón* og *kar* (sbr. e. *telephone* og *car*) fyrir *sími* og *bíll*. Þá hafa orðmyndir og framburðarsérkenni stirðnað eða jafnvel aukist í vestur-íslensku en horfið að mestu eða öllu á Íslandi. Sem dæmi má nefna flámælið sem enn lifir góðu lífi meðal Vestur-Íslendinga.

3.2 SÉRKENNI ÍSLENSKRAR TUNGU

Íslenska er norður-germanskt tungumál sem myndar vestur-norrænu málaættina ásamt færeysku og nýnorsku. Hún er svokallað FSA-tungumál (eðlileg orðaröð frumlag-umsögn-andlag) og hefur sögnina jafnan í öðru (eða fyrsta) sæti setningar. Vegna ríkulegs beygingakerfis er orðaröð hins vegar tiltölulega frjáls; ákveðin orð geta staðið á ýmsum stöðum án þess að merking breytist. Eftirfarandi setningar hafa t. d. sömu merkingu þrátt fyrir að röð frumlags og andlags hafi verið snúið við:

- Hundurinn (nefnifall) beit köttinn (þolfall).
- Köttinn (þolfall) beit hundurinn (nefnifall).

Íslenska er FSA-tungumál þar sem sögnin er jafnan í öðru (eða fyrsta) sæti setningar en orðaröð þó tiltölulega frjáls.

Íslenska er meðal tiltölulega fárra tungumála þar sem frumlag setningar getur staðið í öðrum föllum en nefnifalli – oftast nær þágufalli en einnig þolfalli (og í nokkrum tilfellum eignarfalli). Í eftirfarandi setningum er t. d. fornafnið í fyrstu persónu eintölu alltaf frumlag, þrátt fyrir að standa í þremur mismunandi föllum:

- Ég (nefnifall) las bókina.
- Mig (þolfall) vantar bókina.
- Mér (þágufall) líkar bókin.

Íslenskan er beygingamál og hefur fjögur föll, þrjú kyn og tvær tölur í nafnorðum, fornöfnum, lýsingarorðum og ákveðna (viðskeytta) greininum. Enginn óákveðinn greinir er notaður í málinu. Auk þessa beygjast lýsingarorð bæði veikt (ákveðið) og sterkt (óákveðið). Sagnir beygjast eftir persónu, tölu, tíð, hætti og mynd. Sagt er að íslenskan sé bræðingsmál sem þýðir að einstök ending er oft notuð fyrir fleiri en eina beygingarformdeild. Fjöldi beygingarflokka flækir svo kerfið enn, þannig að margar mismunandi endingar geta staðið fyrir sömu málfræðiformdeild eða formdeildasamsetningu, allt eftir því hver stofninn er.

Orðaforði málsins er að mestu norrænn að uppruna.

Orðaforðinn er að mestu norrænn (germanskur) að uppruna þótt fjölmörg tökuorð hafi slæðst inn í málið á þeim ellefu öldum sem liðið hafa síðan land byggðist. Eftir kristnitöku árið 1000 voru t. d. fjölmörg orð tekin úr latínu og við siðaskiptin árið 1550 jukust áhrif frá þýsku með þýðingum á trúarritum og sálmum. Þá var Ísland undir danskri stjórn frá 1380 til 1944 og áhrif danskrar tungu frá þessum tíma eru augljós. Ýmis dönsk orð voru tekin inn í málið og mörg þeirra urðu hluti af íslensku. Þar má m. a. nefna orð eins og *gardínur* (*gardin* á dönsku) og *viskustykki* (*viskestykke* á dönsku).
Það er opinber stefna að ný orð skuli smíða úr íslenskum efnivið í stað þess að fá lánuð orð úr erlendum málum. Þar sem margs konar hljóðavíxl eru algeng í íslensku má nota þau til þess að mynda nýtt orð af öðru, svo sem *leysni* af *lausn*, og einnig eru hin fjölmörgu viðskeyti málsins notuð til þess að mynda nýtt orð af rótum sem þegar eru til í málinu, svo sem *disk-lingur* af orðinu *diskur*. Algengast er þó að mynda ný orð með samsetningu tveggja eða fleiri sjálfstæðra orða, rétt eins og í *staf-setningar-orða-bók* og *um-hverfis-mála-ráðu-neyti*. Þetta gerir tungumálið bæði líflegt og gagnsætt.

Orðmyndun í íslensku er mjög virk.

Framburður íslensku er tiltölulega gagnsær og að mestu hægt að segja fyrir um hann út frá stafsetningunni. Sá

sem kann þær reglur sem gilda um vensl stafsetningar og framburðar ætti því að geta borið fram ný orð sem verða á vegi hans vandræðalaust, svo framarlega sem hann greinir réttilega orðhlutaskil en þau geta haft áhrif á framburð sumra orða. Reglur um áherslu orða eru einnig mjög einfaldar þar sem aðaláherslan fellur alltaf á fyrsta atkvæði og aukaáhersla kemur svo vanalega á annað hvert atkvæði eftir það, þótt það eigi ekki alltaf við í samsettum orðum.

Ritmálið byggist á latneska stafrófinu en þó eru notaðir í íslensku nokkrir stafir sem ekki þekkjast t. d. í ensku. Þetta eru stafirnir Þ/þ (einungis notaður í íslensku þótt upprunann megi rekja til fornensku), Ð/ð (einnig notaður í færeysku), Æ/æ (einnig notaður í norsku, dönsku og færeysku) og Ö/ö (einnig notaður í sænsku, finnsku, eistnesku, þýsku og ungversku). Að auki eru notaðir í íslensku sex broddstafir fyrir ákveðna sérhljóða: Á/á, É/é, Í/í, Ó/ó, Ú/ú og Ý/ý.

Ritaða málið hefur breyst tiltölulega lítið frá upphafi ritaldar sem gerir Íslendingum það kleift með nokkurri þjálfun að lesa fornískenska texta. Meginbreytingar á stafsetningu á undanförnum áratugum hafa verið niðurfelling setunnar (sem þó er enn notuð í fáeinum eigin- nöfnum og ættarnöfnum eins og *Zóphónías* og *Haralz*) og upptaka *é* í stað *je*.

3.3 NÝLEG ÞRÓUN

Allt frá hernámi Breta og síðar Bandaríkjamanna í heimstyrjöldinni síðari hefur íslenskan orðið fyrir mun sterkari áhrifum frá ensku en dönsku og þau áhrif hafa aukist að mun við innreið tónlistar, kvikmynda og sjónvarpsefnis frá Bretlandi og Bandaríkjunum. Vöxtur netsins hefur einnig aukið áhrif ensku á íslensku, enda eru um 95% þjóðarinnar netvædd.

Áhrif frá ensku eru augljósust í fjölda tökuorða úr ensku í íslensku en fæst þessara orða er þó að finna í orðabókum og þau sjást sjaldan á prenti. Þau eru að auki oft litin hornauga af málræktarmönnum. Notkun þeirra einskorðast því að mestu við talað mál og að auki má finna þau í óopinberum og persónulegum skrifum, svo sem í tölvupósti, á bloggsíðum o.s.frv.

Tökuorð úr ensku eru algeng í daglegu tali en mun minna áberandi í ritmáli.

Ensk áhrif á málkerfið virðast þó óveruleg. Mörg tökuorðanna sem notuð eru hversdagslega fá íslenskar endingar þótt nokkur þeirra beygist ekki. Þar má nefna *næs* (úr e. *nice*), *kúl* (úr e. *cool*), o.s.frv. Stundum er því haldið fram að sumar breytingar í setningagerð og hljóðkerfi íslenskunnar, svo sem hið útvíkkaða framvinduhorf og tvinnhljóðunin á *tj* sem áður eru nefndar, megi rekja til enskra áhrifa, en um það er þó deilt.

Á undanförnum árum hefur mikið verið rætt um svokallað „umdæmistap“ á Íslandi eins og í mörgum öðrum löndum. Íslenskur vinnumarkaður hefur orðið sífellt alþjóðlegri á síðustu árum – íslensk fyrirtæki starfa erlendis og erlend fyrirtæki starfa á Íslandi. Ensk tunga er því hluti af daglegu starfi þessara fyrirtækja og fundir og bréfleg samskipti fara iðulega fram á ensku. Þá er það orðið algengt að ársskýrslur þessara fyrirtækja, vefsíður og annað efni, séu að hluta eða öllu á ensku. Einnig virðist það vera hálfgerð tíska að íslensk fyrirtæki beri enskt nafn, ýmist eingöngu eða að hluta. Dæmi um þetta eru nöfn eins og *Icelandair*, *Actavis*, *Baugur Group* og *Stoðir Invest* [16].

Annað svið atvinnulífsins þar sem ensk tunga er áberandi er upplýsingatækni, en um hana verður betur rætt í næsta aðalkafla.

3.4 ÍSLENSK MÁLRÆKT

Í íslenskri málrækt hefur áhersla löngum verið lögð á bæði varðveislu og eflingu íslenskrar tungu. Þetta má sjá greinilega á þeirri vinnu sem lögð hefur verið í uppbyggingu orðaforðans með starfsemi ýmissa íðorðanefnda. Þær eru vanalega skipaðar sjálfboðaliðum úr ýmsum

fræði- og atvinnugreinum en málræktarsvið Stofnunar Árna Magnússonar í íslenskum fræðum styður við starf þeirra. *Íslensk málnefnd* var stofnuð 1964 [17] en meginhlutverk hennar er að vera stjórnvöldum, og þá einkum mennta- og menningarmálaráðuneytinu, til ráðgjafar um íslenska tungu og íslenska málstefnu auk þess að semja árlega ályktun um stöðu tungunnar. Íslensk málnefnd ber ábyrgð á þeim stafsetningarreglum sem auglýstar eru af menntamálaráðuneytinu og notaðar eru í skólakerfinu. Nefndin hafði frumkvæði að stofnun *Málræktarsjóðs* en hlutverk hans er að „beita sér fyrir og styðja hvers konar starfsemi til eflingar íslenskri tungu og varðveislu hennar“ [18].

Stundum er sagt að allir Íslendingar séu málfræðingar. Bændur og sjómenn, hjúkrunarfræðingar og kennarar hringja í útvarpsstöðvar og Stofnun Árna Magnússonar í íslenskum fræðum til að ræða hnökra á málfari annarra og kvarta undan málvillum. Fólk hefur einlægar áhyggjur af stöðu tungunnar í landinu og heilmiklar umræður fara fram um það hvernig best sé að varðveita málið og jafnvel hvort sú varðveisla sé ómaksins verð.

Íslensk málnefnd er stjórnvöldum til ráðgjafar um íslenska tungu og íslenska málstefnu.

Þó líta flestir Íslendingar á tungumálið sem kjarna íslenskrar menningar og íslenskrar sjálfsmyndar og því hefur mikið starf verið unnið í þeim tilgangi að varðveita það sem best.

Miðstöð íslenskrar málræktar er í *Stofnun Árna Magnússonar í íslenskum fræðum* en meginhlutverk hennar er að „vinna að rannsóknum í íslenskum fræðum og skyldum fræðigreinum, einkum á sviði íslenskrar tungu og bókmennta, að miðla þekkingu á þeim fræðum og varðveita og efla þau söfn sem henni eru falin eða hún á“ [19]. Stofnunin skiptist í nokkrar deildir sem sinna mismunandi þáttum íslensks máls, bókmennta og menningar, svo sem málrækt, orðfræði, máltækni, nafn- og örnefnafræði, handritafræði, þjóðfræði og alþjóðlegum tengslum.

Ríkisútvarpið hefur löngum leikið stórt hlutverk í varðveislu tungunnar, ekki aðeins vegna eigin málstefnu heldur einnig vegna vinsælla útvarpsþátta áður fyrr, eins og *Íslensks máls* og *Daglegs máls* þar sem málfræðingar ræddu um tunguna og orðaforðann, og *Orð skulu standa*, þar sem tvö lið kepptust um að finna rétta merkingu sjaldgæfra orða og hugtaka. Almennt gegna fjölmiðlarnir mikilvægu hlutverki í verndun íslenskrar tungu.

Ríkisútvarpið hefur löngum leikið stórt hlutverk í varðveislu tungunnar.

Tuttugu og tvær útvarpsstöðvar eru í landinu og talað mál í þeim öllum er að mestu leyti á íslensku þótt enskan sé yfirgnæfandi í tónlistinni sem leikin er. Að auki eru í landinu tíu sjónvarpsstöðvar og þótt meiri hluti þess efnis sem sjónvarpað er sé á erlendum tungumálum er staða íslenskunnar sterk [20]. Allt erlent sjónvarpsefni er textað á íslensku – fyrir utan sumt barnaefni sem er talsett – og þegar um beinar útsendingar er að ræða frá erlendum stórviðburðum segir íslenskur þulur vanalega frá því helsta sem er að gerast [21].

Dagur íslenskrar tungu hefur verið haldinn hátíðlegur síðan 1996 á fæðingardegi þjóðskáldsins Jónasar Hallgrímssonar, 16. nóvember, og er honum ætlað að efla umræður um íslenska tungu [22].

3.5 ÍSLENSKA Í MENNTAKERFINU

Íslensk tunga er mikilvægur þáttur í skólakerfinu og nemendur í 1.-4. bekk grunnskóla verja að lágmarki 1.120 mínútum á viku í íslenskt mál og bókmenntir. Í 5.-7. bekk hefur þessi tími minnkað niður í 680 mínútur á viku og síðan 630 mínútur á viku í 8.-10. bekk en það er töluvert minna en aðrar Norðurlandaþjóðir verja

í móðurmálskennslu [23]. Í framhaldsskóla er einnig minni tíma varið til móðurmálskennslu en annars staðar á Norðurlöndunum, eða að lágmarki 20 einingum af þeim 200 sem krafist er til stúdentsprófs [24].

Í PISA-könnununum sem gerðar hafa verið frá árinu 2000 fór lesskilningur íslenskra ungmenna, sérstaklega drengja, stöðugt minnkandi. Í könnuninni 2009 hafði ástandið hins vegar batnað nokkuð og Ísland var þar í ellefta sæti og í svipaðri stöðu og aðrar Norðurlandaþjóðir að Finnum frátöldum [25].

Háskóli Íslands er eini háskólinn þar sem hægt er að taka doktorspróf í íslensku en meistarapróf í málinu er hægt að taka frá Manitobaháskóla í Kanada auk Háskóla Íslands. Þó nokkrir háskólar víða um heim bjóða upp á B.A.-próf í íslensku.

Aðeins tveir af þeim sjö háskólum sem í landinu eru hafa sérstaka málstefnu þar sem íslenska er tilgreind sem opinbert mál háskólans. Enska er sífellt meira notuð í starfi háskólanna þar sem erlendum kennurum hefur fjölgað og þar að auki stefna allir háskólarnir að því að fjölga erlendum nemendum. Vegna þessa fer námskeiðum sem kennd eru á ensku fjölgandi, sem og doktorsritgerðum skrifuðum á því máli. Þá hefur það aukist að íslenskir fræðimenn skrifi fræðigreinar sínar á ensku og námsefni í skólunum er æ meir á enskri tungu [16]. Með því að fjölga íslenskutímum í skólum landsins má bæta íslenskukunnáttu nemenda og búa þá þannig betur undir virka þátttöku í íslensku samfélagi.

Með því að fjölga íslenskutímum í skólum landsins má bæta íslenskukunnáttu nemenda og búa þá þannig betur undir virka þátttöku í samfélaginu.

Máltækni gæti verið hjálpleg í þessu sambandi enda gefur hún möguleika á tölvustuddu tungumálanámi sem gerir nemendum kleift að njóta tungumálsins á skemmtilegan hátt, t. d. með því að tengja orðaforða í ákveðnum texta við skilgreiningar á orðunum eða við hljóðskrá eða myndband með viðbótarupplýsingum, svo sem framburði orðanna.

3.6 ALÞJÓÐLEGIR ÞÆTTIR

Ísland er lítið land og í raun aðeins örríki í samfélagi þjóðanna, og því eru áhrif íslenskra lista, vísinda og fræða erlendis aðeins smávægileg. Fáeinir íslenskir tónlistarmenn hafa náð vinsældum utan landsins, svo sem *Björk*, *SigurRós* og *Gus Gus*, en þar sem tónlist þeirra er að litlu leyti sungin á íslensku gerir hún lítið til þess að auka hróður tungumálsins utan landsteinanna. Það sama má segja um velgengni íslenskra rithöfunda erlendis sem hefur kynnt íslenska menningu fyrir öðrum þjóðum en ekki beinlínis íslenska tungu. Hins vegar hafa vinsældir íslenskra tónlistarmanna og rithöfunda, uppgangur – og fall – íslenskra banka og fyrirtækja erlendis, svo og áherslur Íslands á umhverfisvæna orku vakið athygli annarra þjóða á Íslandi og skilað sér í aukinni umfjöllun um landið í erlendum fjölmiðlum og fjölgun ferðamanna til landsins. Íslendingasögurnar, víkingarnir og íslenski hesturinn eru því ekki lengur einu íslensku fjársjóðirnir sem heilla útlendinga.

Áhugi á íslensku á alþjóðavettvangi fer vaxandi.

Íslensk tunga hefur lítil áhrif á önnur tungumál og aðeins örfá íslensk orð hafa ratað sem tökuorð inn í önnur mál. Þar eru langalgengust orð dregin af eiginnafninu *Geysir* sem í mörgum málum tákna goshver. Þá er enska orðið *eider* tökuorð úr íslensku, komið af orðinu *æður*, og íslenska orðið *tölt* er almennt notað erlendis um fimmta gang íslenska hestsins.

Aukinn áhugi á íslenskri tungu og menningu kemur greinilega fram í vaxandi fjölda þeirra nemenda sem stunda íslenskunám, ýmist á Íslandi eða í öðrum löndum. Við Háskóla Íslands jókst fjöldi erlendra nema í íslenskunámi um nærri 100% milli áranna 2005 og 2007

og árið 2008 bauð Háskólinn í fyrsta sinn upp á námsleið í hagnýtri íslensku ætlaða þeim sem vilja læra tungumálið án þess að leggja áherslu á hinn akademíska þátt námsins. Íslenska er nú kennd í um 40 háskólum utan Íslands og styrkir Ísland 18 þeirra fjárhagslega [16]. Þá er boðið upp á sjálfstæð íslenskunámskeið í fjölmörgum löndum, svo sem í fyrrum Íslendingabyggðum Kanada og Bandaríkjanna, og á milli 300 og 400 manns fara daglega inn á heimasíðu *Icelandic Online* [26].

Staða íslensku myndi væntanlega styrkjast á alþjóðavettvangi ef landið gengi í Evrópusambandið.

Íslensk tunga er hvergi gjaldgeng í alþjóðlegum samskiptum en því hefur verið haldið fram að staða málsins myndi styrkjast á alþjóðavettvangi ef landið gengi í Evrópusambandið [27], þar sem íslenska yrði þar með eitt af opinberum tungumálum sambandsins [28]. Einnig er hægt að nýta máltækni til að bregðast við þeirri ógn sem stafar af ensku með því að þróa vélþýðingar og margmála upplýsingaheimt og hjálpa þannig til við að lágmarka óhagræðið sem felst í því, bæði fyrir einstaklinga og viðskiptalífið, að hafa ekki ensku að móðurmáli.

3.7 ÍSLENSKA Á NETINU

Í júní 2010 höfðu um það bil 95% þjóðarinnar aðgang að netinu [29] og í aldurshópnum 35-44 ára var hlutfallið allt að 100%. Í byrjun maí 2011 voru 197.000, eða 61,8% þjóðarinnar, skráðir notendur Facebook [30].

Næstum allir Íslendingar nota netið.

Árið 2010 voru 25.000 .is lén skráð [31] og um það bil 5.600 lén voru á landinu fyrir utan .is kerfið [32]. Fjöldi vefsetra er talinn í kringum 7.500 en þar eru þó hvorki taldar bloggsíður innan .is léna né vefir á erlendum lénum eins og blogspot.com og wordpress.com.

Netið er orðið svo vinsælt að árið 2010 gerðist það í fyrsta sinn að auglýsendur eyddu meiri peningum í auglýsingar á netinu en í prentmiðlunum [33]. Slíkt hefur reyndar ekki enn gerst á Íslandi en virðist þó stefna í þá átt. Af sjö vinsælustu vefjunum á Íslandi eru þrír fréttamiðlar (*mbl.is*, *visir.is*, *pressan.is*). Netið hefur einnig að miklu leyti tekið við af símaskránni þar sem upplýsingasíðan *ja.is* er fimmti mest notaði vefur landsins. Aðrir vinsælir vefir eru Google, Facebook og YouTube [34] sem allir bjóða nú upp á íslenskt notendaviðmót.
Vöxtur netsins er mikilvægur fyrir máltækni að tvennu leyti. Annars vegar er fjöldi texta á stafrænu formi algjör gullnáma þegar kemur að greiningu á notkun tungumála, og þá sérstaklega þegar safna þarf tölfræðilegum upplýsingum. Hins vegar býður netið upp á fjöldann allan af notkunarsviðum fyrir máltækni.

Vöxtur netsins skiptir miklu máli fyrir máltækni.

