Efnisöflun og efniviður í málrannsóknum

Textasöfn og málheildir

Ásta Svavarsdóttir og Eiríkur Rögnvaldsson

20. Raskráðstefna
Íslenska málfræðifélagsins

28. janúar 2006
Reykjavík

Forsendur

- Þörf á fjölbreyttum efniviði til málrannsókna
- Raunverulegir textar mikilvægir sem grundvöllur rannsókna
- Vaxandi áhersla á talmál; afla þarf gagna um það og gera þau aðgengileg
- Textana þarf að greina og marka svo þeir komi að fullum notum
- Samnýting gagna úr ýmsum verkefnum; kallar á samræmingu í frágangi

Rannsóknar- og þróunarverkefni

- Tilbrigði í setningagerð
  - Höskuldur Þráinsson o.fl.
- Mörkuð íslensk málheild (MÍM)
  - OH; Sigrún Helgadóttir o.fl.
- Íslenskur markari
  - Eiríkur Rögnvaldsson o.fl.
- ÍS-TAL
  - Þórunn Blöndal o.fl.
- Aðkomuorð í Norðurlandamálum (MIN)
  - Helge Sandøy o.fl.

Textasöfn

- Textasöfn
  - Textasafn Orðabókar Háskólans (ca. 50 millj. lesmálsorð úr 1300 verkum)
- Málheildir (corpus)
  - Mörkuð íslensk málheild (MÍM; í smíðum) (25 millj. lesmálsorð úr 900 textum)
  - BNC (British National Corpus) (100 millj. lesmálsorð)
- Ýmis gagnasöfn
  - Gagnasafn Morgunblaðsins, tímarit hjá Lbs.-Hbs. o.fl.
- Veraldarvefurinn
  - Leitarvélar: Google, Embla o.fl.

Hvað er málheild?

- A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language.

  David Crystal, *A Dictionary of Linguistics and Phonetics*, Blackwell, 3rd Edition, 1991.

- A collection of naturally occurring language text, chosen to characterize a state or variety of a language.

  John Sinclair, *Corpus, Concordance, Collocation*, OUP, 1991

Jafnvæg málheild
(balanced corpus)

- Textar af fyrirfram ákveðnu tagi í tilteknum hlutföllum
  - til að tryggja fjölbreytni
  - til að safnið verði dæmigert fyrir málið eða tiltekinn hluta þess
  - yfirleitt (of) lágt hlutfall af talmáli (ca. 10% í BNC)
- Textar yfirleitt markaðir (tagged)
- Málheildum er ætlað fjölbreytilegt hlutverk (rannsóknir, orðabókagerð, tungutækni o.fl.)

Talmálsgögn

- Talmál þarf að hljóðrita og umrita
- Stöðlun nauðsynleg m.t.t. nýtingar, samhæfing æskileg m.t.t. samnýtingar
- Íslenskt talmálsefni, m.a.
  - ÍSTAL: óformleg, sjálfsprottin samtöl
  - Umræður á Alþingi: formlegt talmál
  - Hópviðtöl úr MIN-rannsókninni
  - Þjóðfræðaefni (frásagnir, viðtöl; SÁM)
  - Frásagnir barna og fullorðinna (Hrafnhildur Ragnarsdóttir)

Nýting efniviðarins

- Til að rannsaka mál og málnotkun eins og hún birtist í textunum
  - Tíðni ýmissa fyrirbæra í máli
  - Orðaforði og orðanotkun
  - Setningagerð og beygingar
  - Ýmiss konar tilbrigði í máli
- Til dæmaleitar, t.d. í rannsóknum og við gerð orðabóka, námsefnis o.fl.
- Við tungutæknirannsóknir og -úrlausnir

Dæmaleit í textum

- Einföld textaleit
  - *lang**
  - *langa, langar, langi, langaði, langað*
    - engan okkar langar til að deyja
    - hann heyrði til þeirra langar leiðir
- Kjarnafærsla í aukasetningum
  - *Ég veit að þennan mann þekkir þú ekki*

Trjábankar

- Trjábanki (treebank)
  - Setningafræðilega greind málheild
- Mismunandi aðferðir við greiningu
  - Stundum byggt á ákveðinni teoríu
  - Stundum óháð teoríum
- Mjög tímafrek greining
  - Verður ekki til í fyrirsjáanlegri framtíð

