Eiríkur Rögnvaldsson
Háskóla Íslands

# Málfræðileg mörkun fornískenskum

22. Rask-ráðstefnan
Háskólatorgi
26. janúar 2008

# Efni erindis

- Vélræn mörkun og markarar
  - örstutt kynning
- Mörkun fornískenskra texta
  - þjálfun og handvirk leiðrétting
  - tilraunir með mismunandi greiningarlíkön
- Nýting mörkunar í rannsóknum
  - athuganir á tíðni orðflokka og formdeilda
  - athuganir á setningafræðilegum atriðum
- Vefaðgangur að mörkuðum forntextum

# Verkefni og aðstandendur

- Verkefni:
  - Setningafræðileg nýting vélrænnar greiningar fornískensku
    - Íslensk málnotkun - vefviðmót
- Styrktaraðilar:
  - Rannsóknasjóður Háskóla Íslands
  - Nýsköpunarsjóður námsmanna
- Að verkinu hafa unnið
  - S. Andrea Ásgeirsdóttir
  - Anton Karl Ingason
  - Einar Freyr Sigurðsson
  - Eiríkur Rögnvaldsson
  - Guðlaugur Jón Árnason
  - Sigrún Ammendrup
  - Sigrún Helgadóttir
  - Skúli Bernhard Jóhannsson

# 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 (e. *morphological tagging*)
  - kyn, tala, fall, persóna, háttur, tíð, stig, ákveðni

# Úr grunni *Íslenskrar orðtíðnibókar*

| | | |
|---|---|---|
| f p k e n | hann | hann |
| s f g 3 e þ o | átti | eiga |
| n h e o | afmæli | afmæli |
| a o | í | í |
| n k e o | dag | dagur |
| c | og | og |
| n k e n g | hvolpurinn | hvolpur |
| n k e n - m | Vaskur | Vaskur |
| s f g 3 e þ | var | vera |
| n v e n | afmælisgjöf | afmælisgjöf |

# Forsendur vélrænnar greiningar

- Forsenda vélgreiningar er gott þjálfunarsafn
  - texti sem hefur verið greindur „í höndunum“
  - eftir sama kerfi og vélræna greiningin notar
- Uppbygging þess fer venjulega fram í þrepum
  - lítill hluti textasafns markaður handvirkt
  - markari þjálfaður á þeim hluta
  - stærri hluti safnsins markaður á vélrænan hátt
  - niðurstöður leiðréttar og markari þjálfaður aftur
  - uns viðunandi nákvæmni er náð

# Markarar fyrir íslensku

- Tveir markarar hafa verið þjálfaðir og prófaðir
  - á grunnskrám *Íslenskrar orðtíðnibókar*
  - rúm 500 þúsund orð úr 100 textum af 5 tegundum
  - markamengi (e. *tagset*) um 660 mismunandi mörk
- Gagnamarkarinn (e. *data-driven t.*) *TnT*
  - höfundur Brants, þjálfaður af Sigrúnu Helgadóttur
- Reglumarkarinn (e. *rule-based t.*) *IceTagger*
  - höfundur Hrafn Loftsson

# Mörkun forntexta

- Báðir þessir markarar ná yfir 90% nákvæmni
  - á textum sömu tegundar og þeir eru þjálfaðir á
  - en nýtast þeir við mörkun forntexta?
- Heildartexti 1.651.398
  - *Íslendinga sögur og -þættir* 1.074.731
  - *Sturlunga saga* 283.002
  - *Heimskringla* 250.920
  - *Landnáma (Sturlubók)* 42.745

# Mörkunartilraun

- Tilraun var gerð fyrir tveimur árum
  - allur textinn markaður með *TnT*
- Tilraunin lofaði góðu
  - virtist skila u.þ.b. 88% réttri greiningu
  - margar greiningarvillur fyrirsjáanlegar
    - *er* oftast greint sem sagnmynd en er oft samtenging
    - *og* alltaf greint sem samtenging en er oft atviksorð
    - *okkar*, *ykkar*, *yðar* oft rangt greind

# Handvirk leiðrétting

- Textar valdir til handvirkrar leiðréttingar
  - úr níu sögum í *Sturlungu*
  - sjö heilar sögur og tvö brot
- Samtals 95.194 orð og greinarmerki af 283.002
  - eða þriðjungur heildartextans (33,6%)
- Slík yfirferð er seinleg
  - að hámarki hægt að fara yfir 5.000 orð á dag