Leitarvélar eru án efa mest notaði hugbúnaðurinn á netinu en þær nýta margs konar sjálfvirka málvinnslu eins og við munum sjá í síðari hluta þessa rits. Þar er um að ræða margbrotna máltækni sem er breytileg eftir tungumálum. Í íslensku þarf til dæmis að taka tillit til mismunandi beygingarendinga nafnorða, lýsingarorða og sagna, svo og hljóðavíxla í stofni, eins og t. d. í orðmyndunum *svartur* og *svört*. Notendur netsins geta einnig nýtt máltækni á annan hátt, svo sem með sjálfvirkum þýðingum vefsíðna á mörg tungumál. Þegar litið er á gríðarlegan kostnað við mennska þýðingu þessa efnis vekur furðu hversu lítið hefur verið gert til að þróa slíkan þýðingarbúnað. Ástæðuna má ef til vill rekja til þess hversu margslungin íslensk tunga er í raun, svo og hversu fjölbreytta tækni þarf til að smíða dæmigerðan máltæknibúnað.
Í næsta kafla er að finna yfirlit um máltækni og helstu afurðir hennar en einnig er kynnt mat á stöðu máltækni fyrir íslensku.

4

MÁLTÆKNI FYRIR ÍSLENSKU

Undir máltækni falla m. a. hugbúnaðarkerfi sem hönnuð eru til þess að vinna með mannlegt mál. Tungumál eru bæði rituð og töluð en þótt talmálið hafi þróast á undan og sé þannig eðlilegasta form mállegra samskipta er ritmálið það form sem notað er til geymslu og miðlunar margbrotinna upplýsinga og mestallrar mannlegrar þekkingar. Til að vinna með og framleiða tungumál í þessum mismunandi myndum höfum við annars vegar taltækni og hins vegar textatækni, en hvorttveggja byggist á orðasöfnum, málfræðireglum og merkingarfræði. Þetta þýðir að máltækni tengir tungumálið við mismunandi form þekkingar, óháð því hvernig henni er miðlað (í tali eða texta, sjá mynd 1).

Í öllum samskiptum tengjum við tungumálið öðrum samskiptaháttum og upplýsingamiðlum – tali getur fylgt látbragð og andlitstjáning. Stafrænir textar tengjast myndum og hljóði. Í kvikmyndum getur komið fram bæði talað og ritað mál. Tal- og textatækni skarast því og fléttast saman við margs konar aðra tækni sem greiðir fyrir úrvinnslu fjölhátta samskipta og margmiðlunargagna.

Hér á eftir verður fjallað um meginverksvið máltækni, þ. e. málrýni, vefleit, taltækni og vélþýðingar. Undir þetta fellur verkbúnaður og grundvallartækni eins og:

- stafrýni
- ritstoð
- tölvustutt tungumálanám
- upplýsingaheimt
- útdráttur upplýsinga
- samantekt texta
- spurningasvörun
- talkennsl
- talgerving

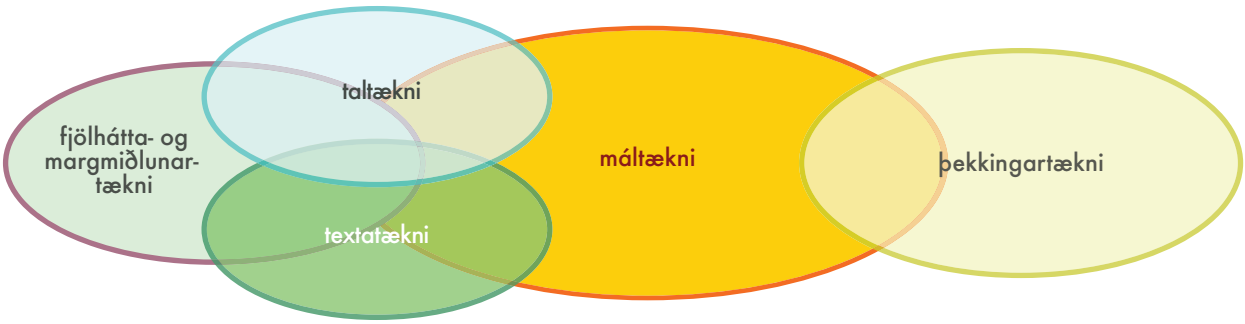
Máltækni er mótað og öflugt rannsóknarsvið og hægt er að vísa á fjölda inngangstexta um sviðið, t. d. [35, 36, 37, 38, 39]. Áður en ofannefndum notkunarsviðum og búnaði verða gerð skil verður högun dæmigerðs máltæknikerfis lýst stuttlega.

4.1 HÖGUN MÁLTÆKNIBÚNAÐAR

Í dæmigerðum hugbúnaði til málvinnslu felast nokkrar einingar sem endurspegla mismunandi þætti tungumálsins. Mynd 2 sýnir mjög einfaldaða byggingu ritvinnslukerfis. Þrjár fyrstu einingarnar snúa að gerð og merkingu ílagstextans:

1. Forvinnsla: hreinsun gagna, afnám sniðs, greining ílagstungumáls, o.s.frv.
2. Málfræðigreining: sögnin fundin, andlög hennar og ákvæðisorð, og setningagerðin greind.
3. Merkingargreining: einræðing orða (fundið út hver er merking orðsins í tilteknu samhengi); greining endurvísunar (t. d. hvaða fornafn vísar til hvaða nafnorðs í setningunni) og staðgengla; og merking setningarinnar sýnd á þann hátt að tölva geti lesið hana.

Eftir greiningu textans geta verkbundnar einingar séð um ýmsar aðrar aðgerðir, svo sem sjálfvirka samantekt



1: Samhengi máltækninnar

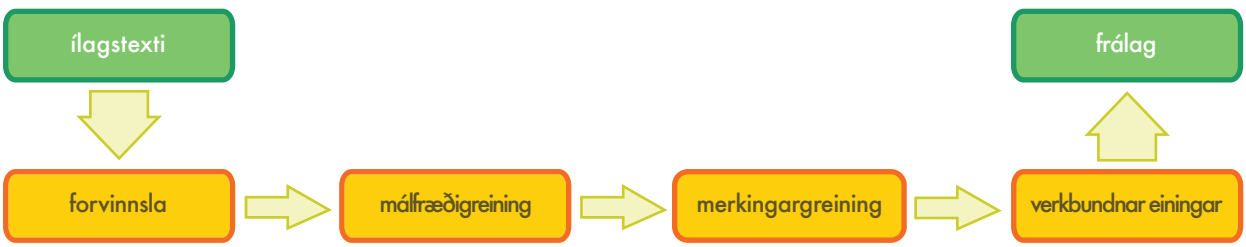
ílagstexta og uppflettingu í gagnagrunni. Þetta er einföldud lýsing á uppbyggingu verkbúnaðarins en gefur þó innsýn í það hversu flókinn máltæknibúnaður er. Að lokinni kynningu á helstu verksviðum máltækninnar verður gefið stutt yfirlit yfir yfirlit yfir núverandi stöðu máltæknirannsókna og máltæknimenntunar, og að lokum drepið á rannsóknarverkefni sem ýmist er lokið eða eru í gangi. Síðan verður gerð grein fyrir mati sérfræðinga á stöðu helstu máltæknitóla og málfanga út frá ýmsum mælikvörðum, s. s. aðgengi, þroska og gæðum. Heildarstaða máltækni fyrir íslensku er svo dregin saman í töflu í lok þessa kafla (mynd 8). Þau hugtök og málföng sem eru feitletruð í textanum er að finna í þessari töflu. Í framhaldi af þessu er máltæknistuðningur við íslensku borinn saman við stuðning við önnur tungumál sem fjallað er um í þessari ritröð.

4.2 HELSTU VERKSVIÐ

Í þessum kafla verður fjallað um mikilvægustu máltæknitól og málföng, og gefið yfirlit yfir máltækni á Íslandi.

4.2.1 Málrýni

Flestir sem hafa unnið með ritvinnslukerfi eins og Microsoft Word vita að í því er stafrýnir sem bendir á stafsetningarvillur og stingur upp á leiðréttingum. Fyrstu stafrýnarnir báru orðin í textanum saman við safn rétt ritaðra orða. Nú er þessi hugbúnaður mun þróaðri. Með því að nota sérhæfð algrím til **málfræðigreiningar** má greina villur í beygingu (svo sem ranga eignarfallsendingu) og setningagerð, eins og þegar sögnina vantar eða þegar ósamræmi er á milli sagnar og frumlags (t. d. *ég *skrifar bréf*). Hins vegar munu fæstir stafrýnar finna villur í eftirfarandi dæmum:



2: Dæmigerð kerfishögun við textavinnslu

3: Málrýni (tölfræðileg; reglubyggð)

- Ég var um þetta leiti á næsta leyti.
- Hún segir að móðir sýn hafi aðra sín á málið.
- Hann þótti hafa stirt stöðu sína.

Til þess að hægt sé að fást við slíkar villur þarf að greina samhengi textans, t. d. þegar ákveða skal hvort lýsingarorð eigi að vera með einu *n*-i (kvenkyn) eða tveim (karlkyn), eins og í eftirfarandi dæmi:

- Hann er farinn.
- Hún er farin.

Greining slíkra villna byggist ýmist á sérstakri **málfræðilýsingu** fyrir hvert tungumál, sem mikinn tíma og sérþekkingu þarf til að fella inn í hugbúnaðinn, eða á tölfræðilegu mállíkani. Slíkt líkan reiknar líkurnar á því að tiltekið orð birtist í ákveðnu umhverfi (t. d. eftir því hvaða orð fara á undan og á eftir). Til dæmis er *hann er farinn* líkleg orðaruna en *hún er farinn* er það ekki. Tölfræðilegu mállíkani af þessu tagi má koma upp á sjálfvirkan hátt með því að nota mikið af (réttum) málgögnum (**málheild**). Báðar aðferðirnar (reglusmíði og tölfræðilíkan) hafa einkum verið þróaðar fyrir ensk málföng og það er ekki auðvelt að yfirfæra þær á íslensku sem hefur sveigjanlegri orðaröð, ótakmarkaða möguleika á samsetningu orða og ríkulegra beygingarkerfi.

Málrýni er ekki bundin við ritvinnslukerfi; hún er líka notuð í ritstoðarkerfum.

Málrýni er ekki bundin við ritvinnslukerfi; hún er líka notuð í ritstoðarkerfum, þ. e. hugbúnaðarumhverfi til að skrifa handbækur og önnur rit samkvæmt ákveðnum stöðlum fyrir flókna upplýsingatækni, heilbrigðisgeirann, verkfræði og fleira. Af ótta við kvartanir og skaðabótakröfur viðskiptavina vegna rangrar notkunar sem rekja má til illskiljanlegra leiðbeininga leggja fyrirtæki sífellt meiri áherslu á gæði tæknilegra leiðbeininga, á sama tíma og þau stefna á alþjóðlegan markað (með þýðingum og staðfærslu). Framfarir í málvinnslu hafa leitt til þróunar á ritstoðarbúnaði sem aðstoðar höfunda tæknilegra leiðbeininga við að velja orð og setningagerð sem samræmist iðnaðarreglum og skorðum fyrirtækja á notkun íðorða.

Stafrýnir hefur verið til fyrir íslensku frá því seint á níunda áratugnum þegar Friðrik Skúlason ehf. (Frisk Software) þróaði stafsetningaforritið *Púka*. Forritið hefur síðan verið uppfært og endurbætt. Það er til fyrir MS Office og er mikið notað. Aðrir stafrýnar hafa einnig verið hannaðir. Árið 2002 þróaði hollenska fyrirtækið Polderland stafrýni fyrir MS Office og einnig er til stafrýnir í opnum hugbúnaði fyrir GNU/Linux forrit, byggður á Aspell. Þessi forrit skoða eingöngu stök orð og ráða því ekki við margar algengar stafsetningarvillur. Frumgerð að samhengisháðum stafrýni hefur verið felld inn í LanguageTool [40] og vinnur með OpenOffice. Sá stafrýnir gæti hugsanlega myndað grunninn að málfræðiryni, en slíkt forrit er ekki til fyrir íslensku.

Fyrir utan stafrýna og ritstoð er málrýning einnig mikilvæg fyrir tölvustutt tungumálanám og henni er líka beitt

við sjálfvirka leiðréttingu á fyrirspurnum sem sendar eru vefleitarvélum eins og tillögukerfi Google *Áttirðu við:*.

4.2.2 Vefleit

Leit á vefnum, svo og á innri netum og í stafrænum bókasöfnum, er væntanlega það svið þar sem máltækni er mest notuð nú á dögum, en er þó fremur skammt á veg komin. Leitarvélin Google, sem kom fram á sjónarsviðið 1998, er nú notuð í 80% allra vefleita í heiminum [41]. Síðan 2004 hefur sögnin *gúgla* verið notuð í íslensku þótt hún hafi ekki enn komist í prentaðar orðabækur. Hvorki leitarviðmót Google né framsetning niðurstaðna hefur tekið grundvallarbreytingum frá fyrstu útgáfu. Í nýjustu útgáfu býður Google reyndar upp á leiðréttingar á ranglega stafsettum orðum og hefur nú bætt við merkingarlegum leitarmöguleikum sem geta bætt nákvæmni leitarinnar með því að greina merkingu orða í samhengi leitarorðsins [42]. Velgengni Google sýnir að með stóru gagnasafni og skilvirkum aðferðum við að lykla gögnin getur tölfræðileg aðferð skilað vel viðunandi niðurstöðum.

Þegar um flóknari upplýsingaleit er að ræða er nauðsynlegt að nýta dýpri málfræðiþekkingu til textatúlkunar. Tilraunir með **orðaföng** eins og tölvutæk samheitasöfn og verufræðileg málföng (s. s. WordNet fyrir ensku og GermaNet fyrir þýsku) hafa sýnt verulega bættan árangur í að finna síður þar sem samheiti við leitarorðið koma fyrir, svo sem *hagnaður*, *arður*, *gróði* og *ábati* eða jafnvel fjarskyldari orð.

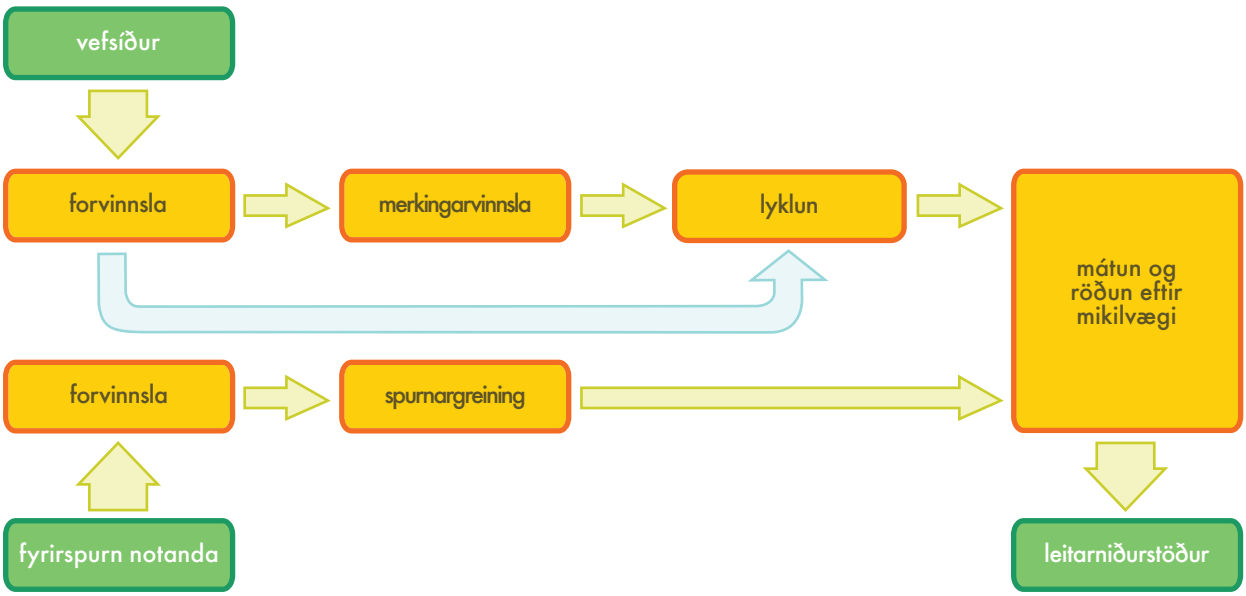
> Næsta kynslóð leitarvéla verður að vera útbúin mun þróaðri máltækni.

Næsta kynslóð leitarvéla verður að vera útbúin mun þróaðri máltækni, einkum til að ráða við leitartexta í formi spurningar eða annars konar setningar í stað einstakra leitarorða. Til að bregðast við fyrirspurninni „Láttu mig fá lista yfir öll fyrirtæki sem voru yfirtekin af öðrum fyrirtækjum síðustu fimm árin“ þarf máltæknikerfið að framkvæma bæði setningagreiningu og **merkingargreiningu** fyrirspurnarinnar og hafa atriðisorðaskrá til að kalla fram viðeigandi skjöl á fljótvirkan hátt. Til að unnt sé að gefa viðunandi svar þarf að beita setningalegri þáttun til greiningar á málfræðilegri formgerð setningarinnar og greina að verið sé að leita að fyrirtækjum sem hafa verið yfirtekin en ekki þeim fyrirtækjum sem tóku yfir önnur fyrirtæki. Þá þarf að skilgreina sambandið *síðustu fimm* ár svo hægt sé að ákvarða við hvaða ár er átt. Að lokum þarf að máta leitarfyrirspurnina við ógrynni af óskipulögðum gögnum svo að finna megi upplýsingarnar sem leitað er að. Þetta er kallað „upplýsingaheimt“ og felur í sér leit að skjölum og vægisröðun þeirra. Til þess að hægt sé að búa til lista yfir fyrirtæki þarf kerfið einnig að þekkja ákveðinn orðastreng í skjali sem nafn fyrirtækis, en það ferli kallast „nafnakennsl“.

Enn meiri ögrun felst í því að máta leitarfyrirspurnina við skjöl á öðrum tungumálum. Þvermála upplýsingaheimt felur í sér sjálfvirka þýðingu leitarfyrirspurnar yfir á öll möguleg tungumál og síðan þýðingu niðurstaðnanna aftur yfir á markmálið.

Nú er gögn í auknum mæli að finna á öðru sniði en sem texta og því er orðin til þörf á þjónustu sem gefur kost á margmiðlunarupplýsingaheimt með því að leita að myndum, hljóði eða myndböndum. Þegar um er að ræða hljóð- og myndbandsskrár þarf sérstök talkennslaeining að breyta tali í texta (eða hljóðritun) sem síðan er hægt að máta við leitarfyrirspurnina.

Í beygingarmálum eins og íslensku er mikilvægt að hægt sé að leita að öllum beygingarmyndum orðs í einu í stað þess að þurfa að leita að hverri mynd sérstaklega. Þetta má gera með aðstoð gagnagrunnsins *Beygingarlýsing íslensks nútímamáls*, *BÍN* [43], sem þróaður hefur verið á Stofnun Árna Magnússonar í íslenskum fræðum. Gagnagrunnurinn hefur að geyma um það bil 280.000 beygingardæmi með meira en 5,8 milljónum beyging-



4: Vefleit

armynda. Hver færsla inniheldur nefnimyndina, orðmyndina, orðflokkinn og beygingarþætti nafnorða, sérnafna, lýsingarorða, sagna og atviksorða.
Fyrir nokkrum árum þróaði fyrirtækið Spurl leitarvélina *Emblu* sem nýtti þennan gagnagrunn. Sama algrím er notað við leit í íslensku símaskránni og á nokkrum öðrum síðum. Google leitarvélin er nú búin svipuðum hæfileikum, en þó ekki eins margþættum.

4.2.3 Talsamskipti

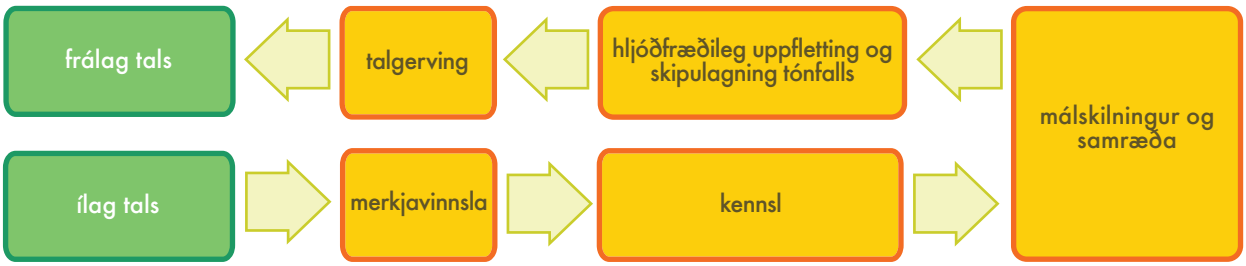
Talsamskipti eru eitt margra verksviða sem byggjast á taltækni, þ. e. tækni til að vinna með talað mál. Talsamskiptatækni er notuð til að smíða viðmót sem gerir notandanum kleift að tala við tölvuna í stað þess að nota tölvuskjáinn, lyklaborð og mús. Nú á dögum nýta fyrirtæki raddstýrð notendaviðmót í ýmiss konar sjálfvirkri og hálfsjálfvirkri símaþjónustu við viðskiptavini, starfsmenn eða viðskiptafélaga. Helstu atvinnugreinar sem nýta slík raddstýrð viðmót eru bankastarfsemi, birgjar, almenningssamgöngur og fjarskiptafyrirtæki. Talsamskiptatækni má t. d. einnig nota í viðmóti leiðsögutækja í bílum og í stað myndræns viðmóts og snertiskjáa sem notendaviðmót í snjallsímum.

