Talmál og tilbrigði

Skráning, úrvinnsla, mörkun og setningafræðileg nýting talmálssafna

Eiríkur Rögnvaldsson og Ásta Svavarsdóttir

Hugvísindaþing
Reykjavík 5. apríl 2008

HÁSKÓLI ÍSLANDS

Efni erindisins

1. Íslensk talmálssöfn
   - yfirlit og skráning
2. Mörkun talmálstexta
   - lemmun, orðtíðni
3. Setningafræðileg nýting mörkunar
   - við leit að ákveðnum atriðum
   - tenging hljóðskráa og textaskráa

Talmálsefni í tilbrigðaverkefninu

- Efninu var safnað fyrir önnur verkefni
  - Þegar var búið að skrá stóran hluta efnisins að meira eða minna leyti
  - Skráning var yfirfarin og leiðrétt og gengið frá efninu á samræmdan hátt innan tilbrigðaverkefnisins
- Gögn úr eðlilegu talmáli samtímans á Íslandi
  - og úr vestur-íslensku
  - auk viðtala um tilbrigði í setningagerð (í vinnslu)

Yfirlit yfir talmálsgögn

- Íslenskur talmálsbanki
  - Rúmlega 50 klst. af stafrænum upptökum úr nútíma-máli (2000-2006) með nákvæmri skráningu; unnið að því að tengja hljóð- og textaskrár (með *Transcriber*)
- Vestur-íslenska
  - U.þ.b. 145 viðtöl við Vestur-Íslendinga frá 1972 (Hallfreður Örn Eiríksson) og 1982 (Gísli Sigurðsson)
    - tölvuskráð eftir eldri umritunum og ekki tengt við hljóðskrár
- Viðtöl um tilbrigði í setningagerð
  - Viðtöl við valda málhafa í tilbrigðaverkefninu; stafrænar upptökur en óvíst um skráningu (í vinnslu)

Íslenskur talmálsbanki

- ÍSTAL
  - Sjálfsprottin, persónuleg samtöl frá 2000; tekin upp við eðlilegar aðstæður víðs vegar um land; fullorðnir Íslendingar, karlar og konur; 31 samtal, alls nærri 20 klst.
    - 185.000 lesmálsorð; 14.000 orðmyndir; 9.000 flettiorð
- MIN: Rannsókn á aðkomuorðum í íslensku
  - Hópviðtöl um erlend áhrif í nútímamáli tekin í Reykjavík 2003; 8 viðtöl með 3 þátttakendum í hverju auk spyrils, alls um 9 klst.
- Umræður á Alþingi
  - Upptökur frá utandagskrárumræðum um ýmis málefni frá 2004–2005 (ekki skrifaðar ræður); 52 ræðumenn, karlar og konur á ýmsum aldri, alls rúmlega 20 klst.
- Samtöl ungs fólks
  - Samtöl um fyrirfram ákveðið efni (ferðalög) frá 2006; 8 samtöl með 2 viðmælendur í hverju – ungt fólk sem talar ýmist við jafnaldra eða eldri manneskju, alls um 4 klst.

Sýnishorn af skráningu

- B: svo þarftu líka að skrifa undir að þú sért samþykk <A>þessu ((hlær))</A>
- A: að ég sé samþykk að þú notir samtalið mitt</B> jájá <B>(x)</B>
- B: ég á</A> ekkert von á því að við tölum um eitthvað sem að
- A: neinei ekki svona neitt sérstakt=
- B: =ekki háalvarlegt alla vega=
- A: =ekki háalvarleg mál
- B: nei
- A: en hvað <TS>segirðu</TS>
- B: bara allt=
- A: =það er svo hryllilega langt síðan að <B>ég</B> hef séð þig <TS>hefurðu verið eitthvað í leikfiminni</TS>
- B: já
- B: voðalega lítið=
- A: =já <B>ég hef</B> svo lítið séð þig ég hef nefnilega verið á öllum mögulegum tímum
- B: ég fór
- B: já ég hef mjög lítið verið núna í einar tvær þrjár vikur ég hef bara ekki mátt vera að því
- A: nei ((raddir í fjarska))
- B: svo átti ég nú ekki <Á>kort</Á> í eina viku eða tíu daga
- A: já
- B: en ég á nú orðið kort ég keypti mér sko gatakort <HM>tuttuguogfjögra</HM> tíma

Mörkun ritmálstexta

- Þrír ólíkir markarar hafa verið þjálfaðir
  - á ritmálstextum af fimm tegundum
- Markamengið er hátt í 700 mörk
  - nafnorð geta haft 48 mismunandi mörk
  - sagnir geta haft 106 mismunandi mörk
  - lýsingarorð geta haft 120 mismunandi mörk
- TnT gaf bestu niðurstöðurnar
  - 90,36% rétta mörkun

Tilreiðsla textans

- Áður en textinn var markaður þurfti að
  - henda út öllum skýringum og athugasemdum
    - ((hlær)), (x), <TS>, <Á>, <HM> ((raddir í fjarska))
  - henda út táknun mælanda
    - A:, <A>, </A>
  - setja inn punkt þegar nýr mælandi tók við
    - á undan A:, B: o.s.frv.