# Endurþjálfun markara

- Markarinn þjálfaður aftur
  - fyrst á leiðréttum þriðjungi *Sturlungu*
  - svo á *Íslenskri orðtíðnibók* + þriðjungi *Sturlungu*
- Markarinn svo keyrður aftur á textana
  - með báðum nýju greiningarlíkönunum
- Greiningin síðan yfirfarin
  - niðurstöður þriggja líkana bornar saman

# Nákvæmni greiningar

- Nákvæmni greiningar með þrem líkönum
  - Íslensk orðtíðnibók 88,05%
  - Sturlunga 91,73%
  - Íslensk orðtíðnibók + Sturlunga 92,65%
- Þetta er allgóður árangur
  - betri en fæst með *TnT* fyrir nútímamál
  - þar er nákvæmnin 90,44%

# Árangur bættur

- Hægt er að ná betri árangri
  - með því að keyra líkönin þrjú á textann
  - gefa sér að greining sé rétt ef þeim ber saman
  - en skoða orð þar sem líkönin eru ósammála
- Líkönum ber saman í 84,55% tilvika
  - í 80,88% tilvika réttilega, 3,68% tilvika ranglega
- Í 15,45% tilvika ber líkönum ekki saman
  - þau tilvik þarf að skoða

# Greiningardæmi

- 

| | ÍO | | Stu | | ÍO+Stu | |
|---|---|---|---|---|---|---|
| Skipið | nheng | | nheng | | nheng | |
| varð | sfg3eþ | | sfg3eþ | | sfg3eþ | |
| lítið | lhensf | | lhensf | | lhensf | |
| til | ae | | ae | | ae | |
| skutanna | nvfeg | x | nkfeg | | nvfeg | x |
| en | c | | c | | c | |
| breitt | aa | x | aa | x | aa | x |
| um | ao | | ao | | ao | |
| miðbyrðið | nheog | | nheog | | nheog | |

# Munur líkana

- Líkön þjálfuð á Sturlungu gefa betri útkomu
  - gefum okkur að greining sé rétt beri þeim saman
  - þá er niðurstaðan röng í 4,28% tilvika
  - tæplega 71 þúsund orð
- Skoðum dæmi þar sem þessi líkön greinir á
  - það eru 6,48% dæma
  - u.þ.b. 107 þúsund orð
  - mánaðarvinna í yfirferð

# Greiningardæmi

- | | ÍO | | Stu | | ÍO+Stu | |
|---|---|---|---|---|---|---|
| Steinn | nken-m | | nken-m | | nken-m | |
| hét | sfg3eþ | | sfg3eþ | | sfg3eþ | |
| sá | faken | | faken | | faken | |
| er | sfg3en | x | ct | | sfg3en | x |
| fyrir | aþ | | aþ | | aþ | |
| þeim | fphfþ | | fakeþ | x | fakeþ | x |
| var | sfg3eþ | | sfg3eþ | | sfg3eþ | |

# Viðunandi nákvæmni

- Hægt væri að ná 95% nákvæmni í greiningu
  - á öllu textasafninu
    - *Íslendinga sögum, Sturlungu, Heimskringlu, Landnámu*
  - með samanburði af þessu tagi
- 95% nákvæmni dugir í flestum tilvikum
  - margar villur sem eftir standa smávægilegar
  - villur í einu greiningaratriði, t.d. falli, tölu, kyni

# Tilgangur mörkunar

- Til hvers er þetta?
  - hvaða gagn er hægt að hafa af þessari greiningu?
- Hægt að skoða tíðni málfræðilegra formdeilda
  - og bera saman við nútímamál
- Hægt að leita að setningafræðilegum atriðum
  - mörkin eru vissulega beygingarleg
  - en geyma þó ýmsar setningafræðilegar upplýsingar

# Markaður texti

- En c þessi faven er sfg3en frásögn nven til ae þess fphee að c þeir fpkfn voru sfg3fþ heljar-skinn nhfo kallaðir sþgkfn að c einn tfkeo tíma nkeo er c Hjör nken-m konungur nken skyldi svg3eþ sækja sng konungastefnu nveo var sfg3eþ drottning nven hans fpkee ólétt lvensf og c varð sfg3eþ hún fpven léttari lvenvm meðan c konungur nken var sfg3eþ úr aþ landi nheþ og c fæddi sfg3eþ hún fpven tvo tfkfo sveina nkfo . .

