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)

Editors
Georg Rehm
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: georg.rehm@dfki.de

Hans Uszkoreit
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: hans.uszkoreit@dfki.de

ISSN 2194-1416 ISSN 2194-1424 (electronic)
ISBN 978-3-642-30173-5 ISBN 978-3-642-30174-2 (eBook)
DOI 10.1007/978-3-642-30174-2
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012939477

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-