

Dansk Sprognævn

Halvautomatisk udvælgelse af lemmakandidater til en nyordsordbog

Jakob Halskov
jhalskov@dsn.dk

retskrivning sb., -en, -er, f
sms. retskrivnings-, fx ret-
skrivningssystem.
retslig (el. retlig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, rets-
lægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

En kort oversigt

- Sproglige nydannelser og nye ord
- Veje til nye ord: menneske kontra maskine
- En halvautomatisk vej: Dansk Sprognævns Ordtrawler
- Evaluering af Ordtrawleren
 - Eksperiment #1: Hvor meget får systemet med (*recall*)?
 - Eksperiment #2: Hvad er systemets træfrate (*precision*)?
- Foreløbige konklusioner
- Perspektiver
 - Diakrone frekvensprofiler: Kan de bruges?
 - Byg og overvåg dit eget webkorpus (GlossaNet)

Sproglige nydannelser og nye ord

- På det semantiske niveau
 - Nye ord som refererer til nyt indhold (fx *klimacertifikat*)
 - Eksisterende ord som får nyt indhold (fx *blæksprutte* i betydningen *altnuligmand*)
 - Eksisterende ord som erstatter eksisterende indhold (fx *sort* for *neger*)
- På det fraseologiske niveau
 - Nye flerordsudtryk (fx *få enderne til at mødes*)
- På det syntaktiske niveau
 - Ny valens (fx *dumpe en eksamen* for *dumpe til eksamen*)
- På det fonetiske niveau
- På det ortografiske niveau

retskrivning sb., -en, -er, -e
 sms. retskrivnings-, fx retskrivningssystem.
 retslig (et. rellig) adj., -t.
 retslæge sb., -n, -i.
 retslægeråd sb., -et, retslægeråd, bf. pl. -ene.
 retslærd adj., itk. d.s.
 retsløs adj., -t.

Fra excerpt til ordbog

The screenshot shows the ILEX v 2.1 software interface. On the left, a list of words is shown under 'ordsamlingen'. The main window displays the entry for 'burkini' with its grammatical classification and source information. On the right, a browser window shows a search result for 'nye ord på nettet' with a search box containing 'burkini' and a 'Søg nu' button. Below the browser, there are two bullet points providing context for the word 'burkini'.

ordsamlingen

Documents

Opslag Design

burkini
 burkiner
 burkini
 burkinsk
 burkuller
 burlaks
 burleytobak
 burlije
 Burma
 burmakat
 burmakilling
 burmakvinde
 burmarubiner
 burmeserkillling
 burner
 burn in
 burn-in
 burning
 burn-in-system
 burnout
 burn out-konkurrence
 burnoutvelour
 burnrate
 burnt out-syndrom
 burnus
 burre
 burrebånd

burkini / ordsamlingen (Dokument)

```

<artikel>
<opslagsord>burkini (Slå op i Ordtrawler)
<hom>
<ordklasse>
<sb/>
<ordets_brug>
<tom>
<domaene>
<1. belæg>
<asset>509/A00030509-1
dato: 29.03.2007
<aktuel_ordform>burkinien
<belægstekst>
<afsnit>.. med opfindelsen
klædt. 20-årige Medda Laa
alting.
<kildehenvisning>
<genre>Nyhedsstof
<kilde>Pol
<kilde_fortsat>1:1
dato: 19.03.2007
<koder>JNJ
<excerptist>JNJ
<2. belæg>
<asset_ny>2009/02/0410334
dato: 04.02.2009
  
```

NOiD_www.PNG @ 100% (RGB)

http://www.nyeorddansk.dk/noid/noid. ☆ glossanet

Mest besøgte Igang med Firefox Seneste nyheder

Nye ord på nettet Materialer og udgivelser GlossaNet --- Online pre...

NYE ORD I DANSK på nettet fra 1955 til i dag **Avanceret søgning**

simpel søgning

burkini **Søg nu**

• Vis resultatet prioriteret • Vis resultatet kronologisk

Hjælp

1 artikel.

burkini (Burqini ®) sb. (2007)
 heldækkende badedragt til muslimske kvinder

- Det er ikke ligefrem en fornøjelse at boltre sig i bølgerne ikklædt burka ... Takket være et program søsat for at lokke unge muslimske kvinder til stranden er det prekære problem nu løst med en kreativ opfindelse: burkinien *Politiken* 19.3.2007
- En del muslimske kvinder foretrækker mere anstændigt strandtøj end at ligge topløs og iført et par små bikini-trusser. Og det nye badedøj til muslimske kvinder - burkinien - sælger godt *Kristeligt Dagblad* 13.6.2008

