Eiríkur Rögnvaldsson

Sambúð tungu og tækni

Ráðstefna í Þjóðarbókhlöðu
á degi íslenskrar tungu
16. nóvember 1999

Efni fyrirlestursins

- Hvað er tungutækni?
  - margvísleg tengsl tungumáls og tölvutækni
- Íslenska og upplýsingatækni
  - staða og framtíð íslensks máls innan upplýsingatækninnar
- Starfshópur um tungutækni
  - gerð grein fyrir hópnum og helstu tillögum hans
- Forgangsverkefni í íslenskri tungutækni
  - nánari lýsing á þeim verkum sem þarf að vinna á næstunni
- Framtíðin

Hvað er tungutækni?

- **Tungutækni** er rúmlega ársgamalt nýyrði
  - um það sem nefnist ‘language technology’ á ensku
  - einnig ‘language engineering’ – **tungumálaverkfræði**
- Ýmiss konar samvinna tungumáls og tölvutækni
- Þessi samvinna er tvenns konar:
  - annars vegar nýting tölvutækni í þágu tungumálsins
  - hins vegar notkun tungumálsins innan tölvutækninnar

Hagnýting tölvutækni í þágu tungumálsins

- Tölvutækni má nýta á ýmsan hátt til að auðvelda mönnum að nota tungumálið

- Þar má nefna
  - forrit til leiðréttingar á stafsetningu og málfari
  - vélrænar þýðingar
  - tölvuorðabækur af ýmsu tagi
  - talgervla
  - ýmiss konar kennsluforrit
  - hjálpartæki handa fötluðum

Notkun tungumálsins innan tölvutækninnar

- Tungumálið gegnir sívaxandi hlutverki innan tölvutækninnar

- Þar má nefna
  - leit í gagnabönkum
    - spurningar bornar fram í samfelldu, eðlilegu máli í stað þess að nota takmarkaðan orðaforða á fastmótaðan hátt
  - stjórn ýmiss konar tækja
    - talað er við tæki á venjulegu máli og þeim stjórnað þannig með rödd og tungumáli í stað þess að ýta á takka

Stafar íslensku ógn af upplýsingatækninni?

- Þrjú einkenni upplýsingatækninnar skipta máli
  – þegar áhrif hennar á íslenska tungu eru metin

- Hún er (eða er að verða)
  - ✶ mikilvægur þáttur
  - ✸ í daglegu lífi
  - ✹ alls almennings

➔ Þess vegna verður hún að vera á íslensku!
  – að öðrum kosti er tungan feig

Þrengt notkunarsvið móðurmálsins

- Hvað gerist ef móðurmálið er ekki gjaldgengt á sviði
  - sem er mikilvægt
  - í daglegu lífi
  - alls almennings?

- Hvað gerist ef það er ekki nothæft
  - innan nýrrar tækni
  - í því sem er nýtt og spennandi
  - á sviðum þar sem nýsköpun á sér stað
  - og þar sem ný atvinnutækifæri bjóðast?

Tungumál í hættu

- Við slíkar aðstæður hefst dauðastríð tungumálsins
  - móðurmálið verður víkjandi
  - aðeins hæft til heimabrúks
  - en ekki til neinna alvarlegra hluta
- Unga kynslóðin sér þá ekki tilgang í að læra málið
  - heldur leggur áherslu á að tileinka sér enskuna sem best
- Hvað er þá til ráða?

Tveir kostir í stöðunni:

- Við gætum hafnað tækninni en haldið tungunni
  - látið eiga sig að tileinka okkur ýmsar nýjungar
  - úr því að tungumálið er ekki gjaldgengt á þessu sviði
- ➔ Þessi kostur er ekki raunhæfur!

- Við gætum fórnað tungunni en fylgst með tækninni
  - notað ensku í allri upplýsinga- og tölvutækni
  - úr því að íslenska er ekki nothæf á því sviði
- ➔ Þessi kostur er óviðunandi!

