

Korpusbaseret udvælgelse og prioritering af lemmakandidater

LEDA-oplæg 26. april 2017

Jørg Asmussen

Det Danske Sprog- og Litteraturselskab

Baggrund

Tilføje flere tusinde nye artikler til *Den Danske Ordbog* på ordnet.dk

1. Hvor & hvordan finder man “nye” ord?
2. Hvordan afgør man, om et ord er relevant?
3. Hvordan organiserer man beskrivelsen af relevante ord?

TIDSMASKINEN

- >> 1983 til 2017
- >> 1 milliard løbende ord
- >> ePOS-tagget
- >> Kildeoplysninger
- >> DDO-redaktionens primære kilde
- >> Søgeseystem: CoREST, CQP-baseret
- >> Offentlig version: korpus.dsl.dk/corest

Text Metadata

ID	1401212880
Date	2015-12-30
Group	io_infomedia_2015-12
Source	Weekendavisen
Title	Skandinavisk mørke
Publ.ID	11066
Publisher	Weekendavisen
Issue	1
Section	1
Pages	5
URI	urn:newsml:infomedia.dk:20151231:e56ec64d
Factuality	fact
Purpose	inform
Created	2015
Author	JON HUSTAD
Role	author
Sex	0
Age	adult

CoREST – Hovedvindue

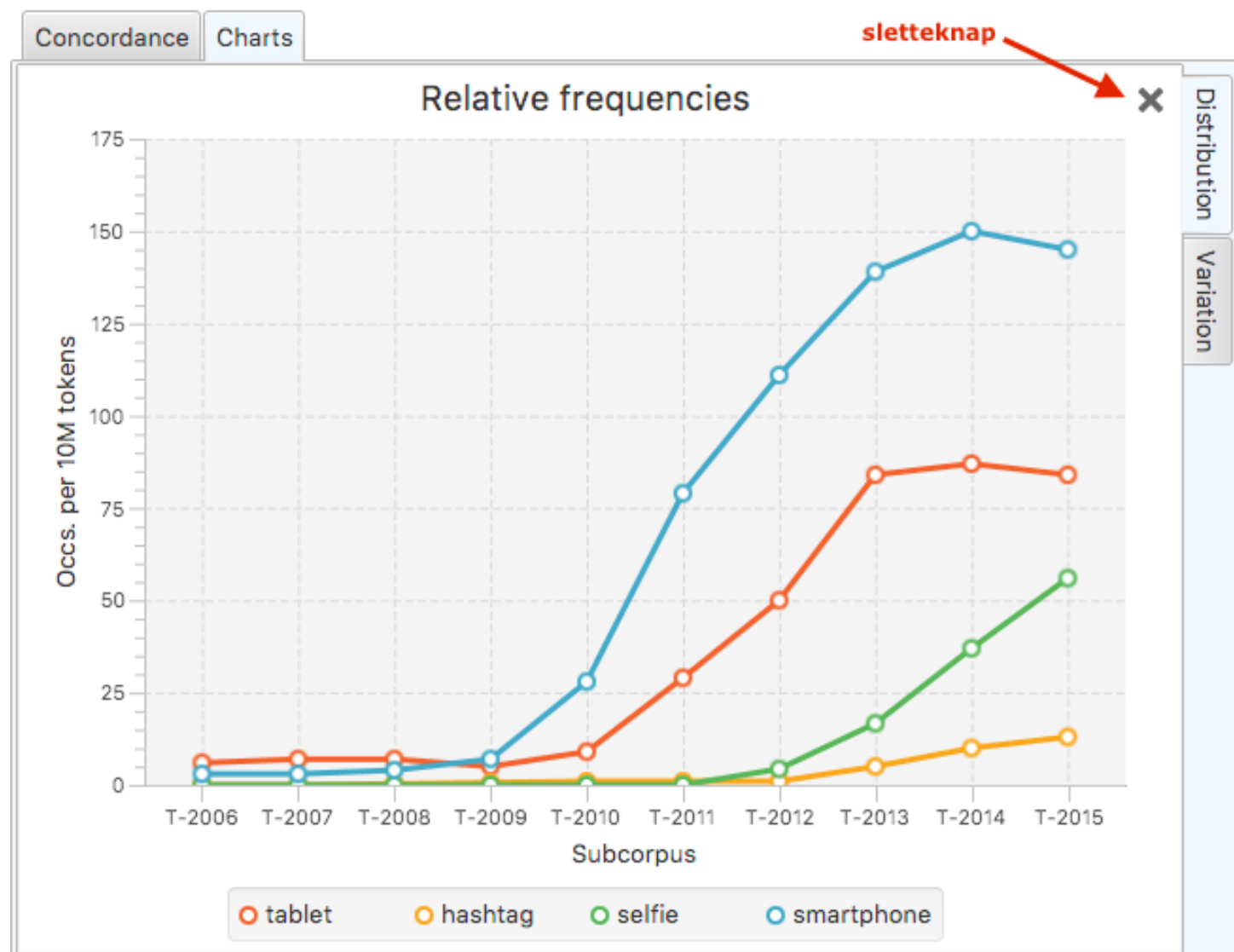
The screenshot displays the CoREST 2017 Standard Edition interface. The top bar includes the title 'CoREST 2017 · Standard Edition' and navigation buttons: 'Korpusvælger', 'Søgeudtryk', 'Søgeknop', and 'Filtre fra/til'. Below this, a dropdown menu shows 'KORPUS-DK' and a search input field contains 'korpus'. A 'Søgefelt' button is next to the input. To the left, a 'Listervælger' button is visible. The main interface is divided into several sections:

- Searches**: A table with 'My Queries' and a single entry 'korpus' with a count of 1.
- Concordance**: A table showing search results. The first row is highlighted, showing the text '...r færdigt, kan du her på siderne søge i Korpus 90.¶Korpus 90 og Korpus 2000 vil på man...'. The word 'Korpus' is highlighted in yellow.
- Sortering**: A dropdown menu showing 'Match sorted' and a page number '1'.
- Sidevælger**: A button for side navigation.
- Text**: A section with metadata for the selected text, including 'Date', 'Group', 'Source', 'Title', and 'Author'.
- Context**: A section showing the context of the selected text, including 'Kontekst' and 'Ordpålysninger'.
- Tags**: A section showing tags for the selected text, including 'Tags' and 'Ordpålysninger'.

The 'Søgeliste' button is located at the bottom left of the interface.

CoREST – Tidsstudier

Korpus **TiDK 2016** (2006–2015)



Frekvensforløb – Demo

Eksempler fra TIDSMASKINEN 1983–2017:

>> cloud

>> \$år 2000 problem.*

>> smartphone.*

>> tsunami.*

>> @fred

>> muslim.*

>> terror.*

Typer af forløb

- >> **brat stigning & brat fald**, evt. tilbagevendende:
år 2000-problem, tsunami
- >> langsom stigning: *cloud, islam-, terror-*
- >> langsomt fald: *fred*
- >> **brat stigning & langsomt fald**: *smartphone*
- >> langsom stigning & brat fald: ?
- >> synkronicitet: *islam-, terror-*

Årets ord – typen *brat stigning*

Årets udbredte ▼	2016 ▼	
Udtryk	Score	
landbrugspakke	22525	^
topskattelettelse	3322	
flygtningekrise	3062	
integrationsydelse	1960	
app	1745	
migrant	1472	
kontanthjælpsloft	1171	
bidragssats	1020	
kupforsøg	1010	
flygtningebarn	880	
grænsekontrol	838	
cannabis	791	▼

Årets påfaldende ▼	2016 ▼	
Udtryk	Score	
læringsplatform	14114	^
ældreminister	12459	
disruption	11825	
tyngdebølge	8280	
migrantkrise	7138	
omprioriteringsbidrag	6951	
shitstorm	5881	
hadprædikant	5112	
visumfrihed	3689	
udlændingeminister	3225	
jobparat	2153	
ambulancedrift	2082	▼

Beregning – Norm & afvigelse

- >> **Kvalitativ** bestemmelse af sproglig normalitet er let: ortografi, grammatik.
- >> **Kvantitativ** bestemmelse er problematisk:
Hvad er det normale antal verber i præsens i en tekst?
- >> Ingen absolut sandhed => **heuristisk** fremgangsmåde
- >> **Konsistens**: Den valgte metode anvendes hver gang og på tekstmasser af ca. samme omfang

Finde årets prominente ord

1. For hvert ord i årgang x beregnes dets relative frekvens
2. For de samme ord beregnes relativ frekvens for alle øvrige år
3. Beregn **overhyppighed** for hvert ord i årgang x ved at dividere resultatet fra (1) med resultatet fra (2)
4. Sorter efter overhyppighed
5. Anvend **frekvensfiltre** efter behov => udbredte vs. påfaldende
6. Vi kunne kalde metoden *Mutual Information*, sort of

Overhyppighed kan bruges mange steder

- >> Årets ord
- >> Månedens ord
- >> Ord i bestemte tekster eller teksttyper, fx *Wikipedia*



Månedens ord ▼ 2016-07 ▼		
Udtryk	Score	
kupforsøg	4927	^
sommerserie	4457	
tidstab	2205	
sidevind	2001	
pointtrøje	1405	
kupmager	1220	
cykelkultur	1214	
bjergenkeltstart	1214	
cykelnation	1029	
nice	1003	
dopingprogram	962	
sportsministerium	849	
konvent	817	
juli	766	
sommerquiz	734	
udenlandsdansker	678	
fængselspersonale	678	
klassement	571	
ferieramme	516	▼

Eksterne kilder: “Nomatch”

- >> No-matches fra DDO-søgelog
 - >> Hyppigst søgte
 - >> Potentielt relevante (lavfrekvente i korpus)
- >> No-matches fra KorpusDK-søgelog