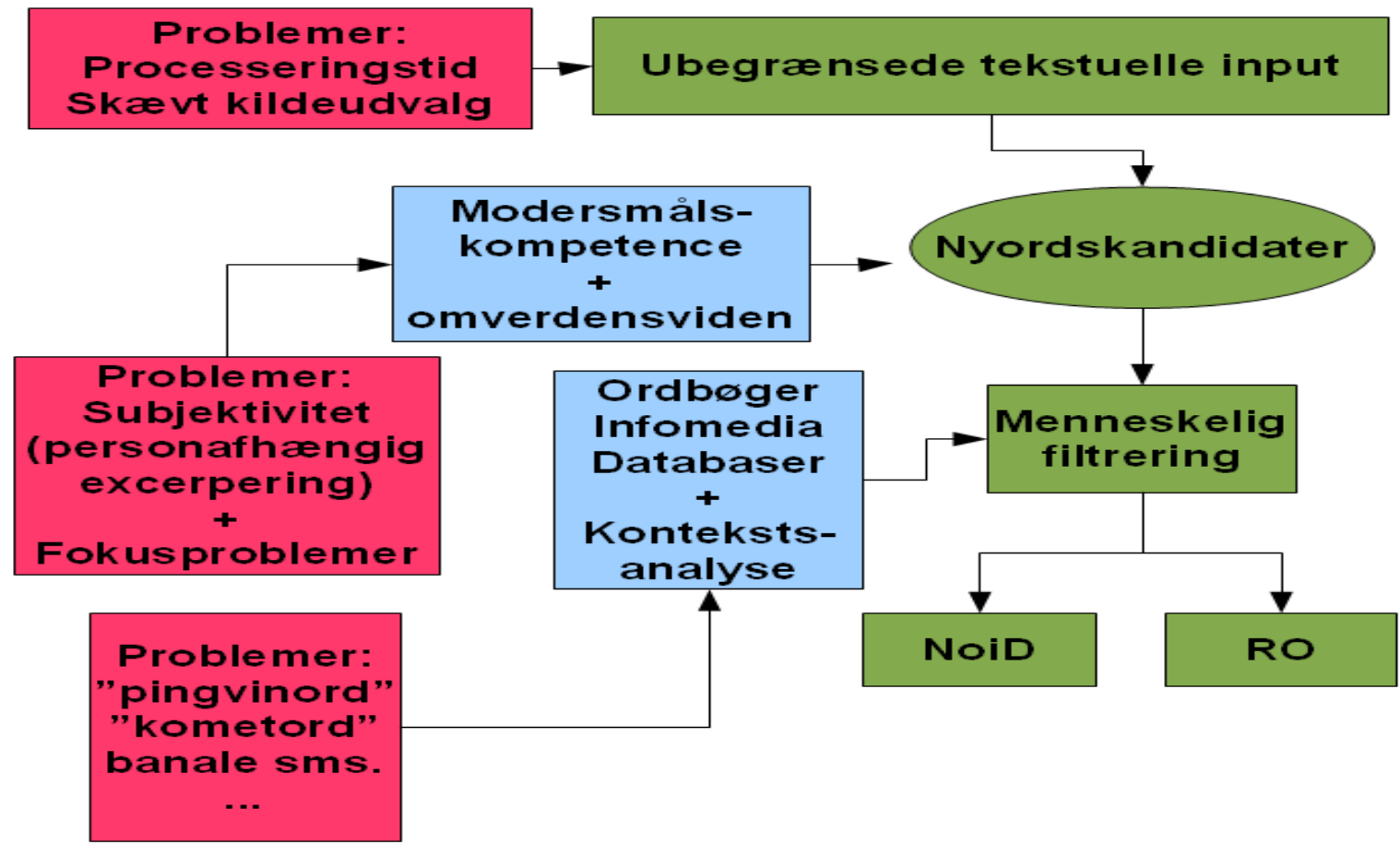
Veje til nye ord

retskrivning sb., -en, -er, i
sms. retskrivnings-, fx ret-
skrivningssystem.
retslig (et. retlig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, rets-
lægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

- Traditionel excerpering
 - Fx avistekst → Ordsamlingen → Nye ord i dansk
- "Spørg folket"
 - Direkte: Nyordsblanket (dsn.dk, sproget.dk, ordnet.dk)
 - Indirekte: Søgelogs (nyeordidansk.dk, retskrivningsordbogen.dk)
- Automatisk korpusanalyse
 - Traditionelle korpuser (fx Infomedia)
 - Nettet som korpus (fx blogovervågning via GlossaNet)