Ílag markarans

svo	að	á	.	ekki	það	verið
þarftu	ég	ekkert	neinei	háalvarleg	er	eitthvað
líka	sé	von	ekki	mál	svo	í
að	samþykk	á	svona	.	hryllilega	leikfiminni
skrifa	að	því	neitt	nei	langt	.
undir	þú	að	sérstakt	en	síðan	já
að	notir	við	.	hvað	að	.
þú	samtalið	tölum	ekki	segirðu	ég	
sért	mitt	um	háalvarlegt	.	hef	
samþykk	jájá	eitthvað	alla	bara	séð	
þessu	.	sem	vega	allt	þig	
.	ég	að	.	.	hefurðu	

Mörkun talmálstexta

- Við bjuggumst við verri niðurstöðum úr mörkun talmáls en ritmáls
  - talmálstextarnir eru gerólíkir ritmálstextum
    - markararnir höfðu reynst vera næmir fyrir mismun textategunda
  - í talmálinu er mikið um hvers kyns „óreglu“
    - ófullkomnar setningar, endurtekningar, mismæli, ósamræmi, …

Útkoma úr mörkun

- ja au það fphen skiptir sfg3en ekki aa máli nheþ kannski aa . . nei au . . svo aa þarftu sfg2en líka aa að cn skrifa sng undir aþ að c þú fp2en sért svg2en samþykk lvensf þessu faheþ . . að c ég fp1en sé svg1en samþykk lvensf að c þú fp2en notir svg2en samtalið nheog mitt feheo jájá au . . ég fp1en á sfg1en ekkert foheo von nveo á aþ því fpheþ að c við fp1fn tölum sfg1fn um ao eitthvað foheo sem ct að aþ . . neinei au ekki aa svona aa neitt foheo sérstakt lheosf . . ekki aa háalvarlegt lhensf alla fokfo vega nkfo . . ekki aa háalvarleg lhfnsf mál nhfn . . nei au . . en c hvað fsheo segirðu sfg2en . . bara aa allt fohen . . það fphen er sfg3en svo aa hryllilega aa langt aa síðan aa að c ég fp1en hef sfg1en séð ssg þig fp2eo hefurðu sfg2en verið ssg eitthvað foheo í aþ leikfiminni nveþg

Niðurstöður úr mörkunartilraun

- Niðurstöðurnar voru furðu góðar
  - um 92,5% orða rétt mörkuð í fyrstu tilraun
- Ástæðurnar sennilega einkum tvær:
  - samtölin varða daglegt líf
    - óþekkt orð því aðeins 4,89% (6,84 í ritmáli)
  - setningar yfirleitt stuttar og einfaldar
    - ekki margir flóknir liðir eða mikil langdræg vensl

Lemmun textans

- Eftir þetta var textinn lemmaður
  - með CST Lemmatizer
  - sem þjálfaður var á *Íslenskri orðtíðnibók*
- Nákvæmni lemmunar var ekki reiknuð
  - en er væntanlega betri en 92,5%
  - stór hluti mörkunarvillna hefur ekki áhrif
- Nú er hægt að skoða orðtíðni í talmáli

Algengustu orð í talmáli

•	1	vera	2	16	hún	11
•	2	að (st)	3	17	einhver	59
•	3	það	6	18	sem	9
•	4	**já**	179	19	**nei**	-
•	5	ég	8	20	svo	28
•	6	og	1	21	en	12
•	7	í	4	22	þá	44
•	8	þessi	15	23	hafa	10
•	9	hann	7	24	fara	29
•	10	**sko**	-	25	með	18
•	11	**bara**	-	26	vita	67
•	12	á	5	27	**hérna**	-
•	13	ekki	13	28	segja	25
•	14	þú	39	29	nú	49
•	15	**svona**	176	30	allur	26

