# Íslensk tungutækni: tilgangur og forsendur

© Eiríkur Rögnvaldsson,

9. október 2002

# Hvað er tungutækni?

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

# Fyrstu tengsl tölva og tungumáls

- Tengsl tölva og tungumáls má rekja aftur til fyrstu ára tölvunnar um miðja 20. öld
- Fljótlega var farið að nota tölvur til að gera ýmiss konar orðaskrár, skoða tíðni orða í mismunandi textum o.s.frv.
- Talsvert var gert að því að leita höfunda texta eða skoða áhrif eins höfundar á annan
  - með því að bera saman orðaforða þeirra og orðtíðni

# *Lítill heimur*

- „Ja, rannsóknarverkefnið mitt var Shakespeare og T.S. Eliot,“ sagði Persse.
- „Þar hefði ég getað orðið þér að liði,“ greip Dempsey frammí. Hann var nýkominn á barinn ásamt Angelicu, sem var undrafögur í skósíðum serk úr vínrauðri bómull ofinni daufu mynstri af öðrum litum. „Þar hefði tölvuvinnsla einmitt verið vel við hæfi,“ hélt Dempsey áfram. „Þú hefðir ekki þurft annað en koma textanum á tölvutækt form og þá hefðirðu getað fengið tölvuna til að gera skrá yfir hvert einasta orð, orðasamband og setningarbyggingu sem er að finna hjá báðum þessum höfundum. Þú hefðir getað reiknað nákvæmlega út áhrif Shakespeares á T.S. Eliot.“

# Tölvuþýðingar

- Á 6. áratug 20. aldar og fram á þann 7. var miklu fé varið í tilraunir með tölvuþýðingar
- Fyrstu forritin þýddu texta orð fyrir orð
  - studdust ekki við málfræðikenningar eða líkön
- 1966 birti bandaríska vísindaakademían „svarta skýrslu“ um tölvuþýðingar
  - þar sem fram kom að þrátt fyrir gífurlegan kostnað hefði árangurinn verið ákaflega lítill

# Máltölvun

- Literary and Linguistic Computing
  - máltölvun
- Hvers kyns notkun tölva við lausn mállegra verkefna
  - talningar orða og bókstafa, tíðniskrár
  - orðstöðulyklar, orðabókagerð
- Ekki þörf á mikilli tölvukunnáttu
  - oft unnið með hjálp tilbúinna forrita eða forritapakka

# Tölvufræðileg málvísindi

- Computational Linguistics
  - tölvufræðileg málvísindi/tölvumálvísindi
- Að setja fram aðferðir (algrím) sem tölvur geta unnið með við greiningu tungumáls
  - undirstaða þess að hægt sé að nota tölvur við vélrænar þýðingar, lemmun, talgreiningu o.fl.
- Þeir sem semja mállýsinguna þurfa að hafa góða hugmynd um það hvernig tölvur vinna

# 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

# Afmörkun tungutækni

- En hvað er þá tungutækni?
  - þýðingar forrita?
  - tölvustudd orðabókargerð?
  - tölvunotkun í tungumálakennslu?
  - tölvustuddar þýðingar?
- Miða má við virka kunnáttu á báðum sviðum
  - nýtingu tölvutækni í þágu tungumálsins
  - eða tungumálsins í þágu tölvutækninnar

# Tæknin 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 og önnur hjálpartæki handa fötluðum
  - ýmiss konar kennsluforrit

# Tungumálið í þágu tækninnar

- Tungumálið gegnir sívaxandi hlutverki
  - innan upplýsingatækninnar
- 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 með rödd og tungumáli í stað þess að ýta á takka

# 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

# Íslensk tungutækni

- Kemur íslensk tungutækni af sjálfu sér
  - eigum við bara að bíða þolinmóð?
- Fáum við íslensk tungutæknitól að utan?
  - það er ólíklegt
  - tungutæknilausnir eru mjög dýrar
  - íslenski markaðurinn alltof lítill
- Sprettur tungutækni af sjálfu sér innanlands?
  - varla – af sömu ástæðum

# Menntun og rannsóknir

- Þekking, menntun, reynsla
  - ekkert nám af þessu tagi hefur verið til á Íslandi
  - engar rannsóknir hafa verið á þessu sviði
  - fáir Íslendingar búa yfir þekkingu og reynslu
- Úr þessu þarf að bæta
  - og um það voru gerðar tillögur í skýrslu starfshóps um tungutækni vorið 1999

# Úr skýrslu starfshóps um tungutækni

- Óráðlegt er að ætla að Íslendingar geti byggt upp öflugt starf á sviði tungutækni án þess að hyggja að fræðilegum undirstöðum slíks starfs. Nauðsynlegt er að fá sem fyrst til starfa vel menntað fólk á sviði íslensks máls og tölvunarfræði sem gerir sér grein fyrir sérkennum íslenskrar málfræði og þörfum íslensks málsamfélags.

# ... og áfram:

- Ef ekki verður byggð upp innlend þekking á þessu sviði innan menntastofnana verðum við um ófyrirsjáanlega framtíð þiggjendur á þessu sviði og höfum miklu minni möguleika á að bregðast við breyttum aðstæðum og nýjungum, og þróa þau tól og tæki sem henta best íslenskum aðstæðum.