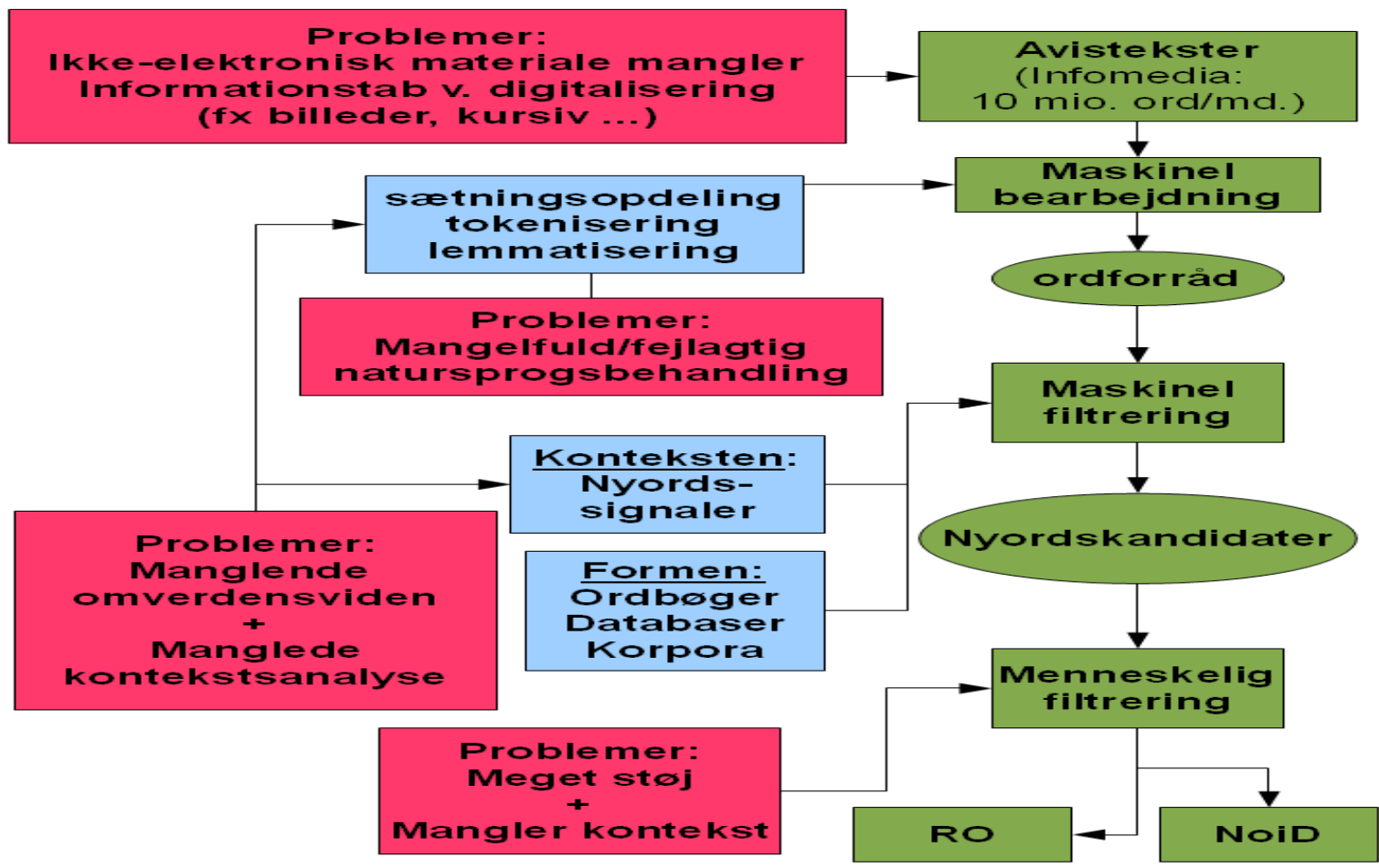
retskrivning sb., -en, -er, f. sms. retskrivnings-, fx retskrivningssystem.
retslig (et. retlig) adj., -t.
retslæge sb., -n, -i.
retslægeråd sb., -et, retslægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