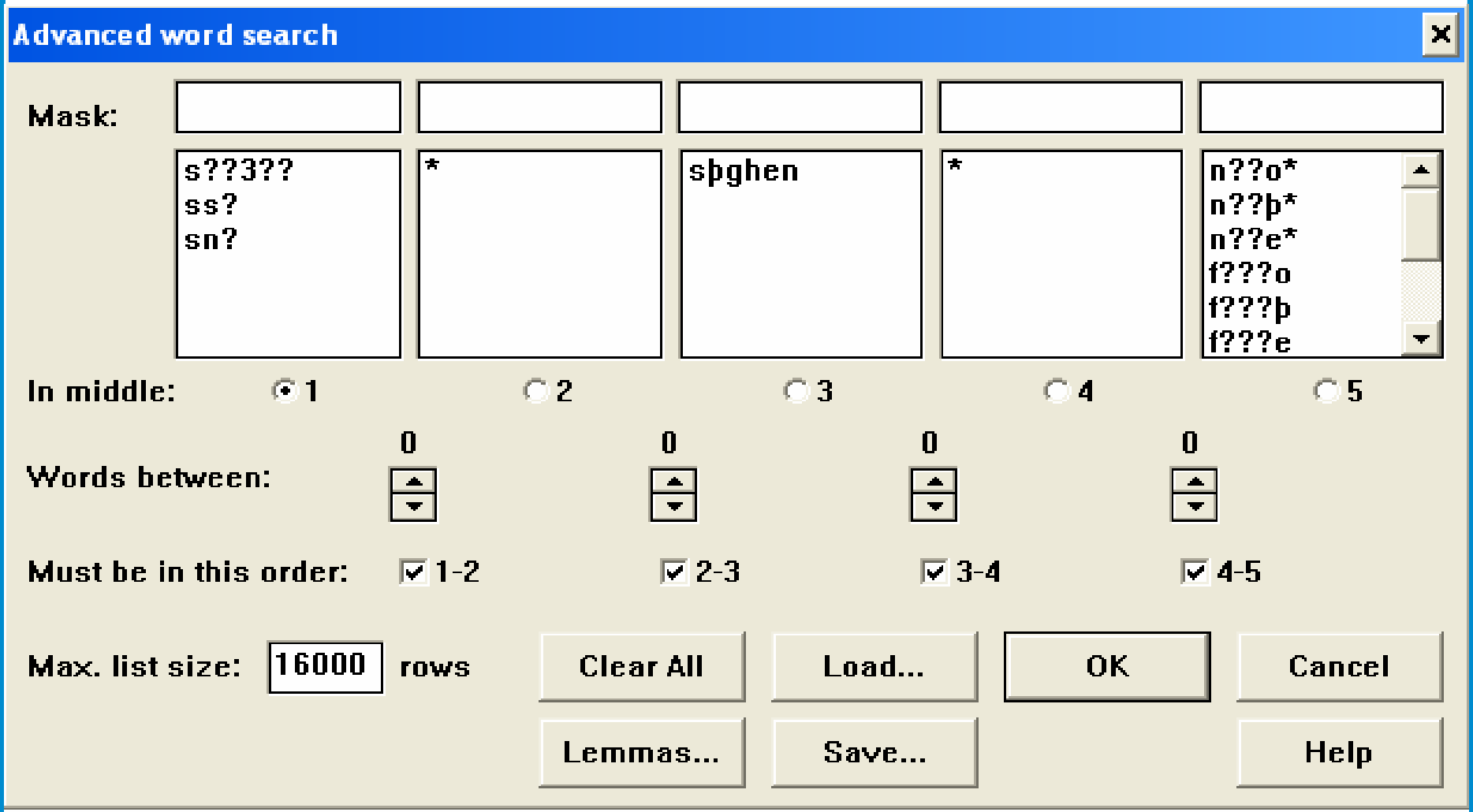
Algengustu orð í ritmáli

•	1	og	6	16	við	41
•	2	vera	1	17	um	40
•	3	að (st)	2	18	með	25
•	4	í	7	19	af	36
•	5	á	12	20	að (fs)	31
•	6	það	3	21	**sig**	56
•	7	hann	9	22	koma	34
•	8	ég	5	23	verða	42
•	9	sem	18	24	fyrir	45
•	10	hafa	23	25	segja	28
•	11	hún	16	26	allur	30
•	12	en	21	27	**svo**	65
•	13	ekki	13	28	sá	20
•	14	til	39	29	fara	24
•	15	þessi	8	30	þegar	47

Setningafræðilegar upplýsingar

• Helgi	nken-m	nefnifall - frumlag
• minn	feken	samræmist nafnorði
• farðu	sbg2en	sögn í persónuhætti
• niður	aa	atviksorð
• og	c	aðaltenging
• skoðaðu	sbg2en	sögn í persónuhætti
• nýja	lkeovf	samræmist nafnorði
• tölvuleikinn	nkeog	aukafall - andlag
• þinn	fekeo	samræmist nafnorði