> Taltækni er notuð til að smíða viðmót sem gerir notandanum kleift að tala við tölvuna í stað þess að nota tölvuskjáinn, lyklaborð og mús.

Talsamskipti byggjast á ferns konar grundvallartækni:

1. Sjálfvirk **talkennsl** ákvarða hvaða orð notandinn segir í tiltekinni segð.
2. Málskilningur greinir setningafræðilega formgerð segðarinnar og túlkar hana út frá viðkomandi kerfi.
3. Samræðustjóri ákvarðar hvað þarf að gera út frá ílagi notandans og möguleikum kerfisins.
4. **Talgerving** breytir svari kerfisins í hljóð sem notandinn nemur.

Eitt erfiðasta viðfangsefni talkennslabúnaðar er að greina rétt þau orð sem notandinn segir. Því þarf annaðhvort að takmarka hugsanlegar segðir notandans við afmarkað mengi lykilorða eða byggja upp mállíkön sem



5: Talsamskiptakerfi

ná yfir stóran hluta segða í eðlilegu máli. Með vélrænum námsaðferðum er líka hægt að koma upp mállíkönum á sjálfvirkan hátt úr **talmálsheildum**, stórum söfnum hljóðskráa með textaumritun. Takmörkun leyfilegra segða leiðir venjulega til þvingaðrar notkunar á talviðmótinu og getur haft þau áhrif að notendur taki því ekki vel; en smíði viðamikils mállíkans, fínstilling þess og viðhald eykur kostnaðinn við kerfið verulega. Raddstýrð notendaviðmót sem nýta mállíkan og gefa notandanum sveigjanleika í því hvernig hann ber fram erindi sitt í byrjun – t. d. heilsa með *Hvað get ég gert fyrir þig*? – eru yfirleitt sjálfvirk og fá jákvæðari viðbrögð notenda.

Fyrirtæki nota yfirleitt upptökur með lestri atvinnumanna til að mynda frálag talviðmótsins. Í stöðluðum segðum þar sem orðalagið er ekki háð tilteknu samhengi eða ákveðnum notanda getur þetta verið fullkomlega nóg til að notandinn sé sáttur. En þegar segðirnar eru breytilegar getur tónfallið orðið óeðlilegt vegna þess að bútar úr mismunandi hljóðskrám eru tengdir saman. Talgervlar eru sífellt að verða betri í því að skila breytilegum segðum sem hljóma eðlilega, en þó má enn bæta þá.

Viðmót á talsamskiptamarkaðnum hafa verið stöðluð umtalsvert á undanförnum áratug að því er snýr að hinum ýmsu tæknieiningum þeirra. Einnig hefur orðið veruleg markaðssamþjöppun fyrirtækja í talkennslum og talgervingu. Á innanlandsmarkaði í G20-löndunum (fjölmennum og efnahagslega sterkum löndum) hafa fimm alþjóðleg fyrirtæki verið ríkjandi, og í Evrópu einkum tvö – Nuance (bandarískt) og Loquendo (ítalskt). Árið 2011 tilkynnti Nuance um yfirtöku Loquendo þannig að markaðssamþjöppunin heldur enn áfram.

Þrír talgervlar fyrir íslensku hafa verið settir á markað. Formendabyggður talgervill var upphaflega gerður í kringum 1990 og annar, byggður á hljóðatvenndum (Snorri), um 2000. Báðir voru einkum notaðir af blindum og sjónskertum en þóttu ekki nógu fullkomnir til notkunar í kerfum og verkbúnaði fyrir almennan markað.

Árið 2005 var búinn til nýr talgervill (Ragga) í samvinnu Háskóla Íslands, Símans og Hex hugbúnaðar. Talgervillinn byggðist á tækni Nuance sem sá um þjálfun hans. Hann hefur verið notaður dálítið í verkbúnaði fyrir almennan markað en mörgum notendum finnst raddgæðin ekki fullnægjandi. Þar sem gæði tiltækra talgervla þykja ekki nógu mikil hefur Blindrafélagið gengist fyrir þróun nýs talgervils í samvinnu við Háskóla Íslands, Háskólann í Reykjavík og pólska hugbúnaðarfyrirtækið Ivona. Þessi talgervill hefur tvær raddir (Karl og Dóru) og verður tilbúinn síðar á þessu ári (2012) [44].

Stakorðagreinir var þróaður fyrir íslensku árið 2003. Hann skilaði góðum árangri í greiningu, eða um 97% nákvæmni. Þá hefur íslenskur stúdent við Tokyo Institute of Technology hannað frumgerð af kerfi fyrir sjálfvirk orðaflaumskennsl í íslensku. Kerfið náði 67,5% orðaná-

kvæmni [45]. Hvorugt þessara kerfa hefur verið notað í verkbúnað fyrir almennan markað. Um mitt ár 2011 hófu Háskólinn í Reykjavík og Máltæknisetur samvinnu við Google um undirbúning að smíði talþekkjara fyrir íslensku [46].

Miklar breytingar má sjá framundan vegna útbreiðslu snjallsíma sem nýs vettvangs fyrir tengsl fyrirtækja og viðskiptavina, í viðbót við venjulega síma, vefinn og tölvupóst. Þessar breytingar munu einnig hafa áhrif á nýtingu taltækninnar. Notkun á raddstýrðu viðmóti venjulegra síma mun fara minnkandi en mikilvægi talaðs máls sem notendavæns samskiptamáta við snjallsíma er sífellt að aukast. Það sem knýr þessa þróun er einkum aukin nákvæmni í talkennslum óháðum mælanda í þeim upplestrarkerfum sem þegar eru í boði sem miðlæg þjónusta fyrir notendur snjallsíma.

4.2.4 Vélþýðingar

Hugmyndina að því að nota tölvur til að þýða mannleg mál má rekja til ársins 1946 og var henni fylgt eftir með umtalsverðu fjármagni til rannsókna á sjötta áratug síðustu aldar og aftur á þeim níunda. Samt sem áður hafa **vélþýðingar** ekki náð að standa undir þeim væntingum um sjálfvirkar þýðingar milli tungumála sem þær gáfu á upphafsárunum.

Einfaldasta gerð vélrænna þýðinga felst í því að skipta út orðum í öðru málinu fyrir orð úr hinu málinu.

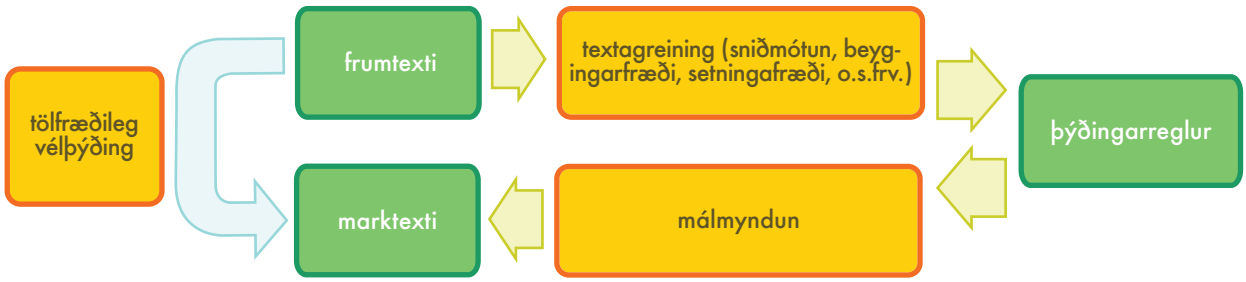
Einfaldasta gerð vélþýðinga felst í því að skipta út orðum í öðru málinu fyrir orð úr hinu málinu. Þetta getur verið gagnlegt á efnissviðum þar sem notað er mjög afmarkað og staðlað mál, svo sem í veðurfregnum. En til þess að þýðing á máli sem er ekki eins staðlað verði viðunandi þarf að fella stærri textaeiningar (orðasambönd, málsgreinar, jafnvel heilar efnisgreinar) sem nákvæmast að samsvarandi einingum í markmálinu. Helstu vandkvæðin felast í því að mannlegt mál er margrætt. Margræðni skapar vanda á mörgum sviðum, svo sem við einræðingu merkingar á orðasviðinu (*villa* er bæði 'mistök' og 'veglegt hús'), og fallstjórn á setningafræðisviðinu, eins og í:

- The woman saw the car and her husband, too.
- Konan sá bílinn og **maðurinn hennar** líka.
- Konan sá bílinn og **manninn sinn** líka.

Ein leið til að búa til vélþýðingarkerfi er að nota málfræðilegar reglur. Þegar þýtt er á milli náskyldra tungumála getur aðferð beinna umskipta verið fýsileg, eins og í dæminu hér að ofan. En reglubyggð kerfi (byggð á málfræðilegri þekkingu) greina oft ílagstextann og skapa táknbyggt millistig sem texti markmálsins er síðan leiddur af. Árangur þessarar aðferðar er undir því kominn að til sé yfirgripsmikið orðasafn með beygingarlegum, setningafræðilegum og merkingarlegum upplýsingum, ásamt stóru safni málfræðireglna sem þjálfaðir málfræðingar hafa smíðað af vandvirkni. Það er langt og þar með dýrt ferli að koma þessum forsendum upp.

Á síðari hluta níunda áratugarins þegar tölvur urðu öflugri og ódýrari jókst áhugi á að nýta tölfræðileg líkön í vélþýðingum. Tölfræðileg líkön byggjast á greiningu tvímála málheilda, svo sem Europarl **hliðstæðu málheildarinnar**, sem hefur að geyma þingskjöl Evrópuþingsins á 21 Evrópumálum. Ef nóg er af gögnum virka tölfræðilegar vélþýðingar nægilega vel til þess að gefa nokkurn veginn rétta merkingu texta á erlendu tungumáli með því að skoða samhliða texta og greina líkleg orðamynstur. Ólíkt þekkingarknúnum kerfum skila tölfræðilegar (eða gagnaknúnar) vélþýðingar oft málfræðilega röngu frálagi. Kosturinn við gagnaknúnar vélþýðingar er sá að þær eru ekki eins mannaflsfrekar, og einnig geta þær ráðið við ýmis málleg sérkenni (s. s. málshætti og orðtök) sem fara forgörðum í þekkingarknúnu kerfunum.

Styrkleikar og veikleikar þekkingarknúinna og gagnaknúinna vélþýðinga eru á mismunandi sviðum og því



6: Vélþýðingar (tölfræðilegar; reglubyggðar)

einbeita vísindamenn sér núorðið að blönduðum aðferðum sem sameina aðferðafræði beggja. Ein aðferðin er sú að nota bæði þekkingarknúið og gagnaknúið kerfi og láta svo sérstaka valeiningu ákveða hvert sé besta frálag hverrar setningar. Þegar um er að ræða lengri setningar en 12 orð verða niðurstöðurnar þó sjaldnast fullkomnar. Betri aðferð er að sameina bestu hluta hverrar setningar úr mörgum frálögum; þetta getur verið tiltölulega flókið þar sem samsvaranir mismunandi möguleika eru ekki alltaf augljósar og því þarf að samskipa þeim.

Vélþýðingar milli íslensku og annarra mála eru mjög snúnar.

Vélþýðingar milli íslensku og annarra mála eru mjög snúnar. Vegna fjölbreyttra möguleika til smíði samsettra orða er oft erfitt að greina orð og hafa nægilega yfirgripsmikið orðasafn; frjáls orðaröð og sagnaragnir skapa ýmis vandamál í greiningu, og auðugt beygingarkerfi veldur vandkvæðum við að merkja rétt öll beygingaratriði s. s. kyn, fall, tölu, hátt, tíð, o.s.frv.

Þróun vélþýðinga fyrir íslensku hefur ekki orðið ýkja mikil. Stefán Briem, sjálfstætt starfandi fræðimaður, hefur unnið að vélþýðingum síðan snemma á níunda áratugnum og hefur hannað vélþýðingarkerfi fyrir íslensku. Árið 2008 opnaði hann á vefnum ókeypis þjónustu sem býður upp á þýðingar milli íslensku og þriggja annarra tungumála (ensku, dönsku og esperantó) [47]. Hrafn Loftsson, kennari við Háskólann í Reykjavík, og samstarfsmenn hans hafa hannað reglubyggt grófþýðingakerfi úr íslensku á ensku, grundvallað á Apertium-verkvangnum [48]. Forútgáfa er nú á vefnum [49]. Google Translate hefur gefið kost á þýðingum úr og á íslensku síðan 2009. Gæðin voru heldur lítil í byrjun en hafa aukist.

Enn má auka gæði vélþýðingarkerfa verulega. Helstu vandkvæðin felast í aðlögun málfanganna að tilteknum efnissviðum eða notendahópum, og samþættingu tækninnar við vinnuferli sem nú þegar eru búin íðorðagrunni og þýðingarminni. Annað vandamál er að flest núverandi kerfi eru miðuð við ensku og sinna einungis þýðingum milli íslensku og örfárra annarra mála. Þetta leiðir til árekstra í þýðingarflæðinu og þvingar notendur vélrænna þýðinga til að læra á mismunandi orðakótunartól fyrir mismunandi kerfi.

Matskeppnir nýtast vel til að bera saman gæði vélþýðingarkerfa, mismunandi aðferðafræði og frammistöðu þeirra gagnvart mismunandi tungumálapörum. Taflan hér á eftir, sem unnin var innan Euromatrix+ verkefnis Evrópusambandsins, sýnir útkomu allra para milli 22 af 23 opinberum tungumálum Evrópusambandsins. (Írska var ekki með í samanburðinum.) Niðurstöðum er raðað samkvæmt BLEU einkunnakvarða, þar sem hærri einkunn fæst fyrir betri þýðingu [51]. Mennskur þýðandi myndi ná um 80 stigum.

	Markmál – Target language																					
	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
BG	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

7: Vélþýðingar milli 22 Evrópusambandstungumála – Machine translation between 22 EU-languages [50]

Bestu niðurstöðurnar (í grænum og bláum lit) fengust fyrir tungumál sem njóta góðs af umfangsmiklum samhæfðum rannsóknaráætlunum, sem og af tilvist margra samhliða málheilda (t. d. enska, franska, hollenska, spænska og þýska). Þau tungumál sem verr koma út eru merkt með rauðu. Þau skortir annaðhvort slíkar rannsóknaráætlanir eða eru eðlisólík öðrum tungumálum (t. d. ungverska, maltneska og finnska).

Gerð máltæknibúnaðar felur oft í sér fjölda undirverkþátta sem ekki eru alltaf sýnilegir notendunum en gegna þó þýðingarmiklum þjónustuhlutverkum á bak við tjöldin.

4.3 ÖNNUR VERKSVIÐ

Gerð máltæknibúnaðar felur oft í sér fjölda undirverkþátta sem ekki eru alltaf sýnilegir notendunum en gegna þó þýðingarmiklum þjónustuhlutverkum á bak við tjöldin. Þessir verkþættir byggjast allir á mikilvægum rannsóknarefnum sem hafa orðið að sjálfstæðum undirgreinum innan tölvumálvísinda. Spurningasvörun er t. d. virkt rannsóknarsvið og í tengslum við það hafa markaðar málheildir verið byggðar upp og vísindasamkeppnir haldnar. Spurningasvörun felur í sér annað og meira en lykilorðaleit (þar sem leitarvélin svarar með því að skila af sér safni skjala sem gætu varðað efnið) og gerir notendum kleift að spyrja beinskeyttra spurninga sem kerfið svarar á einkvæman hátt. Til dæmis:

Spurning: Hversu gamall var Neil Armstrong þegar hann steig fæti á tunglið?

Svar: 38 ára.

Þótt spurningasvörun sé augljóslega af sömu rót og vefleit er hún nú fyrst og fremst yfirheiti yfir rannsóknarspurningar eins og: hvaða tegundir spurninga eru til og

hvernig á að fást við þær; hvernig á að greina og bera saman þau skjöl sem hugsanlega hafa að geyma svarið (veita þau ósamrýmanleg svör?); og hvernig á að veiða afmarkaðar upplýsingar (svarið) út úr skjali á öruggan hátt án þess að hunsa samhengið.

Þetta tengist upplýsingaútdrætti, sviði sem var sérlega vinsælt og áhrifaríkt á tímum tölfræðibyltingarinnar í tölvumálvísindum snemma á tíunda áratug síðustu aldar. Með upplýsingaútdrætti er reynt að bera kennsl á tilteknar upplýsingaeiningar í tilteknum skjalaflokkum, svo sem að greina helstu þátttakendur í yfirtöku fyrirtækja eins og frá þeim er greint í umfjöllun dagblaða. Annað svið sem hefur verið rannsakað er frásagnir af hryðjuverkum. Þar er helsti vandinn að fella textann að sniðmáti sem tilgreinir brotamann, skotmark, tíma, staðsetningu og afleiðingar atviksins. Slík útfylling efnisbundinna sniðmáta er megineinkenni upplýsingaútdráttar og hann er því annað dæmi um tækni á bak við tjöldin sem myndar vel afmarkað rannsóknarsvið sem síðan þarf að fella inn í viðeigandi verkbúnað.

Flókinn hugbúnaður til textagreiningar og málmyndunar er ekki til fyrir íslensku.

Tvö jaðarsvið sem ýmist geta verið sjálfstæður verkbúnaður eða þjónað sem stoðþættir bak við tjöldin eru samantekt texta og **málmyndun**. Með samantekt er leitast við að draga meginatriði langs texta saman í stuttu máli og er meðal annars boðið upp á slíkt í Microsoft Word. Þar er einkum stuðst við tölfræðilega aðferð til að greina „mikilvæg“ orð í textanum (þ. e. orð sem eru hlutfallslega mun algengari í textanum en í almennri málnotkun) og ákvarða síðan hvaða setningar hafa að geyma hæst hlutfall þessara mikilvægu orða. Þær setningar eru síðan dregnar út úr textanum og settar saman til að mynda samantektina. Í þessari aðferð sem er mjög algeng í búnaði á almennum markaði felst samantektin eingöngu í því að draga setningar úr textanum, og textinn er því skorinn niður í hlutmengi upphaflegra setninga. Önnur aðferð, sem talsvert hefur verið rannsökuð, er sú að mynda nýjar setningar sem ekki koma fyrir í frumtextanum. Þetta krefst dýpri skilnings á textanum og er því mun viðkvæmara. Í flestum tilfellum er textamyndun ekki sjálfstæður búnaður heldur er hún felld inn í viðameiri hugbúnað, svo sem upplýsingakerfi í heilbrigðisþjónustu þar sem upplýsingum um sjúklinga er safnað, þær geymdar og síðan unnið úr þeim. Skýrslugerð er aðeins eitt af mörgum sviðum þar sem samantekt nýtist.

Ekkert af þeim búnaði sem rætt er um í þessum undirkafla er til fyrir íslensku.

4.4 NÁMSLEIÐIR

Máltækni er mjög þverfaglegt svið þar sem saman kemur sérþekking málfræðinga, tölvunarfræðinga, stærðfræðinga, heimspekinga, sálfræðinga, taugafræðinga og fleiri. Hún hefur því ekki öðlast traustan sess í íslensku háskólaumhverfi. Um síðustu aldamót var ekki boðið upp á neinar námsleiðir eða einstök námskeið í máltækni eða tölvumálvísindum í neinum íslenskum háskóla og engar rannsóknir voru í gangi á þessum sviðum.

Haustið 2002 tók Háskóli Íslands upp þverfaglegt meistaranám í máltækni. Um er að ræða tveggja ára nám (120 ECTS einingar) þar sem forkröfur eru B.A.-próf í tungumálum eða málvísindum eða B.Sc.-próf í tölvunarfræði (eða rafmagns- eða hugbúnaðarverkfræði). Árið 2007 var námið endurskipulagt í samvinnu milli íslenskudeildar Háskóla Íslands og tölvunarfræðideildar Háskólans í Reykjavík. Á meðan Norræni máltækniskólinn (Nordic Graduate School of Language Technology – NGSLT) var og hét, á árunum 2004-2009, gátu nemendur einnig tekið einstök námskeið við skóla annars staðar á Norðurlöndunum og í Eystrasaltslöndunum.

Vegna skorts á fé og mannafla hefur ekki verið mögulegt að taka nýja nemendur inn í meistaranámið síðan 2009. Hins vegar er reglulega boðið upp á einstök námskeið í

máltækni, málvinnslu og gagnamálfræði, bæði við Háskóla Íslands og Háskólann í Reykjavík.

4.5 INNLEND VERKEFNI OG VIÐFANGSEFNI

Aðeins um 330.000 manns tala íslensku og það er ekki nóg til þess að standa undir kostnaðarsamri þróun nýrra afurða. Það kostar jafnmikið að smíða máltæknibúnað fyrir íslensku og fyrir tungumál sem hundruð milljóna manna tala. Vegna þessa starfa næstum engin máltæknifyrirtæki á almennum markaði á Íslandi. Friðrik Skúlason ehf. hefur þróað og selt stafrýninn *Púka* en vinnur ekki að neinum nýjum framleiðsluvörum á sviði máltækni. Á síðasta áratug unnu Síminn og hugbúnaðafyrirtækið Hex með Háskóla Íslands að smíði bæði stakorðagreinis og talgervils fyrir íslensku. Hvorugt þessara fyrirtækja vinnur lengur að mál- eða taltækni. Clara er nýlegt fyrirtæki sem þjónustar önnur fyrirtæki sem vilja vita hvað fólki finnst um framleiðsluvörur þeirra og þjónustu. Kerfi Clöru notar merkingargreiningu og sérstaka aðferð við framsetningu gagna til að greina viðhorf fólks á netinu. Verkfæri fyrirtækisins til greiningar á vefsíðum á íslensku kallast *Vaktarinn* [52]. Á fyrsta starfsári var það með 1200 notendur ef með eru taldir þeir sem notuðu þjónustuna ókeypis til reynslu. Clara er eina fyrirtækið á Íslandi sem er að þróa máltæknibúnað sem markaðsvöru.