Fra ord i lange baner til nyt ord: menneske



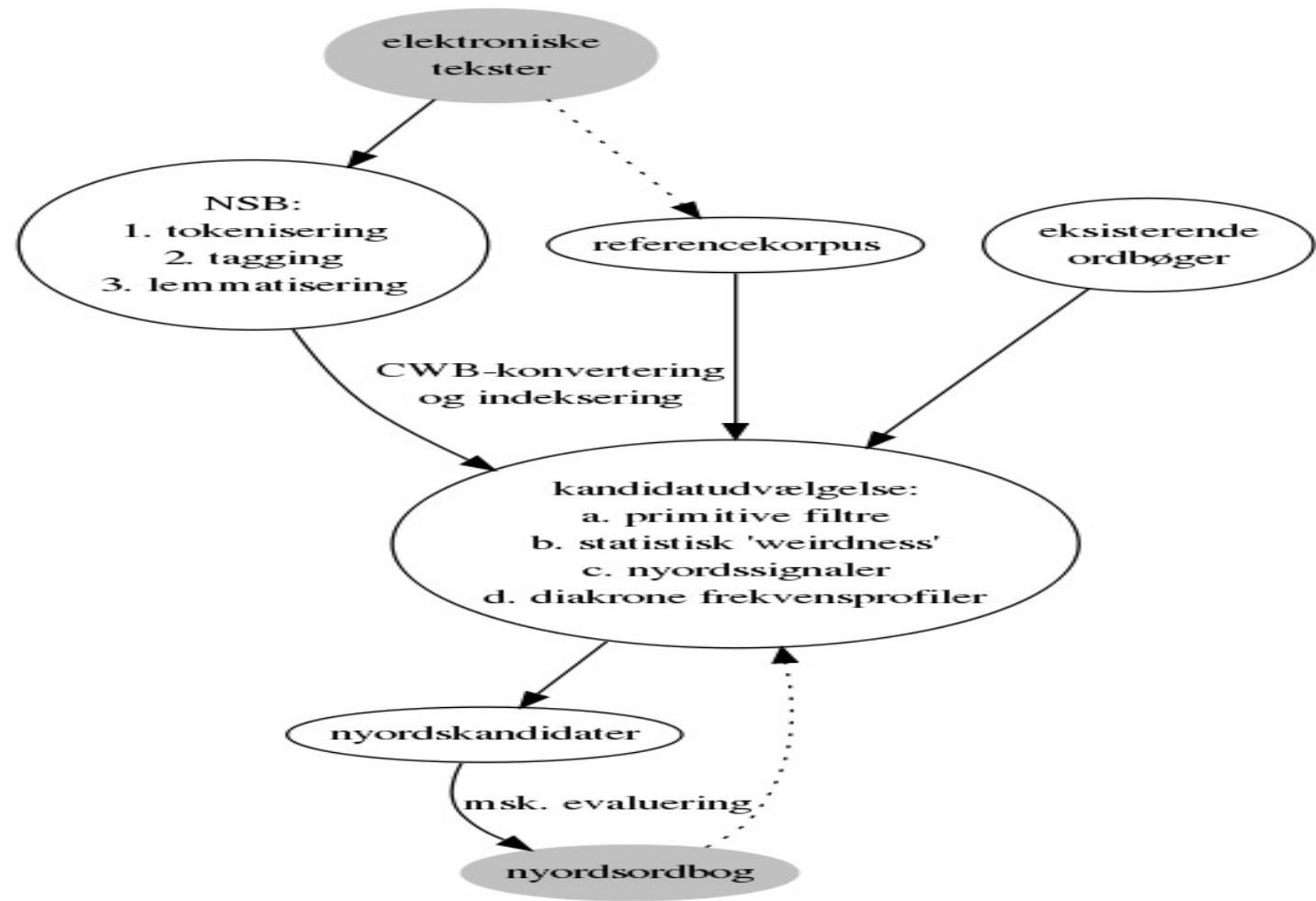
retskrivning sb., -en, -er, f. sms. retskrivnings-, fx retskrivningssystem.
retslig (et. retlig) adj., -t.
retslægeråd sb., -n, -r.
retslægeråd sb., -et, retslægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

Fra ord i lange baner til nyt ord: maskine



retskrivning sb., -en, -er, f
sms. retskrivnings-, fx ret-
skrivningssystem.
retslig (et. retlig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, rets-
lægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

Dansk Sprognævnets Ordtrawler



retskrivning sb., -en, -er, i
sms. retskrivnings-, fx ret-
skrivningssystem.
retslig (et. retlig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, rets-
lægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

a) Primitiv filtrering

Maskinel filtrering af allerede kendte ordformer			
<i>Nr.</i>	<i>Filter</i>	<i>Antal lemmaer</i>	<i>Antal ordformer</i>
1	Retskrivningsordbogen 2001	64.038	399.062
2	I DDO, men ikke i 1	34.960	Samme
3	I Ordsamlingen (sep. 2008), men ikke i 1-2	221.679	Samme
4	I Korpus 90, men ikke i 1-3	?	124.585
5	I Korpus 2000, men ikke i 1-4	?	436.004
I alt		?	1.216.290

retskrivning sb., -en, -er, i sms. retskrivnings-, fx retskrivningssystem.
 retslig (et. retlig) adj., -t.
 retslæge sb., -n, -r.
 retslægeråd sb., -et, retslægeråd, bf. pl. -ene.
 retslærd adj., itk. d.s.
 retsløs adj., -t.

b) Statistisk "weirdness" (Evert, 2004)

	$V = v$	$V \neq v$	
$U = u$	O_{11}	O_{12}	$= R_1$
$U \neq u$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

observed frequencies

$$\text{odds-ratio} = \log\left(\frac{O_{11} * O_{22}}{O_{12} * O_{21}}\right)$$

	$V = v$	$V \neq v$	
$U = u$	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$	
$U \neq u$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$	

expected frequencies

$$\text{log-like-ratio} = 2 * \sum_{ij} (O_{ij} * \log\left(\frac{O_{ij}}{E_{ij}}\right))$$

retskrivning sb., -en, -er, i
sms. retskrivnings-, fx ret-
skrivningssystem.
retslig (et. retlig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, rets-
lægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

Eksempel: rejsevaccinationer

U=Jyllandsposten, V=rejsevaccinationer: 5

U=Jyllandsposten, V≠rejsevaccinationer: 75.000-5

U≠Jyllandsposten*, V=rejsevaccinationer: 0

U≠Jyllandsposten*, V≠rejsevaccinationer: 28.000.000-0

Undgå at dividere med 0: Læg 0,5 til alle fire tal

$\text{Log}((5,5*28000000,5)/(74995,5*0,5)) = \log(4107) \approx 8.32$

* (dvs. Korpus 2000)

retskrivning sb., -en, -er, i sms. retskrivnings-, fx retskrivningssystem.
 retslig (et. retlig) adj., -t.
 retslæge sb., -n, -r.
 retslægeråd sb., -et, retslægeråd, bf. pl. -ene.
 retslærd adj., itk. d.s.
 retsløs adj., -t.

