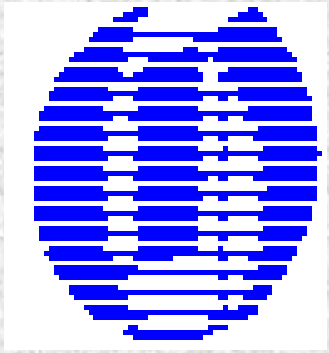


Eiríkur Rögnvaldsson

Setningagerð í textasöfnum – greining og leit

Tungutækni og orðabækur
17. febrúar 2006

Efni erindisins

- Orðagrunnar með setningarlegum upplýsingum
- Leit að setningagerðum í málfræðilega mörkuðum texta
- Málfræðilegri mörkun breytt í setninga-fræðilega mörkun
- Setningafræðileg þáttun – full þáttun og hlutaþáttun

Úr grein í *Orði og tungu* 4

- Kröfur til setningafræðilegrar lýsingar í orðabók:
  - Fjöldi röklíða (áhrifsgildi) sagna þarf að koma skýrt fram
  - Hvaða rökliðir eru skyldubundnir og hverjum má sleppa
  - Hvort og hvenær forsetningarliðir geta komið í stað andlaga
  - Fallstjórn sagna þarf að taka sérstaklega fram
  - Miðmynd þarf mjög oft að vera sjálfstætt flettiorð
  - Lýsingarháttum þarf að gera mun hærra undir höfði
  - Skerpa þarf notkun táknunarinnar ÓP
  - Takmarkanir á setningarstöðu orða verða að koma fram

Úr skýrslu starfshóps um tungutækni

- Koma þarf upp fullgreindu orðasafni (með málfræðilegri og merkingarlegri greiningu) til nota í áframhaldandi vinnu.
  - Orðasafn með grunnorðaforða íslenskunnar (nokkrum tugum þúsunda orða) er forsenda ýmiss konar vinnu í tungutækni. Í þessu orðasafni þurfa að vera sem nákvæmastar upplýsingar um hvert orð; framburð þess, orðflokk, beygingu, setningarstöðu, merkingu, stílgildi o.s.frv. Slíkar upplýsingar koma að gagni við gerð málfræðileiðréttingarforrita, vélrænar þýðingar, leit í gagnabönkum o.fl.

STO - SprogTeknologisk Ordbase

- STO er en samling ordbogsdata, oprindeligt udarbejdet til maskinel anvendelse i sprogteknologiske værktøjer.
- STO-basen indeholder 81.000 opslagsord, alle med detaljerede oplysninger om stavning, bøjning og for substantivers vedkommende også sammensætning.
- Af disse er 45.000 opslagsord desuden forsynet med oplysninger om deres syntaktiske konstruktioner (mulighed for at indgå i sætninger).

Eiginleikar STO

- The most important features of the STO lexicon are, beside its size, being theoretically well-founded and empirically supported.
- The descriptions are very detailed, each piece of information is labelled explicitly and precisely, and any item is easily accessible as the entire lexicon is structured and stored in a relational database.
- Thus, it is straightforward to extract, for example, all syntactic frames of a lemma, or all lemmas sharing the same syntactic frames or a particular syntactic construction, for research purposes.

Hvað höfum við?

- ÍSLEX-grunnur Orðabókar Háskólans
  - 50.000 flettur
- Sagnagreining Orðabókarinnar
  - ófullgerð
- Sagnaflokkun í *Íslenskri orðabók*
  - umfangsmikil en ekki fullkomin
- Greining í öðrum orðabókum
  - einkum bókum Jóns Hilmars Jónssonar
- Er hægt að nýta þessi gögn
  - til að koma upp orðagrunni fyrir tungutækni?

Setningafræðileg dæmaleit í textum

- Getum við leitað að tiltekinni setningagerð
  - í textum án nokkurrar mörkunar?
- 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*

