Máltækni og málföng fyrir íslensku

Eiríkur Rögnvaldsson

Efni erindisins

- Máltækni – hvað er það?
  - tækni í þágu tungumáls, tungumál í þágu tækni
- www.málföng.is – nýr vefur fyrir íslensk málföng
  - tvö ný íslensk málgögn, MÍM og IcePaHC
- META-NORD verkefnið
  - máltækniskýrslur, staða íslenskrar máltækni
- Málræktarþing 27. apríl
  - Máltækni fyrir alla

Hvað er máltækni?

- **Máltækni** 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

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
  - samskipti við upplýsingaveitur
    - 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

Málföng

- **Málföng** er einnig ungt nýyrði
  - fyrir ‘language resources’
  - hugbúnaður (máltæki) og gagnasöfn (málgögn)
- Komið hefur verið upp nýrri vefsíðu
  - þar sem íslensk málföng eru skráð
  - sum þeirra eru einnig vistuð þar
  - en krækjur eru á önnur
- http://www.málföng.is/

Mörkuð íslensk málheild, MÍM

- Í vinnslu hjá Stofnun Árna Magnússonar
  - u.þ.b. 25.000.000 lesmálsorð
  - úr 900–1.000 textabútum
    - hver allt að 40.000 lesmálsorð, a.m.k. 10% sleppt
- Málheildin er skráð með stöðluðu sniði
  - XML-útgáfa af TEI-sniði fyrir málheildir
    - hverjum textabút fylgja bókfræðilegar upplýsingar
    - hverri orðmynd fylgir mark og nefnimynd

Efni

- Allt efni frá 2000 og síðar
- Textinn verður markaður vélrænt

Textaflokkar í Markaðri íslenskri málheild	fjöldi orða	%
Ræður fluttar á Alþingi	250.000	1,0
Blogg	1.964.495	7,6
Dagblöð (Morgunblaðið, Fréttablaðið)	7.222.133	27,9
Dómar	316.134	1,2
Efni til upplestrar	222.872	0,9
Fréttir útvarps og sjónvarps	287.554	1,1
Skýrslur og greinargerðir af vefsetrum ráðuneyta	1.658.618	6,4
Frumvörp og lög af vef Alþingis	747.914	2,9
Lokaritgerðir háskólastúdenta	485.165	1,9
Stúdentsprófsritgerðir í íslensku	178.949	0,7
Af vefsetrum fyrirtækja, samtaka og stofnana	1.594.504	6,2
Textavarp	42.520	0,2
Safnaðarblöð	6.472	0,0
Texti um tónlist	24.357	0,1
Prentuð tímarit af ýmsu tagi	2.243.084	8,7
Vefmiðlar	243.750	0,9
Veftímarit	145.399	0,6
Tölvupóstlistar	121.164	0,5
Pistlar af Vísindavef	1.770.184	6,8
Textar úr bókum	5.770.545	22,3
Talmál	574.732	2,2
Samtals	25.870.545	100

Hvað má lesa út úr málheildinni?

- Upplýsingar um
  - tíðni orðflokka, orða, beygingarmynda
  - orðasambönd
  - setningagerð
  - merkingu
- Nú eru notuð flókin tölfræðilíkön
  - til að vinna upp þessar upplýsingar
  - og hagnýta þær

Nýting

- Stór málheild er nauðsynleg fyrir
  - þróun þýðingarforrita
  - orðabókagerð
  - þróun máltæknitóla
    - t.d. fyrir talgreiningu og talgervingu
  - þróun hjálparforrita með ritvinnslu
    - t.d. forrita sem leiðbeina um stafsetningu og málfræði
- En eykur líka almenna þekkingu á málinu

Sögulegi íslenski trjábankinn

- Tvenns konar nýting
  - ætlaður bæði fyrir máltækni og málfræðirannsóknir
- Spönnun í tíma
  - rúm 800 ár – varla raunhæft í öðrum tungumálum
- Stærð
  - ein milljón orða – aðeins tveir trjábankar stærri
- Aðgengi
  - algerlega opinn og ókeypis, notkun án takmarkana

