HÁSKÓLI ÍSLANDS

Íslensk talgreining: Efniviður og úrvinnsla

Málfræðispjall 21. nóvember 2003

Eiríkur Rögnvaldsson

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

Niðurstöður starfshóps um tungutækni

- 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

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

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

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.

Hjal

- Fyrir ári tóku nokkrir aðilar sig saman
  - og ákváðu að reyna að koma upp hljóðsafni
  - sem nýta mætti í íslenska talgreiningu
- Þátttakendur í verkinu voru fimm
  - Háskólinn, Landssíminn, Hex, Nýherji, Grunnur
- Sótt var um styrk úr Tungutæknisjóði
  - sem veitti 14,8 milljónir til verksins, *Hjals*

Forsvarsmenn þátttakenda

- Frá undirritun samnings, 31. mars 2003

Stjórn og starfsmenn

- Formaður stýrihóps:
  - Sæmundur Þorsteinsson, Landssímanum
- Verkefnisstjóri:
  - Helga Waage, Hex
- Umsjón með málfræðilegum hluta:
  - Eiríkur Rögnvaldsson, Háskóla Íslands
- Starfsmenn – stúdentar í tungutækni:
  - Björn Kristinsson, Geir Gunnarsson, Jón Pétur Friðriksson, Valdís Ólafsdóttir

Markmið og eðli greiningar

- Markmið:
  - söfnun hráefnis í hljóðsafn
  - til að gera tölvum kleift að skilja íslenskt talmál
- Stakorðagreining
  - greining einstakra orða
  - ekki greining formgerðar eða merkingar
- Slík greining er vissulega takmörkuð
  - en nýtist þó mjög vel á mörgum sviðum

Forsendur: tækni og hráefni

- Samstarf við tungutæknifyrirtækið *ScanSoft*
  - sem sá um þjálfun talgreinisins
  - með tækni sem er óháð tungumálum
  - og hefur verið beitt á tæp 50 mál
- Hráefni við gerð íslensks talgreinis
  - upptökur með framburði 2000 Íslendinga
  - hljóðritun á þessum upptökum
  - hljóðritað safn algengustu orðmynda málsins

Hljóðritunarkerfi

- International Phonetic Alphabet (IPA)
  - alþjóðlegt kerfi; notar mikinn fjölda tákna
    - óþjált vegna takmarkana lyklaborðs og ýmissa forrita
- Tvö kerfi gerð til notkunar í tölvum
  - ARPAbet
    - ameríska; notar ASCII (< 128)
  - SAMPA
    - evrópskt ; notar ASCII (< 128)

Íslenskt SAMPA

- *ScanSoft* notar SAMPA
  - því þurfti að koma upp íslenskum SAMPA-staðli
  - velja hljóðtákn fyrir íslensk málhljóð
- Ákveðið var að víkja sem minnst frá hefðinni
  - í vali tákna fyrir einstök hljóð
- Hljóðritun er líka að flestu leyti hefðbundin
  - lengd samhljóða er þó ekki táknuð sérstaklega

Samanburður SAMPA og IPA

SAMPA	*IPA*	*Orð*	*Hljóðritun*
– f	f	finna	/**f**Ina/
– v	v	vera	/**v**E:ra/
– D	ð	viður	/vI:**D**Yr/
– T	θ	þunnur	/**T**YnYr/
– s	s	sofa	/**s**O:va/
– j	j	játa	/**j**au:da/
– C	ç	hjóla	/**C**ou:la/
– G	ɣ	saga	/sa:**G**a/
– x	x	rækta	/rai**x**da/
– h	h	halda	/**h**alda/

Nauðsyn mikils hráefnis

- Hvers vegna þessi fjöldi þátttakenda?
  - framburður sama orðs getur verið mismunandi
  - bæði milli manna og hjá sama málhafa
- Talgreinirinn styðst eingöngu við hljóðbylgjur
  - í greiningu á hljóðum og orðum
  - hefur ekki stuðning af setningagerð eða merkingu
- Því þarf fjölda dæma um hvert hljóðasamband
  - til að koma upp traustu greiningarlíkani

