Eiríkur Rögnvaldsson
Háskóla Íslands

Staða íslenskrar tungutækni

Hádegisverðarfundur Skýrslutæknifélagsins
um íslensku í tölvugeiranum
15. nóvember 2004

Hvað er tungutækni?

- **Tungutækni** er ungt nýyrði
  - fyrir hugtakið ‘language technology’
    - eða ‘language engineering’
- Samvinna tungumáls og tölvutækni
  - í einhverjum hagnýtum tilgangi
- Tvær hliðar samvinnunnar:
  - notkun tækninnar í þágu tungumálsins
  - notkun tungumálsins í þágu tækninnar

Þrjár merkingar orðsins *tungutækni*

- Orðið *tungutækni* hefur þrjár merkingar
  - vissulega nátengdar, en þó aðskildar
- Þverfagleg fræðigrein
  - sem byggist á málvísindum og tölvunarfræði
- Hugbúnaður og tæki
  - sem byggjast á fræðilegum rannsóknum
- Iðnaðarstarfsemi
  - þar sem fengist er við gerð tungutæknitóla

Forsendur fyrir íslenskri tungutækni

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

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

Áætlaður kostnaður

	MKR
• Þróunarmiðstöð	25-50
• Rannsókna- og þróunarsjóður	150
• Styrkir til stærri alþjóðlegra verkefna	30
• Stutt hagnýtt nám í máltækni	10
• Meistaranám í tungutækni	10

Alls 225-250

á ári í 4-5 ár

–

Hvað hefur fengist?

	MKR
• Fjáraukalög 2000	40
• Fjárlög 2001	64,5
• Fjárlög 2002	0
• Fjárlög 2003	15
• Fjárlög 2004	13,5
	Alls 133 MKR

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; ...)*
- Windows XP og Microsoft Office er komið á íslensku
  - og einnig ýmis önnur forrit
  - en óljóst hvaða útbreiðslu þýðingarnar fá

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.*
- Hér hefur staðan batnað
  - m.a. með aukinni útbreiðslu Unicode
- Nú er hægt að nota íslenska stafi í GSM
  - með takmörkunum þó

3. Málgreining

- *Unnið verði að þróun málgreiningar fyrir íslensku, með það að markmiði að geta greint íslenskan texta í orðflokka og setningarliði.*
- Tvö verkefni á þessu sviði hafa verið styrkt af Tungutæknisjóði:
  - málfræðilegur markari (grammatical tagger)
  - vélræn íslensk setningagreining

3.1 Textaheild – 3.2 Orðasafn

- *Koma þarf upp stórri tölvutækri textaheild með íslenskum textum af sem fjölbreyttustum toga til að byggja áframhaldandi vinnu á.*
- Vinna við slíka textaheild er nýhafin

- *Koma þarf upp fullgreindu orðasafni (með málfræðilegri og merkingarlegri greiningu) til nota í áframhaldandi vinnu.*
- Ekkert slíkt orðasafn er til
  - þótt til sé hráefni sem vinna mætti út frá

4. Hjálparforrit við ritun

- *Til verði góð hjálparforrit við ritun texta á íslensku, s.s. orðskiptiforrit, stafsetningar-leiðréttingarforrit, málfarsleiðréttingarforrit o.fl.*
- Nýtt forrit til stafsetningarleiðréttingar hefur verið unnið á vegum Microsoft
- Málfarsleiðréttingaforrit eru ekki til enn
  - en forvinna að slíku forriti er í gangi

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.*
- Talgervill Infovox hefur verið endurbættur
  - byggist nú á fullkomnari tækni en áður
  - er þó langt frá því að vera nógu góður
- Undirbúningsvinna að nýjum talgervli er hafin

6. Talgreining

- *Unnið verði að þróun talgreiningar fyrir íslensku, með það að markmiði að til verði forrit sem geti túlkað eðlilegt íslenskt tal.*
- Háskólinn og fjögur fyrirtæki stóðu að *Hjali*
  - íslenskri stakorðagreiningu
- Íslenskur talgreinir er nú til og virkar vel
  - en langt er í greiningu samfellds máls

7. Vélrænar þýðingar

- *Unnið verði að þróun forrita til vélrænna þýðinga milli íslensku og annarra tungumála, m.a. til að auðvelda leit í gagnabönkum.*
- Hér hefur lítið gerst
  - einstöku tilraunir hafa þó verið gerðar
  - ýmsir hafa unnið með þýðingarminni
  - en engin nothæf þýðingarforrit eru á leiðinni

8. Ábyrgðaraðilar

- *Ákveðnum aðilum (stofnunum eða fyrir-tækjum) verði falin ábyrgð á einstökum verkefnum.*
- Sett var á fót verkefnisstjórn í tungutækni
  - sem átti að hafa yfirlit yfir stöðu mála í landinu
  - ýta verkefnum af stað og samræma aðgerðir
- Þetta skilaði góðum árangri
  - en verkefnisstjórnin verður lögð niður um áramót

Tungutækniáætlunin á enda

- Tungutækniáætlunin hefur skilað sínu
  - menntun á sviði tungutækni er hafin
  - Íslendingar farnir að fara í nám erlendis
  - gagnasöfn hafa verið byggð upp
  - ýmsum verkefnum verið ýtt af stað
- En íslensk tungutækni er ekki orðin sjálfbær
  - nú þegar tungutækniáætlunin er á enda
  - og einmitt þyrfti meira fé í rannsóknir og þróun

Ógnar upplýsingatæknin tungunni?

- Þrjú einkenni upplýsingatækni skipta máli
  - þegar áhrif hennar á íslenska tungu eru metin
- Hún 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 ef móðurmálið er ekki gjaldgengt á sviði
  - sem er mikilvægt
  - í daglegu lífi
  - alls almennings?
- Hvað ef það er ekki nothæft
  - í nýrri tækni og öðru sem er nýtt og spennandi
  - á sviðum þar sem nýsköpun á sér stað
  - og þar sem ný atvinnutækifæri bjóðast?

Tungutækni fyrir málnotendur

- Tungutækni snýst ekki bara um málvernd
  - einnig um þjónustu og sjálfsvirðingu
- Eigum við að sitja við sama borð og aðrir
  - eða eigum við að sitja skör lægra?
- Við eigum kröfu á að geta notað móðurmálið
  - sem víðast, við sem fjölbreyttastar aðstæður
- Allt annað er uppgjöf

Fordæmi Eista

- Eistar eru smáþjóð eins og við
  - aðeins um ein milljón talar eistnesku
- Þeir hafa gert áætlun um þróun tungutækni
  - Estonian HLT Roadmap for 2004-2011
- Þeir eru núna á svipuðu stigi og við
  - en þeirra tungutækniáætlun er að byrja
  - okkar að enda
- Ætlum við að láta hér við sitja?

Þakka ykkur fyrir áheyrnina

eirikur@hi.is