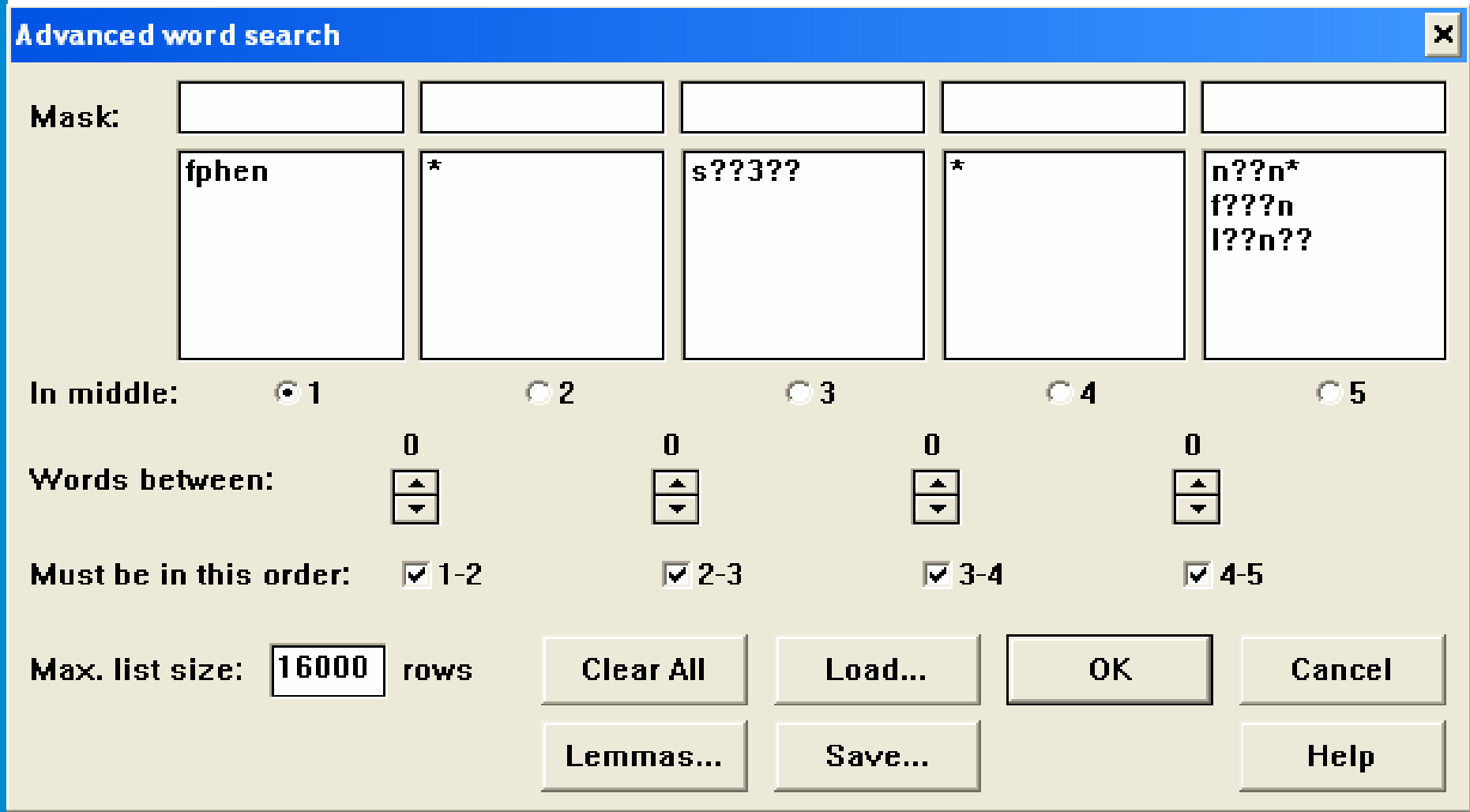
Ný þolmynd


Advanced word search
Mask:
s??3??
ss?
sn?
*
spghen
*
n??o*
n??þ*
n??e*
f???o
f???þ
f???e
In middle:
1
2
3
4
5
Words between:
0
0
0
0
Must be in this order:
1-2
2-3
3-4
4-5
Max. list size:
16000
rows
Clear All
Load...
OK
Cancel
Lemmas...
Save...
Help

Leitarniðurstöður

- Þetta mynstur fyndi dæmi eins og
  - *(Það) var barið mig*
  - *(Það) hefur verið barið mig*
  - *(Það) mun verða barið mig*
- En aðeins eitt dæmi fannst
  - *það var lokað tjaldstæðinu á Þingvöllum*
- Þetta er þó sennilega annars eðlis