Árið 2000 setti íslenska ríkið af stað sérstakt máltækniátak með það fyrir augum að styðja stofnanir og fyrirtæki í því að búa til grundvallargögn fyrir íslenska máltækni. Þetta frumkvæði leiddi til nokkurra verkefna sem hafa haft mjög mikil áhrif á máltækni á Íslandi. Helstu afurðir máltækniátaksins eru eftirfarandi [2]:

- Gagnagrunnur með beygingarlýsingu íslensks nútímamáls
- Málfræðilega mörkuð málheild með 25 milljónum orða
- Þjálfunarsafn fyrir gagnastýrða málfræðilega mörkun
- Talgervill
- Stakorðagreinir
- Betrumbættur stafrýnir

Þegar máltækniátakinu lauk árið 2004 ákváðu fræðimenn frá þremur stofnunum (Háskóla Íslands, Háskólanum í Reykjavík og Stofnun Árna Magnússonar í íslenskum fræðum) sem höfðu tekið þátt í flestum verkefnum máltækniátaksins að sameinast um stofnun Máltækniseturs með það að markmiði að vinna áfram að verkefnum sem þegar voru komin af stað. Aðalhlutverk Máltækniseturs er að:

- vera upplýsingaveita um íslenska máltækni og reka vefsetur í því skyni (http://maltaeknisetur.is);
- stuðla að samstarfi háskóla, stofnana og fyrirtækja um máltækniverkefni;
- skipuleggja og samhæfa háskólakennslu á sviði máltækni;
- taka þátt í norrænu, evrópsku og alþjóðlegu samstarfi á sviði máltækni;
- standa fyrir og eiga aðild að rannsóknar- og þróunarverkefnum á sviði máltækni;
- halda utan um ýmiss konar hráefni og afurðir á sviði máltækni;
- halda máltækniráðstefnur með þátttöku fræðimanna, fyrirtækja og almennings;
- beita sér fyrir eflingu íslenskrar máltækni á öllum sviðum.

Á undanförnum árum hafa fræðimenn Máltækniseturs átt frumkvæðið að nokkrum nýjum verkefnum sem hafa verið styrkt að hluta til af Rannsóknasjóði og Tækniþróunarsjóði. Mikilvægasta afurð þessara verkefna er opni hugbúnaðurinn IceNLP (málfræðilegi markarinn IceTagger, hlutaþáttarinn IceParser, og lemmunarforritið Lemmald) [53], sem hægt er að nota á vefnum

(http://nlp.cs.ru.is). Árið 2009 fékk Máltæknisetur háan þriggja ára öndvegisstyrk frá Rannís til verkefnisins „Hagkvæm máltækni utan ensku – íslenska tilraunin“. Innan þessa verkefnis var unnið áfram að því að byggja upp grunnstoðir íslenskrar máltækni.
Eins og hér hefur komið fram hafa margvísleg verkefni leitt til þróunar ýmissa máltæknitóla og málfanga fyrir íslensku. Hér á eftir er gefið yfirlit yfir núverandi stöðu íslenskrar máltækni.

4.6 AÐGENGI AÐ MÁLTÆKNITÓLUM OG MÁLFÖNGUM

Á mynd 8 er gefið yfirlit yfir stöðu íslenskrar máltækni og máltæknibúnaðar. Einkunnir máltæknitóla og málfanga eru byggðar á mati helstu sérfræðinga á sviðinu sem gáfu einkunnir á skalanum frá 0 (mjög lágt) til 6 (mjög hátt) út frá sjö viðmiðum.
Meginniðurstöður fyrir íslensku eru eftirfarandi:

- Íslenska stendur þokkalega hvað varðar einföldustu grunnforsendur máltækninnar í búnaði og málföngum, svo sem textagreiningu og málheildum.
- Einnig eru til einstöku gögn og búnaður með takmarkaða virkni á sviðum eins og talgervingu, talkennslum, vélþýðingum, talmálsheildum, hliðstæðum málheildum og orðagögnum.
- Háþróaður máltæknibúnaður og málföng, svo sem til textatúlkunar og málmyndunar, er ekki til.

Um síðustu aldamót var íslensk máltækni varla til. Þetta breyttist eftir 1999, þegar sérstakur starfshópur skilaði skýrslu um máltækni til menntamálaráðherra [3]. Í þessari skýrslu voru gerðar tillögur um ýmsar aðgerðir til að koma íslenskri máltækni á laggirnar. Starfshópurinn áætlaði að það myndi kosta u.þ.b. einn milljarð króna (sem þá jafngilti um 10 milljónum evra) að gera íslenska máltækni sjálfbæra. Þegar því marki væri náð ætti markaðurinn að geta tekið við þar eð hann hefði aðgang að opnum málföngum sem hefði verið komið upp á vegum máltækniáætlunar ríkisstjórnarinnar og yrðu afhent á jafnréttisgrundvelli til allra sem hygðust nýta þau í markaðsvörum.
Það verður að benda á að heildarfjármagnið sem veitt var til máltækniáætlunarinnar frá 2000–2004 var aðeins um 1/8 af þeirri upphæð sem áðurnefndur starfshópur taldi að þyrfti til [2]. Það þarf því ekki að koma á óvart að íslensk máltækni er enn á bernskuskeiði. 330.000 málnotendur eru ekki nægilegur fjöldi til að standa undir kostnaðarsamri þróun á nýjum vörum. Um þessar mundir vinna nánast engin íslensk fyrirtæki að máltækni vegna þess að þau sjá enga hagnaðarvon í henni. Því er ákaflega mikilvægt að halda áfram opinberum stuðningi við íslenska máltækni enn um sinn.

4.7 SAMANBURÐUR TUNGUMÁLA

Máltæknistuðningur er mjög mismunandi milli málsamfélaga. Til að bera saman stöðuna milli mála er í þessum kafla sett fram mat sem byggist á tveimur verkbúnaðarsviðum (vélþýðingum og talvinnslu), einni gerð baklægrar tækni (textagreiningu) og grundvallarmálföngum sem þarf til smíði máltæknibúnaðar. Málunum var raðað á fimm bila kvarða.

1. Afburðagóður stuðningur
2. Góður stuðningur
3. Sæmilegur stuðningur
4. Brotakenndur stuðningur
5. Lítill sem enginn stuðningur

Máltæknistuðningur var metinn út frá eftirfarandi viðmiðunum:

Talvinnsla: Gæði fyrirliggjandi talkennslatækni, gæði fyrirliggjandi talgervingartækni, yfirgrip sviða, fjöldi og

	Magn	Aðgengi	Gæði	Yfirgrip	Þroski	Sjálfbærni	Aðlögunarhæfni
Máltækni: tól, tækni og verkbúnaður							
Talkennsl	1	1	1	1.5	1	0	1
Talgerving	1	1	2.5	2.5	2	1	1
Málfræðigreining	2	5.5	4	3	3.5	3.5	3
Merkingargreining	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Málmyndun	0	0	0	0	0	0	0
Vélþýðingar	1	4	1	1.5	1.5	1.5	2
Málföng: tilföng, gögn og þekkingargrunnar							
Málheildir	1.5	4	3	2.5	2.5	4.5	3
Talmálsheildir	1	2	1.5	1.5	1	1.5	1.5
Hliðstæðar málheildir	1	1	1	0.5	1	1	1
Orðaföng	1	2	2.5	2.5	2	2	2
Málfræðilýsingar	1	4	2.5	2	2.5	2.5	2

8: Staða máltæknistuðnings við íslensku

stærð fyrirliggjandi talmálsheilda, magn og fjölbreytni fyrirliggjandi talbúnaðar.

Vélþýðingar: Gæði fyrirliggjandi vélþýðingartækni, fjöldi tungumálapara sem vélþýðing tekur til, yfirgrip málfræðiatriða og svið málsins, gæði og stærð fyrirliggjandi samhliða málheilda, magn og fjölbreytni vélþýðingarbúnaðar.

Textagreining: Gæði og yfirgrip fyrirliggjandi textagreiningartækni (beyging og orðmyndun, setningagerð, merkingar), yfirgrip málfræðiatriða og svið málsins, magn og fjölbreytni fyrirliggjandi verkbúnaðar, gæði og stærð fyrirliggjandi (markaðra) textamálheilda, gæði og yfirgrip fyrirliggjandi orðafanga (svo sem WordNet) og málfræðilýsinga.

Málföng: Gæði og stærð fyrirliggjandi textamálheilda, talmálsheilda og samhliða málheilda, gæði og yfirgrip fyrirliggjandi orðafanga og málfræðilýsinga.

Á myndum 9 til 12 sést að íslenska er í lægsta klasanum hvað varðar öll tól og málföng sem um ræðir. Hún er þar á sömu slóðum og önnur tungumál sem fáir tala, svo sem írska, lettneska, litháíska og maltneska. Þessi tungumál eru langt að baki stórþjóðamálum eins og t. d. þýsku og frönsku. En jafnvel málföng og máltæknitól fyrir þau tungumál ná hvorki sömu gæðum né yfirgripi og hliðstæð föng og tól fyrir ensku, sem er í fararbroddi á nær öllum sviðum máltækninnar. Þó eru enn fjölmargar eyður í enskum málföngum hvað varðar hágæða búnað.

4.8 NIÐURSTÖÐUR

Í þessari ritröð hefur verið gerð mikilvæg upphafsatlaga að því að meta máltæknistuðning fyrir 30 Evrópumál, og gera ítarlegan samanburð á þessum málum. Með því að greina eyður, þarfir og skort er evrópskt máltæknisamfélag og aðrir hagsmunaaðilar nú í stöðu til þess að skipuleggja meiri háttar rannsóknar- og þróunaráætlun sem miðast að því að Evrópa verði raunverulega margmála með stuðningi tækninnar.

Hér hefur komið fram geysimikill innbyrðis munur á Evrópumálum. Þótt ágætur hugbúnaður og málföng sé til fyrir sum tungumál og verksvið eru grundvallareyður á þessum sviðum í öðrum málum (venjulega þeim „smærri"). Mörg tungumál skortir grunntækni til textagreiningar og nauðsynleg málföng til að þróa slíka tækni. Önnur hafa grundvallarbúnað og málföng en hafa ekki burði til að ráðast í merkingarlega vinnslu. Þess vegna er enn þörf á víðtæku átaki til að ná því metnaðarfulla markmiði að koma upp hágæða vélþýðingum milli allra Evrópumála.

Fyrir lítið málsamfélag og lítið rannsóknarumhverfi eins og það íslenska er samvinna lífsnauðsyn – ekki bara innanlands heldur einnig alþjóðleg. Þess er að vænta að þátttaka Íslands í META-NORD og META-NET muni gera mögulegt að þróa, staðla og gera aðgengileg ýmis mikilvæg málföng og stuðla þannig að vexti og viðgangi íslenskrar máltækni.

Langtímamarkmið META-NET er að koma upp hágæða máltækni fyrir öll tungumál. Þetta krefst þess að allir hagsmunaaðilar – í stjórnmálum, rannsóknum, viðskiptum, og samfélaginu öllu – sameini krafta sína. Tæknin sem út úr þeirri samvinnu kemur mun hjálpa til við að brjóta múra og smíða brýr milli Evrópumála og skapa þannig pólitíska og efnahagslega einingu úr menningarlegum fjölbreytileika.

Afburðagóður stuðningur	Góður stuðningur	Sæmilegur stuðningur	Brotakenndur stuðningur	Lítill/enginn stuðningur
	Enska	Finnska Franska Hollenska Ítalska Portúgalska Spænska Tékkneska Þýska	Baskneska Búlgarska Danska Eistneska Galisíska Gríska Írska Katalónska Norska Pólska Serbneska Slóvakíska Slóvenska Sænska Ungverska	**Íslenska** Króatíska Lettneska Litháíska Maltneska Rúmenska

9: Talvinnsla: Staða máltæknistuðnings við 30 Evrópumál

Afburðagóður stuðningur	Góður stuðningur	Sæmilegur stuðningur	Brotakenndur stuðningur	Lítill/enginn stuðningur
	Enska	Franska Spænska	Hollenska Ítalska Katalónska Pólska Rúmenska Ungverska Þýska	Baskneska Búlgarska Danska Eistneska Finnska Galisíska Gríska Írska **Íslenska** Króatíska Lettneska Litháíska Maltneska Norska Portúgalska Serbneska Slóvakíska Slóvenska Sænska Tékkneska

10: Vélþýðingar: Staða máltæknistuðnings við 30 Evrópumál

Afburðagóður stuðningur	Góður stuðningur	Sæmilegur stuðningur	Brotakenndur stuðningur	Lítill/enginn stuðningur
	Enska	Franska Hollenska Ítalska Spænska Þýska	Baskneska Búlgarska Danska Finnska Galisíska Gríska Katalónska Norska Pólska Portúgalska Rúmenska Slóvakíska Slóvenska Sænska Tékkneska Ungverska	Eistneska Írska **Íslenska** Króatíska Lettneska Litháíska Maltneska Serbneska

11: Textagreining: Staða máltæknistuðnings við 30 Evrópumál

Afburðagóður stuðningur	Góður stuðningur	Sæmilegur stuðningur	Brotakenndur stuðningur	Lítill/enginn stuðningur
	Enska	Franska Hollenska Ítalska Pólska Spænska Sænska Tékkneska Ungverska Þýska	Baskneska Búlgarska Danska Eistneska Finnska Galisíska Gríska Katalónska Króatíska Norska Portúgalska Rúmenska Serbneska Slóvakíska Slóvenska	Írska **Íslenska** Lettneska Litháíska Maltneska

12: Málföng: Staða máltæknistuðnings við 30 Evrópumál

UM META-NET

META-NET er öndvegisnet fjármagnað af Evrópusambandinu að hluta til. Þátttakendur í því eru nú 54 rannsóknasetur í 33 Evrópulöndum. META-NET stendur að Tæknibandalagi um margmála Evrópu (Multilingual Europe Technology Alliance, META), sem er sístækkandi samfélag evrópskra fræðimanna og stofnana á sviði máltækni. META-NET fóstrar tæknilega undirstöðu raunverulegs margmála evrópsks upplýsingasamfélags sem:

- gerir kleift að eiga samskipti og samvinnu þvert á tungumál;
- tryggir öllum Evrópubúum jafnan aðgang að upplýsingum og þekkingu óháð því hvaða mál þeir tala;
- byggir á netvæddri upplýsingatækni og eflir virkni hennar.

Netið styður við einingu Evrópu í einum stafrænum markaði og upplýsingarými. Það ýtir undir og kemur á framfæri margmála tækni fyrir öll Evrópumál. Sú tækni raungerir sjálfvirkar þýðingar, samningu efnis, upplýsingavinnslu og þekkingarstjórnun fyrir fjölbreyttan búnað og margvísleg efnissvið. Hún gerir líka kleift að nota auðskiljanlegt mállegt notendaviðmót á margs konar tækni allt frá heimilistækjum og ökutækjum til tölva og vélmenna.

META-NET var sett af stað 1. febrúar 2010 og hefur þegar staðið fyrir ýmsum aðgerðum innan þriggja aðgerðaáætlana sinna; META-VISION, META-SHARE og META-RESEARCH.

META-VISION fóstrar kvikt og áhrifamikið samfélag hagsmunaaðila sem sameinast um sameiginlega sýn og útfærða rannsóknarstefnu (strategic research agenda, SRA). Megináherslan í þessu starfi er að byggja upp samræmt og samtengt máltæknisamfélag í Evrópu með því að leiða saman fulltrúa dreifðra og fjölbreyttra hópa hagsmunaaðila. Þessi hvítbók var samin í tengslum við bækur fyrir 29 önnur tungumál. Hin sameiginlega tæknisýn var þróuð í þremur rýnihópum sem höfðu með sér verkaskiptingu. Tækniráði META var komið á fót til að ræða og undirbúa rannsóknarstefnuna sem byggð er á þessari sýn, í náinni samvinnu við allt máltæknisamfélagið.

META-SHARE skapar opinn og dreifðan vettvang til að skiptast á gögnum og deila þeim. Net jafnrétthárra gagnabrunna mun hýsa málleg gögn, tól og vefþjónustu sem allt verður skjalað með hágæða lýsigögnum og skipulagt í stöðluðum flokkum. Auðvelt verður að nálgast gögnin og leita í þeim í heild. Þarna verða bæði opin og ókeypis málföng sem og gögn með takmörkuðum aðgangi sem greiða verður fyrir notkun á.

META-RESEARCH smíðar brýr til skyldra tæknisviða. Í þessu starfi er leitast við að nýta framfarir á öðrum sviðum og einbeita sér að nýskapandi rannsóknum sem geta eflt máltækni. Einkum er stefnt að því að gera brautryðjendarannsóknir í vélþýðingum, safna gögnum, útbúa gagnasett og skipuleggja málföng til að nota við mat; gera skrár um tól og aðferðir; og skipuleggja vinnustofur og þjálfun fyrir þátttakendur í máltæknisamfélaginu.

office@meta-net.eu – http://www.meta-net.eu

1

EXECUTIVE SUMMARY

Information technology changes our everyday lives. We typically use computers for writing, editing, calculating, and information searching, and increasingly for reading, listening to music, viewing photos and watching movies. We carry small computers in our pockets and use them to make phone calls, write emails, get information and entertain ourselves, wherever we are. How does this massive digitisation of information, knowledge and everyday communication affect our language? Will our language change or even disappear? What are the Icelandic language's chances of survival?

Many of the world's 6,000 languages will not survive in a globalised digital information society. It is estimated that at least 2,000 languages are doomed to extinction in the decades ahead. Others will continue to play a role in families and neighbourhoods, but not in the wider business and academic world. The status of a language depends not only on the number of speakers or books, films and TV stations that use it, but also on the presence of the language in the digital information space and software applications.

In this context, Icelandic is not very well off. At the end of the 20th century, Icelandic language technology was virtually non-existent. There was a relatively good spell checker, a not-so-good speech synthesiser, and that was all. There were no programs or even individual courses on language technology or computational linguistics at any Icelandic university or college, there was no ongoing research in these areas, and no Icelandic software companies were working on language technology [2].

Things started to change after a specially appointed Expert Group delivered a white paper on Language Technology to the Minister of Education, Science and Culture in 1999 [3]. In this white paper, several actions to establish Icelandic language technology were proposed. In 2000, the Government launched a special Language Technology Programme, with the aim of supporting institutions and companies in creating basic resources for Icelandic language technology work. This initiative resulted in a number of projects which have laid the groundwork for Icelandic language technology [2].

After the Language Technology Programme ended in 2004, researchers from three institutes (University of Iceland, Reykjavik University, and the Árni Magnússon Institute for Icelandic Studies), who had been involved in most of the projects funded by the programme, decided to join forces in a consortium called the Icelandic Centre for Language Technology (ICLT) [4], in order to follow up on the tasks of the programme. Since 2005, the ICLT researchers have initiated several new projects which have been partly supported by the Icelandic Research Fund and the Icelandic Technical Development Fund.

The present report reveals that despite considerable achievements in the last decade, it is only with respect to the most basic tools and resources such as tokenisers, part-of-speech taggers, morphological analysers/generators, syntactic parsers, reference corpora, and syntax corpora, that the situation for Icelandic is reasonably good. When it comes to advanced fields like sentence and text semantics, advanced discourse

processing, information retrieval, language generation, summarisation, dialogue management, semantics and discourse corpora, ontological resources, etc., no tools and resources exist for Icelandic. Thus, it is clear that we still have a long way to go to ensure the future of Icelandic as a full-fledged player in the modern – and future – European information society.

Information and communication technology are now preparing for the next revolution. After personal computers, networks, miniaturisation, multimedia, mobile devices and cloud-computing, the next generation of technology will feature software that understands not just spoken or written letters and sounds but entire words and sentences, and supports users far better because it speaks, knows and understands their language. Forerunners of such developments are the free online service Google Translate that translates between 57 languages, IBM's supercomputer Watson that was able to defeat the US-champion in the game of "Jeopardy", and Apple's mobile assistant Siri for the iPhone that can react to voice commands and answer questions in English, German, French and Japanese.

The next generation of information technology will master human language to such an extent that human users will be able to communicate using the technology in their own language. Devices will be able to automatically find the most important news and information from the world's digital knowledge store in reaction to easy-to-use voice commands. Language-enabled technology will be able to translate automatically or assist interpreters; summarise conversations and documents; and support users in learning scenarios. For example, it will help immigrants to learn the Icelandic language and integrate more fully into the country's culture.

The next generation of information and communication technologies will enable industrial and service robots (currently under development in research laboratories) to faithfully understand what their users want them to do and then proudly report on their achievements. This level of performance means going way beyond simple character sets and lexicons, spell checkers and pronunciation rules. The technology must move on from simplistic approaches and start modelling language in an all-encompassing way, taking syntax as well as semantics into account to understand the drift of questions and generate rich and relevant answers.

