Íslensk máltækni – fortíð og framtíð

Eiríkur Rögnvaldsson

Hugvísindaþing

14. mars 2009

Staðan 1999

- Fyrir 10 árum var íslensk máltækni varla til
- Við höfðum
  - ágætan stafrýni (ritvilluleitarforrit), *Púka*
  - nothæfan talgervil
- Við höfðum ekki
  - námsleiðir eða einstök námskeið í máltækni
  - rannsóknir á sviði íslenskrar máltækni
  - fyrirtæki sem ynnu að þróun máltæknibúnaðar

Starfshópur um tungutækni

- Haustið 1998 var skipaður starfshópur
  - á vegum menntamálaráðherra, Björns Bjarnasonar
- Hópinn skipuðu
  - Rögnvaldur Ólafsson formaður,
    - Eiríkur Rögnvaldsson, Þorgeir Sigurðsson
- Verkefni hópsins voru
  - að gera úttekt á stöðu máltækni á Íslandi
  - að gera tillögur um eflingu íslenskrar máltækni

Forsendur máltækni

- *Tungutækni – skýrsla starfshóps*
  - menntamálaráðuneytið, apríl 1999
- Þrjár meginstoðir íslenskrar máltækni
  - menntað fólk
  - málsöfn
  - málgreiningarforrit
- Áhugi fyrirtækja þarf að vera fyrir hendi
  - og líka stuðningur hins opinbera

Álit starfshópsins

- Íslensk máltækni sprettur ekki af sjálfu sér
  – vegna smæðar málsamfélagsins og markaðarins
- Nauðsynlegt er að hefja sem fyrst átak
  – til að skjóta stoðum undir íslenska máltækni
- Ríkið verður að hafa forgöngu um þetta átak
  – og bera meginkostnaðinn á fyrstu stigum þess
- Æskilegast er að markaðurinn taki síðan við
  – en getur ekki borið þróunarkostnaðinn í upphafi

Megintillögur starfshópsins

- Byggð verði upp sameiginleg gagnasöfn, málsöfn, sem geti nýst fyrirtækjum sem hráefni í afurðir
- Fé verði veitt til að styrkja hagnýtar rannsóknir á sviði máltækni
- Fyrirtæki verði styrkt til þess að þróa afurðir máltækni
- Menntun á sviði máltækni og málvísinda verði efld

Tungutækniáætlunin

- Í framhaldi af skýrslunni setti menntamála-ráðuneytið af stað tungutækniáætlun
  - til að styrkja stofnanir og fyrirtæki til að byggja upp grunngögn og búnað fyrir máltækni
- Til verkefnisins var varið 133 milljónum kr.
  - á árunum 2000-2004
- U.þ.b. 1/8 af því sem starfshópurinn taldi þurfa
  - 225-250 m.kr. árlega í 4-5 ár – u.þ.b. milljarður

Helstu afurðir áætlunarinnar

- Endurbættur stafrýnir, *Púki*
- Beygingarlýsing íslensks nútímamáls
- Þjálfunarlíkan fyrir málfræðilegan markara
- Talgreinir (stakorðagreinir)
- Talgervill, *Ragga*
- Mörkuð málheild, 25 milljónir orða
- Beygingar- og málfræðigreinikerfi (lauk ekki)

Norræn samvinna

- Nordic Language Technology Research Programme (2001-2004) – ýmis net
- Nordic Graduate School of Language Technology (NGSLT, 2004-2009)
- Northern European Association for Language Technology (NEALT, stofnað 2006)
- Þátttaka í margvíslegum umsóknum
  - sem fæstar hafa hlotið brautargengi

Meistaranám

- Meistaranám í máltækni hófst við HÍ 2002
  - þverfaglegt nám
  - nemendur úr íslensku og tölvunarfræði
- Námið var endurvakið haustið 2007
  - nú í samvinnu HÍ og HR
  - auk þess sem námskeið í NGSLT eru nýtt
- Framhaldið er þó ótryggt
  - einkum vegna þess að NGSLT er að hætta