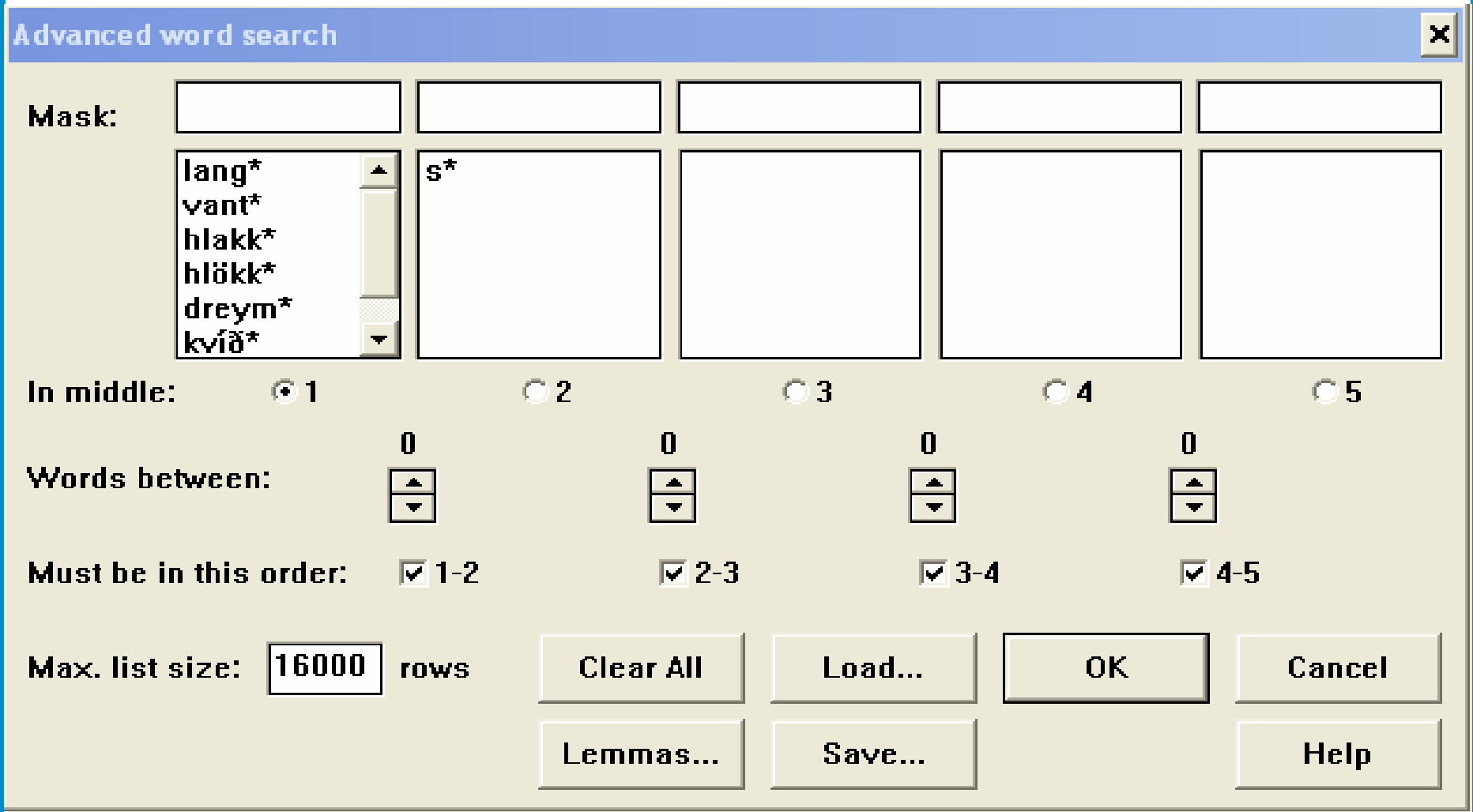
það-leppur með áhrifssögn



Leitarniðurstöður

- Ekkert hafði fundist í orðstöðulykli
  - þegar leitað var þar að dæmum
- Með þessu móti fundust nokkur dæmi
  - *það átti enginn skap saman*
  - *það þekkja allir Rósu*
  - *það vita allir hver Rósa er*
  - *það heldur enginn að þú sért hommi*