# Tíðni nafnorða eftir föllum

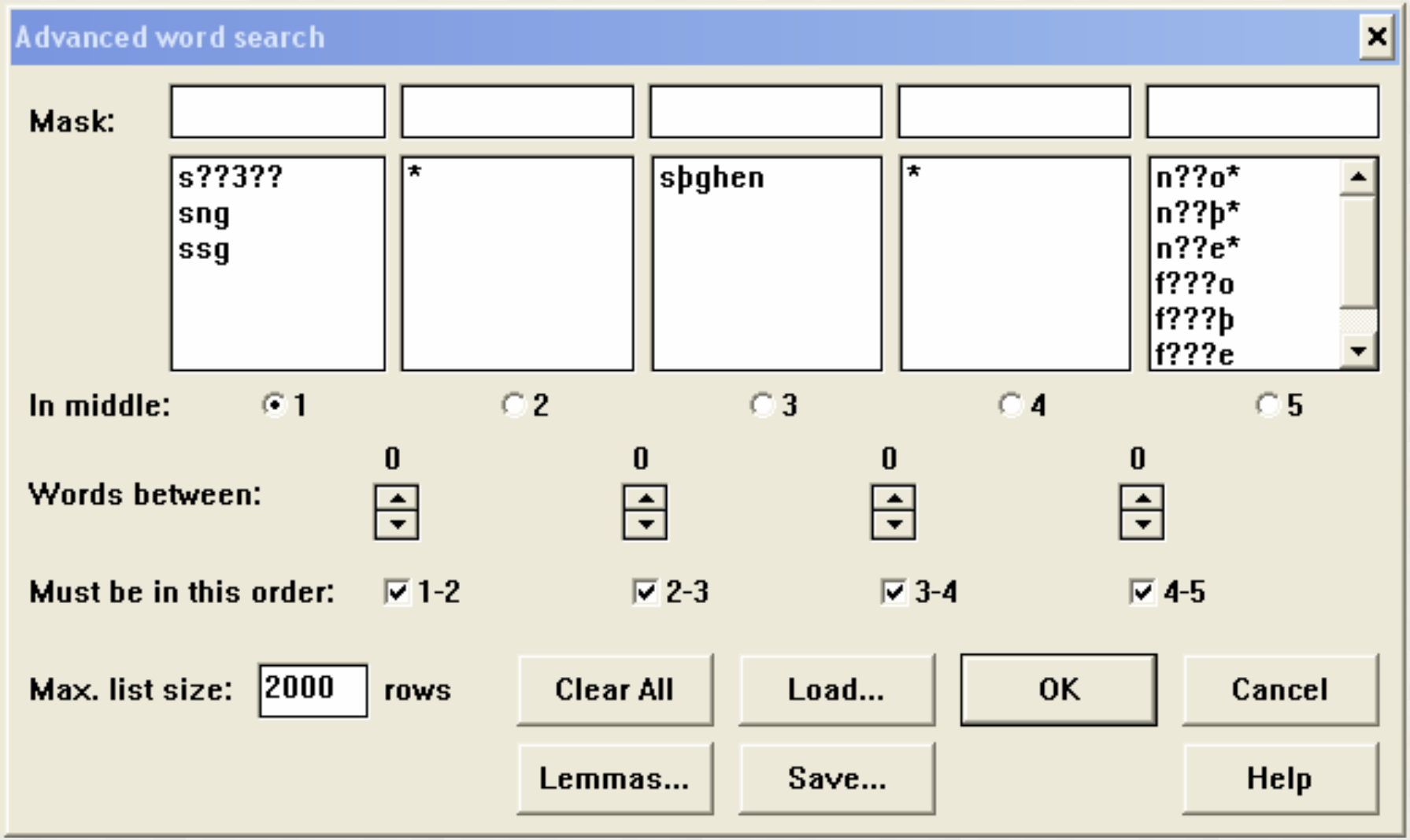
- *Sturlunga*
  - | | samnöfn | öll |
    |---|---|---|
    | nf. | 27,8 | 38,9 |
    | þf. | 38,3 | 27,2 |
    | þgf. | 20,6 | 19,2 |
    | ef. | 13,3 | 14,7 |
- Samtals
  - 13.045 23.205

- *Íslensk orðtíðnibók*
  - | | samnöfn | öll |
    |---|---|---|
    | nf. | 27,6 | 31,2 |
    | þf. | 30,9 | 27,9 |
    | þgf. | 30,2 | 29,0 |
    | ef. | 11,3 | 11,9 |
- Samtals
  - 102.788 122.623