Textategundir og orðafjöldi

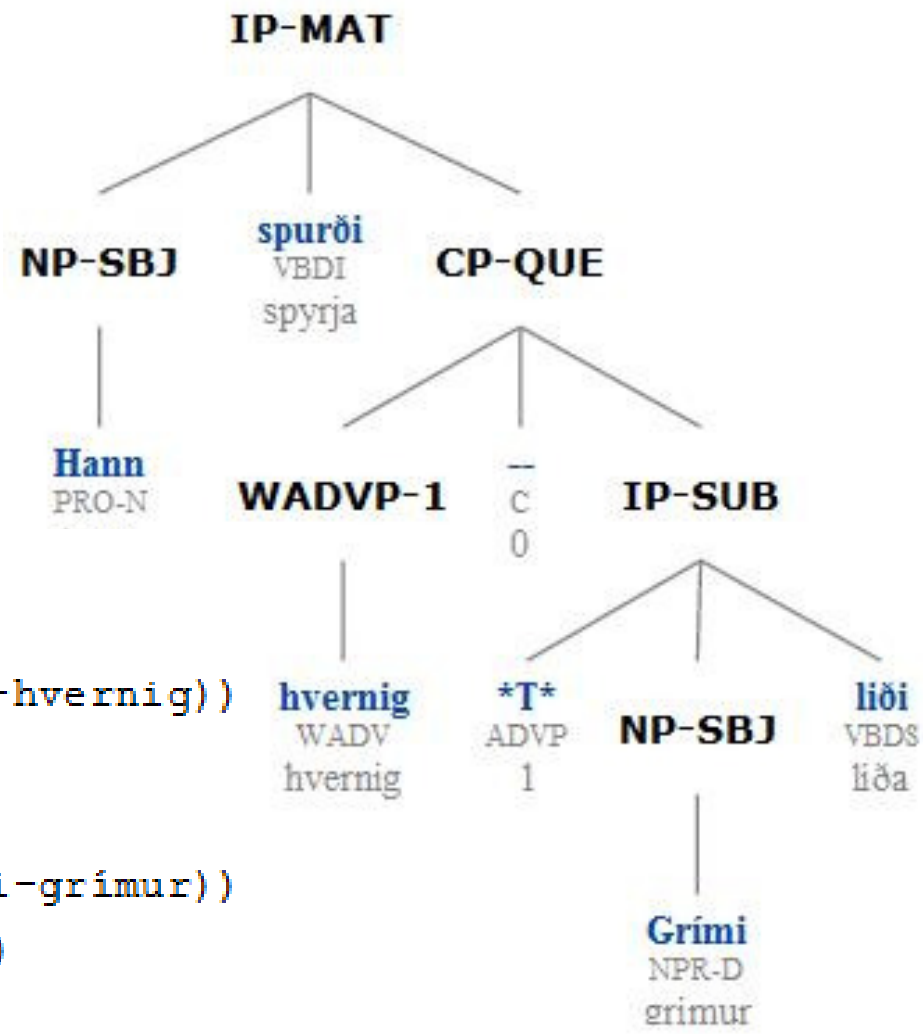
Öld	Frásagnir	Trúarrit	Ævisögur	Fræði	Lög	Samtals
12.	0	40871	0	4439	0	45310
13.	93463	21196	0	0	6183	120842
14.	77370	21315	0	0	0	98685
15.	111560	0	0	0	0	111560
16.	35733	60464	0	0	0	96197
17.	46281	28134	52997	0	0	127412
18.	63322	22963	22099	0	0	108384
19.	100362	20370	0	3268	0	124000
20.	103921	21234	0	0	0	125155
21.	43102	0	0	00	0	43102
Samtals	675114	236547	75096	7707	6183	1000647

Leitað í trjábankanum

- Hægt er að sækja trjábankann í heild
  - ásamt leitarhugbúnaði fyrir Windows og Linux
    - http://www.linguist.is/icelandic_treebank/Download
- Leitað er með forritinu CorpusSearch
  - sem notar sérstakt fyrirspurnamál
  - og gerir kleift að setja fram flóknar fyrirspurnir
- Bankinn hefur einnig verið settur inn í INESS
  - safn trjábanka við Háskólann í Bergen
  - hægt að skoða setningar og leita að formgerðum