# Þetta svið á sér víða langa hefð

- Computational linguistics
  - í enskumælandi löndum
- Datalingvistik
  - á Norðurlöndum
- Mikill vöxtur hefur verið í þessum greinum
  - samfara örri þróun í tungutækni sem iðngrein
- En jafnframt hafa áherslur breyst

# Aukin áhersla á hagnýtingu

- Greinar með áherslu á hagnýtingu í ýmiss konar tækjum og tólum hafa komið upp
  - við hlið hefðbundinna akademískra greina
- Language technology
  - í stað eða við hlið Computational Linguistics
- Sprogteknologi/språkteknologi
  - í stað eða við hlið Datalingvistik

# Tungutæknieiningar

- Gagnasöfn og greiningartæki
  - nýtt sem hráefni í tungutæknitól
- Langflest verkefni innan tungutækni byggjast á einhvers konar mállegum gagnasöfnum
- Þrenns konar söfn skipta mestu máli:
  - orðasöfn
  - textasöfn
  - hljóðsöfn

# Tungutæknisöfn og orðabækur

- Tvenns konar munur
  - á rafrænum orðabókum og tungutæknisöfnum
- Tungutæknisöfnin þurfa að vera ítarlegri
  - stafsetning, orðflokkur, beyging, merking
  - setningareiginleikar, orðastæður, stílgildi ...
- Tungutæknisöfnin þurfa að vera stöðluð
  - allar upplýsingar settar fram á samræmdan hátt

# Málheildir og gagnsemi þeirra

- Málheild (e. *corpus*)
  - safn valinna texta sett saman eftir föstum reglum
  - um efnisflokka, kyn og aldur höfunda o.s.frv.
- Stórar málheildir eru grundvallarforsenda fyrir þróun ýmissa tungutæknitóla
  - leiðréttingarforrita
  - þýðingarforrita
  - samræðukerfa (e. *dialogue systems*)

# Mörkun texta

- Mörkun (e. *tagging*)
  - að merkja einingar í texta á kerfisbundinn hátt
    - bókstafi, orð, setningar; sérnöfn; erlend orð; o.s.frv.
- Orðflokksmörkun (e. *PoS tagging*)
  - *Gamla*<lo> *konan*<no> *mætti*<so> *þessum*<fn> *tveim*<to> *drengjum*<no> *í*<fs> *morgun*<no>
- Málfræðimörkun
  - kyn, tala, fall, persóna, háttur, tíð, stig, ákveðni

# Mörkun og málfarsleiðrétting

- Málfarsleiðrétting er útilokuð án greiningar:
  - villur felast sjaldan í notkun óleyfilegra mynda
    - *föður**s*** í stað *föður*
    - *k**ey**ptu* í stað *k**au**ptu*
  - fremur í að nota réttar myndir á röngum stöðum
    - *Ég hitti syst**i**r þína* > *syst**u**r*
    - *vegna þeirrar tilhneiging**u*** > *tilhneiging**ar***
    - *fjöldi manna kom**u*** > *kom*
    - ***mér** langar* > ***mig** langar*

# Stafsetning og vélrænar þýðingar

- Sama gildir um stafsetningarleiðréttingu
  - margar villur finnast aðeins með málgreiningu
    - *það er kominn morgu**n*** > *morgun**n***
    - *ég hitti Kristi**nn*** > *Kristi**n***
    - *hann er fari**n*** > *fari**nn***
- Vélrænar þýðingar krefjast málgreiningar
  - annars eru þær bara uppfletting í orðasafni
    - *hot spring river this book* (*hver á þessa bók*)

# Tilgangur

- Er rétt að verja stórfé
  - í uppbyggingu og þróun íslenskrar tungutækni?
- Er ekki best að bíða
  - og sjá hverju fram vindur?
- Þrenns konar réttlæting fyrir tungutækni
  - nýsköpun þekkingar
  - verndun og varðveisla tungumálsins
  - virðing og samkeppnisstaða málnotenda

# 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?

# Tungumál í hættu

- Við þær aðstæður hefst dauðastríð tungunnar
  - móðurmálið verður víkjandi
  - aðeins hæft til heimabrúks
  - en ekki til neinna alvarlegra hluta
- Ungt fólk 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

- Að hafna tækninni en halda tungunni
  - látið eiga sig að tileinka okkur ýmsar nýjungar
  - fyrst tungumálið er ekki gjaldgengt á þessu sviði
- Þessi kostur er ekki raunhæfur
- Að fórna tungunni en fylgjast með tækninni
  - nota ensku í upplýsinga- og tölvutækni
  - úr því að íslenska er ekki nothæf á því sviði
- Þessi kostur er óviðunandi

# – og sá þriðji:

- Að hefjast handa
  - gera átak á sviði tungutækni
  - gera íslensku nothæfa innan upplýsingatækninnar
- Það er eini valkostur okkar
  - ef við viljum halda áfram að nota íslensku
  - á öllum sviðum þjóðlífsins
- Annars verður málið fljótlega forngripur
  - dauðadæmt og gæti dáið út á fáum áratugum

# 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

# Tákn og tungumál

- Við munum aldrei hafa allt á íslensku
  - hvað með R, N, P á gírstönginni í bílnum okkar?
  - þetta stendur fyrir *rear*, *neutral*, *park*
  - en fyrir okkur eru þetta bara tákn, óháð tungumáli
- Mál í virkri notkun getur aldrei orðið tákn
  - á sama hátt – orðin slitna ekki frá tungumálinu
- Þess vegna verður málið að vera íslenska
  - að öðrum kosti verðum við málfarslega undirokuð

# Ég þakka áheyrnina

- eirikur@hi.is