Tungutæknisetur

- Icelandic Center for Language Technology
  - ICLT, stofnað 2005
- Aðstandendur:
  - Málvísindastofnun Háskóla Íslands
  - Tölvunarfræðideild Háskólans í Reykjavík
  - Stofnun Árna Magnússonar í íslenskum fræðum
- Setrinu er ætlað að vera samstarfsvettvangur
  - um rannsóknir, þróun og kennslu í máltækni

Verkefni setursins

- Hlutverki sínu gegnir setrið m.a. með því að:
  - vera upplýsingaveita um íslenska máltækni og reka vefsetur í því skyni
  - 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
  - eiga frumkvæði að og taka þátt í rannsóknaverkefnum á sviði máltækni
  - eiga frumkvæði að og taka þátt í hagnýtum verkefnum á sviði máltækni
  - halda utan um ýmiss konar hráefni og afurðir á sviði máltækni
  - halda árlega ráðstefnu með þátttöku fræðimanna, fyrirtækja og almennings
  - beita sér fyrir eflingu íslenskrar máltækni á öllum sviðum

Helstu afurðir 2005-2009

- Frá 2005 hafa ýmsar afurðir verið þróaðar
  - styrktar af Rannsóknasjóði og Tækniþróunarsjóði
- Málfræðilegur reglumarkari, *IceTagger*
- Setningafræðilegur hlutaþáttari, *IceParser*
- Textaskimi
- Lemmunarforrit, *Lemmald*
- Samhengisháð ritvilluleit

Erindi og greinar

- Erindi og veggspjöld á ráðstefnum
  - norrænum, evrópskum, alþjóðlegum
    - FinTAL, GoTAL, SLTC, NoDaLiDa, LREC, FLAIRS, NAACL-HLT, EACL, Interspeech, o.fl.
- Greinar um máltækni
  - í ritrýndum tímaritum, innlendum og alþjóðlegum
    - Orð og tunga, Íslenskt mál, Language Resources and Evaluation, Nordic Journal of Linguistics
  - og í ritrýndum ráðstefnuritum

Staðan 2009

- Íslensk máltækni hefur orðið til þennan áratug
  - menntun á sviði máltækni er í boði
  - þátttaka í norrænni samvinnu hefur verið veruleg
  - mikilvæg gagnasöfn hafa verið byggð upp
  - ýmis grundvallarhugbúnaður hefur verið þróaður
  - máltæknirannsóknir eru komnar af stað
- Sviðið fékk nýlega mikilsverða viðurkenningu
  - þriggja ára öndvegisstyrk Rannís, alls 43,5 m.kr.

Verkefnið

- *Hagkvæm máltækni utan ensku*
  - *íslenska tilraunin*
- *Viable Language Technology Beyond English*
  - *Icelandic as a Test Case*
- Þverfaglegt rannsóknarverkefni
  - meginmarkmið að þróa vísindalegar máltækniaðferðir sem henta auðlindalitlum tungumálum, einkum beygingamálum

Aðstandendur

- Verkefnisstjóri
  - Eiríkur Rögnvaldsson
- Aðrir þátttakendur
  - Hrafn Loftsson
  - Kristín Bjarnadóttir
  - Matthew Whelpton
- Samstarfsaðilar
  - Mikel L. Forcada
  - Anthony Kroch
- Nýdoktor
  - Joel Wallenberg
- Doktorsnemar
  - Anna Nikulásdóttir
  - Sigrún Helgadóttir
- Meistaranemar
  - Anton Karl Ingason
  - Martha Dís Brandt
  - NN

Aðferðafræði

- Að markmiðunum verður unnið með því að
  - endurbæta rannsóknaraðferðir og laga að íslensku
  - nýta sérkenni íslenskunnar til að þróa nýjar hagkvæmar aðferðir sem gera kleift að byggja upp tól og gögn á einfaldari hátt en áður
  - nýta þverfaglega þekkingu rannsóknarhópsins, reynslu hans úr fyrri verkefnum og samstarf við framúrskarandi erlenda vísindamenn til að tengja á frjóan hátt aðferðir ólíkra fræðigreina