Setning í IcePaHC

- Úttak úr INESS →
- Úttak úr CorpusSearch
  ↓




```
( (IP-MAT (NP-SBJ (PRO-N Hann-hann))
        (VBDI spurði-spyrja)
        (CP-QUE (WADVP-1 (WADV hvernig-hvernig))
              (C 0)
              (IP-SUB (ADVP *T*-1)
                    (NP-SBJ(NPR-D Grími-grímur))
                    (VBDS liði-líða))))
(ID 1888.GRIMUR.NAR-FIC,.301))
```

Aðgengi og nýting

- Bankinn er opinn og ókeypis til niðurhals
  - ekki háður neinum leyfum
  - flestir textar löngu komnir úr höfundarrétti
  - en leyfi höfunda var fengið fyrir nýlegum textum
- Hver sem er má gera hvað sem er við bankann
  - við teljum að það sé allra hagur
  - bankinn hefur nú þegar verið nýttur í rannsóknum
  - og verður vonandi nýttur í máltækni á næstunni

META-NORD

- META-NORD er verkefni sem Norðurlönd og Eystrasaltslönd standa að. Það er eitt þriggja systurverkefna innan META-NET sem hafa að markmiði að efla málleg gagnasöfn sem nýst geti í margvíslegum máltækniverkefnum
- Megintilgangur verkefnanna er að skapa tækni-legar forsendur fyrir margmála upplýsingasam-félagi í Evrópu þar sem allir geti notað móður-mál sitt við öflun og úrvinnslu upplýsinga

Markmið ESB með verkefnunum

- The overall objective is to ease and speed up the provision of online services centred around computer-based translation and cross-lingual information access and delivery
  - The focus is on assembling, linking across languages, and making widely available the basic language resources (models, tools and datasets) used by developers, professionals and researchers to build specific products and applications
  - Emphasis is placed on less-resourced languages

Þátttakendur í META-NORD

- Stjórnandi verkefnisins er Tilde í Riga
  - aðrir þátttakendur háskólar í Reykjavík, Gautaborg, Bergen, Kaupmannahöfn, Helsinki, Tartu og Vilnius
- Íslenski þátttakandinn er Máltæknisetur
  - samstarfsvettvangur Málvísindastofnunar HÍ, tölvunarfræðideildar HR og Árnastofnunar
- Fé til verkefnisins er u.þ.b. 380 milljónir króna
  - þar af rúmar 34 milljónir (á gengi dagsins) til Íslands

Verkefni META-NORD á Íslandi

- Vefsíða – vefir.hi.is/metanord
- Efling máltækni á Íslandi
  - samning málskýrslu á íslensku og ensku
  - vitundarvakning, kynning, útbreiðsla, áframhald
- Söfnun og uppbygging íslenskra málfanga
  - athugun og frágangur á réttindamálum og leyfum
  - val og söfnun málfanga og viðræður við rétthafa
  - lýsing, uppfærsla, stöðlun og frágangur málfanga

Tegundir málgagna

- Áhersla lögð á þrjár tegundir málgagna:
  - orðanet
    - venslaflokkuð orðasöfn
  - trjábanka
    - setningagreindar málheildir
  - (íð)orðasöfn
    - einmála og margmála
- Þessi gögn á að tengja milli mála eftir föngum
  - til að þau nýtist sem best í margmála verkefnum

Opin gögn og íslensk málstefna

- Gögnin verða aðgengileg á www.málföng.is
  - og einnig í META-SHARE, gagnabrunni META-NET
  - með stöðluðum leyfum s.s. CC, GPL, LGPL o.fl.
- Íslensk málstefna samþykkt á Alþingi 12.3. 2009
  - „[...] málleg gagnasöfn og hugbúnaður til að vinna með íslenskt mál verði gerð opin og frjáls eftir því sem kostur er (sbr. stefnu stjórnvalda um frjálsan og opinn hugbúnað)“