Not all European languages are equally well prepared for this future. This report presents an evaluation of the status of language technology support for 30 European languages, based on four key areas: machine translation, speech processing, text analysis, as well as basic resources needed for building language technology applications. The languages were grouped into five clusters. Unsurprisingly, Icelandic is in the bottom cluster for all of the tools and resources listed. It compares well with other languages with a small number of speakers, such as Irish, Latvian, Lithuanian, and Maltese. These languages lag far behind large languages like German and French, for instance. But even language technology resources and tools for those languages clearly do not yet reach the quality and coverage of comparable resources and tools for the English language, which is in the lead in almost all language technology areas.

What needs to be done in order to ensure the future of the Icelandic language in the information society? In 1999, the Language Technology Expert Group estimated that it would cost around one billion Icelandic krónas (which then amounted to about ten million Euros) to make Icelandic language technology self-sustained. After that, the free market should be able to take over, since it would have access to public resources that would have been created by the government-funded Language Technology Programme, and that would be made available on an equal basis to everyone who was going to use these resources in their commercial products [3].

Even though the Language Technology Programme was successful and had a great impact on the development of Icelandic language technology, the fact remains that its total budget from 2000–2004 was only around 1/8 of the sum that the expert group estimated would be needed [2]. It should therefore come as no surprise that Icelandic language technology is still in its infancy. 330,000 speakers are simply to few to sustain costly development of new products. At present, almost no companies are working in the language technology area because they do not see it as profitable. Continued public support for Icelandic language technology is necessary in order to guarantee exploitation of the tools already developed and the knowledge and experience of researchers and companies which has already been accrued.

The Icelandic language is not in imminent danger, even from the prowess of English language computing. However, the whole situation could change dramatically when a new generation of technologies really starts to master human languages effectively. Through improvements in machine translation, language technology will help in overcoming language barriers, but it will only be able to operate between those languages that have managed to survive in the digital world. If there is adequate language technology available, then it will be able to ensure the survival of languages with very small populations of speakers. If not, even 'larger' languages will come under severe pressure. If Icelandic is to survive as a viable national language in the developed world, it must be able to meet IT demands. Consequently, investment in language technology must form an essential part of its language preservation policy.

META-NET's vision is high-quality language technology for all languages that supports political and economic unity through cultural diversity. This technology will help tear down existing barriers and build bridges between Europe's languages. This requires all stakeholders – in politics, research, business, and society – to unite their efforts for the future.

This white paper series complements the other strategic actions taken by META-NET. Up-to-date information such as the current version of the META-NET vision paper [5] or the Strategic Research Agenda (SRA) can be found on the META-NET web site: http://www.meta-net.eu.

2

LANGUAGES AT RISK: A CHALLENGE FOR LANGUAGE TECHNOLOGY

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in information and communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

The digital revolution is comparable to Gutenberg's invention of the printing press.

After Gutenberg's invention, real breakthroughs in communication were accomplished by efforts such as Luther's translation of the Bible into vernacular language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled exchanges across languages;
- the creation of editorial and bibliographic guidelines assured the quality of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology has helped to automate and facilitate many processes:

- desktop publishing software has replaced typewriting and typesetting;
- Microsoft PowerPoint has replaced overhead projector transparencies;
- e-mail allows documents to be sent and received more quickly than using a fax machine;
- Skype offers cheap Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- web search engines provide keyword-based access;
- online services like Google Translate produce quick, approximate translations;
- social media platforms such as Facebook, Twitter and Google+ facilitate communication, collaboration, and information sharing.

Although these tools and applications are helpful, they are not yet capable of supporting a fully-sustainable, multilingual European society in which information and goods can flow freely.

2.1 LANGUAGE BORDERS HOLD BACK THE EUROPEAN INFORMATION SOCIETY

We cannot predict exactly what the future information society will look like. However, there is a strong likelihood that the revolution in communication technology is bringing together people who speak different languages in new ways. This is putting pressure both on individuals to learn new languages and especially on developers to create new technology applications to ensure mutual understanding and access to shareable knowledge. In the global economic and information space, there is increasing interaction between different languages, speakers and content thanks to new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter, YouTube, and, recently, Google+) is only the tip of the iceberg.

The global economy and information space confronts us with different languages, speakers and content.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognise that it is in a language that we do not understand. According to a recent report from the European Commission, 57% of Internet users in Europe purchase goods and services in non-native languages; English is the most common foreign language followed by French, German and Spanish. 55% of users read content in a foreign language while 35% use another language to write e-mails or post comments on the Web [6]. A few years ago, English might have been the lingua franca of the Web—the vast majority of content on the Web was in English—but the situation has now drastically changed. The amount of online content in other European (as well as Asian and Middle Eastern) languages has exploded. Surprisingly, this ubiquitous digital linguistic divide has not gained much public attention; yet, it raises a very pressing question: Which European languages will thrive in the networked information and knowledge society, and which are doomed to disappear?

2.2 OUR LANGUAGES AT RISK

While the printing press helped step up the exchange of information in Europe, it also led to the extinction of many European languages. Regional and minority languages were rarely printed and languages such as Cornish and Dalmatian were limited to oral forms of transmission, which in turn restricted their scope of use. Will the Internet have the same impact on our modern languages? Europe's approximately 80 languages are one of our richest and most important cultural assets, and a vital part of this unique social model [7]. While languages such as English and Spanish are likely to survive in the emerging digital marketplace, many European languages could become irrelevant in a networked society. This would weaken Europe's global standing, and run counter to the strategic goal of ensuring equal participation for every European citizen regardless of language.

According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society [8].

The variety of languages in Europe is one of its richest and most important cultural assets.

2.3 LANGUAGE TECHNOLOGY IS A KEY ENABLING TECHNOLOGY

In the past, investments in language preservation focussed primarily on language education and translation. According to one estimate, the European market for translation, interpretation, software localisation and website globalisation was €8.4 billion in 2008 and is expected to grow by 10% per annum [9]. Yet this figure covers just a small proportion of current and future needs in communicating between languages. The most compelling solution for ensuring the breadth and depth of language usage in Europe tomorrow is to use appropriate technology, just as we use technology to solve our transport and energy needs among others.

Language technology targeting all forms of written text and spoken discourse can help people to collaborate, conduct business, share knowledge and participate in social and political debate regardless of language barriers and computer skills. It often operates invisibly inside complex software systems to help us already today to:

- find information with a search engine;
- check spelling and grammar in a word processor;
- view product recommendations in an online shop;
- follow the spoken directions of a navigation system;
- translate web pages via an online service.

Language technology consists of a number of core applications that enable processes within a larger application framework. The purpose of the META-NET language white papers is to focus on how ready these core enabling technologies are for each European language.

Europe needs robust and affordable language technology for all European languages.

To maintain our position in the frontline of global innovation, Europe will need language technology, tailored to all European languages, that is robust and affordable and can be tightly integrated within key software environments. Without language technology, we will not be able to achieve a really effective interactive, multimedia and multilingual user experience in the near future.

2.4 OPPORTUNITIES FOR LANGUAGE TECHNOLOGY

In the world of print, the technology breakthrough was the rapid duplication of an image of a text using a suitably powered printing press. Human beings had to do the hard work of looking up, assessing, translating, and summarising knowledge. We had to wait until Edison to record spoken language – and again his technology simply made analogue copies.

Language technology can now simplify and automate the processes of translation, content production, and knowledge management for all European languages. It can also empower intuitive speech-based interfaces for household electronics, machinery, vehicles, computers and robots. Real-world commercial and industrial applications are still in the early stages of development, yet R&D achievements are creating a genuine window of opportunity. For example, machine translation is already reasonably accurate in specific domains, and experimental applications provide multilingual information and knowledge management, as well as content production, in many European languages.

As with most technologies, the first language applications such as voice-based user interfaces and dialogue systems were developed for specialised domains, and often exhibit limited performance. However, there are huge market opportunities in the education and entertainment industries for integrating language technologies into games, edutainment packages, libraries, simu-

lation environments and training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just some of the application areas in which language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology helps overcome the "disability" of linguistic diversity.

Language technology represents a tremendous opportunity for the European Union. It can help to address the complex issue of multilingualism in Europe – the fact that different languages coexist naturally in European businesses, organisations and schools. However, citizens need to communicate across the language borders of the European Common Market, and language technology can help overcome this final barrier, while supporting the free and open use of individual languages. Looking even further ahead, innovative European multilingual language technology will provide a benchmark for our global partners when they begin to support their own multilingual communities. Language technology can be seen as a form of "assistive" technology that helps overcome the "disability" of linguistic diversity and makes language communities more accessible to each other. Finally, one active field of research is the use of language technology for rescue operations in disaster areas, where performance can be a matter of life and death: Future intelligent robots with cross-lingual language capabilities have the potential to save lives.

2.5 CHALLENGES FACING LANGUAGE TECHNOLOGY

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. Widely-used technologies such as the spelling and grammar correctors in word processors are typically monolingual, and are only available for a handful of languages.

Technological progress needs to be accelerated.

Online machine translation services, although useful for quickly generating a reasonable approximation of a document's contents, are fraught with difficulties when highly accurate and complete translations are required. Due to the complexity of human language, modelling our tongues in software and testing them in the real world is a long, costly business that requires sustained funding commitments. Europe must therefore maintain its pioneering role in facing the technological challenges of a multiple-language community by inventing new methods to accelerate development right across the map. These could include both computational advances and techniques such as crowdsourcing.

2.6 LANGUAGE ACQUISITION IN HUMANS AND MACHINES

To illustrate how computers handle language and why it is difficult to program them to process different tongues, let's look briefly at the way humans acquire first and second languages, and then see how language technology systems work.

Humans acquire language skills in two different ways. Babies acquire a language by listening to the real interactions between their parents, siblings and other family members. From the age of about two, children produce

their first words and short phrases. This is only possible because humans have a genetic disposition to imitate and then rationalise what they hear.

Learning a second language at an older age requires more cognitive effort, largely because the child is not immersed in a language community of native speakers. At school, foreign languages are usually acquired by learning grammatical structure, vocabulary and spelling using drills that describe linguistic knowledge in terms of abstract rules, tables and examples.

Humans acquire language skills in two different ways: learning from examples and learning the underlying language rules.

Moving now to language technology, the two main types of systems 'acquire' language capabilities in a similar manner. Statistical (or 'data-driven') approaches obtain linguistic knowledge from vast collections of concrete example texts. While it is sufficient to use text in a single language for training, e. g., a spell checker, parallel texts in two (or more) languages have to be available for training a machine translation system. The machine learning algorithm then "learns" patterns of how words, short phrases and complete sentences are translated.

This statistical approach usually requires millions of sentences to boost performance quality. This is one reason why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, and services such as Google Search and Google Translate, all rely on statistical approaches. The great advantage of statistics is that the machine learns quickly in a continuous series of training cycles, even though quality can vary randomly.

The second approach to language technology, and to machine translation in particular, is to build rule-based systems. Experts in the fields of linguistics, computational linguistics and computer science first have to encode grammatical analyses (translation rules) and compile vocabulary lists (lexicons). This is very time consuming and labour intensive. Some of the leading rule-based machine translation systems have been under constant development for more than 20 years. The great advantage of rule-based systems is that the experts have more detailed control over the language processing. This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. However, due to the high cost of this work, rule-based language technology has so far only been developed for a few major languages.

The two main types of language technology systems acquire language in a similar manner.

As the strengths and weaknesses of statistical and rule-based systems tend to be complementary, current research focusses on hybrid approaches that combine the two methodologies. However, these approaches have so far been less successful in industrial applications than in the research lab.

As we have seen in this chapter, many applications widely used in today's information society rely heavily on language technology, particularly in Europe's economic and information space. Although this technology has made considerable progress in the last few years, there is still huge potential to improve the quality of language technology systems. In the next section, we describe the role of Icelandic in European information society and assess the current state of language technology for the Icelandic language.

3

THE ICELANDIC LANGUAGE IN THE EUROPEAN INFORMATION SOCIETY

3.1 GENERAL FACTS

Approximately 330,000 people have Icelandic as their first language. Most of them live in Iceland [10], with other speakers of the language primarily being Icelanders living abroad [11], in places such as Scandinavia, Mainland Europe and North America. Additionally, Icelandic is the first language of some second and third generation Icelanders in Canada and the United States [12], but most of these speakers are seventy years or older by now. With increased immigration to the country the number of second-language speakers of Icelandic has grown, although it is still considerably small.

Icelandic is the language of government and administration, all levels of the school system, and the language of business and general day-to-day interactions in Iceland.

The Icelandic Parliament (Alþingi) has just passed a law in which Icelandic is given the status of official language [13]. It is the language of government and administration, all levels of the school system, and the language of business and general day-to-day interactions in Iceland. The language has only very minor dialectal variations with the most common being the pronunciation of intervocalic stops, which are aspirated in the north but non-aspirated in other parts of the country, in words like *æpa* [scream], *vita* [know], and *taka* [take]. It is the only phonetic dialectal variant that is still strong, as others are slowly disappearing, such as the pronunciation of voiced sonorants before stops in words like *úlpa* [coat], *svampur* [sponge], *vanta* [lack]; pronunciation of monophthongs before *ng/nk* in words like *banki* [bank]; and the so-called *hv*-pronunciation where an unvoiced velar fricative is used instead of a velar aspirated stop in words like *hvar* [where] [14]. However, a new dialectal variation may be emerging – the affrication of *tj* in words like *tjald* [tent] [15]. Dialectal differences in the syntax are minimal and most are not geographically conditioned. On the other hand, a few syntactic innovations seem to be gaining ground among younger speakers, especially the 'new passive', such as *það var barið mig* [it was hit me] instead of *ég var barin(n)* [I was hit]; and the 'extended progressive' *vera að* (be in the process of), such as *ég er ekki að skilja þetta* [I am not understanding this], *þeir voru að spila mjög vel* [they were playing really well]. These constructions are usually not accepted by middle-aged and older speakers.

The Icelandic language spoken in the immigrant communities of North America could be categorised as a separate dialect (or dialects), as the vocabulary evolved differently in these communities. The words for 'telephone' and 'car' are for instance *telefón* and *kar* in Canadian Icelandic but *sími* and *bíll* in the language spoken in Iceland. Similarly, certain Icelandic words and pronunciation patterns became fossilised or even increased in Canada while they disappeared in Iceland. An example of this is the flámæli (where the front vowels *i/u* and *e/ö*, respectively, get mixed up).

3.2 PARTICULARITIES OF THE ICELANDIC LANGUAGE

Icelandic is a Germanic language, more specifically Northern Germanic, and belongs to the branch of Insular Scandinavian languages along with Faroese. It is a SVO (subject-verb-object) language with a strong V2 rule that requires the verb to appear in the second (or first) position of the sentence. However, because of the rich inflectional system word order is relatively free; certain words can be moved around without the meaning of the sentence being lost. The following sentences both convey the same meaning even though the order of the subject and object has been switched:

- Hundurinn (nominative) beit köttinn (accusative) [the dog bit the cat].
- Köttinn (accusative) beit hundurinn (nominative) [the dog bit the cat].

Icelandic is a SVO language with the finite verb in the second (or first) position of the sentence, but the word order is however relatively free.

Icelandic is among the relatively few languages in which the grammatical subject of a sentence can stand in other cases than the nominative – most often in the dative, but also in the accusative (and in a few cases in the genitive). Thus, the first person singular pronoun is the subject of all the following sentences, even if it stands in a different case in each of them:

- Ég las bókina [I (nominative) read the book].
- Mig vantar bókina [I (accusative) need the book].
- Mér líkar bókin [I (dative) like the book].

Icelandic is a highly inflected (synthetic) language with four cases, three genders, and two numbers for nouns, pronouns, adjectives and the definite (suffixed) article, while no indefinite article exists in the language. Additionally, adjectives decline in both weak (definite) and strong (indefinite) form. Verbs in the language are conjugated for person, number, tense, mood and voice. The language is fusional, such that a single ending usually stands for more than one morphological category. The inflectional system is further complicated by a great number of inflectional and conjugational classes, such that the same morphological category, or combination of categories, is represented by a number of different endings depending on the stem.

The vocabulary of Icelandic is mostly of Germanic origin.

The vocabulary is mostly of Norse (Germanic) origin, even though numerous loan words have entered the language through the eleven centuries since the country was settled. After Christianity was adopted in the year 1000, a number of words were borrowed from Latin. The reformation in 1550 brought with it influence from German through the translation of religious books and psalms. Iceland was under Danish rule from 1380 till 1944 and the influence of the Danish language on the vocabulary of Icelandic was considerable. Danish words were adopted and many of them became a natural part of the language. These are words such as *gardínur* from Danish *gardin* [curtains] and *viskustykki* from Danish *viskestykke* [dish towel].

Icelandic is extremely productive when it comes to coining new words.

The official policy is to coin words for new things and concepts from domestic material instead of borrowing foreign (international) words or terms. Icelandic has a

number of sound alternations which can be used to derive new words from existing ones, such as *leysni* [solubility] from *lausn* [solution], and also productive suffixes which can be combined with existing roots to coin new words, such as *disk-lingur* [diskette] from *diskur* [disk]. However, the most common method for coining new words is to use compounding, such as *stafsetningar-orða-bók* for [spelling dictionary] and *umhverfis-mála-ráðu-neyti* for [ministry of environment]. This makes the language both vivid and transparent.

The pronunciation of the language is fairly transparent, in the sense that the projection from the spelling to the pronunciation obeys almost unexceptional rules. Thus, a speaker who knows the rules should be able to pronounce relatively accurately any new words that (s)he reads – provided, admittedly, that (s)he detects possible morpheme boundaries which can affect the pronunciation of some letter combinations. The stress rule is also very simple as the main stress of the word is always on the first syllable, usually with additional stress on every second syllable after that, although this does not always hold for compound words.

Written Icelandic is based on the Latin alphabet but nevertheless uses a few of its own characters not used in, for instance, English. These are the character Þ/þ (only used in Icelandic, although it originated in Old English), Ð/ð (also used in Faroese), Æ/æ (also used in Norwegian, Danish and Faroese) and Ö/ö (also used in Swedish, Finnish, Estonian, German, and Hungarian). In addition, Icelandic uses the accented vowels Á/á, É/é, Í/í, Ó/ó, Ú/ú and Ý/ý.

The written language has changed considerably little since Old Norse, which makes it possible for Icelanders to read Old Norse texts with a little practice. The main changes in spelling in recent decades have been the abolishment of z from the language (except in proper names like *Zóphónías* and family names like *Haralz*), as well as confirming the use of *é* instead of the former variant *je*.

3.3 RECENT DEVELOPMENTS

Since the British, and later American, occupation of Iceland during World War II, Icelandic has been influenced more by English than Danish, and these influences have only become stronger with the influx of British and American music, movies and television shows into the country. The rise of the Internet has also lead to the increased use of English, with about 95% of the population online.

The influence of English is most obvious in an increasing number of loan words. However, the majority of these words are usually considered to be substandard – they are not found in dictionaries, are rarely seen in print, and are frowned upon by language authorities. Their use is mostly confined to the spoken language and unofficial or personal writings like e-mail, weblogs, etc.

English loan words are common in the spoken language but much less so in the written language.

On the other hand, English influence on the linguistic system itself appears to be negligible. Many of the English loan words used in everyday language get Icelandic inflectional endings, although some of them remain uninflected, such as *næs* [nice], *kúl* [cool], etc. It has been claimed that some emerging developments in the syntax and the phonology of Icelandic, such as the 'extended progressive' and the affrication of *tj* mentioned above, can be traced to English influence, but this is disputed.

In recent years, the concept of 'domain loss' has been a major issue in the linguistic discussion in Iceland as in many other countries. The Icelandic job market has become increasingly more international in the last few years, with Icelandic companies working outside of the country as well as international companies working within it. This has lead to the greater use of English in

the daily life and running of these companies, with correspondence and meetings taking place in English. Year-end reports, websites and other materials are also often published in English, as well as in Icelandic, and in some cases only in English. It has also been a growing trend for Icelandic companies to bear English names, either in full or half. These are names such as *Icelandair*, *Actavis*, *Baugur Group* and *Stoðir Invest* [16].

Another domain in which English has become predominant is that of information technology, which will be discussed in the next chapter.

3.4 OFFICIAL LANGUAGE PROTECTION IN ICELAND

The official goal of language planning in Iceland has long been that of preservation and strengthening. This has particularly been clear in the development of vocabulary, realised in organised terminological work, mostly carried out voluntarily by specialists in various fields and supported by the Árni Magnússon Institute for Icelandic Studies. The *Icelandic Language Council* (Íslensk málnefnd) was formed in 1964 [17].

The Icelandic Language Council has the role to provide the Government with counselling regarding the Icelandic language.

Its role is to provide the Government with counselling regarding the Icelandic language, making suggestions to the Minister of Education regarding language policy, as well as to provide yearly reports on the status of Icelandic. The Icelandic Language Council is responsible for the spelling rules that are published by the Minister of Education and used in the school system. It has established the *Icelandic Language Cultivation Fund* (Málræktarsjóður), whose goal it is to promote and support any kind of activities that strengthen and preserve the Icelandic language [18].

It is sometimes said that everyone in Iceland is a linguist. Farmers and fishermen to nurses and teachers call in to daily radio talk shows to discuss the latest nuances in the language and complain about blunders in speech. People worry about the status of the language in the country and huge debates take place over how to preserve the language and whether it is even worth the fight. However, most Icelanders see the language as the centre of Icelandic culture and Icelandic identity, and various efforts have been made in order to preserve the language as well as is possible.