# Tíðni nafnorða eftir kynjum

- *Sturlunga*
  - | | kk. | kv. | hk. |
    |---|---|---|---|
    | nf. | 50,5 | 22,1 | 14,1 |
    | þf. | 19,3 | 34,5 | 43,7 |
    | þgf. | 15,2 | 28,7 | 19,2 |
    | ef. | 14,9 | 13,9 | 13,4 |
- Samtals
  - 14.667 4.214 4.325
- *Íslensk orðtíðnibók*
  - | | kk. | kv. | hk. |
    |---|---|---|---|
    | nf. | 37,8 | 31,8 | 21,7 |
    | þf. | 25,1 | 28,7 | 31,2 |
    | þgf. | 24,5 | 28,6 | 35,0 |
    | ef. | 12,6 | 10,9 | 12,1 |
- Samtals
  - 47.353 39.232 35.124

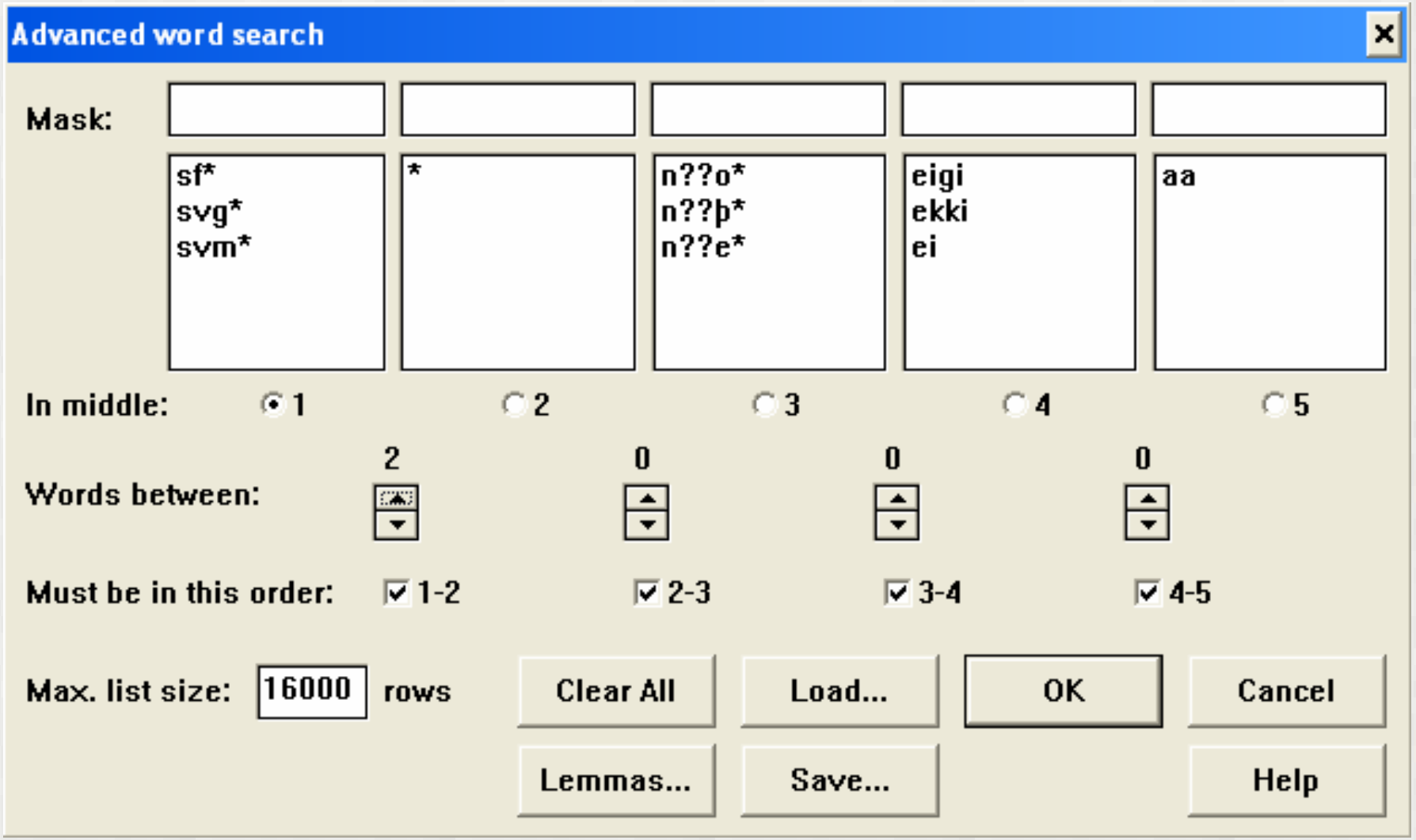
# Nýja þolmyndin



# Dæmi sem leitin skilar

- Þessi dæmi sýna sennilega ekki nýja þolmynd
  - Eftir það **var lokið þinginu**
  - Og er þetta **var sagt Snorra goða**
  - Í því **var sleppt blámanninum**
  - og **var skipt landinu** í helminga
  - Eftir það **var lýst áverkunum**
  - **Var skipt verkum** með húskörlum
  - Í því **var lokið stofuhurðinni**
  - og **var komið griðum** á