c) Nyordssignaler

... behandlingsgarantien, **som den kaldes** i folkemunde ... (Nordjyske Stiftstidende 28/9/08)

Signalstreng	Frekvens (ppm)	Fremfinder:	Antaget effektivitet
Citationstegn I (")	832	NP'er, VP'er	Høj genkaldelse, lav træfrate
Citationstegn II (')	427	NP'er, VP'er	ditto
så kaldt(e)	155	NP'er	Moderat genkaldelsesrate, moderat træfrate
som den det de hedder	8	NP'er, VP'er	Lav genkaldelse, højere træfrate
som den det de kaldes	0,8	NP'er, VP'er	Meget lav genkaldelse, høj træfrate

retskrivning sb., -en, -er, i
sms. retskrivnings-, fx ret-
skrivningssystem.
retslig (et. retlig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, rets-
lægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

1. eksperiment (*recall*)

177 korte avisartikler (ca. 75.000 løbende ord fra Jyllands-Posten den 20. september 2008) hvori en praktikant ved Dansk Sprognævn manuelt har opmærket i alt 252 nydannelser.

Nye ord: 208 (fx "basisteam", "terrormanual")

Nye udtryk*: 33 (fx "at gå kort", "unge til unge-mentorer")

Ny betydning: 11 (fx "retorik")

Guldstandarden er efterfølgende blevet revideret af en seniorexcerptist (Pia Jarvad). Nyordskandidaterne er vurderet i forhold til excerpteringens mål, i dette tilfælde en nyordsbog.

*Kræver store mængder data (+ statistik & avanceret NLP) at fremfinde maskinelt => problem ift. recall

retskrivning sb., -en, -er, f
sms. retskrivnings-, fx ret-
skrivningssystem.
retslig (el. retlig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, rets-
lægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

Hvad finder Ordtrawleren?

Rang (log-odds)	f(Jyp.)	f(K2000)	kandidat	støjtype
1	5	0	vasopressin	fagsprog
2	5	0	rejsevaccinationer	banal sms.
3	5	0	rejserådgivning	banal sms.
4	5	0	ABX	proprium
5	4	0	vurderingspris	gammel
...				
13	3	0	klangmassage	o.k.
...				
19	6	1	24-års	fragment
20	2	0	voksenhandicapområdet	o.k

retskrivning sb., -en, -er, f.
 sms. retskrivnings-, fx retskrivningssystem.
 retslig (et. retlig) adj., -t.
 retslæge sb., -n, -r.
 retslægeråd sb., -et, retslægeråd, bf. pl. -ene.
 retslærd adj., itk. d.s.
 retsløs adj., -t.

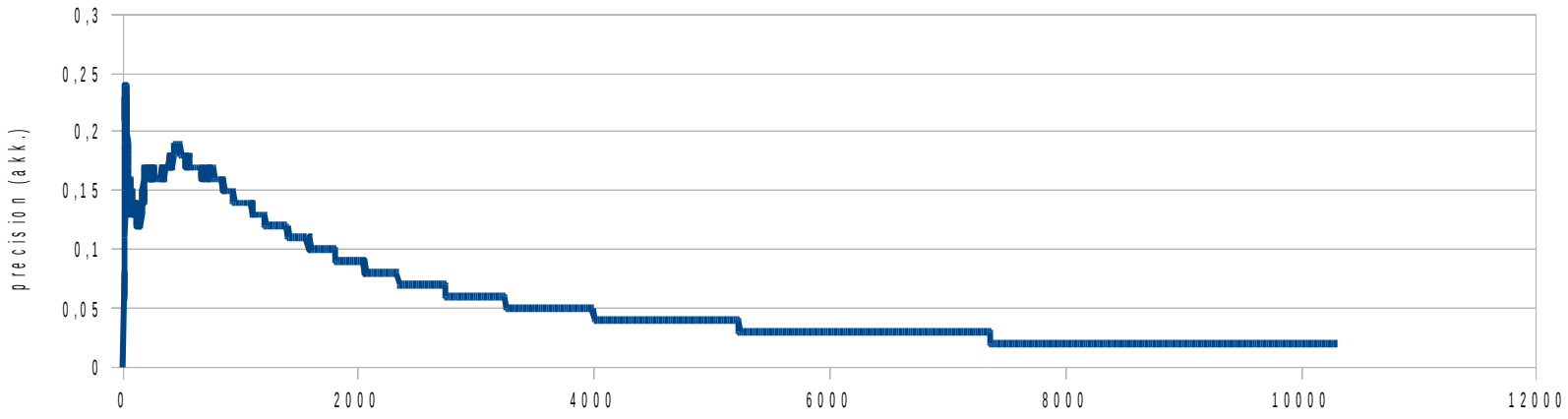
Maskine mod menneske: primitivt filter (208) eller signaler (219)