– og sá þriðji:

- Við verðum að hefjast handa!
  - gera átak á sviði tungutækni
  - gera íslensku nothæfa innan tölvutækninnar
- ➔ Það er eini valkostur okkar
  - ef við viljum halda áfram að nota íslensku
  - á öllum sviðum þjóðlífsins

- Annars verður málið fljótlega rykfallinn forngripur
  - sem ber dauðann í sér
  - og gæti dáið út á fáum áratugum

Starfshópur um tungutækni

- Starfshópur á vegum menntamálaráðuneytisins
  - tók til starfa haustið 1998 og skilaði áliti vorið 1999
- Hópinn skipuðu:
  - Rögnvaldur Ólafsson
    - dósent í eðlisfræði; formaður starfshópsins
  - Eiríkur Rögnvaldsson
    - prófessor í íslenskri málfræði
  - Þorgeir Sigurðsson
    - starfsmaður Staðlaráðs; verkfræðingur og íslenskufræðingur
  - Sigurður H. Pálsson
    - málfræðingur og tölvufræðingur; starfaði með hópnum

Niðurstöður starfshópsins

- Nauðsynlegt er að hefja sem fyrst átak
  – til að skjóta stoðum undir íslenska tungutækni
- Ríkið verður að hafa forgöngu um þetta átak
  – og bera meginkostnaðinn af því á fyrstu stigum þess
- Æskilegast er að markaðurinn taki síðan við
  – en hann getur ekki borið þróunarkostnaðinn í upphafi

Tillögur 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 tungutækni
- Fyrirtæki verði styrkt til þess að þróa afurðir tungutækni
- Menntun á sviði tungutækni og málvísinda verði efld

Kostnaður

	MKR
Þróunarmiðstöð	25-50
Rannsókna- og þróunarsjóður	150
Sérstakur styrkur til stærri alþjóðlegra verkefna	30
Stutt hagnýtt nám í máltækni	10
Meistaranám í máltölvun	10
Alls	**225-250**

– á ári í 4-5 ár

Forgangsverkefni í íslenskri tungutækni

- Meginmarkmið Íslendinga hlýtur að vera að unnt verði að nota íslenska tungu, ritaða með réttum táknum, sem víðast innan tölvu- og fjarskiptatækninnar

- Það er mikið verkefni að gera íslensku gjaldgenga á öllum sviðum, við allar aðstæður. Því verður að leggja megináherslu á þá þætti sem varða daglegt líf og starf alls almennings, eða munu gera það á næstu árum

1. Þýðing tölvuforrita

- Helstu tölvuforrit á almennum markaði verði á íslensku (Windows, Word, Excel; Netscape, Internet Explorer; Eudora; …)
  - Á þessu sviði hefur orðið afturför á undanförnum árum; fyrir tíu til fimmtán árum voru helstu ritvinnsluforrit á íslensku. Nauðsynlegt er að ýta á eftir því að Windows-stýrikerfið og önnur helstu forrit frá Microsoft verði íslenskuð, en einnig þarf að þýða ýmis önnur forrit sem almenningur notar hversdagslega, s.s. vefskoðara, póstforrit o.fl.

2. Íslenskir bókstafir

- Unnt verði að nota íslenska bókstafi (áéíóúýðþæö ÁÉÍÓÚÝÐÞÆÖ) við allar aðstæður; í tölvum, GSM-símum, textavarpi og öðrum tækjum sem almenningur notar.
  - Á þessu sviði hefur orðið mikil framför á undanförnum árum, og þar hafa íslenskir staðlamenn og málverndarmenn unnið gott starf. Þar eru þó ýmsar blikur á lofti. Þannig eru íslenskir stafir t.d. ekki í stafatöflu GSM-síma, sem er alvarlegt, ekki síst í ljósi þess að tengsl tölvutækni og fjarskiptatækni eru sífellt að aukast.

3. Málgreining

- Unnið verði að þróun málgreiningar fyrir íslensku, með það að markmiði að geta greint íslenskan texta í orðflokka og setningarliði.
  - Með málgreiningu (parsing) er átt við vélræna greiningu texta í orðflokka, setningarliði og setningar. Slík greining er mikilvæg fyrir gerð málfræðileiðréttingarforrita, þýðingarforrita o.fl. Mörg tungumál eiga forrit til nokkuð fullkominnar málgreiningar, en lítið er um slíkt hér á landi.

3.1 Textaheild

- Koma þarf upp stórri tölvutækri textaheild með íslenskum textum af sem fjölbreyttustum toga til að byggja áframhaldandi vinnu á.
  - Til að unnt sé að útbúa forrit sem vinna með tungumál þurfa að liggja fyrir miklar og nákvæmar upplýsingar um málið og notkun þess. Einn meginþátturinn í öflun slíkra upplýsinga felst í því að koma upp sem stærstri textaheild sem hafi að geyma tölvutæka íslenska texta. Úr þessum textum þarf síðan að vinna margs konar upplýsingar sem nauðsynlegar eru til að hægt sé að skrifa forrit til hvers kyns vinnu með málið. Gerð textaheildar af þessu tagi, og úrvinnsla úr henni, er forsenda markvissrar vinnu í íslenskri tungutækni.