Íslenskur markari

- Málfræðilegur markari (tagger)
  - Orðabók Háskólans – Sigrún Helgadóttir
- Byggist á *Íslenskri orðtíðnibók*
  - 100 textar – 500 þúsund orð
- 662 mismunandi greiningarstrengir
  - 45 í *Penn Treebank Tagset*

Málfræðileg greining

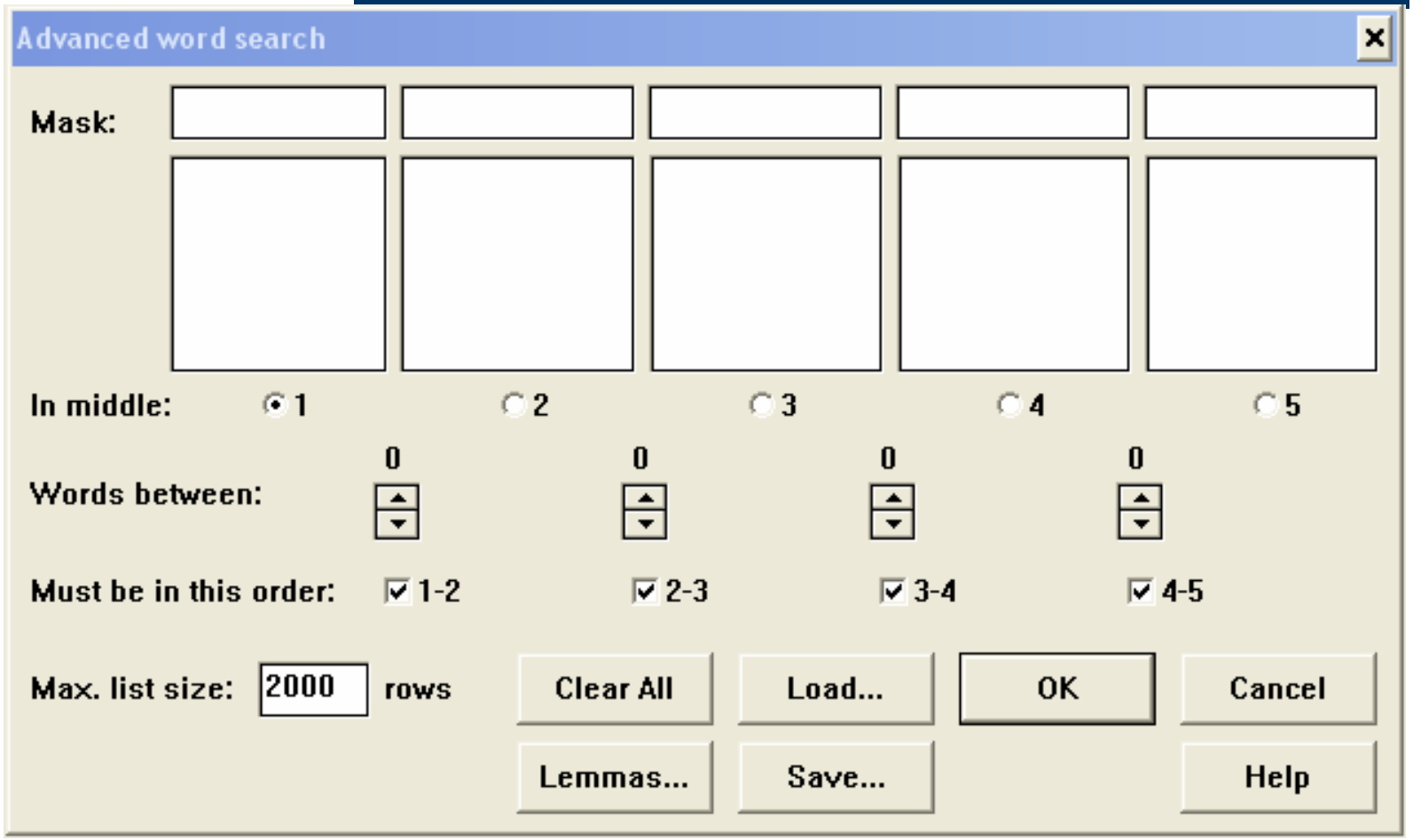
Orð	*mark*
ég	fp1en
stökk	sfg1eþ
á	aa
eftir	aþ
stræto	nkeþ
og	c
veifaði	sfg1eþ

Orð	*mark*
,	,
vagnstjórinn	nkeng
sá	sfg3eþ
mig	fp1eo
og	c
stoppaði	sfg3eþ
.	.