Textablöð

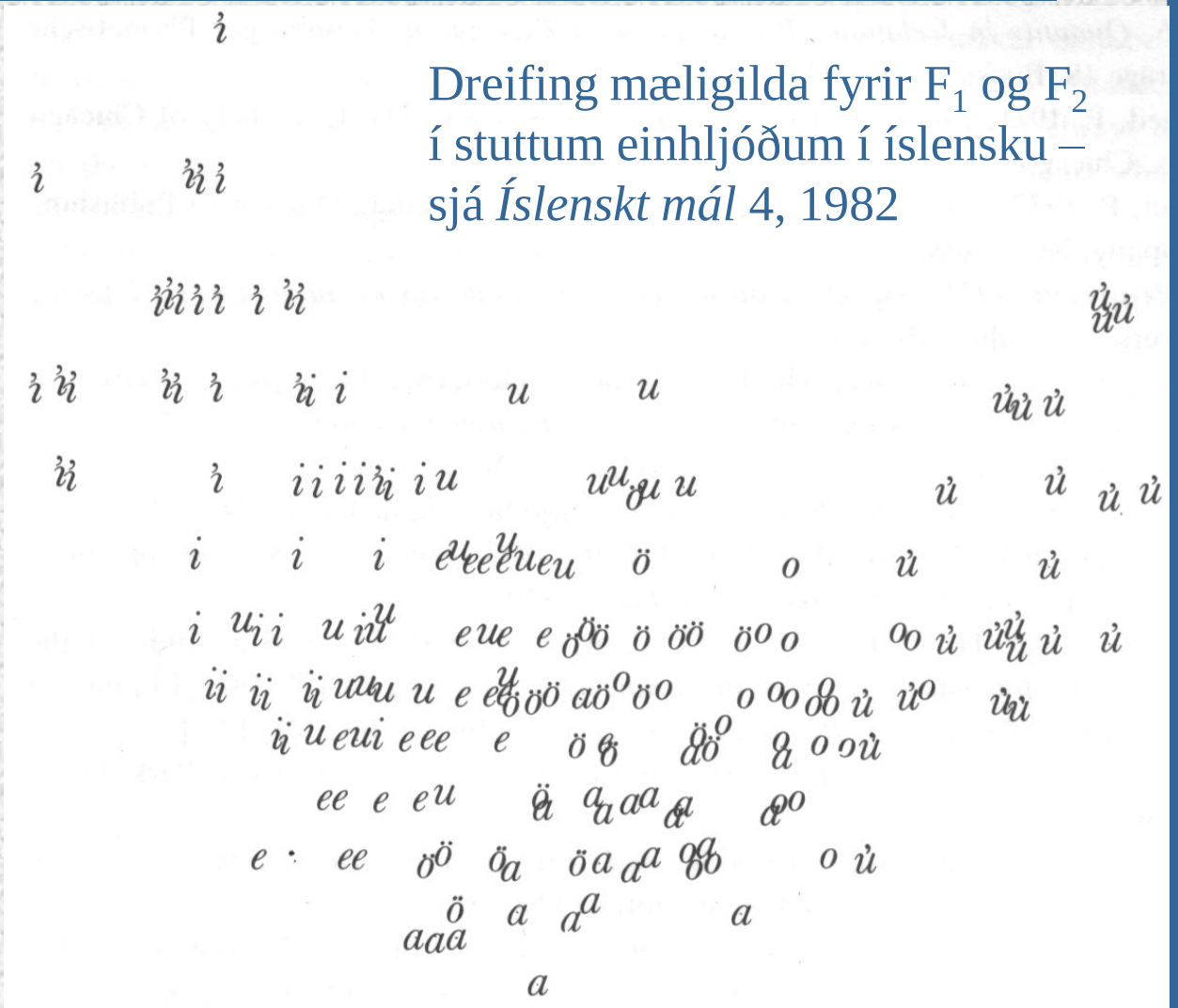
- Útbúin voru 1000 mismunandi textablöð
  - sem þátttakendur lásu upp í síma
    - aðeins 2-3 lásu sama blaðið
- Á blöðunum voru ýmis algeng orð
  - sem tengjast líklegu notkunarsviði talgreinisins
    - dagsetningar, tölur, manna- og staðanöfn, ...
- Einnig voru þar heilar setningar
  - valdar eftir hljóðasamböndum sem þær geymdu
    - til að fá framburð allra hljóðasambanda málsins