Nyordskandidater i Jyllandsposten den 20/9 2008 (14431 types)				
	#kandidater	F-score	recall	precision
Menneske	208	1	100%	100%
Maskine (inkl. proprier)	1061	0,22	69% (142/208)	13% (142/1061)
Maskine (eks. proprier)	589	0,31	60% (124/208)	21% (124/589)
Maskine (inkl. prop. u. Korpus 2000)	1498	0,20	84% (174/208)	12% (174/1498)
Maskine (eks. prop. u. Korpus 2000)	878	0,28	73% (152/208)	17% (149/878)
Maskine (kun nyordssignaler, inkl. proprier)	15	0,06	3% (6/219)	40% (6/15)

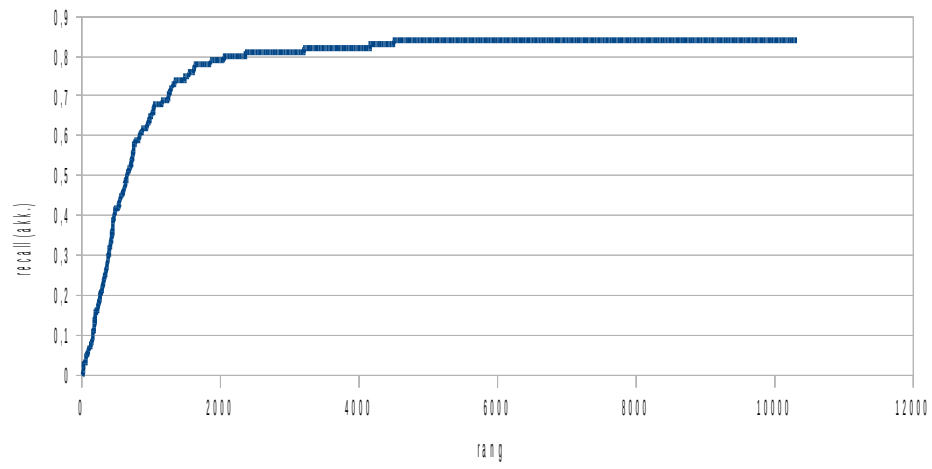
retskrivning sb., -en, -er, i sms. retskrivnings-, fx retskrivningssystem.
retslig (et. retlig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, retslægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

Maskine mod menneske II: statistisk sort. mod Korpus 2000 (219)

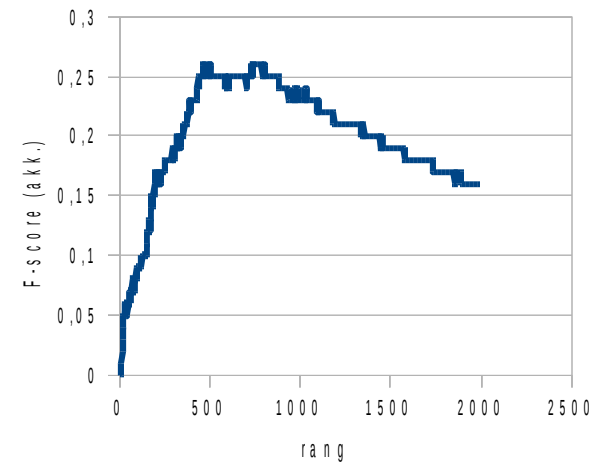
Log-odds (alle types undt. navne)



Log-odds (alle types undt. navne)



Log-odds (alle types undt. navne)



retskrivning sb., -en, -er, i
sms. retskrivnings-, fx ret-
skrivningssystem.
retslig (et. retlig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, rets-
lægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

2. eksperiment (*precision*)

Ca. 96,7 millioner løbende ord fra 55 forskellige danske dagblade i Infomediabasen (9. oktober 2007 - 11. oktober 2008).

Ordtrawlerens konfiguration: nyordssignaler kombineret med primitiv filtrering og udelukkelse af "singletonner".

Resultat: 1784 nyordskandidater

Manuel evaluering af de 200 mest og 200 mindst frekvente kandidater.

Resultat: To excerpter anså i alt 180 af de 400 ord for relevante nok til at indgå i Ordsamlingen. Enighed om 152 af de 180 (enighedsgrad på 84,4 %). Træfrate ml. 38 % og 45 %.