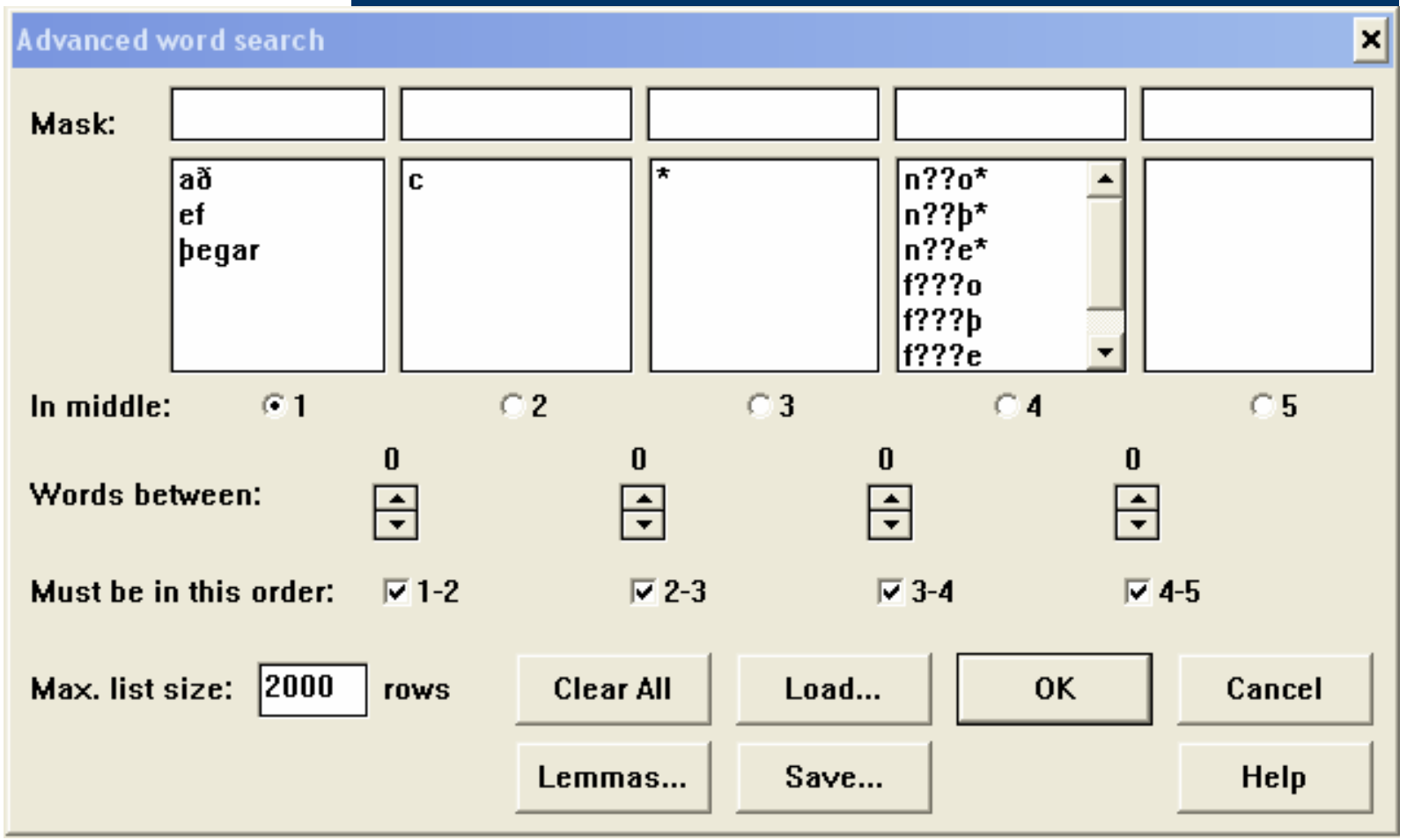
Mörk í samfelldum texta

- Greiningarstrengur á eftir hverju orði
  - ég fp1en stökk sfg1eþ á aa eftir aþ stræto nkeþ og c veifaði sfg1eþ , , vagnstjórinn nkeng sá sfg3eþ mig fp1eo og c stoppaði sfg3eþ . . ég fp1en tautaði sfg1eþ takk au og c brosti sfg1eþ til ae hans fpkee um ao leið nveo og c ég fp1en lét sfg1eþ miðann nkeog detta sng . .