The staple for these efforts is the *Árni Magnússon Institute for Icelandic Studies* (Stofnun Árna Magnússonar í íslenskum fræðum) whose main role is to conduct research in the field of Icelandic language and literature, and to disseminate knowledge in those areas, as well as to protect and develop its manuscript collections [19]. Within the institution there are several departments, each focusing on different aspects of Icelandic language, literature and culture, such as language planning, vocabulary studies, lexicography, language technology, place names and name studies, manuscript studies, folkloristics and international outreach.

The National Radio has long played a role in preserving the language.

The Icelandic National Broadcasting Service has long played a role in preserving the language, not only because of its own language policy but also because of popular radio programmes like *Íslenskt mál* [Icelandic language], where linguists discuss the language, often with a particular emphasis on vocabulary, with listeners of the programme; and *Orð skulu standa* [words shall stand], a quiz show where two teams compete in finding the right meaning of rare words and sayings. In general

the media plays a significant role in the preservation of the language.

There are 22 radio stations in the country, all of which broadcast primarily in Icelandic, even though the music played is more commonly in other languages, particularly English. In addition there are 10 television stations, and even though a considerable percentage of the broadcast material is in languages other than Icelandic, the status of Icelandic is unquestioned [20]. All foreign-language television shows are subtitled in Icelandic, except for some material intended for children, which is more often dubbed. When live-events are being broadcast in other languages, Icelandic-speaking commentators will recap the main highlights [21].

Icelandic Language Day has been celebrated since 1996, on November 16 each year (the birthday of the poet Jónas Hallgrímsson), and is intended to promote the Icelandic language [22].

3.5 LANGUAGE IN EDUCATION

Icelandic language is an important, mandatory part of the Icelandic school system and grades 1-4 spend the minimum of 1,120 minutes per week on Icelandic language and literature. In grades 5-7 and 8-10 this has been lowered to 680 and 630 minutes per week, respectively, which is considerably less than the other Scandinavian countries spend educating on their mother tongues [23]. In middle school less time is spent on the mother tongue than in the other Scandinavian countries, amounting to a minimum of 20 credits out of the 200 that are required for graduation [24].

In the PISA studies that have been conducted since 2000 there has, until recently, been a steady decline in the reading comprehension of Icelandic teenagers, particularly among boys. In the latest study from 2009, however, the situation has improved again and Iceland is now sitting in 11th place, on a par with the other Scandinavian countries, excepting Finland [25].

The University of Iceland is the only university that offers a Ph.D. in Icelandic, although a master's degree can be obtained at the University of Manitoba in Canada, in addition to the University of Iceland. Several universities offer a B.A. degree in Icelandic.

Of the seven universities and colleges in the country, only two have a special language policy where Icelandic is clearly stated as the language of the university. In fact, English is increasingly being used in higher education institutions, since more and more are employing foreign professors and all of the institutions aim towards attracting more foreign students. This has resulted in a growing number of classes being taught in English and more doctoral dissertations being written in English. Additionally, more Icelandic scholars write their articles in English, which has resulted in an increase in the use of English-language class material in universities [16].

Increasing the amount of Icelandic language teaching in schools would help students with their language skills.

Increasing the amount of Icelandic language teaching in schools is one possible step towards providing students with the language skills required for active participation in society. Language Technology can make an important contribution here by offering so-called computer-assisted language learning (CALL) systems, which allow students to experience language in a playful way, for example, by linking special vocabulary in electronic text to comprehensible definitions or to audio or video files supplying additional information, e.g., the pronunciation of a word.

3.6 INTERNATIONAL ASPECTS

As a small country and a small player on the world stage the influence of Icelandic art, science or scholarly works

in other countries is minimal. A small number of Icelandic musicians have gained popularity outside of the country, such as *Björk*, *Sigur Rós* and *Gus Gus*, but as their music is more or less sung in English it has only a minimal effect on how much the language is heard outside of Iceland. Similarly the success of Icelandic writers abroad has exposed the artists to other cultures, but not the language. However, the popularity of Icelandic musicians and writers, the success – and eventual failure – of Icelandic banks and business companies abroad, as well as Iceland's focus on green energy has increased the exposure of the country to other nations, which leads to more discussions in foreign newspapers and more tourism. The sagas, the Vikings and the Icelandic horse are no longer the only Icelandic treasures that interest the outside world.

There is an increased international interest in Icelandic language.

When looking at Icelandic in the international context it is clear that it has very limited influence on other languages. Only a very few loanwords from Icelandic have found their way into other languages, with the most common being *geyser* – so in English, French, Galician, Italian, and similar words in many other languages. The word *eider* in English is also a borrowing from the Icelandic word *æður* and within the riding community the Icelandic word *tölt* is used for the fifth gait of the Icelandic horse.

The increased international interest in Icelandic language and culture can be seen by the rising numbers of students studying Icelandic, either in Iceland or abroad. At the University of Iceland the number of students studying Icelandic as a foreign language increased by almost 100% between the years 2005 and 2007 and in 2008 the university added courses in applied Icelandic, an option for those that want to learn the language without the academic aspect. Courses in Icelandic are now taught in 40 universities outside of the country, 18 of which are supported financially by the Icelandic government [16]. Furthermore, Icelandic community classes are offered in various countries, such as in the former Icelandic settlements in Canada and the United States, and about 300–400 people access the web course Icelandic Online daily [26].

The status of Icelandic would probably be strengthened internationally if the country joined the EU.

It goes without saying that Icelandic cannot be used anywhere in international communication. It has been claimed that the status of Icelandic would be strengthened internationally if the country joined the EU [27], since this would give Icelandic the status of an official EU language [28]. Language Technology can address the challenge of English from a different perspective by offering services like Machine Translation or cross-lingual information retrieval to foreign language text and thus help diminish personal and economic disadvantages naturally faced by non-native speakers of English.

3.7 ICELANDIC ON THE INTERNET

In June 2010, approximately 95% of the population had access to the Internet [29] and the percentage amongst those 35-44 years of age was 100%. At the beginning of May 2011, about 197,000, or 61.8% of the nation, had a Facebook account [30].

In 2010 there were 25,000 registered .is domains [31] and about 5,600 domains existed in the country outside of the .is system [32]. The number of websites is

estimated around 7,500, although that would not include numerous blog sites hosted on .is domains, as well as on foreign servers such as blogspot.com, and wordpress.com.

The Internet has been gaining such popularity that in 2010, for the first time, advertisers spent more money on advertisements on the Internet than in the printed media [33]. This has not yet happened in Iceland but the trend seems nevertheless to be heading that way. Of the seven most used websites in Iceland there are three online newspapers (*mbl.is*, *visir.is*, *pressan.is*). The Internet has also partly taken over the role of the phonebook with the information site *ja.is* being the fifth most used website in Iceland. Other commonly used websites were Google, Facebook and YouTube [34]. An Icelandic interface is available for all three.

Almost all Icelanders have access to the Internet.

For Language Technology, the growing importance of the Internet is important in two ways. On the one hand, the large amount of digitally available language data represents a rich source for analysing the usage of natural language, in particular by collecting statistical information. On the other hand, the Internet offers a wide range of application areas for Language Technology. The most commonly used web application is certainly Web Search, which involves the automatic processing of language on multiple levels, as we will see in more detail the second part of this paper. It involves sophisticated Language Technology, differing for each language. For Icelandic, this comprises for instance taking into account different inflectional endings of nouns, adjectives and verbs, and different stem forms like *svartur* [black, masculine] and *svört* [black, feminine].

The growing importance of the Internet is important for Language Technology.

But Internet users and providers of web content can also profit from Language Technology in less obvious ways, e.g., if it is used to automatically translate web contents from one language into another. Considering the high costs associated with manually translating these contents, comparatively little usable Language Technology is developed and applied, compared to the anticipated need. This may be due to the complexity of the Icelandic language and the number of technologies involved in typical Language Technology applications.

In the next chapter, we will present an introduction to Language Technology and its core application areas as well as an evaluation of the current situation of Language Technology support for Icelandic.

4

LANGUAGE TECHNOLOGY SUPPORT FOR ICELANDIC

Language technology is used to develop software systems designed to handle human language and are therefore often called "human language technology". Human language comes in spoken and written forms. While speech is the oldest and in terms of human evolution the most natural form of language communication, complex information and most human knowledge is stored and transmitted through the written word. Speech and text technologies process or produce these different forms of language, using dictionaries, rules of grammar, and semantics. This means that language technology (LT) links language to various forms of knowledge, independently of the media (speech or text) in which it is expressed. Figure 1 illustrates the LT landscape.

When we communicate, we combine language with other modes of communication and information media – for example speaking can involve gestures and facial expressions. Digital texts link to pictures and sounds. Movies may contain language in spoken and written form. In other words, speech and text technologies overlap and interact with other multimodal communication and multimedia technologies.

In this section, we will discuss the main application areas of language technology, i. e., language checking, web search, speech interaction, and machine translation. These applications and basic technologies include

- spelling correction
- authoring support
- computer-assisted language learning
- information retrieval
- information extraction
- text summarisation
- question answering
- speech recognition
- speech synthesis

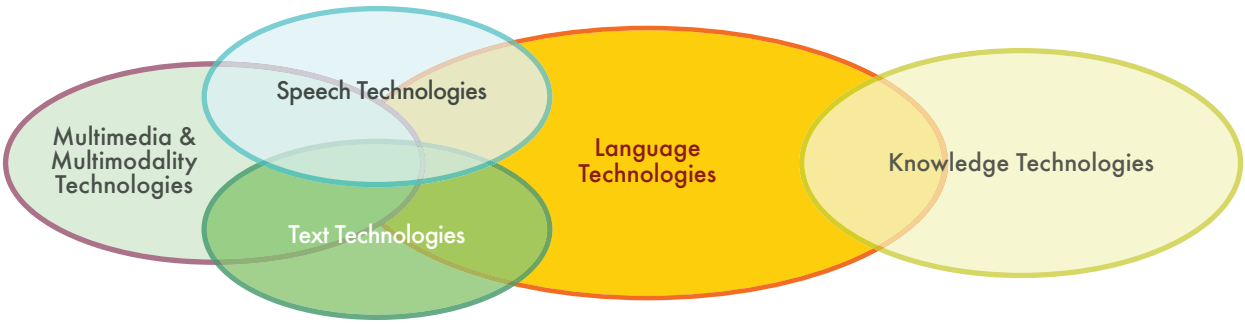
Language technology is an established area of research with an extensive set of introductory literature. The interested reader is referred to the following references: [35, 36, 37, 38, 39].

Before discussing the above application areas, we will briefly describe the architecture of a typical LT system.

4.1 APPLICATION ARCHITECTURES

Software applications for language processing typically consist of several components that mirror different aspects of language. While such applications are typically very complex, figure 2 shows a highly simplified architecture of a typical text processing system. The first three modules handle the structure and meaning of the text input:

1. Pre-processing: cleans the data, analyses or removes formatting, detects the input languages, and so on.
2. Grammatical analysis: finds the verb, its objects, modifiers and other parts of speech; detects the sentence structure.



1: Language technologies

3. Semantic analysis: performs disambiguation (i. e., computes the appropriate meaning of words in a given context); resolves anaphora (i. e., which pronouns refer to which nouns in the sentence); represents the meaning of the sentence in a machine-readable way.

After analysing the text, task-specific modules can perform other operations, such as automatic summarisation and database look-ups.

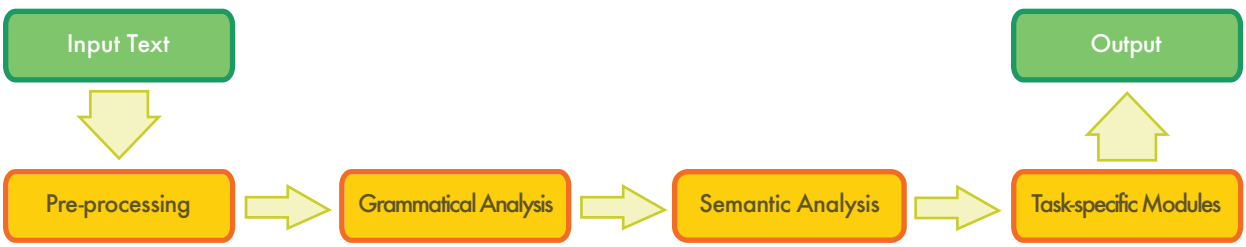
In the remainder of this section, we firstly introduce the core application areas for language technology, and follow this with a brief overview of the state of LT research and education today, and a description of past and present research programmes. Finally, we present an expert estimate of core LT tools and resources for Icelandic in terms of various dimensions such as availability, maturity and quality. The general situation of LT for the Icelandic language is summarised in figure 7 (p. 60) at the end of this chapter. This table lists all tools and resources that are boldfaced in the text. LT support for Icelandic is also compared to other languages that are part of this series.

4.2 CORE APPLICATION AREAS

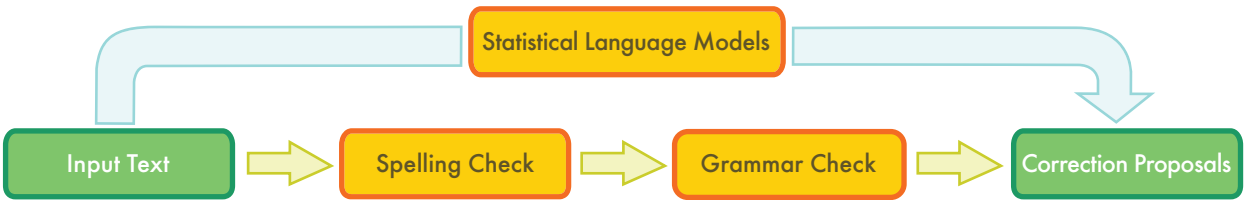
In this section, we focus on the most important LT tools and resources, and provide an overview of LT activities in Iceland.

4.2.1 Language Checking

Anyone who has used a word processor such as Microsoft Word knows that it has a spell checker that highlights spelling mistakes and proposes corrections. The first spelling correction programs compared a list of extracted words against a dictionary of correctly spelled



2: A typical text processing architecture



3: Language checking (top:statistical; bottom:rule-based)

words. Today these programs are far more sophisticated. Using language-dependent algorithms for **grammatical analysis**, they detect errors related to morphology (e. g., plural formation) as well as syntax–related errors, such as a missing verb or a conflict of verb-subject agreement (e. g., *she *write a letter*). However, most spell checkers will not find any errors in the following text [54]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.

Handling these kinds of errors usually requires an analysis of the context. For example: deciding if an Icelandic adjective should be written with a single (feminine) or double (masculine) 'n', as in

- Hann er **farinn**.
  [He is gone.]
- Hún er **farin**.
  [She is gone.]

This type of analysis either needs to draw on language-specific **grammars** laboriously coded into the software by experts, or on a statistical language model. In this case, a model calculates the probability of a particular word as it occurs in a specific position (e. g., between the words that precede and follow it). For example: *Hann er farinn* is a probable word sequence whereas *Hún er farinn* is not. A statistical language model can be automatically created by using a large amount of (correct) language data, a **text corpus**. Most of these two approaches have been developed around data from English. Neither approach can transfer easily to Icelandic because the language has a flexible word order, unlimited compound building and a richer inflection system. Language checking is not limited to word processors; it is also used in "authoring support systems", i. e., software environments in which manuals and other types of technical documentation for complex IT, healthcare, engineering and other products, are written. To offset customer complaints about incorrect use and damage claims resulting from poorly understood instructions, companies are increasingly focusing on the quality of technical documentation while targeting the international market (via translation or localisation) at the same time. Advances in natural language processing have led to the development of authoring support software, which helps the writer of technical documentation to use vocabulary and sentence structures that are consistent with industry rules and (corporate) terminology restrictions.

Language checking is not limited to word processors but also applies to authoring systems.

A spell checker for Icelandic has been available since Frisk Software developed the spell-checking program *Púki* in the late 1980s. Since then the program has been improved and updated. It is available for MS Office and

is widely used. Other spell checkers have also been developed. In 2002, the Dutch company Polderland developed a spell-checking program for the MS Office package. An open source spell checker for Icelandic also exists, that can be used with GNU/Linux applications and is based on Aspell. These programs are word-based, and hence cannot cope with many common spelling errors. A prototype of a context-sensitive spell checker has been integrated into LanguageTool [40] and works with OpenOffice. This spell checker could possibly lay ground for a basic grammar checker, although no such tool exists for Icelandic. Besides spell checkers and authoring support, language checking is also important in the field of computer-assisted language learning. And language checking applications also automatically correct search engine queries, as found in Google's *Did you mean...* suggestions.

4.2.2 Web Search

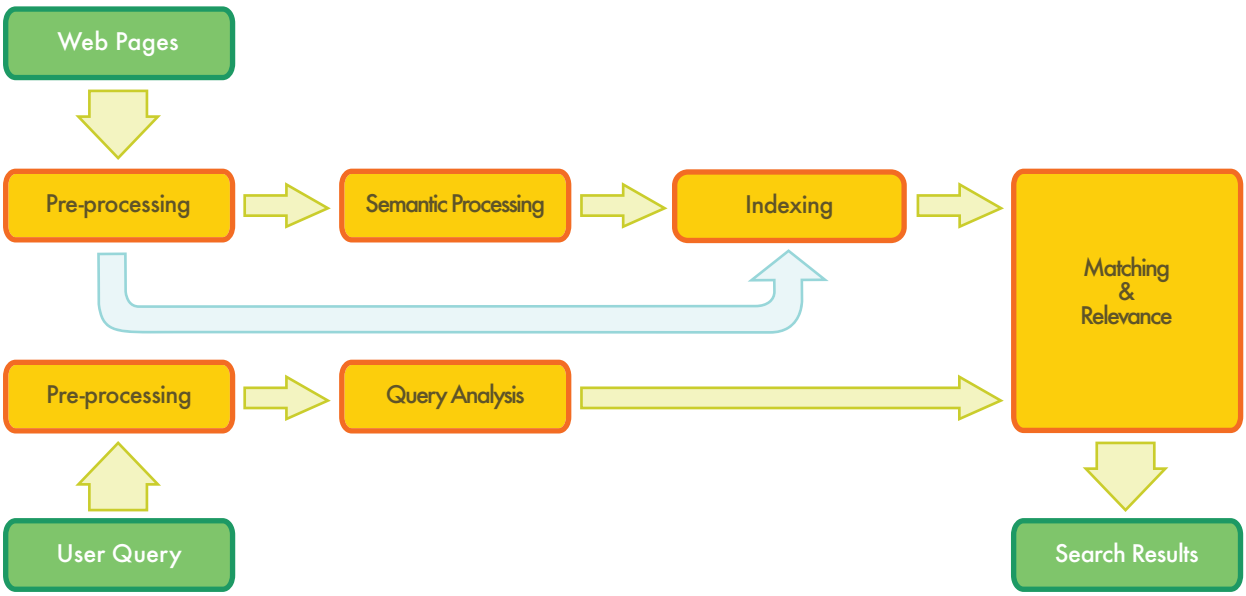
Searching the Web, intranets or digital libraries is probably the most widely used yet largely underdeveloped language technology application today. The Google search engine, which started in 1998, now handles about 80% of all search queries [41]. Since 2004, the verb *gúg(g)la* is commonly used in Icelandic, even though it has not made its way into printed dictionaries. The Google search interface and results page display has not significantly changed since the first version. However, in the current version, Google offers spelling correction for misspelled words and incorporates basic semantic search capabilities that can improve search accuracy by analysing the meaning of terms in a search query context [42]. The Google success story shows that a large volume of data and efficient indexing techniques can deliver satisfactory results using a statistical approach to language processing.

For more sophisticated information requests, it is essential to integrate deeper linguistic knowledge to facilitate text interpretation. Experiments using **lexical resources** such as machine-readable thesauri or ontological language resources (e. g., WordNet for English or GermaNet for German) have demonstrated improvements in finding pages using synonyms of the original search terms, such as *hagnaður* [profit], *arður* [dividend], *gróði* [profit] and *ábati* [gain], or even more loosely related terms.

The next generation of search engines will have to include much more sophisticated language technology.

The next generation of search engines will have to include much more sophisticated language technology, especially to deal with queries consisting of a question or other sentence type rather than a list of keywords. For the query, *Give me a list of all companies that were taken over by other companies in the last five years*, a syntactic as well as **semantic analysis** is required. The system also needs to provide an index to quickly retrieve relevant documents. A satisfactory answer will require syntactic parsing to analyse the grammatical structure of the sentence and determine that the user wants companies that have been acquired, rather than companies that have acquired other companies. For the expression *last five years*, the system needs to determine the relevant range of years, taking into account the present year. The query then needs to be matched against a huge amount of unstructured data to find the pieces of information that are relevant to the user's request. This process is called information retrieval, and involves searching and ranking relevant documents. To generate a list of companies, the system also needs to recognise a particular string of words in a document represents a company name, using a process called named entity recognition.

A more demanding challenge is matching a query in one language with documents in another language.



4: Web search

Cross-lingual information retrieval involves automatically translating the query into all possible source languages and then translating the results back into the user's target language.