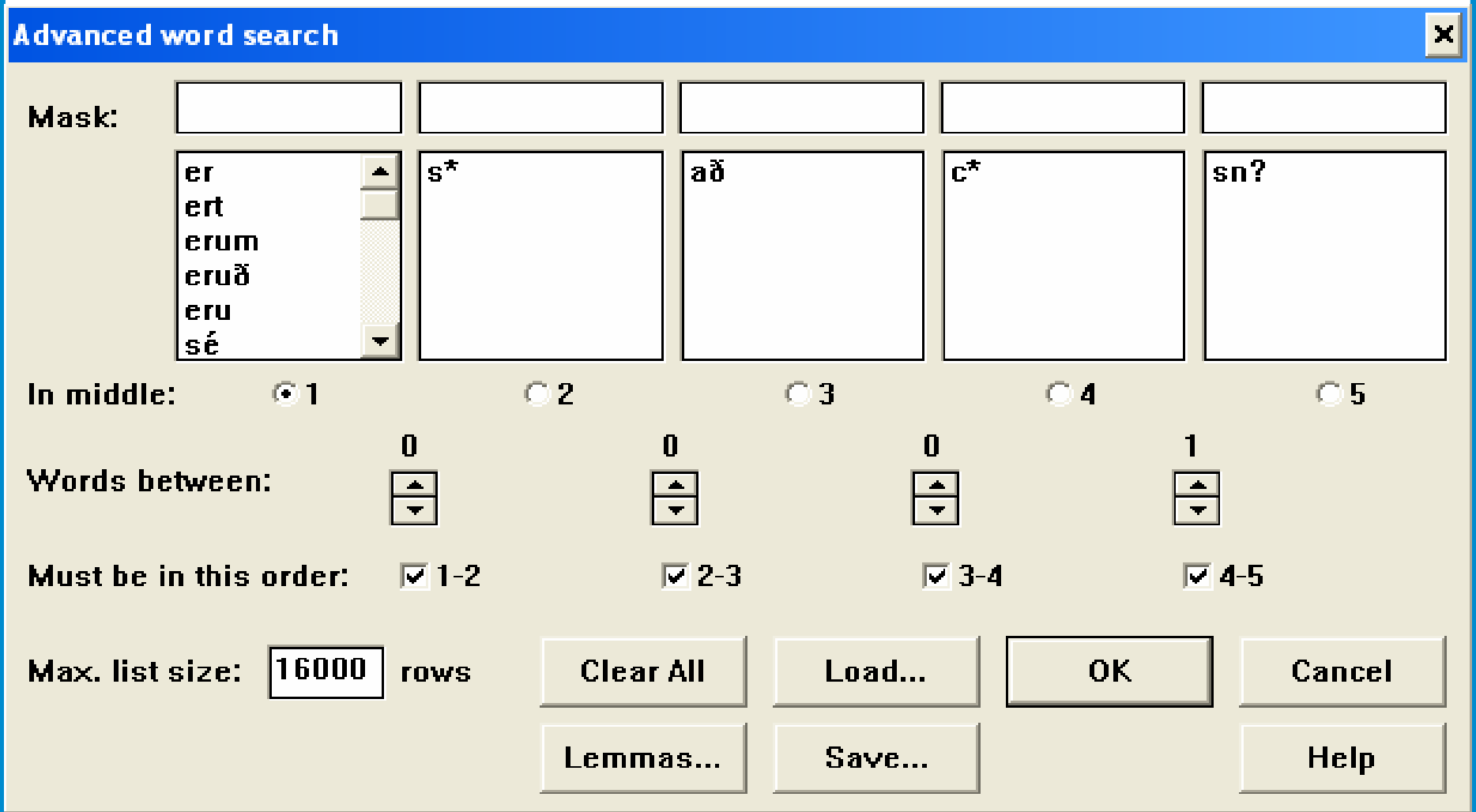
langa, vanta, kvíða, dreyma, hlakka



Leitarniðurstöður – 116 dæmi

- Aðeins tvö dæmi sýna þágufallssýki
  - *já við vildum hafa þetta bara á hreinu af því að sko við vorum að setja sko Pétur **Pétri vantar** bleiur*
  - *mamma sko þegar mamma þegar maður er (bú-) þegar maður fær sér í glas og er einn heima hjá sér með einhverjum karli sem maður er búinn að búa með í hundrað ár þá **langar manni** að fara eitthvað annað þegar maður er búinn að fá sér í glas* ((hátur))

Framvinduhorf



Leitarniðurstöður – 720 dæmi

1.	já au ég fp1en líka aa ég fp1en nefnilega aa	var	sfg3eþ að cn tala sng um ao þetta faheo því fpheþ að c sko aa ég fp1en
2.	já au svo aa ég fp1en	var	sfg1eþ að cn hugsa sng um ao það fpheo að cn vera sng þá aa bara aa
3.	r sfg1en að cn gera sng þarna aa sko aa pipar nvfo en c ég fp1en	er	sfg1en að cn nota sng lamba nhfe .
4.	já au það fphen var sfg3eþ	verið	ssg að cn kynna sng þetta fahen fullt lhensf af aþ nýju lheþsf drasli nhe
5.	nna sng núna aa úti aa í aþ Nóatúni nheþ-ö það fphen var sfg3eþ	verið	ssg að cn kynna sng það fphen er sfg3en eins aa og c pítsa nven í aþ l
6.	þ bara aa borðað ssg það fphen þurftir sfg3en ekkert fohen að cn	vera	sng að cn krydda sng og c hreinsa sng og c þá aa fannst sfm3eþ mér fp
7.	st sfm1en nú aa ekki aa mjög aa mikið lheosf með ao hvað fsheo	er	sfg3en að cn gerast snm .
8.	g c hann fpken hætti sfg3eþ að cn bera sng út aa sá faken sem ct	var	sfg3eþ að cn bera sng út ao Dagblaðið nheogs hérna aa og c við fp1fn
9.	Beggi nken-m	var	sfg3eþ að cn bera sng út aa .
10.	rslu nveo aðra foveo tölvugreinina nveog ensku nveo hann fpken	er	sfg3en að cn vonast snm til ae að cn ná sng níu tfhfo í aþ ensku nveþ í
11.	ér fp1eþ finnst sfm3en náttúrulega aa dálítið aa skítt lhensf að cn	vera	sng að cn fara sng í ao hægferð nveo .
12.	en c ég fp1en	er	sfg1en að cn vona sng sko aa nei au tæplega aa .
13.	g fimm tfkfn held sfg1en ég fp1en svo aa hefur sfg3en hann fpken	verið	ssg að cn taka sng próf nheo samfélagsfögin nhfog eða c samfélagsfra
14.	ið ssg að cn taka sng próf nheo núna aa hann fpken hefur sfg3en	verið	ssg að cn taka sng alveg aa rosalega aa fín lvensf próf nheo hann fpke
15.	el aa undir ao þetta faheo og c ég fp1en vildi svg1eþ ekkert fohen	vera	sng að cn særa sng hann fpkeo með aþ því fpheþ að cn spyrja sng han
16.	a fahen væri svg3eþ svo aa takmarkað lhensf sem c hægt lhensf	væri	svg3eþ að cn gera sng það fphen væri svg3eþ jú au hægt lhensf að cn
17.	n sagði sfg3eþ að c hann fpken hefði svg3eþ reyndar aa alltaf aa	verið	ssg að cn kvarta sng en c hún fpven sagði sfg3eþ að c það fphen var s
18.	kemur sfg3en til aa baka aa altalandi lvfnof og c þegar c ég fp1en	er	sfg1en að cn hjálpa sng henni fpveþ í aþ samræmdu lhfþvf prófi nheþ þ
19.	nfljót lhfnsf að cn grípa sng tungumál nhfo eins aa og c þau fphfn	eru	sfg3fn að cn ýta sng því fpheþ frá aþ sér fpkeþ og c taka sng upp aa an
20.	fakeo dag nkeo anna nvfe vinnufélagi nken minn feken hún fpven	er	sfg3en að cn fara sng til ae Kaupmannahafnar nvee-ö akkúrat aa þenna
21.	já au hún fpven	er	sfg3en að cn fara sng þá fpkfo .
22.	g er sfg3en það fpheo ekki aa á ao morgun nkeo sem ct þeir fpkfn	eru	sfg3fn að cn setja sng við ao samningaborðið nheog bara aa alveg aa l
23.	hann fpken	er	sfg3en að cn verða sng það fpheo núna aa fjórtánda sng eftir aþ viku n
24.	fg3en alltaf aa þegar c að aa þegar c að aa þegar c að c ég fp1en	er	sfg1en að cn stússast snm þá aa er sfg3en þetta fahen alveg aa sama
25.	heyrðu sbg2en hvað fshen	er	sfg3en að cn frétta sng af aþ Einari nkeþ-m .
26.	a systkini nhfo mín fehfo sem c búa sfg3fn í aþ Danmörku nveþ-ö	eru	sfg3fn að cn pirra sng sig fpkeo yfir ao eitthvað foheo sem c þeim fpkfþ
27.	ng að aþ svona lheþof pífulaki nheþ Svanhvít nven-m systir nven	var	sfg3eþ að cn kaupa sng sér fpkfþ hérna aa amerískt lheosf svona lheo
28.	hann fpken	er	sfg3en að cn reportera sng .
29.	ég fp1en	var	sfg1eþ að cn hugsa sng um ao ferðina nveog líka aa norður aa sko aa
30.	jájá au hún fpven	var	sfg3eþ að cn tala sng um ao það fpheo .
31.	era sng sko aa ég fp1en veit sfg1en ekki aa hvað fsheo þeir fpkfn	eru	sfg3fn að cn hugsa sng það faheo eru sfg3fn einhverjir fokfn jú au það
32.	ði nkeþ-ö væri svg3eþ svona aa ja au hafi svg3en eitthvað fohen	verið	ssg að cn tala sng um ao stærðfræðikeppnina nveog hann fpken vildi s
33.	jú au þú fp2en	ert	sfg2en að cn suða sng núna aa .