Hljóð og hljóðasambönd

- Hljóð verða fyrir áhrifum frá umhverfinu
  - því þarf að greina öll hugsanleg hljóðasambönd
- Líkan *ScanSoft* byggist á tví- og þrístæðum
  - því þurfti að tryggja að þær kæmu allar með
- Dæmi: *valur*
  - tvístæður: #v – va – al – lu – ur – r#
  - þrístæður: #va – val – alu – lur – ur#

Skörun einstakra hljóða

- Mismunur /a/ eftir umhverfi
  - *valur*
  - *gat*
  - *safi*
  - *maður*



Dreifing mæligilda fyrir F_1 og F_2 í stuttum einhljóðum í íslensku – sjá *Íslenskt mál* 4, 1982

Val setninga

- Hljóðmynstur fengust úr *Íslenskri rímorðabók*
  - alls 182 mynstur
  - tæpar 800 tví- og þrístæður
- Setningar valdar úr textasafni eftir tvístæðum
  - tryggt að nægilegur fjöldi allra kæmi með
    - a.m.k. 500 dæmi um þær helstu, 50 um sjaldgæfar
- Valdar voru 5-12 orða setningar
  - samtals 1433 notaðar af tæplega 90 þúsund

Dæmi um mynstur

- V eitthvert sérhljóð
- V* öll sérhljóð
- J* eitthvert af j/i/í/e/æ
- J- annað en j/i/í/e/æ
- C+ einfaldur eða tvöfaldur samhljóði
- $ hvaða strengur sem er – má aðeins hafa eitt sérhljóð

- V k J-
- # k J*
- V* k J*
- V* k k J-
- V* k k J*
- # k l
- # k n
- $ k r
- V k r
- V k s
- # k V*
- V k+ l
- V k+ n
- # k V*

Hljóðritun og þjálfun

- Upptökurnar voru síðan skráðar
  - með venjulegri stafsetningu
  - aðeins hljóðritað ef framburður var „óvenjulegur“
    - `fimmta<fImta>`
  - og upptökur og skráning sent til *ScanSoft*
- Einnig var hljóðritað með SAMPA
  - safn með 50.000 algengustu orðmyndum málsins
  - völdum úr *Morgunblaðinu*, skáldsögum og *Ístal*

Dæmi úr orðaskrá

félag
félaga
félagana
félaganna
félagar
félagarnir
félagasamtaka
félagasamtök
félagasamtökum
félagi
félagið
félaginn
félagins
félaginu
félags
félagsfræði
félagsfræðingur
félagsfund
félagsfundi
félagsgjöld
félagsheimili
félagsheimilið
félagsheimilinu
félagsins
félagsleg
félagslega
félagslegan
félagslegar
félagslegra
félagslegrar
félagslegri
félagslegs
félagslegt
félagslegu
félagslegum
félagslið
félagsliða
félagsliðum
félagslíf
félagslífi
félagslífið
félagslynd
félagslyndur
félagsmaður
félagsmanna
félagsmál
félagsmála
félagsmálaráðherra
félagsmálaráðuneytið
félagsmálaráðuneytinu
félagsmálaráðuneytisins
félagsmálastjóri
félagsmálastofnun
félagsmálum
félagsmenn
félagsmiðstöð
félagsmiðstöðinni
félagsmiðstöðva
félagsmiðstöðvar
félagsmönnum
félagsráðgjafa
félagsráðgjafar
félagsráðgjafi
félagsráðgjafinn