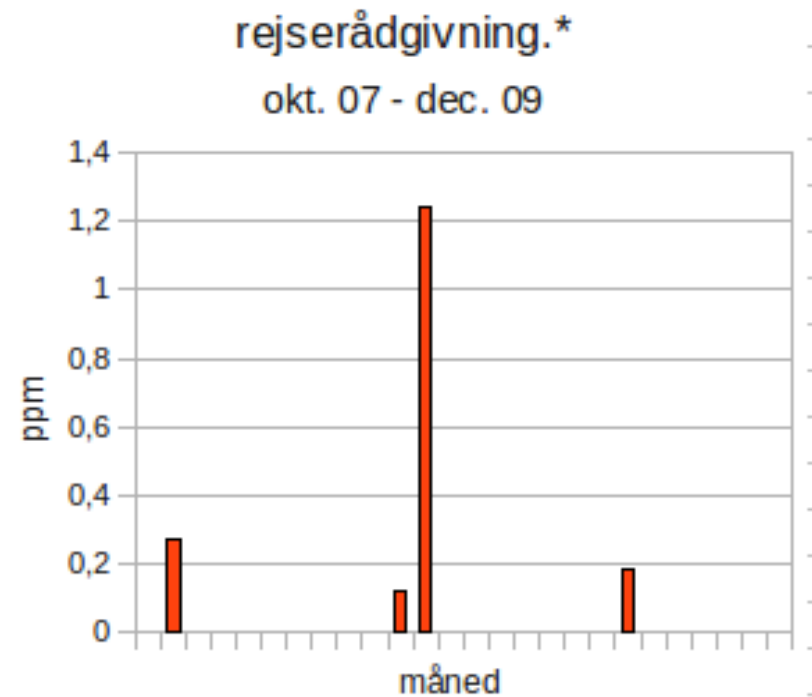
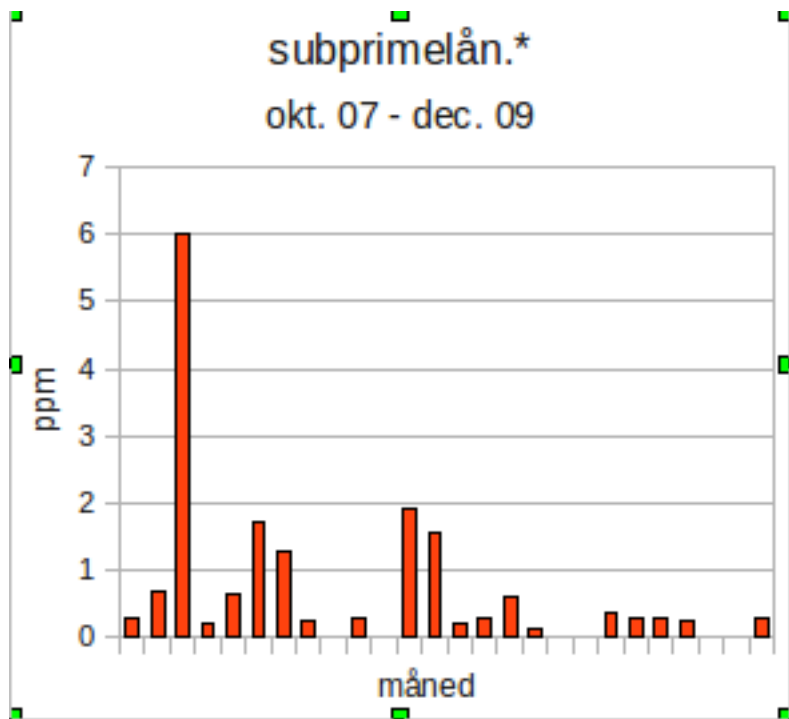
retskrivning sb., -en, -er, i
sms. retskrivnings-, fx ret-
skrivningssystem.
retslig (et. retlig) adj., -t.
retslæge sb., -n, -i.
retslægeråd sb., -et, rets-
lægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

Foreløbige konklusioner

- Kollokerende **nyordssignaler** som relevanskriterium øger systemets træfrate betydeligt, men reducerer genkaldelsesraten voldsomt.
- Er der adgang til store mængder data (fx www), så er reduktionen i genkaldelsesrate mindre væsentlig.
- Primære **støjkilder**
 - Bøjningsformer hvor lemmaet ikke er nyt (28,8 %)
 - Banale sammensætninger (28,1 %)
 - Fagsprog (15,3 %)
 - Stavefejl (8,7 %)
 - Proprier (4, 2%)
- Forbedringsforslag
 - Brug af god **lemmatizer** (ikke bare udfoldet RO)
 - Inddragelse af kandidaternes **diakrone frekvensprofiler**

retskrivning sb., -en, -er, f.
sms. retskrivnings-, fx retskrivningssystem.
retslig (el. retlig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, retslægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

Diakrone frekvensprofiler



Kometord (Jarvad, 1995)
eller varigt tilskud?

Banal sammensætning. Lav
relativ frekvens. Lav spredning.

retskrivning sb., -en, -er, sms. retskrivnings-, fx retskrivningssystem.
 retslig (et. retlig) adj., -t.
 retslæge sb., -n, -i.
 retslægeråd sb., -et, retslægeråd, bf. pl. -ene.
 retslærd adj., itk. d.s.
 retsløs adj., -t.

Byg dit eget webkorpus: GlossaNet

Actions Menu

- Wizard**
- Web resources
- My websites List
- My RSS feeds
- NLP resources
- My tasks
- Concordances

Glossa

- Home
- Forum
- My account
- News**
- Information
- Our websites
- Our RSS feeds
- Publications

Links

- CENTAL
- UCL
- Unitex
- Amazon**

Home > My RSS feeds

Add a RSS feed.