Dæmum raðað eftir sögnum

	F	H	J	L	N	P	R	T	V	W	X
1	sá	að	þeir	voru	að	auglýsa	fjóra	eða	fimm	tfkfo sko aa Volvo nken-m	
2	heyrðu	já	við	vorum	að	ákveða	það	bara	að	cn reyna sng að cn hittas	
3	undir	segulbandið	ég	var	að	benda	henni	á	að	cn það fphen væri svg3eþ	
4			Beggi	var	að	bera	út				
5			Felix	var	að	biðja	mig	um	það	fpheo á ao laugardagskvö	
6		jú	ég	er	að	bíða	eftir	henni			
7	á	Kaffibrennslunni	og	var	að	bíða	eftir	Lilju	og	c þá aa eiginlega aa sko	
8	sem	sagt	þú	ert	að	bjarga	mér				
9		já	þeir	eru	að	bjóða					
10			hún	var	að	bjóða	þér	það	þú	fp2en sagðir sfg2eþ nei a	
11	sem	er	alltaf	verið	að	bjóða	manni				
12	ekki	enn	þá	verið	að	bjóða	það				
13	meina	það	ég	var	að	borða	náttúrulega	bara	kökur	nvfo og c svoleiðis aa me	
14		þetta	lið	er	að	borga	sjötíuþúsund	á	mánuði	nkeþ fyrir ao að cn sýnas	
15			þú	ert	að	borga	þú	veist	bara	aa afnotagjöld nhfo .	
16	sagðir	sko	þú	ert	að	borga	í	raun	og	c veru nveþ s aa hvað fsh	
17		já	þú	ert	að	borga	með	gjaldinu			
18		já	þú	ert	að	borga	fullt	fyrir	börnin	nhfog .	
19	bara	snjór	sem	er	að	bráðna					
20			það	er	að	breytast	sko				
21	mér	veltast	það	er	að	brjótast	um	í	mér	fp1eþ líka aa .	
22	hverfið	sem	er	verið	að	byggja	upp				
23	sig	þegar	ég	var	að	byrja	í	haust	sko	aa í aþ þessari faveþ kenn	
24	fyrsta	þegar	við	vorum	að	byrja	í	deildinni	var	sfg3eþ það fpheo ekki aa	
25	nú	e	e	er	að	deyja					
26	já	hann	hefur	verið	að	drekka	í	sig	kjark	nkeo sko aa þá aa .	
27	vita	ef	ég	er	að	drekkja	ykkur	kaffi			
28	það	er	alltaf	verið	að	dæma	einhverja	sígarettugæja			
29	hafi	nú	ekki	verið	að	eltast	við	stelp	varstu	sfg2eþ að aþ því fpheþ .	
30	svo	núna	er	verið	að	endurflytja	þetta				
31	það	sem	ég	var	að	experimenta	með	þú	veist	sfg2en alveg aa sama fbh	
32	kex	sem	þeir	voru	að	éta	sko				
33	ég	bekki	sem	er	að	fara	til	útlanda	bennan	fakeo dag nkeo anna nvfe	

