Hjal: vélræn íslensk talgreining

Hugvísindaþing 31. október 2003

Eiríkur Rögnvaldsson

Aðstandendur verksins

- Þátttakendur:
  - Háskóli Íslands
  - Landssíminn
  - Hex
  - Nýherji
  - Grunnur gagnalausnir
- Styrktaraðili:
  - Tungutæknisjóður

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:
  - að gera tölvum kleift að skilja íslenskt talmál
  - talkennsl, talgreining (speech recognition)
- 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 sér um þjálfun talgreinisins
  - með tækni sem er óháð tungumálum
  - og hefur verið beitt á meira en 40 mál
- Hráefni við gerð íslensks talgreinis
  - upptökur með framburði 2000 Íslendinga
  - hljóðritun á þessum upptökum
  - hljóðritað safn algengustu orðmynda málsins

Nauðsyn mikils hráefnis

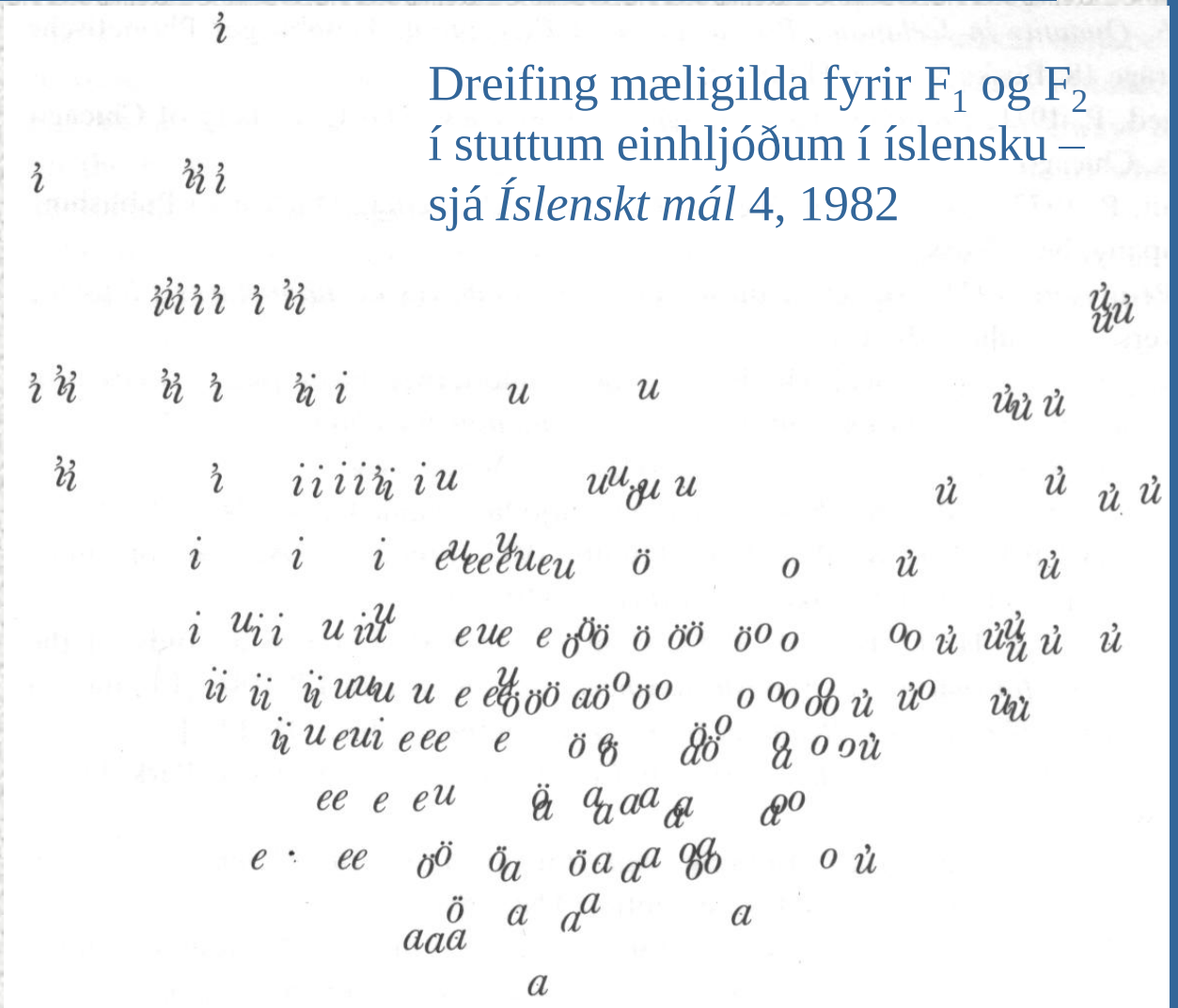
- Hvers vegna þessi fjöldi?
  - framburður sama orðs getur verið mismunandi
  - bæði milli manna og hjá sama málhafa
- Talgreinirinn styðst eingöngu við hljóðbylgjur
  - í greiningu á hljóðum og orðum
  - hefur ekki stuðning af setningagerð eða merkingu
- Því þarf fjölda dæma um hvert hljóðasamband
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

Textablöð

- Útbúin voru sérstök textablöð
  - sem þátttakendur lásu upp í síma
- 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

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

Hljóð og hljóðasambönd

- Hljóð verða fyrir áhrifum frá umhverfinu
  - því þarf að greina öll hugsanleg hljóðasambönd
- Tví- og þrístæður eru greindar
  - tvístæður í *valur*
    - #v – va – al – lu – ur – r#
  - þrístæður í *valur*
    - #va – val – alu – lur – ur#
- Nokkur mynstur
  - V b b
  - # b j
  - # b l
  - # b r
  - # b V*
  - V d d
  - # d j
  - # d r
  - # d v

Hljóðritun og þjálfun

- Upptökurnar voru síðan hljóðritaðar
  - með hljóðritunarkerfinu SAMPA
  - og upptökur og hljóðritun sent til *ScanSoft*
- *ScanSoft* sér um þjálfun sjálfs talgreinisins
  - greiningarbúnaður ber saman upptökur og hljóðritun
  - býr til líkan um samsvörun hljóðbylgna og hljóða/hljóðasambanda

Hljóðum raðað saman í orð

- Síðan þarf að raða hljóðunum saman í orð
  - til að vita hvað sagt var
- Þar styðst talgreinirinn við hljóðritað orðasafn
  - með 30-40 þúsund algengustu orðmyndum málsins
- Greiningarstrengur er borinn saman við safnið
  - leitað að samsvarandi streng þar
  - athugað hvort um íslenskt orð geti verið að ræða

Villur í greiningu

- Oft finnst strengurinn í orðasafninu
  - þá er gert ráð fyrir því að rétt greining sé fundin
- Oft 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

Og svo ...

- Hljóðritun lauk í lok ágúst
  - þá var hún send til *ScanSoft* ásamt upptökunum
- Síðan hefur þjálfun talgreinisins staðið yfir
  - og er nú lokið

- Við bíðum spennt eftir að heyra árangurinn!

Þakka ykkur fyrir áheyrnina

eirikur@hi.is