My RSS feeds

Websites

Type

Categories

Language

Country

Website	Name	Type	Category	Lang.	Country
----	humanisme.dk (Rune Engelbreth)	Blog	Auto	~ Other	Denmark
----	DanskPolitikKommentarer	Blog	Auto	~ Other	Denmark
----	http://feeds.shevy.dk/	Blog	Auto	~ Other	Denmark
----	Om livet i al almindelighed kommentarer	Blog	Auto	~ Other	Denmark
----	http://www.hverkenfuglellerfisk.dk/	Blog	Auto	~ Other	Denmark
----	http://tveskov.com/blog	Blog	Auto	~ Other	Denmark
----	Etcetera	Blog	Auto	~ Other	Denmark
----	http://bizzen.blogs.business.dk/	Blog	Auto	~ Other	Denmark
----	http://medieblogger.dk/	Blog	Auto	~ Other	Denmark
----	Reuters bureau	Blog	Auto	~ Other	Denmark
----	http://www.spacemermaid1001.dk/	Blog	Auto	~ Other	Denmark
----	http://moccapiqen.dk/	Blog	Auto	~ Other	Denmark
----	http://www.larsbachmann.dk/	Blog	Auto	~ Other	Denmark
----	http://piabau.blogspot.com/	Blog	Auto	~ Other	Denmark
----	http://www.hovedetpaabloggen.dk/	Blog	Auto	~ Other	Denmark
----	Uden relevans	Blog	Auto	~ Other	Denmark
----	Søren Pind	Blog	Auto	~ Other	Denmark
----	http://huskebloggen.blogspot.com/	Blog	Auto	~ Other	Denmark
----	http://abctema.blogspot.com/	Blog	Auto	~ Other	Denmark

20
Page 1 of 3
Displaying 1 to 20 of 47 items

retskrivning sb., -en, -er, -e
 sms. retskrivnings-, fx retskrivningssystem.
 retslig (et. retlig) adj., -t.
 retslæge sb., -n, -r.
 retslægeråd sb., -et, retslægeråd, bf. pl. -ene.
 retslærd adj., itk. d.s.
 retsløs adj., -t.

Og overvåg det!

Welcome Jakob Halskov
 Log out

Actions Menu

Wizard

+ Web resources

+ NLP resources

My tasks

Concordances

Glossa

Home

Forum

My account

News

+ Information

Our websites

Our RSS feeds

Publications

Links

CENTAL

UCL

University of Copenhagen

[Home](#) > Concordances

Download CSV
Download HTML

Delete filtered concordances

Suivez vos concordances à l'aide de ce flux RSS

Concordances

Task

Concordance date

Task	Date	Left	Middle	Right
Danish blc	2010-01-28	ointen? Modstandere af Hedegaard, Langballe og Kra	som de siger	. I øvrigt mener jeg, at de tre med flid undgår at sig
Danish blc	2010-01-28	ptærede sikkert for ikke at tale om udfarende både inter	som det hedder	, i deres nye rolle. Efterhånden må det vel også ve
Danish blc	2010-01-28	ler unge, at Fyn og Østjylland ikke er repræsenteret o	som det hedder	på forskersprog. Det er den slags parametre prof
Danish blc	2010-01-28	# 6. jan 2010 23:18, julie_a -----	som man siger	. Hun sov, så trak hun vejret langsommere og men
Danish blc	2010-01-28	en ihjel? http://avisen.dk/jalousi-formentlig-aarsag-til-d	såkaldte	"æresdrab", overvejende begået i det store tyrkisk
Danish blc	2010-01-28	man har ganske godt styr på den del af isbjerget, som	såkaldte	æresforbrydelser i Danmark. Der har været 349 o
Danish blc	2010-01-28	r Langballe går endnu længere i sine beskyldninger m	såkaldte	æresdrab) – og i øvrigt vender det blinde øje til onl
Extract ne	2010-01-28	t). Hvis en pædofil en snes gange havde voldttaget bø	såkaldt	gode venner i årtier beskytter en morder mod at bl
Danish blc	2010-01-27	ptærede sikkert for ikke at tale om udfarende både inter	som det hedder	, i deres nye rolle. Efterhånden må det vel også ve
Danish blc	2010-01-27	ointen? Modstandere af Hedegaard, Langballe og Kra	som de siger	. I øvrigt mener jeg, at de tre med flid undgår at sig
Danish blc	2010-01-27	individets indbyrdes placering til mobning? Personligt f	såkaldte	problemer som især lærer har følt sig kaldet til at k
Danish blc	2010-01-27	ve spillestil. Tre år holdt gruppen sammen, inden Cave	såkaldte	Gothic -genre, og siges at have inspireret vidt fors

Tak for opmærksomheden!

retskrivning sb., -en, -er, f.
sms. retskrivnings-, fx ret-
skrivningssystem.
retslig (et. rellig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, rets-
lægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.

Halskov, Jakob; Pia Jarvad (2010). "Human versus automated extraction of neologisms for lexicography". *Elexicography in the 21st century*. Cahiers du Cental, Vol. 6. Université catholique de Louvain.

Halskov, Jakob; Pia Jarvad (2009). "Om menneskers og maskiners tilgang til excerpering af sproglige nydannelser – en diskussion og en systemevaluering". *Nyt fra Sprognævnet 2009/4*. Dansk Sprognævn.

Jarvad, Pia (1995). *Nye ord – hvorfor og hvordan?* Gyldendal, København.

Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.d.-afhandling, Stuttgart Universitet.