Deiling gagna

- Íslensk málföng eru flest gerð fyrir opinbert fé
  - sjaldnast komið upp í gróðaskyni
  - yfirleitt ekki mikil söluvara
- Það er allra hagur að gera þau aðgengileg
  - við fáum aðgang að gögnum annarra á móti
  - rannsóknir á íslensku máli stóraukast
  - þróun hugbúnaðar og málgagna fyrir íslensku vex

Málskýrslur um 30 Evrópumál

- 80-90 síðna skýrsla (hvítbók) fyrir hvert mál
  - hver skýrsla á viðkomandi máli og ensku
  - flestar í prentun og koma út á næstu vikum
- Í skýrslunum er gerð grein fyrir
  - málsamfélaginu og stöðu og hlutverki málsins í því
  - máltæknirannsóknum og máltækniiðnaði í landinu
  - hlutverki máltækni (afurða og þjónustu) í landinu
  - stöðu máltækni í samanburði við önnur Evrópumál

In everyday communication, Europe's citizens, business partners and politicians are inevitably confronted with language barriers. Language technology has the potential to overcome these barriers and to provide innovative interfaces to technologies and knowledge. This white paper presents the state of language technology support for the Icelandic language. It is part of a series that analyzes the available language resources and technologies for 31 European languages. The analysis was carried out by META-NET, a Network of Excellence funded by the European Commission. META-NET consists of 54 research centres in 33 countries, who cooperate with stakeholders from economy, government agencies, research organisations and others. META-NET's vision is high-quality language technology for all European languages.

Evrópubúar standa sífellt frammi fyrir tungumálaþröskuldum í daglegum samskiptum sín á milli – í viðskiptum, stjórnmálum og hversdagslífinu. Máltækni getur nýst til að yfirstíga þá þröskulda og skapa nýjan aðgang að tækni og þekkingu. Þessi hvítbók kynnir stöðu máltæknistuðnings við íslensku. Hún er hluti af ritröð þar sem gerð er grein fyrir tiltækum málföngum og máltækni fyrir 31 Evróputungumál. Að greiningunni stendur META-NET, sem er öndvegisnet fjármagnað af Evrópusambandinu. Innan META-NET eru 54 rannsóknarsetur í 33 löndum sem starfa með hagsmunaaðilum úr viðskiptalífinu, opinberum stofnunum, rannsóknarsetrum og ýmsum öðrum. Framtíðarsýn META-NET er að til verði hágæða máltækni fyrir öll evrópsk tungumál.

Máltækni er afar mikilvægur stuðningur við alls kyns málrannsóknir og styrkur við þá viðleitni að íslensk tunga verði notuð á öllum sviðum þjóðfélagsins í samræmi við opinbera íslenska málstefnu.
– Dr. Guðrún Kvaran (prófessor, formaður Íslenskrar málnefndar)

This is an excellent overview of the current state of language technology in Europe. This is a call to action for decision makers in countries that want their citizens to participate in the 21st century on equal footing with native English speakers.
– Helga Waage (Co-Founder and CTO at Mobilitus)

www.meta-net.eu

Springer

META-NET

White Paper Series

THE ICELANDIC LANGUAGE IN THE DIGITAL AGE

Hvítbókaröð

ÍSLENSK TUNGA Á STAFRÆNNI ÖLD

Eiríkur Rögnvaldsson
Kristín M. Jóhannsdóttir
Sigrún Helgadóttir
Steinþór Steingrímsson

Mat á stöðu máltækni

- Lagt var mat á stöðu hvers máls á sviði
  - talvinnslu (talgervingar og talgreiningar)
  - vélþýðinga
  - textagreiningar (mörkunar, þáttunar o.fl.)
  - málgagna (orðasafna, málheilda, trjábanka o.fl.)
- Gefnar voru einkunnir fyrir hvern þátt
  - frá 0 upp í 6
- Síðan var málunum raðað í fimm klasa
  - eftir einkunnum í þessum fjórum þáttum