3.2 Orðasafn

- Koma þarf upp fullgreindu orðasafni (með málfræðilegri og merkingarlegri greiningu) til nota í áframhaldandi vinnu.
  - Orðasafn með grunnorðaforða íslenskunnar (nokkrum tugum þúsunda orða) er forsenda ýmiss konar vinnu í tungutækni. Í þessu orðasafni þurfa að vera sem nákvæmastar upplýsingar um hvert orð; framburð þess, orðflokk, beygingu, setningarstöðu, merkingu, stílgildi o.s.frv. Slíkar upplýsingar koma að gagni við gerð málfræðileiðréttingarforrita, vélrænar þýðingar, leit í gagnabönkum o.fl.

4. Hjálparforrit við ritun

- Til verði góð hjálparforrit við ritun texta á íslensku, s.s. orðskiptiforrit, stafsetningarleiðréttingarforrit, málfarsleiðréttingarforrit o.fl.
  - Til eru allgóð íslensk orðskiptiforrit, og einnig forrit til stafsetningarleiðréttingar (Púki Friðriks Skúlasonar o.fl.). Forrit til málfarsleiðréttinga aðstoða notendur við að útrýma beygingarvillum, rangri orðaröð og klúðurslegri setningaskipan. Slík forrit eru aftur á móti engin til fyrir íslensku, en væru mjög þörf.

5. Íslenskur talgervill

- Til verði góður íslenskur talgervill sem geti lesið upp íslenskan texta með skýrum og auðskiljanlegum framburði og eðlilegu tónfalli og sem sé skiljanlegur án þjálfunar.
  - Undanfarin ár hefur verið á markaðnum íslenskaður talgervill frá sænska fyrirtækinu Infovox. Þessi talgervill er byggður á tækni sem nú þykir úrelt. Ljóst er að framburði hans er um margt ábótavant, en þó hefur hann gagnast sumum mjög vel. Nauðsynlegt er að vinna áfram að því að útbúa fullkominn íslenskan talgervil.

6. Talgreining

- Unnið verði að þróun talgreiningar fyrir íslensku, með það að markmiði að til verði forrit sem geti túlkað eðlilegt íslenskt tal.
  - Með talgreiningu (speech recognition) er átt við það að tölvur skilji talað mál. Mjög miklar framfarir hafa orðið á þessu sviði upp á síðkastið. Líklegt er að talgreining muni skipta miklu máli á ýmsum sviðum í framtíðinni, t.d. við upplýsingaleit og stjórn ýmiss konar tækja. Því er mjög mikilvægt að hefja skipulega vinnu að þróun talgreiningar fyrir íslensku.

7. Vélrænar þýðingar

- Unnið verði að þróun forrita til vélrænna þýðinga milli íslensku og annarra tungumála, m.a. til að auðvelda leit í gagnabönkum.
  - Vélrænar þýðingar eiga sér langa sögu, en hafa gengið misjafnlega. Á seinustu árum hafa þó komið fram þýðingarforrit sem virka allvel, a.m.k. á afmörkuðum sviðum. Líklegt er að mikilvægi vélrænna þýðinga muni aukast verulega á næstu árum, t.d. í sambandi við leitir í gagnabönkum o.fl.

8. Ábyrgðaraðilar

- Ákveðnum aðilum (stofnunum eða fyrirtækjum) verði falin ábyrgð á einstökum verkefnum.
  - Færa má rök að því að það hafi staðið allri þróun á þessu sviði fyrir þrifum að enginn aðili hefur borið ótvíræða ábyrgð á því að Íslendingar fylgdust þar með. Meðal þeirra sem eðlilegt er að standi að þeim verkefnum sem hér er lýst má nefna Málvísindastofnun Háskólans, Íslenska málstöð, Orðabók Háskólans, Staðlaráð Íslands og Póst- og fjarskiptastofnunina, en einnig er eðlilegt og nauðsynlegt að einkafyrirtæki taki þátt í verkefnunum.

… og hvað svo?

- Tillögur starfshópsins hafa verið til athugunar
  - hjá verkefnisstjórn um upplýsingasamfélagið
    - Krafan um íslenska tungu í tölvuheiminn er skýr og ótvíræð. Samningur við Microsoft, sem gerður var á síðastliðnum vetri, felur í sér viðurkenningu á þeirri kröfu. Næsta stórverkefni verður að beita hinni nýju tungutækni í þágu íslenskunnar til að tryggja stöðu hennar á sviði tölvu- og upplýsingatækni. (*Úr stefnuræðu forsætisráðherra*)
- Framhaldið ræðst af vilja stjórnvalda
  - en ekki síður áhuga stofnana, fyrirtækja og almennings