Verkþættir

- Málvísindalegum og tölfræðilegum aðferðum
  - verður stefnt saman og látnar vinna í sameiningu
  - til að skapa nýja þekkingu og opna nýja möguleika
- Verkefnið skiptist í þrjá tengda verkþætti
  - sem gerð verður grein fyrir seinna í málstofunni

1) Merkingarnám og merkingarnet
2) Vélrænar grófþýðingar
3) Þáttunaraðferðir og uppbygging trjábanka

BLARK

- Litið er á þetta sem lið í íslensku BLARK
  - Basic LAnguage Resource Kit
- Tiltekin gögn og máltæknibúnaður
  - sem þurfa að vera til fyrir hvert tungumál
  - eigi málið að vera nothæft í upplýsingatækni
- Ýmsar þjóðir vinna að uppbyggingu BLARK
  - t.d. Eistar sem hafa gert metnaðarfulla áætlun

„Vismansrapporten“

- Skýrsla Norrænu ráðherranefndarinnar 2006
  - Norðurlönd leiðandi á sviði máltækni árið 2016
- Í skýrslunni var lögð áhersla á
  - stofnun NEALT og vinnuhópa á vegum þess
  - samningu BLARK-skýrslna fyrir einstök ríki
  - norrænt fé í samvinnu um menntun og þjálfun
  - að einstök ríki styrki hagnýt rannsóknarverkefni með þátttöku háskóla og fyrirtækja

Aðgerðaáætlun og eftirfylgni

- Þegar BLARK-skýrslur lægju fyrir yrði
  - norrænu fé veitt til gerðar máltæknibúnaðar
  - norrænu og innlendu fé veitt til uppbyggingar málheilda, trjábanka og orðasafna
- Ekkert hefur verið gert með skýrsluna
  - sótt hefur verið um fé til norræns meistaranáms
  - og til uppbyggingar rannsóknarinnviða
  - en ekkert fengist

Íslensk málstefna

- Íslensk tunga verði nothæf – og notuð – á öllum þeim sviðum innan tölvu- og upplýsingatækninnar sem varða daglegt líf alls almennings
  - viðmót algengs hugbúnaðar þarf að vera íslenskt
  - til þarf að vera ýmiss konar hugbúnaður sem liðsinnir og leiðbeinir notendum við notkun íslensks máls
  - unnt á að vera að nota íslensku sem samskiptamál við ýmiss konar tölvu- og tæknibúnað

Aðgerðir

- Gerð verði áætlun um stuðning við uppbyggingu og þróun íslenskrar máltækni næsta áratug.
- Stöðugt verði unnið að uppbyggingu og eflingu mállegra gagnasafna sem eru forsenda fyrir þróun og smíði margs kyns máltæknibúnaðar.
- Málleg gagnasöfn og hugbúnaður til að vinna með íslenskt mál verði gerð opin og frjáls eftir því sem kostur er.
- Hugbúnaður til að lagfæra og leiðrétta íslenskt málfar verði gerður og kominn í notkun innan þriggja ára.
- Nothæf þýðingarforrit milli íslensku og valinna erlendra mála, a.m.k. ensku, verði gerð innan fimm ára.
- Íslenskur talgervill og talgreinir sem gerðir voru á vegum tungutækniátaks menntamálaráðuneytisins verði endurbættir og lagaðir að nýjustu tækni.
- Unnið verði markvisst að þróun mállegra samskiptakerfa milli manns og tölvu fyrir íslensku.

Horfur

- Alþingi ályktar
  - að samþykkja tillögur Íslenskrar málnefndar að íslenskri málstefnu sem opinbera stefnu í málefnum íslenskrar tungu.
- Tillagan samþykkt
  - með 36 samhljóða atkvæðum 12. mars
- Framtíðarhorfur ættu því að vera bjartar ☺
  - þótt kreppan geti vissulega raskað þeim ☹


Þökk fyrir áheyrnina

eirikur@hi.is