Nomatch-lister ▼ efterspurgt ▼		
Udtryk	Score	
skisme	736	^
roomie	708	
ligament	694	
misogyn	681	
horribilis	532	
stigningstal	519	
genitalia	516	
vitæ	457	
flexe	456	
laissez	455	
excitere	429	
marquis	429	
valbirk	408	
outro	404	
langbortistan	398	
anførelsestegn	397	
præfix	396	▼

Eksterne kilder: Diverse

- >> DSN's & DSL's fælles indbretningsbase
- >> Ord i DDO-betydningsdefinitioner, som ikke kan slås op i DDO
- >> SpORDhundeord
- >> Årets nye opslagsord i Retskrivningsordbogen
- >> Ord fra gammel lemmakandidatbase

Lemmakandidatbasen *Semafor*

- >> Alle ord fra de nævnte lister
- >> Ordlister fra det hidtidige arbejde med DDO
 - >> DDO1: færdige artikler i DDO pr. 2015-12-31
 - >> Ord i ufærdige artikler
 - >> Ordforslag fra arbejdet med Begrebsordbogen
 - >> Redaktionelle ordforslag

Demo af Semafor

CoREST 2017 · DSL Edition

-TIDSMASKINEN %vragdel

Searches Matches Tags DDO

DDO 2 Today

Word	Who	Weight
vækstminister	IHW	1487
storebæltstæppe	LTJ	544
vragdel	MMS	544
elevløn	IHW	201
arkivstudie	JJ	123
telefonmast	JJ	99
toksikolog	JJ	99
gruppeportræt	JJ	98
resistensudvikling	JJ	97
polyrytmisk	JJ	96
førstegangsaflevering	JJ	76

Concordance Charts

T-2014 20

78 matches Match sorted 1

Text	Context	Tags
...kkerhed, siger til tv-stationen NBC, at hver eneste	vragdel	er vigtig for efterforskningen - og at det også har bety...
...visen The Guardian har et billede af en perforeret	vragdel,	som ligeledes stammer fra flyets forreste del.¶David G...
...det ikke har været muligt at finde nogen som helst	vragdele,	og der er ikke blevet modtaget nogen livstegn fra fly...
...e.¶MH017 blev flået fra hinanden i luften, og såvel	vragdele	som lig er derfor spredt over et stort område omkring ...
...delige spor efter den voldsomme ulykke. Ødelagte	vragdele	og splintret glas fra en bilrude vidnede om det frontal...
...edningsvest eller ej. Så enkelt er det.¶De første	vragdele	fra det forulykkede fly blev spottet tirsdag ud for Born...
...llers have fløjet ud over havet i håbet om at spotte	vragdele	fra MH370, som forsvandt i begyndelsen af marts me...
...delvis kunne forklare, hvorfor det er svært at finde	vragdele,	da flyet i så fald vil være spredt over et meget stort o...
...ellem bombeattentaterne og de to moskébrande.¶	Vragdele	og de første lig fra det forsvundne AirAsia-fly blev i g...
...e Ocean. Trods en gigantisk eftersøgning er ingen	vragdele	fundet.¶I juli blev MH17 skudt ned over det østlige Uk...

Text +

Date 2014-07-21

Group io_infomedia_2007

Source Jyllands-Posten

Title Håb om at eksperter endelig får adgang til vrak af MH017

Publisher Jyllands-Posten

Author HENRIK THOMSEN Jyllands-Postens

Context +

med ting, de finder - og en frivillig besluttede derfor at overlade en pung til en BBC-journalist, som tilfældigt kom forbi. Deborah Hersman, formand for det amerikanske råd for transportsikkerhed, siger til tv-stationen NBC, at hver eneste vragdel er vigtig for efterforskningen - og at det også har betydning for opklaringen af, hvor de landede.

Tags +

Deborah NP:su#:::---
Hersman NP:su#:::---
formand NC:siuc:::---
for T:-:---:---
det PM:s-un:::---
amerikanske AC:sdu#:::---
råd NC:siun:::---
for T:-:---:---
transports... NC:siuc:::---
siger VF:---:sa:---
til T:-:---:---
tv NW:---:---:---
stationen NC:sduc:::---
NBC NP:su#:::---
at CS:---:---:---

Append mode

Add user tag

entryId 53005505

Weight 544

Sources AarOrdTyp-2014 bOrd-medium

Resumé af *Semafor*-demo

- >> Ord fra alle kilder samlet i én base
- >> Kun ord, som ikke har en artikel i DDO, vises umiddelbart
- >> Ord er vægtet efter korpusfrekvens
- >> Ord kan reserveres i basen
- >> Ord kan foreslås i basen
- >> **Fremtid:** Vægtning ud fra en kombination af korpusfrekvens og kilderepræsentation

Konklusion

1. Hvor & hvordan finder man “nye” ord?
 - >> **Overhyppigheder, heuristiske filtre**
2. Hvordan afgør man, om et ord er relevant?
 - >> **Korpusfrekvens, kilderepræsentation**
3. Hvordan organiserer man beskrivelsen af relevante ord?
 - >> **Lemmakandidatbasen Semafor**