Talvinnsla

Afburðagóður stuðningur	Góður stuðningur	Sæmilegur stuðningur	Brotakenndur stuðningur	Lítill/enginn stuðningur
	Enska	Finnska Franska Hollenska Ítalska Portúgalska Spænska Tékkneska Þýska	Baskneska Búlgarska Danska Eistneska Galisíska Gríska Írska Katalónska Norska Pólska Serbneska Slóvakíska Slóvenska Sænska Ungverska	**Íslenska** Króatíska Lettneska Litháíska Maltneska Rúmenska

Vélþýðingar

Afburðagóður stuðningur	Góður stuðningur	Sæmilegur stuðningur	Brotakenndur stuðningur	Lítill/enginn stuðningur
	Enska	Franska Spænska	Hollenska Ítalska Katalónska Pólska Rúmenska Ungverska Þýska	Baskneska Búlgarska Danska Eistneska Finnska Galisíska Gríska Írska **Íslenska** Króatíska Lettneska Litháíska Maltneska Norska Portúgalska Serbneska Slóvakíska Slóvenska Sænska Tékkneska

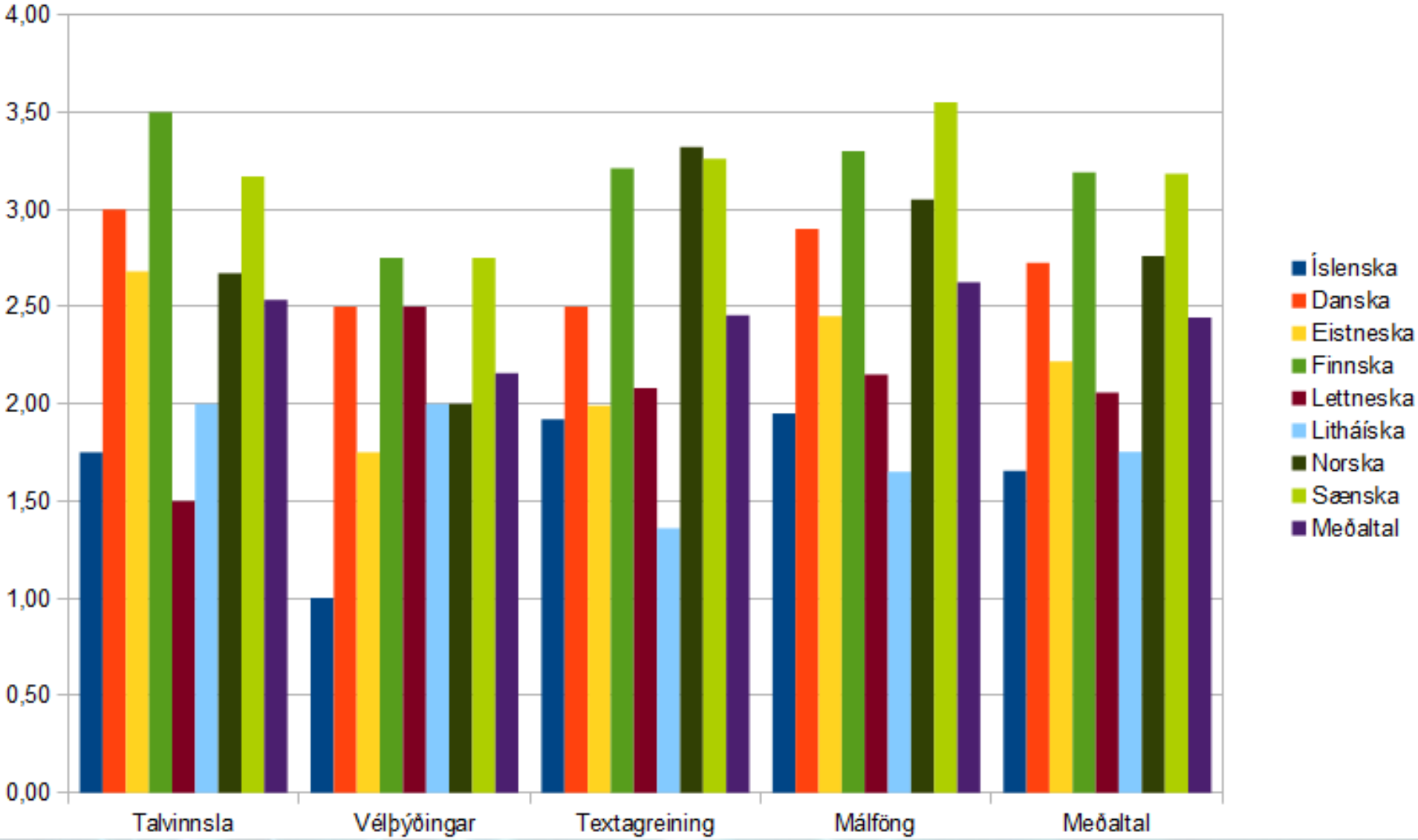
Textagreining

Afburðagóður stuðningur	Góður stuðningur	Sæmilegur stuðningur	Brotakenndur stuðningur	Lítill/enginn stuðningur
	Enska	Franska Hollenska Ítalska Spænska Þýska	Baskneska Búlgarska Danska Finnska Galisíska Gríska Katalónska Norska Pólska Portúgalska Rúmenska Slóvakíska Slóvenska Sænska Tékkneska Ungverska	Eistneska Írska **Íslenska** Króatíska Lettneska Litháíska Maltneska Serbneska