Íslenskt framburðarorðasafn

- Slík framburðarsöfn eru mjög mikilvæg
  - í talkennslum (automatic speech recognition, ASR)
  - og talgervingu (text-to-speech, TTS)
- Einn framburður var valinn sem aðalafbrigði
  - en helstu mállýskur einnig sýndar
    - **banki** `sEDlabauJ0J_I`
    - **banki<sEDlabauJcI>** `sEDlabauJcI`
    - **banki<sEDlabaJ0J_I>** `sEDlabaJ0J_I`

Þjálfun talgreinis

- *ScanSoft* sá um þjálfun sjálfs talgreinisins
  - greiningarbúnaður ber saman upptökur og skráningu
  - býr til líkan um samsvörun hljóðbylgna og hljóða
- Síðan þarf að raða hljóðunum saman í orð
  - til að vita hvað sagt var
- Við þetta er beitt þekktum aðferðum
  - m.a. HMM (Hidden Markov model)

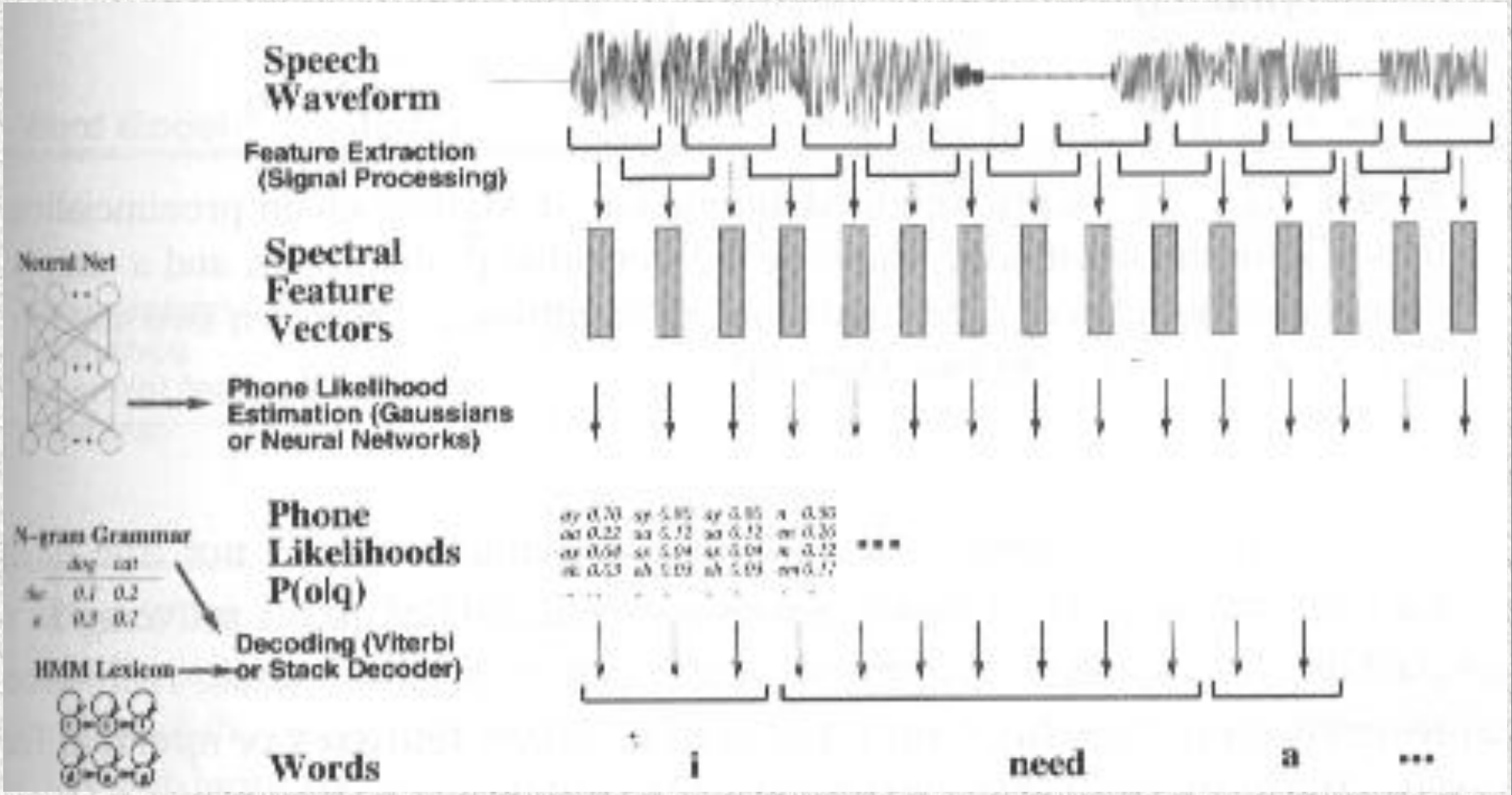
„Noisy channel“ líkanið

- Reynt er að reikna út áhrif „hávaða“ á orðið
  - hávaðinn er þá hvers konar „truflun“ á leiðinni
    - einstaklingsbundinn framburðarmunur
    - breytileiki í framburði (mállýskur, óskýrmæli ...)
    - áhrif orða í umhverfinu
    - ytri áhrif (suð í síma, umhverfishljóð ...)

Að finna rétta orðið

- Reikna þarf út líkindi þess að
  - tiltekið orð komi fram sem tiltekinn hljóðastrengur
  - tiltekið hljóð komi fram sem tilteknir eðlisfræðilegir þættir
- Leitað er að orðum í framburðarorðasafni
  - sem gætu samsvarað hljóðastrengnum
  - og stuðst við mállíkan og afkótara (decoder)

Einfaldað líkan af talgreini



Villur í greiningu

- Oft finnst strengurinn í orðasafninu
  - þá er gert ráð fyrir því að rétt greining sé fundin
- Stundum finnst strengurinn þó ekki
  - en annar mjög svipaður þess í stað