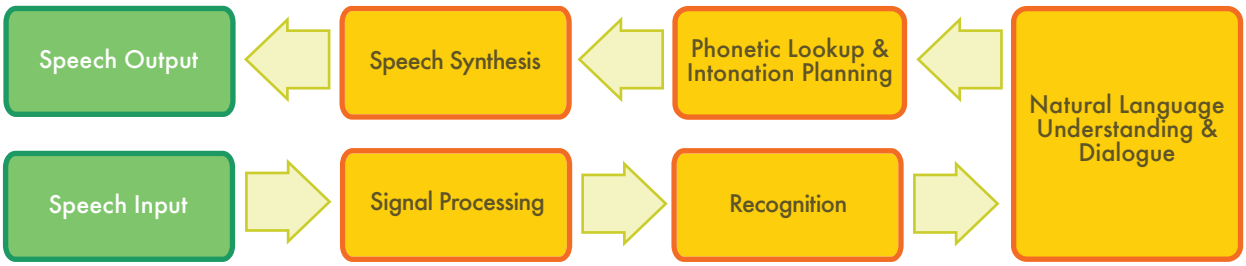
Now that data is increasingly found in non-textual formats, there is a need for services that deliver multimedia information retrieval by searching images, audio files and video data. In the case of audio and video files, a speech recognition module must convert the speech content into text (or into a phonetic representation) that can then be matched against a user query.

For inflectional languages like Icelandic, it is important to be able to search for all the inflectional forms of a word at once, instead of having to enter each different form separately. This can be done with the aid of a comprehensive full form database of modern Icelandic inflections, *BÍN* [43], which has been developed at the Árni Magnússon Institute for Icelandic Studies. The database contains about 280,000 paradigms, with over 5.8 million inflectional forms. Each entry contains lemma, the word form, the word class and morphological features, for common nouns, proper nouns, adjectives, verbs, and adverbs.

A few years ago, the private company Spurl developed a search engine, *Embla*, which made use of the database. The same algorithm is used for search in the Icelandic telephone directory and a few other sites. A similar feature, albeit not as sophisticated, has now been integrated into Google. At present, there is no large scale Icelandic language search engine project/product aside from Google's Icelandic interface.

4.2.3 Speech Interaction

Speech interaction is one of many application areas that depend on speech technology, i. e., technologies for processing spoken language. Speech interaction technology is used to create interfaces that enable users to interact in spoken language instead of using a graphical display, keyboard and mouse. Today, these voice user interfaces (VUI) are used for partially or fully automated telephone services provided by companies to customers, employees or partners. Business domains that



5: Speech-based dialogue system

rely heavily on VUIs include banking, supply chain, public transportation, and telecommunications. Other uses of speech interaction technology include interfaces to car navigation systems and the use of spoken language as an alternative to the graphical or touchscreen interfaces in smartphones.

Speech interaction technology comprises four technologies:

1. Automatic **speech recognition** (ASR) determines which words are actually spoken in a given sequence of sounds uttered by a user.
2. Natural language understanding analyses the syntactic structure of a user's utterance and interprets it according to the system in question.
3. Dialogue management determines which action to take given the user input and system functionality.
4. **Speech synthesis** (text-to-speech or TTS) transforms the system's reply into sounds for the user.

One of the major challenges of ASR systems is to accurately recognise the words a user utters. This means restricting the range of possible user utterances to a limited set of keywords, or manually creating language models that cover a large range of natural language utterances. Using machine learning techniques, language models can also be generated automatically from **speech corpora**, i. e., large collections of speech audio files and text transcriptions. Restricting utterances usually forces people to use the voice user interface in a rigid way and can damage user acceptance; but the creation, tuning and maintenance of rich language models will significantly increase costs. VUIs that employ language models and initially allow a user to express their intent more flexibly — prompted by a *How may I help you?* greeting — tend to be automated and are better accepted.

Companies tend to use utterances re-recorded by professional speakers for generating the output of the voice user interface. For static utterances where the wording does not depend on particular contexts of use or personal user data, this can deliver a rich user experience. But more dynamic content in an utterance may suffer from unnatural intonation because different parts of audio files have simply been strung together. Through optimisation, today's TTS systems are getting better at producing natural-sounding dynamic utterances.

Speech interaction is the basis for interfaces that allow a user to interact with spoken language.

Interfaces in speech interaction have been considerably standardised during the last decade in terms of their various technological components. There has also been strong market consolidation in speech recognition and speech synthesis. The national markets in the G20 countries (economically resilient countries with high populations) have been dominated by just five global play-

ers, with Nuance (USA) and Loquendo (Italy) being the most prominent players in Europe. In 2011, Nuance announced the acquisition of Loquendo, which represents a further step in market consolidation.

Three speech synthesisers have been developed for Icelandic. A formant-based speech synthesiser was originally made around 1990 and another one, based on diphone techniques, around 2000. These synthesisers were used mostly by the blind and visually impaired, while their quality was far from satisfactory for use in commercial applications for the general public.

In 2005, a new text-to-speech system was made under cooperation between the University of Iceland, Iceland Telecom and Hex Software. The system was trained by Nuance and uses their technology. This system has been used in commercial applications to some extent, but many users do not find its voice quality satisfactory. As the existing TTS systems are lacking in quality to their main users, the Icelandic Organisation of Blind and Partially Sighted is now planning to develop a new TTS system in cooperation with the University of Iceland, Reykjavik University, and the Polish Ivona software company. If everything goes as planned, this system will be ready for use later this year (2012) [44].

An isolated word speech recogniser for Icelandic was developed in 2003. The performance turned out to be quite satisfying, or at least 97% word accuracy. An Icelandic student at the Tokyo Institute of Technology has developed a prototype of a system for automatic continuous speech recognition for Icelandic. This system reached up to 67.5% word accuracy [45]. Neither of these systems has been put to use in commercial applications. In the middle of 2011, Reykjavik University and the Icelandic Centre for Language Technology started cooperating with Google on preparatory work for developing a speech recogniser for Icelandic [46].

Looking ahead, there will be significant changes, due to the spread of smartphones as a new platform for managing customer relationships, in addition to fixed telephones, the Internet and e-mail. This will also affect how speech interaction technology is used. In the long term, there will be fewer telephone-based VUIs, and spoken language apps will play a far more central role as a user-friendly input for smartphones. This will be largely driven by stepwise improvements in the accuracy of speaker-independent speech recognition via the speech dictation services already offered as centralised services to smartphone users.

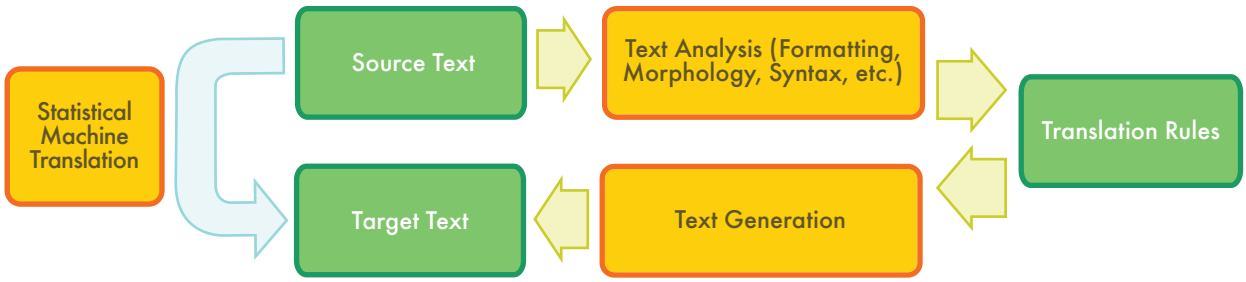
4.2.4 Machine Translation

The idea of using digital computers to translate natural languages can be traced back to 1946 and was followed by substantial funding for research during the 1950s and again in the 1980s. Yet **machine translation** (MT) still cannot meet its initial promise of across-the-board automated translation.

At its basic level, Machine Translation simply substitutes words in one natural language with words in another language.

The most basic approach to machine translation is the automatic replacement of the words in a text written in one natural language with the equivalent words of another language. This can be useful in subject domains that have a very restricted, formulaic language such as weather reports.

However, in order to produce a good translation of less restricted texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language. The major difficulty is that human language is ambiguous. Ambiguity creates challenges on multiple levels, such as word sense disambiguation at the lexical level (a *jaguar* is a brand of car or an animal) or the assignment of case on the syntactic level, for example:



6: Machine translation (left:statistical; right:rule-based)

- Konan sá bílinn og **maðurinn hennar** líka.
- Konan sá bílinn og **manninn sinn** líka.
- The woman saw the car and her husband, too.

One way to build an MT system is to use linguistic rules. For translations between closely related languages, a translation using direct substitution may be feasible in cases such as the above example. However, rule-based (or linguistic knowledge-driven) systems often analyse the input text and create an intermediary symbolic representation from which the target language text can be generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information, and large sets of grammar rules carefully designed by skilled linguists. This is a very long and therefore costly process.

In the late 1980s when computational power increased and became cheaper, interest in statistical models for machine translation began to grow. Statistical models are derived from analysing bilingual text corpora, **parallel corpora**, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 21 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text by processing parallel versions and finding plausible patterns of words. Unlike knowledge-driven systems, however, statistical (or data-driven) MT systems often generate ungrammatical output. Data-driven MT is advantageous because less human effort is required, and it can also cover special particularities of the language (e. g., idiomatic expressions) that are often ignored in knowledge-driven systems.

The strengths and weaknesses of knowledge-driven and data-driven machine translation tend to be complementary, so that nowadays researchers focus on hybrid approaches that combine both methodologies. One such approach uses both knowledge-driven and data-driven systems, together with a selection module that decides on the best output for each sentence. However, results for sentences longer than, say, 12 words, will often be far from perfect. A more effective solution is to combine the best parts of each sentence from multiple outputs; this can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

Machine Translation is particularly challenging for the Icelandic language.

Machine translation is particularly challenging for the Icelandic language. The potential for creating arbitrary new words by compounding makes dictionary analysis and dictionary coverage difficult; free word order and split verb constructions pose problems for analysis; and extensive inflection is a challenge for generating words

with proper markings for gender, case, number, mood, tense, etc.

The development in the area of MT for Icelandic has been limited. Stefán Briem, an independent researcher, has been engaged in MT since the early 1980s, resulting in his development of MT systems for Icelandic. In 2008 he launched a free web-based service, which offers translations between Icelandic and three other languages (English, Danish and Esperanto) [47]. Hrafn Loftsson, a researcher at Reykjavik University, and his associates have been developing a rule-based shallow transfer translation system from Icelandic to English, based on the Apertium platform [48]. A preliminary version is available online [49]. Google Translate has offered translations to and from Icelandic since 2009. The quality of the translations was rather poor in the beginning, but is getting better.

There is still a huge potential for improving the quality of MT systems. The challenges involve adapting language resources to a given subject domain or user area, and integrating the technology into workflows that already have term bases and translation memories. Another problem is that most of the current systems are English-centred and only support a few languages from and into Icelandic. This leads to friction in the translation workflow and forces MT users to learn different lexicon coding tools for different systems.

Evaluation campaigns help to compare the quality of MT systems, the different approaches and the status of the systems for different language pairs. Figure 7 (p. 23), which was prepared during the Euromatrix+ project, shows the pair-wise performances obtained for 22 of the 23 EU languages (Irish was not compared). The results are ranked according to a BLEU score, which indicates higher scores for better translations [51]. A human translator would normally achieve a score of around 80 points.

The best results (in green and blue) were achieved by languages that benefit from a considerable research effort in coordinated programmes and the existence of many parallel corpora (e. g., English, French, Dutch, Spanish and German). The languages with poorer results are shown in red. These languages either lack such development efforts or are structurally very different from other languages (e. g., Hungarian, Maltese and Finnish).

4.3 OTHER APPLICATION AREAS

Building language technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but they provide significant service functionalities "behind the scenes" of the system in question. They all form important research issues that have now evolved into individual sub-disciplines of computational linguistics. Question answering, for example, is an active area of research for which annotated corpora have been built and scientific competitions have been initiated. The concept of question answering goes beyond keyword-based searches (in which the search engine responds by delivering a collection of potentially relevant documents) and enables users to ask a concrete question to which the system provides a single answer. For example:

> *Question: How old was Neil Armstrong when he stepped on the moon?*
>
> *Answer: 38.*

While question answering is obviously related to the core area of web search, it is nowadays an umbrella term for such research issues as which different types of questions exist, and how they should be handled; how a set of documents that potentially contain the answer can be analysed and compared (do they provide conflicting answers?); and how specific information (the answer) can be reliably extracted from a document without ignoring the context.

Question answering is in turn related to information extraction (IE), an area that was extremely popular and influential when computational linguistics took a statistical turn in the early 1990s. IE aims to identify specific pieces of information in specific classes of documents, such as the key players in company takeovers as reported in newspaper stories. Another common scenario that has been studied is reports on terrorist incidents. The task here consists of mapping appropriate parts of the text to a template that specifies the perpetrator, target, time, location and results of the incident. Domain-specific template-filling is the central characteristic of IE, which makes it another example of a "behind the scenes" technology that forms a well-demarcated research area, which in practice needs to be embedded into a suitable application environment.

Language technology applications often provide significant service functionalities behind the scenes of larger software systems.

Text summarisation and **text generation** are two borderline areas that can act either as standalone applications or play a supporting role. Summarisation attempts to give the essentials of a long text in a short form, and is one of the features available in Microsoft Word. It mostly uses a statistical approach to identify the "important" words in a text (i. e., words that occur very frequently in the text in question but less frequently in general language use) and determine which sentences contain the most of these "important" words. These sentences are then extracted and put together to create the summary. In this very common commercial scenario, summarisation is simply a form of sentence extraction, and the text is reduced to a subset of its sentences. An alternative approach, for which some research has been carried out, is to generate brand new sentences that do not exist in the source text.

For the Icelandic language, research in most text technologies is much less developed than for the English language.

This requires a deeper understanding of the text, which means that so far this approach is far less robust. On the whole, a text generator is rarely used as a stand-alone application but is embedded into a larger software environment, such as a clinical information system that collects, stores and processes patient data. Creating reports is just one of many applications for text summarisation.
None of the technologies discussed in this section exist for Icelandic.

4.4 EDUCATIONAL PROGRAMMES

Language Technology is a highly interdisciplinary field, involving the expertise of linguists, computer scientists, mathematicians, philosophers, psycholinguists, and neuroscientists, among others. As such, it has not yet gained a firm ground in Icelandic higher education institutions. At the turn of the century, there were no programmes or even individual courses on Language Technology or Computational Linguistics at any Icelandic university or college, and there was no ongoing research in these areas.
In the fall of 2002, the University of Iceland launched an interdisciplinary Master's programme in Language Technology. This is a two-year programme (120 ECTS credits), which admits students with either a B.A. degree in the humanities (languages and linguistics) or a B.Sc. degree in computer science (or electrical or software engineering). In 2007, the programme was relaunched, now as a joint programme between the Department of Icelandic at the University of Iceland and the School of Computer Science at Reykjavik University. During the operation period of the Nordic Graduate School of

Language Technology (NGSLT) from 2004-2009 students could also take individual courses at other Nordic and Baltic universities. Due to lack of resources, both financial and human, it has not been possible to enroll any new students in the Master's programme since 2009. However, individual courses on Language Technology, Natural Language Processing and Corpus Linguistics are routinely offered at both the University of Iceland and Reykjavik University.

4.5 NATIONAL PROJECTS AND INITIATIVES

There are only about 330,000 people speaking Icelandic, and this is not enough to sustain costly development of new products. It costs just as much to build language resources for Icelandic as for languages with hundreds of millions of speakers. As a result, the number of commercial companies in the language technology industry in Iceland is close to zero. Frisk Software has developed and sells the spell-checking program *Púki*, but does not work on any new products. In the last decade, the companies Iceland Telecom and Hex Software worked with the University of Iceland in developing both an individual word speech recogniser and a text-to-speech system for Icelandic. Neither of these companies works on language or speech technology any longer.

Clara, a recent start-up company, provides service to companies that want to know what people think of their products and services. Clara's system uses semantic analysis and data presentation methods to analyse the attitudes online. Their tool for analysing Icelandic language websites is called *Vaktarinn* [52]. In its first year it had over 1200 users, counting non-paying trial users. Clara is the only company in Iceland currently developing revenue-generating LT products.

In 2000, the Icelandic Government launched a special Language Technology Programme with the aim of supporting institutions and companies in creating basic resources for Icelandic language technology. This initiative resulted in several projects which have had profound influence on the field in Iceland. The main direct products of the LT Programme are the following [2]:

- A full-form morphological database of Modern Icelandic inflections
- A balanced morphosyntactically tagged corpus of 25 million words
- A training model for data-driven PoS taggers
- A text-to-speech system
- An isolated word speech recogniser
- An improved spell checker

After the Language Technology Programme ended in 2004, researchers from three institutes (University of Iceland, Reykjavik University, and the Árni Magnússon Institute for Icelandic Studies), who had been involved in most of the projects funded by the programme, decided to join forces in a consortium called the Icelandic Centre for Language Technology (ICLT), in order to follow up on the tasks of the programme. The main roles of the ICLT are to

- serve as an information centre on Icelandic LT by running a website (http://iclt.is);
- encourage cooperation on LT projects between universities, institutions and commercial companies;
- organise and coordinate university education in LT;
- participate in Nordic, Europeaan and international cooperation on LT;
- initiate and participate in R&D projects in LT;
- keep track of resources and products in the field of Icelandic LT;
- organise LT conferences for researchers, companies and the public;
- support the growth of Icelandic LT in all possible manners.

Since 2005, the ICLT researchers have initiated several new projects which have been partly supported by the Icelandic Research Fund and the Icelandic Technical Development Fund. The most important product of these projects is the open source IceNLP package (IceTagger, IceParser and Lemmald) [53] which is also available as an online service (http://nlp.cs.ru.is). In 2009, the ICLT received a relatively large three year Grant of Excellence from the Icelandic Research Fund to the project 'Viable Language Technology beyond English – Icelandic as a test case'. Within that project, further basic LT resources for Icelandic are being developed.

As we have seen, previous programmes have led to the development of a number of LT tools and resources for the Icelandic language. In the following section, the current state of LT support for Icelandic is summarised.

4.6 AVAILABILITY OF TOOLS AND RESOURCES

Figure 7 provides a rating for language technology support for the Icelandic language. This rating of existing tools and resources was generated by leading experts in the field who provided estimates based on a scale from 0 (very low) to 6 (very high) using seven criteria.

The key results for Icelandic language technology can be summed up as follows:

- Icelandic stands reasonably well with respect to the most basic language technology tools and resources, such as text analysis and text corpora.
- There exist also individual products with limited functionality in fields such as speech synthesis, speech recognition and machine translation, speech corpora, parallel corpora, and lexical resources.
- However, tools and resources for more advanced language technology such as text interpretation and language generation, simply do not exist.

At the end of the 20th century, Icelandic Language Technology was virtually non-existent. Things started to change in 1999, after a specially appointed Expert Group delivered a white paper on Language Technology to the Minister of Education, Science and Culture [3]. In this white paper, several actions to establish Icelandic Language Technology were proposed. The expert group estimated that it would cost around one billion Icelandic krónas (which then amounted to about ten million Euros) to make Icelandic LT self-sustained. After that, the free market should be able to take over, since it would have access to public resources that would have been created by the government-funded Language Technology Programme, and that would be made available on an equal basis to everyone who was going to use these resources in their commercial products.

It must be pointed out that the total budget of the Language Technology Programme from 2000-2004 was only around 1/8 of the sum that the expert group estimated would be needed [2]. It should therefore come as no surprise that Icelandic LT is still in its infancy. 330,000 speakers are simply to few to sustain costly development of new products. At present, almost no companies are working in the LT area because they do not see it as profitable. It is thus extremely important to continue public support for Icelandic LT for some time, but given the current financial situation, it does not seem likely that such support will be coming from the state budget any time soon.

4.7 CROSS-LANGUAGE COMPARISON

The current state of LT support varies considerably from one language community to another. In order to compare the situation between languages, this section will present an evaluation based on two sample application areas (machine translation and speech processing)

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	1	1	1	1.5	1	0	1
Speech Synthesis	1	1	2.5	2.5	2	1	1
Grammatical analysis	2	5.5	4	3	3.5	3.5	3
Semantic analysis	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Text generation	0	0	0	0	0	0	0
Machine translation	1	4	1	1.5	1.5	1.5	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	1.5	4	3	2.5	2.5	4.5	3
Speech corpora	1	2	1.5	1.5	1	1.5	1.5
Parallel corpora	1	1	1	0.5	1	1	1
Lexical resources	1	2	2.5	2.5	2	2	2
Grammars	1	4	2.5	2	2.5	2.5	2

7: State of language technology support for Icelandic

and one underlying technology (text analysis), as well as basic resources needed for building LT applications. The languages were categorised using the following five-point scale:

1. Excellent support
2. Good support
3. Moderate support
4. Fragmentary support
5. Weak or no support

LT support was measured according to the following criteria:

Speech Processing: Quality of existing speech recognition technologies, quality of existing speech synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications.

Machine Translation: Quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications.

Text Analysis: Quality and coverage of existing text analysis technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources (e. g., WordNet) and grammars.

Resources: Quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars.

Figures 8 to 11 show that Icelandic is in the bottom cluster for all of the tools and resources listed. It com-

pares well with other languages with a small number of speakers, such as Irish, Latvian, Lithuanian, and Maltese. These languages lag far behind large languages like German and French, for instance. But even LT resources and tools for those languages clearly do not yet reach the quality and coverage of comparable resources and tools for the English language, which is in the lead in almost all LT areas. And there are still plenty of gaps in English language resources with regard to high quality applications.

4.8 CONCLUSIONS

In this series of white papers, we have made an important effort by assessing the language technology support for 30 European languages, and by providing a high-level comparison across these languages. By identifying the gaps, needs and deficits, the European language technology community and its related stakeholders are now in a position to design a large scale research and development programme aimed at building a truly multilingual, technology-enabled communication across Europe.