Leitarmynstur



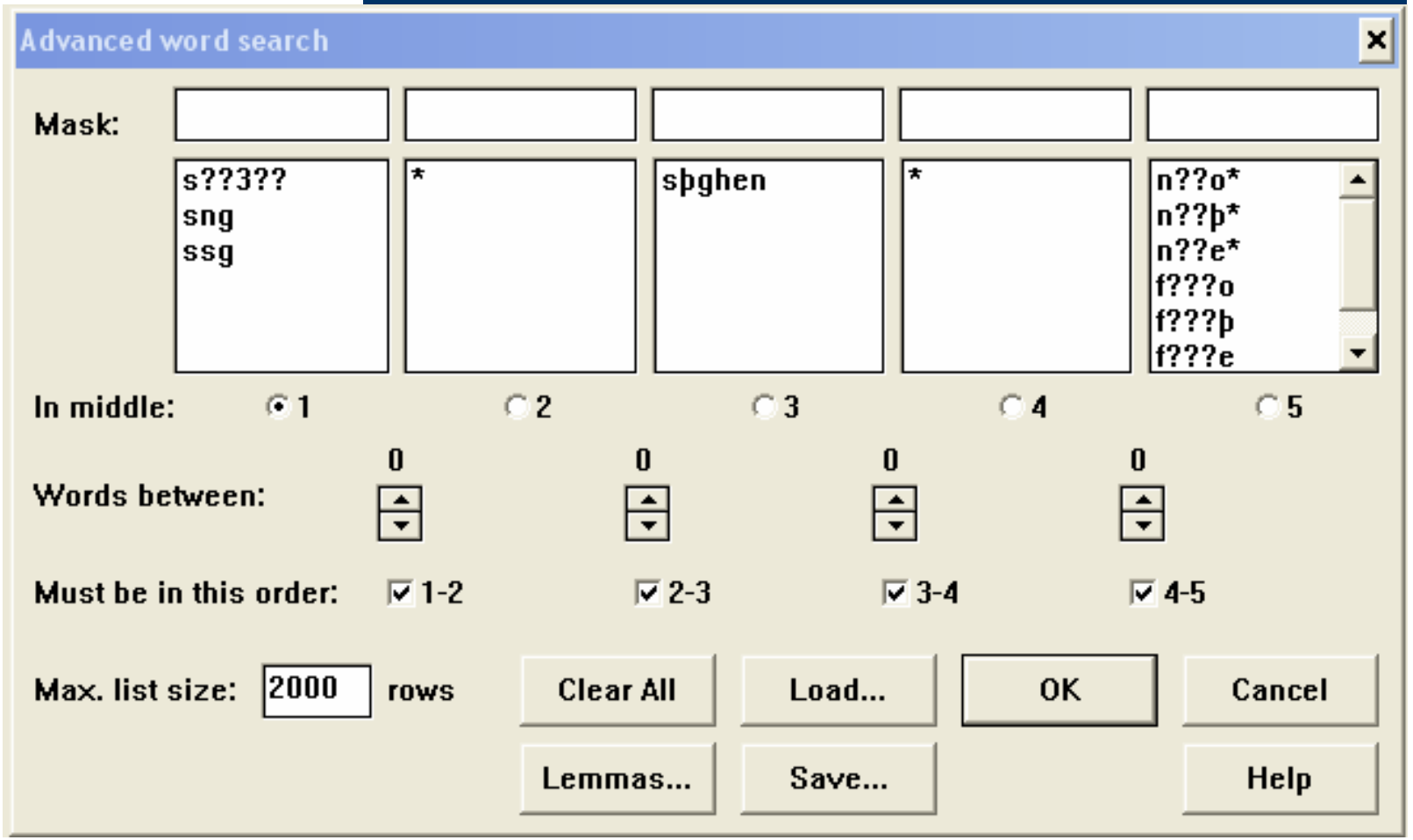
Kjarnafærsla í aukasetningum



Hverju skilar leitin?

- Örfá dæmi um kjarnafærslu
  - að það geri ég líka
  - að þetta máttu þeir ekki
  - ef hlutum fylgir misjafn réttur
- Flest dæmin sýna aukafallsfrumlög
  - að hana væri ekki að dreyma
  - ef henni bauð svo við að horfa
  - þegar þess var þörf

Nýja þolmyndin



Efniviður og túlkun rannsóknarniðurstaðna

- Hversu almennar ályktanir er hægt að draga af niðurstöðum rannsóknar?
- Samanburður innan íslensku, t.d.
  - Milli tímabila
  - Milli landsvæða
  - Milli talmáls og ritmáls
  - Milli kynslóða
  - Milli textategunda, málsniðs o.s.frv.
- Samanburður við önnur mál