Málfræðileg greining

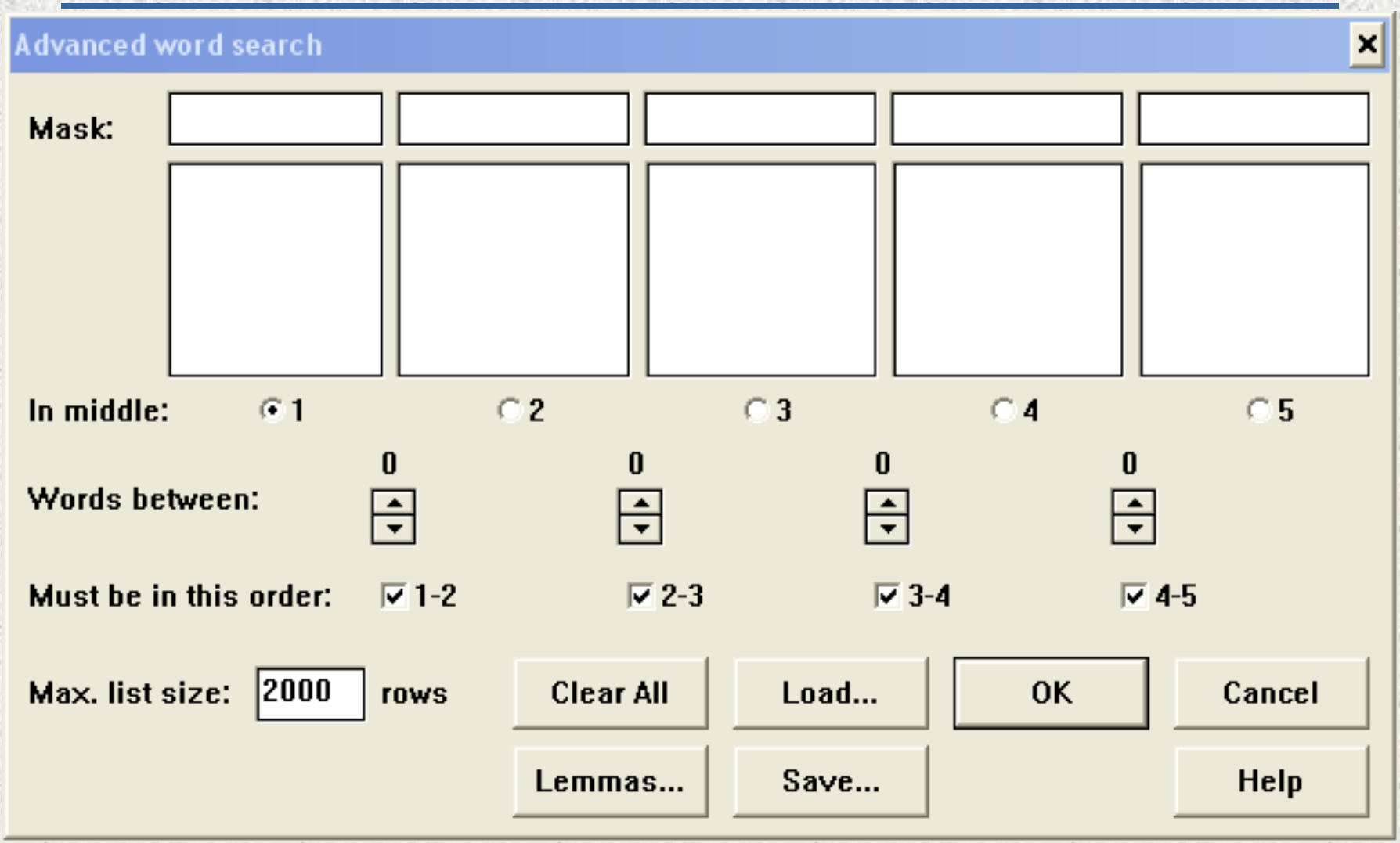
Orð	*mark*
ég	fp1en
stökk	sfg1eþ
á	aa
eftir	aþ
strætó	nkeþ
og	c
veifaði	sfg1eþ