The results of this white paper series show that there is a dramatic difference in language technology support between the various European languages. While there are good quality software and resources available for some languages and application areas, others, usually smaller languages, have substantial gaps. Many languages lack basic technologies for text analysis and the essential resources. Others have basic tools and resources but the implementation of for example semantic methods is still far away. Therefore a large-scale effort is needed to attain the ambitious goal of providing high-quality language technology support for all European languages, for example through high quality machine translation.

For a small language community and a small research environment such as the Icelandic one, it is vital to cooperate, not only on the national level but also internationally. It is to be hoped that Iceland's participation in META-NORD and META-NET will make it possible to develop, standardise and make available several important LT resources and thus contribute to the growth of Icelandic language technology.

The long term goal of META-NET is to enable the creation of high-quality language technology for all languages. This requires all stakeholders - in politics, research, business, and society - to unite their efforts. The resulting technology will help tear down existing barriers and build bridges between Europe's languages, paving the way for political and economic unity through cultural diversity.

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch Finnish French German Italian Portuguese Spanish	Basque Bulgarian Catalan Danish Estonian Galician Greek Hungarian Irish Norwegian Polish Serbian Slovak Slovene Swedish	Croatian **Icelandic** Latvian Lithuanian Maltese Romanian

8: Speech processing: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	French Spanish	Catalan Dutch German Hungarian Italian Polish Romanian	Basque Bulgarian Croatian Czech Danish Estonian Finnish Galician Greek **Icelandic** Irish Latvian Lithuanian Maltese Norwegian Portuguese Serbian Slovak Slovene Swedish

9: Machine translation: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Dutch French German Italian Spanish	Basque Bulgarian Catalan Czech Danish Finnish Galician Greek Hungarian Norwegian Polish Portuguese Romanian Slovak Slovene Swedish	Croatian Estonian **Icelandic** Irish Latvian Lithuanian Maltese Serbian

10: Text analysis: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch French German Hungarian Italian Polish Spanish Swedish	Basque Bulgarian Catalan Croatian Danish Estonian Finnish Galician Greek Norwegian Portuguese Romanian Serbian Slovak Slovene	**Icelandic** Irish Latvian Lithuanian Maltese

11: Speech and text resources: State of support for 30 European languages

ABOUT META-NET

META-NET is a Network of Excellence partially funded by the European Commission. The network currently consists of 54 research centres in 33 European countries [55]. META-NET forges META, the Multilingual Europe Technology Alliance, a growing community of language technology professionals and organisations in Europe. META-NET fosters the technological foundations for a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- grants all Europeans equal access to information and knowledge regardless of their language;
- builds upon and advances functionalities of networked information technology.

The network supports a Europe that unites as a single digital market and information space. It stimulates and promotes multilingual technologies for all European languages. These technologies support automatic translation, content production, information processing and knowledge management for a wide variety of subject domains and applications. They also enable intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots. Launched on 1 February 2010, META-NET has already conducted various activities in its three lines of action META-VISION, META-SHARE and META-RESEARCH.

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. The present White Paper was prepared together with volumes for 29 other languages. The shared technology vision was developed in three sectorial Vision Groups. The META Technology Council was established in order to discuss and to prepare the SRA based on the vision in close interaction with the entire LT community.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, the action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community.

office@meta-net.eu – http://www.meta-net.eu

TILVÍSANIR REFERENCES

[1] Aljoscha Burchard, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit, and Martin Volk. *Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012.

[2] Eiríkur Rögnvaldsson, Hrafn Loftsson, Kristín Bjarnadóttir, Sigrún Helgadóttir, Matthew Whelpton, Anna Björk Nikulásdóttir, and Anton Karl Ingason. Icelandic Language Resources and Technology: Status and Prospects, 2009. http://dspace.utlib.ee/dspace/bitstream/handle/10062/9670/Icelandic%20language%20resources.pdf;jsessionid=A7320810CB6EA717510D0460EADE8C5B?sequence=1.

[3] Menntamálaráðuneytið (Ministry of Education, Science, and Culture). Skýrsla um tungutækni (Report on Language Technology), 1999. http://brunnur.stjr.is/mrn/utgafuskra/utgafa.nsf/xsp/.ibmmodres/domino/OpenAttachment/mrn/utgafuskra/utgafa.nsf/F0250A90B6D7F31B002576F00058D4B8/Attachment/tungutaekni.pdf.

[4] Máltæknisetur. http://maltaeknisetur.is.

[5] Aljoscha Burchardt, Georg Rehm, and Felix Sasaki. The Future European Multilingual Information Society – Vision Paper for a Strategic Research Agenda, 2011. http://www.meta-net.eu/vision/reports/meta-net-vision-paper.pdf.

[6] Directorate-General Information Society & Media of the European Commission. User Language Preferences Online, 2011. http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf.

[7] European Commission. Multilingualism: an Asset for Europe and a Shared Commitment, 2008. http://ec.europa.eu/languages/pdf/comm2008_en.pdf.

[8] Directorate-General of the UNESCO. Intersectoral Mid-term Strategy on Languages and Multilingualism, 2007. http://unesdoc.unesco.org/images/0015/001503/150335e.pdf.

[9] Directorate-General for Translation of the European Commission. Size of the Language Industry in the EU, 2009. http://ec.europa.eu/dgs/translation/publications/studies.

[10] Hagstofa Íslands (Statistics Iceland). Mannfjöldi (Census). http://www.hagstofa.is/Hagtolur/Mannfjoldi.

[11] Emilía Dagný Sveinbjörnsdóttir. Hvað búa margir Íslendingar í útlöndum? (How many Icelanders live abroad?), 2010. http://visindavefur.is/?id=53154.

[12] Statistics Canada. Census. http://www12.statcan.ca/census-recensement/index-eng.cfm.

[13] Alþingi (Icelandic Parliament). Lög um stöðu íslenskrar tungu og íslensks táknmáls (Law on the official status of the Icelandic Language and Icelandic Sign Language), 2011. http://www.althingi.is/altext/139/s/1570.html.

[14] Íslenskar mállýskur (icelandic dialects). http://mallyskur.is.

[15] Bjarki M Karlsson. Tvinnhljóð í íslensku (Affricates in Icelandic), 2007. http://fraedi.is/tvinnhljod/.

[16] Menntamálaráðuneytið (Ministry of Education, Science, and Culture). Íslenska til alls (Icelandic for All Purposes), 2009. http://www.islenskan.is/Islenska_til_alls.pdf.

[17] Stofnun Árna Magnússonar í íslenskum fræðum (The Árni Magnússon Institute for Icelandic Studies). Íslensk málnefnd (Icelandic Language Council). http://www.arnastofnun.is/page/arnastofnun_mal_islenskmalnefnd.

[18] Stofnun Árna Magnússonar í íslenskum fræðum (The Árni Magnússon Institute for Icelandic Studies). Málræktarsjóður (The Icelandic Language Cultivation Fund). http://www.arnastofnun.is/page/arnastofnun_mal_malraektarsjodur.

[19] Alþingi (Icelandic Parliament). Lög um Stofnun Árna Magnússonar í íslenskum fræðum (Law on the Árni Magnússon Institute for Icelandic Studies), 2006. http://www.althingi.is/lagas/139a/2006040.html.

[20] Hagstofa Íslands (Statistics Iceland). Útvarp (Radio). http://www.hagstofa.is/Hagtolur/Menningarmal/Utvarp.

[21] Alþingi (Icelandic Parliament). Útvarpslög (Law on Radio and Television), 2000. http://www.althingi.is/lagas/139a/2000053.html.

[22] Menntamálaráðuneytið (Ministry of Education, Science, and Culture). Dagur íslenskrar tungu (Icelandic Language Day). http://www.menntamalaraduneyti.is/menningarmal/dit/.

[23] Menntamálaráðuneytið (Ministry of Education, Science, and Culture). Aðalnámskrá grunnskóla (The National Guide for Compulsory Schools), 2011. http://www.menntamalaraduneyti.is/utgefid-efni/namskrar/nr/3953.

[24] Menntamálaráðuneytið (Ministry of Education, Science, and Culture). Aðalnámskrá framhaldsskóla (The National Guide for Secondary Schools), 2011. http://www.menntamalaraduneyti.is/utgefid-efni/namskrar/nr/3954.

[25] Almar M. Halldórsson, Ragnar F. Ólafsson, Óskar H. Níelsson, and Júlíus K. Björnsson. Íslenskir nemendur við lok grunnskólans. Helstu niðurstöður PISA 2009 rannsóknarinnar um lesskilning og læsi í stærðfræði og náttúrufræði (Icelandic pupils at the end of compulsory school), 2010. http://www.namsmat.is/vefur/rannsoknir/PISA_2009/pisa_2009_island.pdf.

[26] Icelandic online. http://icelandiconline.is.

[27] Gauti Kristmannsson. ESB er sterkasti leikur íslenskrar tungu (Joining the EU would be the strongest move for Icelandic), 2010. http://www.visir.is/article/20101001/FRETTIR01/175424536.

[28] European Commission Enlargement. Commission Opinion on Iceland's application for membership of the European Union, 2010. http://ec.europa.eu/enlargement/press_corner/key-documents/opinion-iceland_2010_en.htm.

[29] Internet World Stats. Internet Users in Europe March 31, 2011, 2011. http://www.internetworldstats.com/stats4.htm#european.

[30] CheckFacebook.com. http://www.checkfacebook.com.

[31] ISNIC. Samanlagður fjöldi léna og fjöldi skráðra léna á ári (Total number of domains and number of registered domains per year). http://www.isnic.is/tolur/index.html.

[32] WebHosting.info. Domain Registries in Iceland. http://www.webhosting.info/registries/country_stats/IS.

[33] mbl.is. Netið fram úr dagblöðum (The Internet ahead of newspapers), 2010. http://www.mbl.is/vidskipti/frettir/2010/12/21/netid_fram_ur_dagblodum/.

[34] Market and Media Research. Mbl.is og Google notaðir af flestum. Visir.is og Facebook fylgja fast á eftir (Mbl.is and Google most visited. Visir.is and Facebook follow closely), 2011. http://www.mmr.is/images/stories/PDF/1101_tilkynning_vefsidur.pdf.

[35] Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne Jekat, Hagen Langer, and Ralf Klabunde, editors. *Computerlinguistik und Sprachtechnologie: Eine Einführung (Computational Linguistics and Language Technology: An Introduction)*. Spektrum Akademischer Verlag, 2009.

[36] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice Hall, 2009.

[37] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[38] Language Technology World (LT World). http://www.lt-world.org.

[39] Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Battista Varile, Annie Zaenen, and Antonio Zampolli, editors. *Survey of the State of the Art in Human Language Technology (Studies in Natural Language Processing)*. Cambridge University Press, 1998.

[40] LanguageTool. Style and Grammar Checker. http://www.languagetool.org.

[41] Spiegel Online. Google zieht weiter davon (Google is still leaving everybody behind), 2009. http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html.

[42] Juan Carlos Perez. Google Rolls out Semantic Search Capabilities, 2009. http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html.

[43] Stofnun Árna Magnússonar í íslenskum fræðum (The Árni Magnússon Institute for Icelandic Studies). Beygingarlýsing íslensks nútímamáls (Morphological description of Modern Icelandic). http://bin.arnastofnun.is.

[44] Blindrafélagið (Icelandic Organization of the Visually Impaired). Nýr íslenskur talgervill í þjóðareign (A new Icelandic Text-to-speech system for the nation). http://www.blind.is/verkefni/talgervlaverkefnid/.

[45] Arnar Thor Jensson, Koji Iwano, and Sadaoki Furui. Language Model Adaptation Using Machine-Translated Text for Resource-Deficient Languages, 2008. http://www.hindawi.com/journals/asmp/2008/573832/ref/.

[46] Almannarómur. Opið safn íslenskra raddsýna (An open corpus of Icelandic speech samples). http://almannaromur.hr.is.

[47] Tungutorg. Vélrænar þýðingar (Machine Translation). http://tungutorg.is.

[48] Apertium. A free/open-source machine translation platform. http://www.apertium.org.

[49] Apertium is en. Translation from Icelandic to English. http://nlp.cs.ru.is/ApertiumISENWeb/index_en.jsp.

[50] Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 Machine Translation Systems for Europe. In *Proceedings of MT Summit XII*, 2009.

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA, 2002.

[52] Clara. Vaktarinn. http://www.vaktarinn.is.

[53] Icenlp. http://icenlp.sourceforge.net.

[54] Jerrold H. Zar. Candidate for a Pullet Surprise. *Journal of Irreproducible Results*, page 13, 1994.

[55] Georg Rehm and Hans Uszkoreit. Multilingual Europe: A challenge for language tech. *MultiLingual*, 22(3):51–52, April/May 2011.

B

META-NET ÞÁTTTAKENDUR META-NET MEMBERS

Austurríki	Austria	Zentrum für Translationswissenschaft, Universität Wien: Gerhard Budin
Belgía	Belgium	Computational Linguistics and Psycholinguistics Research Centre, University of Antwerp: Walter Daelemans
		Centre for Processing Speech and Images, University of Leuven: Dirk van Compernolle
Bretland	UK	School of Computer Science, University of Manchester: Sophia Ananiadou
		Institute for Language, Cognition and Computation, Center for Speech Technology Research, University of Edinburgh: Steve Renals
		Research Institute of Informatics and Language Processing, University of Wolverhampton: Ruslan Mitkov
Búlgaría	Bulgaria	Institute for Bulgarian Language, Bulgarian Academy of Sciences: Svetla Koeva
Danmörk	Denmark	Centre for Language Technology, University of Copenhagen: Bolette Sandford Pedersen, Bente Maegaard
Eistland	Estonia	Institute of Computer Science, University of Tartu: Tiit Roosmaa, Kadri Vider
Finnland	Finland	Computational Cognitive Systems Research Group, Aalto University: Timo Honkela
		Department of Modern Languages, University of Helsinki: Kimmo Koskenniemi, Krister Lindén
Frakkland	France	Centre National de la Recherche Scientifique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur and Institute for Multilingual and Multimedia Information: Joseph Mariani
		Evaluations and Language Resources Distribution Agency: Khalid Choukri
Grikkland	Greece	R.C. "Athena", Institute for Language and Speech Processing: Stelios Piperidis
Holland	Netherlands	Utrecht Institute of Linguistics, Utrecht University: Jan Odijk
		Computational Linguistics, University of Groningen: Gertjan van Noord
Írland	Ireland	School of Computing, Dublin City University: Josef van Genabith
Ísland	Iceland	School of Humanities, University of Iceland: Eiríkur Rögnvaldsson
Ítalía	Italy	Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli": Nicoletta Calzolari
		Human Lang. Technology, Fondazione Bruno Kessler: Bernardo Magnini

Króatía	Croatia	Institute of Linguistics, Faculty of Humanities and Social Science, University of Zagreb: Marko Tadić
Kýpur	Cyprus	Language Centre, School of Humanities: Jack Burston
Lettland	Latvia	Tilde: Andrejs Vasiļjevs
		Institute of Mathematics and Computer Science, University of Latvia: Inguna Skadiņa
Litháen	Lithuania	Institute of the Lithuanian Language: Jolanta Zabarskaitė
Lúxembúrg	Luxembourg	Arax Ltd.: Vartkes Goetcherian
Malta	Malta	Department Intelligent Computer Systems, University of Malta: Mike Rosner
Noregur	Norway	Department of Linguistic, University of Bergen: Koenraad De Smedt
		Department of Informatics, Language Technology Group, University of Oslo: Stephan Oepen
Portúgal	Portugal	University of Lisbon: António Branco, Amália Mendes
		Spoken Language Systems Laboratory, Institute for Systems Engineering and Computers: Isabel Trancoso
Pólland	Poland	Institute of Computer Science, Polish Academy of Sciences: Adam Przepiórkowski, Maciej Ogrodniczuk
		University of Łódź: Barbara Lewandowska-Tomaszczyk, Piotr Pęzik
		Department of Computer Linguistics and Artificial Intelligence, Adam Mickiewicz University: Zygmunt Vetulani
Rúmenía	Romania	Research Institute for Artificial Intelligence, Romanian Academy of Sciences: Dan Tufiș
		Faculty of Computer Science, University Alexandru Ioan Cuza of Iași: Dan Cristea
Serbía	Serbia	University of Belgrade, Faculty of Mathematics: Duško Vitas, Cvetana Krstev, Ivan Obradović
		Pupin Institute: Sanja Vranes
Slóvakía	Slovakia	Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences: Radovan Garabík
Slóvenía	Slovenia	Jožef Stefan Institute: Marko Grobelnik
Spánn	Spain	Barcelona Media: Toni Badia, Maite Melero
		Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: Núria Bel
		Aholab Signal Processing Laboratory, University of the Basque Country: Inma Hernaez Rioja
		Center for Language and Speech Technologies and Applications, Universitat Politècnica de Catalunya: Asunción Moreno
		Department of Signal Processing and Communications, University of Vigo: Carmen García Mateo

Sviss	Switzerland	Idiap Research Institute: Hervé Bourlard
Svíþjóð	Sweden	Department of Swedish, University of Gothenburg: Lars Borin
Tékkland	Czech Republic	Institute of Formal and Applied Linguistics, Charles University in Prague: Jan Hajič
Ungverjaland	Hungary	Research Institute for Linguistics, Hungarian Academy of Sciences: Tamás Váradi
		Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics: Géza Németh, Gábor Olaszy
Þýskaland	Germany	Language Technology Lab, DFKI: Hans Uszkoreit, Georg Rehm
		Human Language Technology and Pattern Recognition, RWTH Aachen University: Hermann Ney
		Department of Computational Linguistics, Saarland University: Manfred Pinkal

Hátt í 100 sérfræðingar í máltækni – fulltrúar þeirra landa og tungumála sem taka þátt í META-NET – ræddu og samræmdu meginniðurstöður og ábendingar hvítbókanna á fundi META-NET í Berlín í Þýskalandi 21.-22. október 2011. – About 100 language technology experts – representatives of the countries and languages represented in META-NET – discussed and finalised the key results and messages of the White Paper Series at a META-NET meeting in Berlin, Germany, on October 21/22, 2011.

HVÍTBÓKARÖÐ META-NET

THE META-NET WHITE PAPER SERIES

Baskneska	Basque	euskara
Búlgarska	Bulgarian	български
Danska	Danish	dansk
Eistneska	Estonian	eesti
Enska	English	English
Finnska	Finnish	suomi
Franska	French	français
Galisíska	Galician	galego
Gríska	Greek	ελληνικά
Hollenska	Dutch	Nederlands
Írska	Irish	Gaeilge
Íslenska	Icelandic	íslenska
Ítalska	Italian	italiano
Katalónska	Catalan	català
Króatíska	Croatian	hrvatski
Lettneska	Latvian	latviešu valoda
Litháíska	Lithuanian	lietuvių kalba
Maltneska	Maltese	Malti
Norska – bókmál	Norwegian Bokmål	bokmål
Nýnorska	Norwegian Nynorsk	nynorsk
Portúgalska	Portuguese	português
Pólska	Polish	polski
Rúmenska	Romanian	română
Serbneska	Serbian	српски
Slóvakíska	Slovak	slovenčina
Slóvenska	Slovene	slovenščina
Spænska	Spanish	español
Sænska	Swedish	svenska
Tékkneska	Czech	čeština
Ungverska	Hungarian	magyar
Þýska	German	Deutsch

In everyday communication, Europe's citizens, business partners and politicians are inevitably confronted with language barriers. Language technology has the potential to overcome these barriers and to provide innovative interfaces to technologies and knowledge. This white paper presents the state of language technology support for the Icelandic language. It is part of a series that analyzes the available language resources and technologies for 31 European languages. The analysis was carried out by META-NET, a Network of Excellence funded by the European Commission. META-NET consists of 54 research centres in 33 countries, who cooperate with stakeholders from economy, government agencies, research organisations and others. META-NET's vision is high-quality language technology for all European languages.

Evrópubúar standa sífellt frammi fyrir tungumálaþröskuldum í daglegum samskiptum sín á milli – í viðskiptum, stjórnmálum og hversdagslífinu. Máltækni getur nýst til að yfirstíga þá þröskulda og skapa nýjan aðgang að tækni og þekkingu. Þessi hvítbók kynnir stöðu máltæknistuðnings við íslensku. Hún er hluti af ritröð þar sem gerð er grein fyrir tiltækum málföngum og máltækni fyrir 31 Evróputungumál. Að greiningunni stendur META-NET, sem er öndvegisnet fjármagnað af Evrópusambandinu. Innan META-NET eru 54 rannsóknarsetur í 33 löndum sem starfa með hagsmunaaðilum úr viðskiptalífinu, opinberum stofnunum, rannsóknarsetrum og ýmsum öðrum. Framtíðarsýn META-NET er að til verði hágæða máltækni fyrir öll evrópsk tungumál.

„Máltækni er afar mikilvægur stuðningur við alls kyns málrannsóknir og styrkur við þá viðleitni að íslensk tunga verði notuð á öllum sviðum þjóðfélagsins í samræmi við opinbera íslenska málstefnu."
– Dr. Guðrún Kvaran (prófessor, formaður Íslenskrar málnefndar)

„This is an excellent overview of the current state of language technology in Europe. This is a call to action for decision makers in countries that want their citizens to participate in the 21st century on equal footing with native English speakers."
– Helga Waage (co-founder and CTO of Mobilitus)