  - þá er líklegt að villa hafi verið gerð í greiningu
- Koma þarf upp reglum um líklegar villur
  - t.d. ekki ólíklegt að önghljóðum sé ruglað saman
    - *soða* fyrir *sofa* væri „eðlileg“ villa

Setningafræði í talgreiningu

- Ekki er öruggt að greining sé rétt
  - þótt strengurinn finnist í orðasafninu
    - *liður* gæti verið villa fyrir *lifur*, þótt *liður* sé til
- Hér kæmu setningarlegar upplýsingar að gagni
  - umhverfið myndi oft skera úr vafaatriðum
- Þær þarf að vinna úr fullgreindri málheild
  - sem ekki er til fyrir íslensku
  - en nauðsynlegt er að koma upp

Samræðusamhengi

- Setningarlegt samhengi er ekki tiltækt
  - því verður að skapa annað samhengi í staðinn
- Talgreinirinn styðst við samræðusamhengi
  - handrit að samræðum er skrifað
  - notandanum lögð orð í munn til að velja á milli
  - kerfið látið hlusta eftir þeim orðum
  - og öðrum skyldum sem gætu komið í staðinn
    - samheitum, öðrum beygingarmyndum

Mikilvægi handrits að samræðum

- Notkun stakorðagreinis er þjónustumiðuð
  - því er hægt að skilgreina orð til að hlusta eftir
  - og auðvelda þannig greininguna til muna
- Því skiptir miklu máli að handritið sé gott
  - notandinn leiddur þægilega áfram
  - hugsað fyrir öllum eðlilegum viðbrögðum hans
  - beðið um staðfestingu á að rétt sé greint

Íslenski talgreinirinn

- Íslenski talgreinirinn er tilbúinn
  - hefur verið prófaður talsvert og virkar vel
  - greinir rétt í yfir 95% tilvika
- Á næstu vikum verður farið að nýta hann
  - í ýmiss konar þjónustusímum
- Gerið svo vel að prófa
  - hringið í síma 595-6606