# Andlagsfærsla



# Dæmi sem leitin skilar

- Þessi dæmi virðast sýna andlagsfærslu
  - Nú leita þeir um skóginn og finna **Gísla eigi**
  - En þó viljum vér **þenna kost eigi**
  - og fann hann **Snorra ekki** í þessi ferð
  - fyrir því drap eg **Þórð eigi** og alla skipshöfn hans
  - er hann dræpi **Þórð eigi** og förunauta hans
  - og fundu **Þórð eigi** sem von var að
  - Gunnar hallmælti **Hallgerði ekki** um

# Birting á vefnum

- Til stendur að birta textana á vefnum
  - ásamt greiningu
  - í forritinu Xaira frá British National Corpus
- Þá verður hægt að fletta upp í þeim og leita
  - skoða nákvæma greiningu
  - og leita að ýmiss konar orðaröðum og mynstrum
- Þetta verður vonandi á næstu mánuðum

# Birting greiningar í Xaira

| | | | | | | |
|---|---|---|---|---|---|---|
| Engin | fn. | ófn. | kvk. | et. | nf. | |
| skilgreining | no. | kvk. | et. | nf. | | |
| er | so. | frh. | gm. | 3pers. | et. | nt. |
| altæk | lo. | kvk. | et. | nf. | sb. | fr.st |
| og | st. | | | | | |
| frá | fs. | þgf. | | | | |
| þessari | fn. | áfn. | kvk. | et. | þgf. | |
| eins og | st. | | | | | |
| öðrum | fn. | ófn. | kvk. | ft. | þgf. | |
| eru | so. | frh. | gm. | 3pers. | ft. | nt. |
| undantekningar | no. | kvk. | ft. | nf. | | |

# Dæmi úr orðstöðulykli í Xaira

| | | |
|---|---|---|
| klóm lögreglunnar . Sérstaklega ef | **menn** | eru með kjaft . Palli leikur |
| ár og hvar fæddist þessi | **maður** | ? Vísan hvarf af skjánum og |
| var sem sat ágæt og | **maður** | hitti fljótt helling af fólki en |
| eitthvert fjall og svo komu | **menn** | með asna handa honum svo að hann |
| er það lang algengast , þegar | **maður** | étur köku , að ekkert sérstakt |
| « rannsóknarnefnd » , skipuð fremstu | **mönnum** | SS-liðsins í fjármálum , iðnaði , njósnum |
| vegna þess hve miklu fleiri | **menn** | voru í áhöfnum þeirra . Vestmannaeyingarnir |
| fyrir svefninn . Svefn er | **manninum** | nauðsynlegri en næring , því hann |
| sem sagt vanist því , að | **menn** | hverfi af sjónarsviðinu einn af öðrum |
| og virða að vettugi dýr og | **menn** | , sem hafa hreiðrað um sig í |
| og gestkvæmt , venjulega 20 | **manns** | við hádegisverðarborð . Nokkrir starfsmenn |
| frá Laxalóni er enginn venjulegur | **maður** | . Hann er einn af framsýnustu |
| kynni við hann láta engan | **mann** | ósnortinn . Ragnar hefur hlustað mikið |
| er talið að tíundi hver | **maður** | sé að einhverju leyti fatlaður og |
| dottinn . « Hvers vegna læra | **menn** | heimspeki ? » spurði hann . « Til |
| ekkert vita . Nú hafmeyjuna , | **maður** | , sagði einn af yngri strákunum . |
| Creasy . Guido horfði á | **manninn** | meðan andlitsdrættir hans skýrðust . Fimm |
| moskítóflugum og bit þeirra býr | **mönnum** | endalausar martraðir . Í nóttinni má |
| er stöðugt , » sagði gamli | **maðurinn** | við hann . « Það er stöðugra |
| tal saman . Ungi ljóshærði | **maðurinn** | í svissneska sláinu var ótrúlega fús |

# Niðurstöður

- Unnt er að aðlaga markara fyrir nútímamál
  - til að greina fornískensku með viðunandi árangri
  - hægt að bæta árangur með mismunandi líkönum
- Slík mörkun er gagnleg til að
  - skoða tíðni málfræðilegra formdeilda
  - leita að setningafræðilegum atriðum
- Birting markaðs texta á vef verður gagnleg
  - við leit að orðamynstrum og setningagerðum

# Þakka ykkur fyrir áheyrnina

eirikur@hi.is