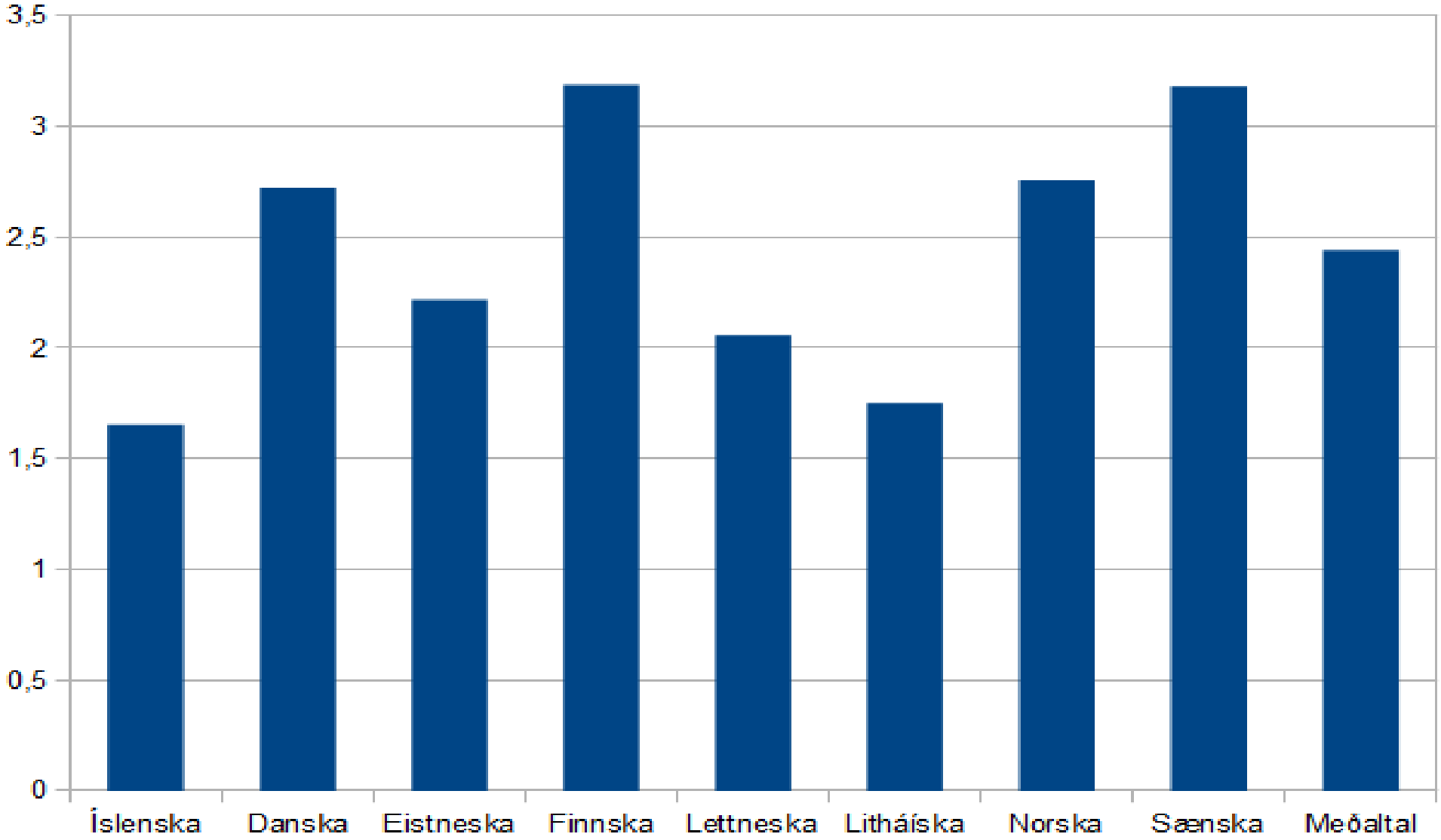
Málgögn

Afburðagóður stuðningur	Góður stuðningur	Sæmilegur stuðningur	Brotakenndur stuðningur	Lítill/enginn stuðningur
	Enska	Franska Hollenska Ítalska Pólska Spænska Sænska Tékkneska Ungverska Þýska	Baskneska Búlgarska Danska Eistneska Finnska Galisíska Gríska Katalónska Króatíska Norska Portúgalska Rúmenska Serbneska Slóvakíska Slóvenska	Írska **Íslenska** Lettneska Litháíska Maltneska

Staða META-NORD málanna



Heildarsamanburður



Niðurstöður

- Íslenska er ekki í bráðri hættu, þrátt fyrir yfirburði enskunnar í máltækni og tölvumálvísindum.
- Á hinn bóginn gæti staðan gerbreyst á svipstundu þegar ný kynslóð tækninnar fer fyrir alvöru að ráða við mannlegt mál á skilvirkan hátt.
- Eigi íslenska að vera lífvænleg þjóðtunga í þróuðum heimi verður hún að geta staðið undir kröfum upplýsingatækninnar.
- Fjárfesting í máltækni verður því að vera grunn-þáttur í framkvæmd íslenskrar málstefnu.

Máltækni fyrir alla, 27. apríl

13.00 – Setning ráðstefnunnar: Guðrún Kvaran, formaður Íslenskrar málnefndar