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

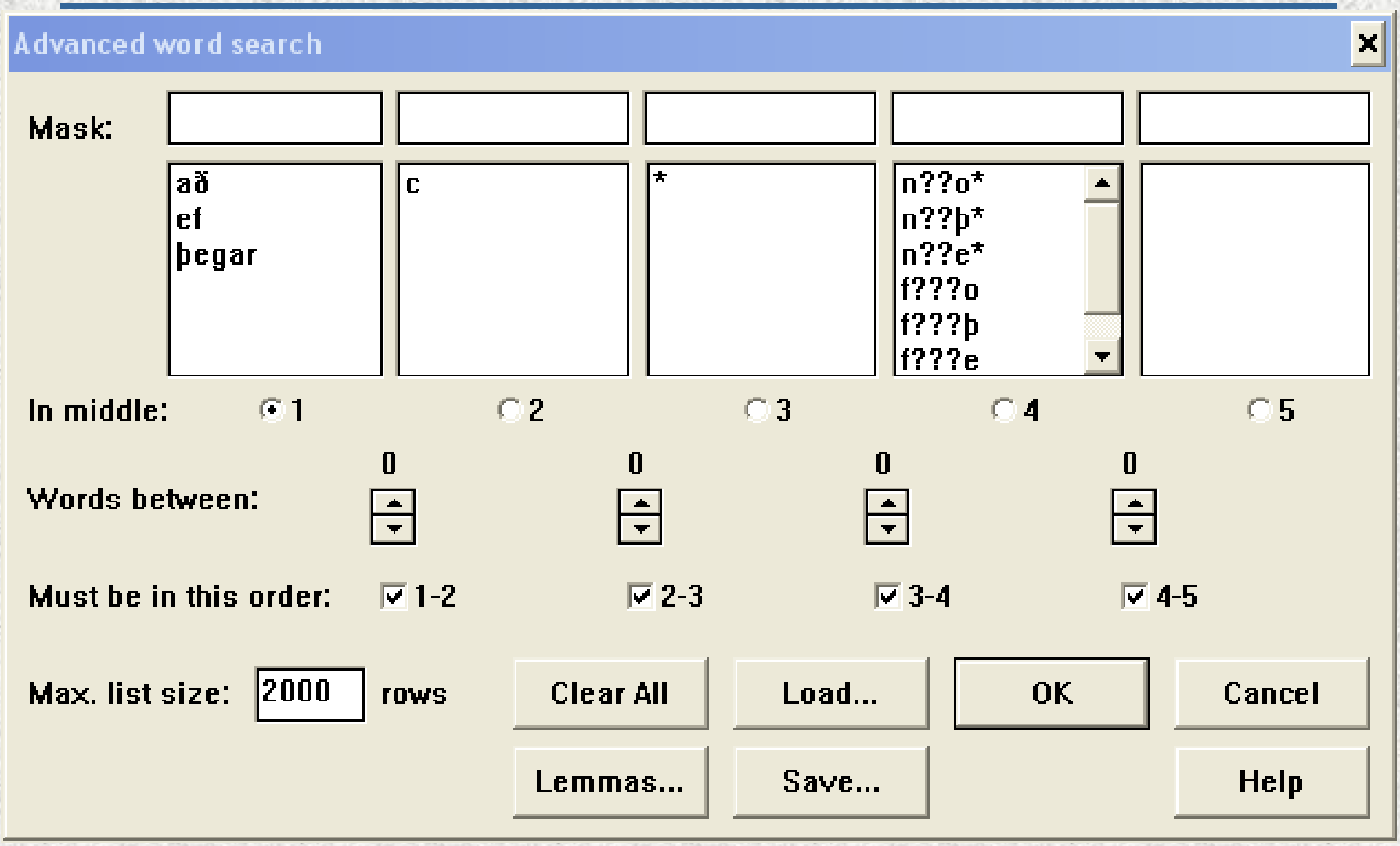
Mörk í samfelldum texta

- Greiningarstrengur á eftir hverju orði
  - ég fp1en stökk sfg1eþ á aa eftir aþ strætó 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 . .
- Mörk meðhöndluð á sama hátt og orð
  - þægilegt er að nota forritið *WinCord*
  - en ýmis önnur koma til greina

Leitarmynstur



Kjarnafærsla í aukasetningum



Hverju skilar leitin?

- Leitin skilar allmörgum dæmum
  - auðvelt að sía frá þau sem ekki eiga við
- Flest dæmin sýna aukafallsfrumlög
  - að hana væri ekki að dreyma
  - ef henni bauð svo við að horfa
  - þegar þess var þörf
- Þó eru örfá dæmi um kjarnafærslu
  - að það geri ég líka
  - að þetta máttu þeir ekki
  - ef hlutum fylgir misjafn réttur

Trjábankar

- Trjábanki (treebank)
  - setningafræðilega greind málheild
    - Penn Treebank fyrstur og þekktastur
- Mismunandi aðferðir við greiningu
  - stundum byggt á ákveðinni teoríu
    - t.d. HPSG eða Dependency Grammar
  - stundum (reynt að hafa) óháð teoríum
- Mjög tímafrek greining
  - en mjög gagnleg í ýmsum tungutækniverkefnum

Setningafræðileg mörk sett inn

– Helgi	nken-m
– minn	feken
– farðu	sbg2en
– niður	aa
– og	c
– skoðaðu	sbg2en
– nýja	lkeovf
– tölvuleikinn	nkeog
– þinn	fekeo

- Bætið ‘F’ við öll beygjanleg orð
  - sem hafa ‘n’ (fyrir *nefnifall*) í markinu
- Bætið ‘A’ við öll beygjanleg orð
  - sem hafa ‘o’ (*þolfall*) eða ‘þ’ (*þágufall*) í markinu