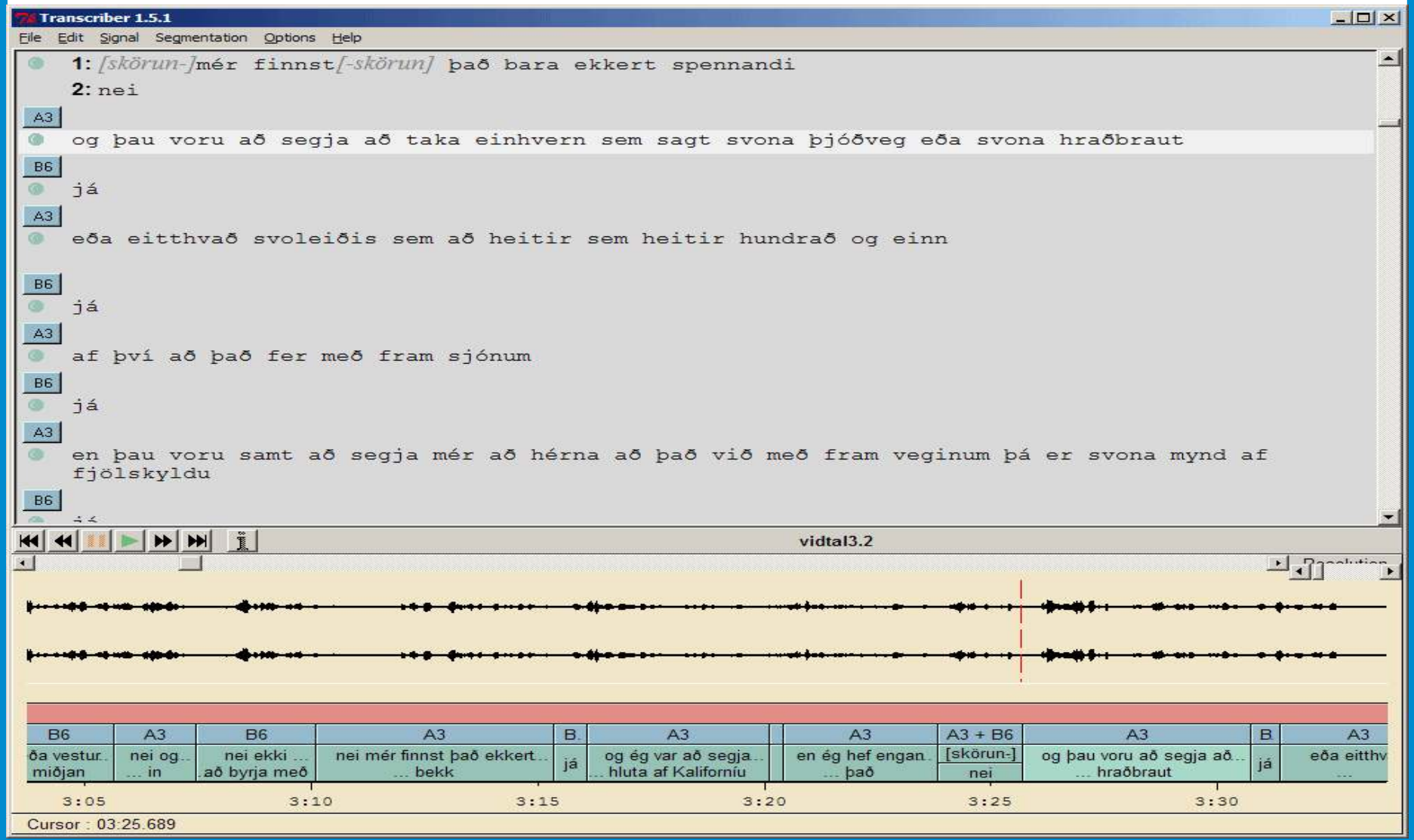
Leit í hljóðskrám

- Stundum er nauðsynlegt að hlusta á dæmin
  - til að átta sig á setningafræðilegu gildi þeirra
- Hægt er að leita í umritunarskránum (textanum)
  - en talsverð vinna að finna réttan stað í hljóðskránum, t.d. til að kanna áherslur og tónfall
- Hægt að tengja hljóð- og textaskrár saman
  - með kóðun í þar til gerðum forritum, t.d. *Transcriber*
  - Þá er hægt að finna dæmi í hljóðskránum á einfaldan hátt með leit í umritunarskránum

Hvernig virkar *Transcriber*?

- Sérstakt umritunarforrit fyrir talað mál
  - var nýtt í verkefninu *Hvernig tala ungir Íslendingar í upphafi 21. aldar?* (Ásta Svavarsdóttir og Sigrún Ammendrup; styrkt af NSN 2006) (Samtöl ungs fólks)
- Forritið ræður við ýmis form hljóðskráa
  - ef þær eru stafrænar, t.d. .mp3 og .wav
- Forritið kóðar skrárnar sjálfkrafa í xml
  - og tengir þannig saman tiltekna staði í hljóðskránni og samsvarandi umritunarskrá
- Unnið er að því að flytja allar umritunarskrár úr talmálsbankanum í *Transcriber*

Dæmi úr *Transcriber*



Niðurstöður

- Talmálsefnið er mikilvæg heimild um tilbrigði í íslenskri setningagerð
- Mörkun textans auðveldar leit að ýmsum setningafræðilegum atriðum í textanum
- Tenging hljóðs og texta gerir kleift að finna og hlusta á tiltekin atriði í hljóðskrám