13.05 – Menntamálaráðherra Katrín Jakobsdóttir

13.15 – Íslensk máltækni í evrópsku samhengi - META-NORD og META-NET: Eiríkur Rögnvaldsson, stjórnandi META-NORD á Íslandi

13.30 – Del eller dø?: Sabine Kirchmeier-Andersen, framkvæmdastjóri dönsku málnefndarinnar

14.05 – Íslenska er málið: Tölvur og íslensk málstefna: Haraldur Bernharðsson, varaformaður Íslenskrar málnefndar

14.25 – Gagnasöfn frá sjónarhorni notandans: Kristín M. Jóhannsdóttir, málfræðingur hjá Háskóla Íslands

14.50 – Kaffi og veggspjaldasýning

15.20 – Talgervlar og tungumál sem fáir tala: Kristinn Halldór Einarsson, formaður Blindrafélagsins

15.40 – Almannarómur - Söfnun á íslensku talmáli fyrir talgreiningu: Jón Guðnason, lektor við Háskólann í Reykjavík

16.00 – Samhengisháð ritvilluvörn: Jón Friðrik Daðason, meistaranemi í tölvunarfræði við Háskóla Íslands

16.20 – "Hér er ég, um ég, frá ég" - Mikilvægi fallbeygingar í leitarvélum: Jón Eðvald Vignisson, tæknistjóri Clara

16.40 – Ráðstefnuslit

META NORD