Reglur sem breyta þáttum

- Breytið marki orðs úr ‘F’ í ‘S’ ef so. *vera* fer næst á undan orðinu og nefnifallsorð þar á undan
- Breytið marki orðs úr ‘A’ í ‘P’ ef forsetning fer næst á undan
- Breytið marki orðs úr ‘A’ í ‘F’ ef sögn sem tekur aukafallsfrumlag stendur næst á eftir því
- Flokkið öll orð sem standa saman og bera sama setningafræðilegt mark í einn hóp (með ‘+’)

Útkoma

– F	Helgi	nken-m
– +	minn	feken
– SF	farðu	sbg2en
– X	niður	aa
– F	og	c
– SF	skoðaðu	sbg2en
– A	nýja	lkeovf
– +	tölvuleikinn	nkeog
– +	þinn	fekeo

Nýtt verkefni

- Nú er að hefjast nýtt verkefni
  - a.n.l. svipað þessu
- Hlutaþáttun íslensks texta
  - verkefni styrkt af Rannsóknasjóði
  - verður unnið á þessu ári
- Þrír þátttakendur
  - Eiríkur Rögnvaldsson, HÍ
  - Hrafn Loftsson, HR – upphafsmaður verksins
  - Sigrún Helgadóttir, OH

Setningafræðileg þáttun

- Þáttun (parsing)
  - greining setninga í liði og hlutverk
- Setningagreiningu er oftast skipt í tvo yfirflokka
  - Annars vegar er um að ræða fulla þáttun (e. full parsing; deep parsing), þar sem búið er til fullkomið **þáttunartré** (e. parse tree) fyrir sérhverja setningu,
  - og hins vegar hlutaþáttun (e. partial parsing; shallow parsing) þar sem setningar eru greindar í setningarhluta án þess að krefjast þess að sérhver hluti passi inn í víðtæka þáttun (e. global parse)

Full þáttun

- Helsta vandamálið við fulla þáttun er að stærð lausnamengisins getur vaxið með veldishraða því yfirleitt reynir þáttarinn að búa til allar mögulegar greiningar á málfræðilega tækri setningu
- Þar sem markmiðið er að búa til fullkomið þáttunartré fyrir sérhverja setningu þá hafnar jafnframt þáttarinn stundum réttum greiningum á hluta setningar á þeim forsendum að viðkomandi hluti passi ekki inn í hið víðtæka þáttunartré

Hlutaþáttun

- Í mörgum máltæknikerfum er nægjanlegt að greina setningar í setningarhluta eða setningarliði, t.d. nafnliði, án þess að krefjast þess að liðirnir passi inn í víðtækt þáttunartré
- Þetta getur átt við á sviðum eins og í **upplýsingaútdrætti** (e. information extraction) eða **textaútdrætti** (e. text summarization) þar sem greining setningarliða er mikilvægari en full þáttun

Mismunandi þáttun setningar

- Full þáttun – mismunandi greiningar:
  - $[_{S}\ [_{NL}$ Margir$]\ [_{SL}$ kysstu $[_{NL}$ Maríu $[_{FL}$ á $[_{NL}$ skrifstofunni$]]]]]$
  - $[_{S}\ [_{NL}$ Margir$]\ [_{SL}$ kysstu $[_{NL}$ Maríu$]]\ [_{FL}$ á $[_{NL}$ skrifstofunni$]]]$
- Hlutaþáttun – ein greining:
  - $[_{NL}$ Margir$]\ [_{SL}$ kysstu$]\ [_{NL}$ Maríu$]\ [_{FL}$ á $[_{NL}$ skrifstofunni$]]$
- Setningarliðirnir ekki felldir saman í eitt tré

Kostir hlutaþáttunar

- Full þáttun
  - nákvæmari og sýnir alla möguleika, en:
  - frek á tíma og reiknigetu
  - viðkvæm fyrir villum í inntaki
- Hlutaþáttun
  - sýnir ekki formgerðina eins nákvæmlega, en:
  - skilar greiningu þrátt fyrir villur í inntaki
  - hentar því vel t.d. fyrir texta á netinu

Setningagreining og orðabækur

- Setningafræðileg greining og upplýsingar
  - gagnast bæði notendum og höfundum orðabóka
- Notendum t.d. í sambandi við fyllilið sagna
  - andlög eða forsetningarliðir, fall, o.s.frv.
- Höfundum t.d. við merkingargreiningu
  - náin tengsl eru milli formgerðar og merkingar
- Því þurfum við þrjú hjálpartæki:
  - orðagrunn með setningafræðilegum upplýsingum
  - málheild með setningafræðilegri greiningu
  - setningafræðilegan þáttara