NyS

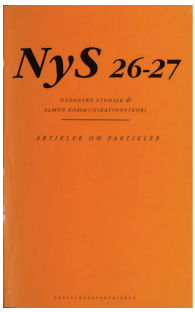
Titel: Tyrannocorpus Rex

Forfatter: Peter Juel Henrichsen

Kilde: *NyS – Nydanske Studier & Almen kommunikationsteori* 26+27. *Artikler om partikler*, 2000, s. 225-245

Udgivet af: Dansklærerforeningen

URL: www.nys.dk



Søgbarhed

Artiklerne i de ældre NyS-numre (NyS 1-36) er skannet og OCR-behandlet. OCR står for 'optical character recognition' og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

TYRANNOCORPUS REX

Nogle indledende korpuslingvistiske undersøgelser af den danske del af www

AF PETER JUEL HENRICHSEN

INDLEDNING

Www er det ultimative tekstkorpus. I det danske afsnit af http-protokollen har man umiddelbart søgeadgang til mere end tre millioner www-dokumenter – en tekstmasse som allerede nu overgår Den Danske Ordbogs tekstmateriale med mindst en faktor 3.[1] Alene størrelsen og genrebredden gør www unik som korpuslingvistisk ressource; læg dertil aktualiteten og tilgængeligheden, og vi står med et forskningsmæssigt potentiale som skal udnyttes, jo før jo hellere.

Internettet har ført til dannelse af nye tekstgenrer, midt i skellet mellem skriftsprog og mundtligt sprog. Indtil for nylig var dette skel af principiel og uoverstigelig art; skriftsproget var lig med envejskommunikation og det mundtlige sprog lig med tovejskommunikation. Der fandtes undtagelser hvor genrerne lånte stiltræk af hinanden, som fx stiv foredragsstil og døves konversationshæfter; men undtagelserne var få og uvæsentlige, og de støttede dermed forestillingen om skrift og tale som absolutte modpoler. Konserves og fastfood. Med Internettet er der opstået en ny genre, det hurtige skrivesprog kaldet *chat* som på få år har udviklet sig væk fra begge sine rødder og nu viser så selvstændige træk at det næppe længere kan forstås som en simpel mellemproportional.

- versaler forsvundet
- interpunktion forsvundet
- konventionaliserede reduktioner i ortografi og syntaks
- konventionaliserede replikker
- ekstralingvistiske jingles, "<g>" etc.
- symbolsk mimik, ;-)

Internettet har i tilgift, udover sine egne genrer, enhver af de gamle genrer repræsenteret til overmål: sagprosa, reklame, fiktion, poesi, historisk tekst,

dagbøger, børnetekst, etc. etc. Nettet er dermed et oplagt studieobjekt for filologen og lingvisten; men allermest oplagt for *korpus*lingvisten, som netop har specialiseret sig i at behandle store tekstmængder.

Derfor er det kreperligt at en bestemt egenskab ved www får korpusundersøgelser til at tage sig håbløse ud ved første blik. Problemet er at www-korpuset er svært at undersøge kvantitativt; de internationale søgemaskiner som giver adgang til www-dokumenterne in toto, giver slet ikke den kontrol over søgningens parametre som man er vant til fra søgetjenester som fx Den Danske Ordbogs. AltaVista tillader fx kun søgning efter fuldt specificerede ord; ingen trunkering, wildcards, regulære udtryk, etc. Årsagen ligger i tekstmassens enorme dimensioner som udelukker egentlig fuldtekstsøgning (selvom AltaVistas programmører er de rene magikere til at *simulere* fuldtekstsøgning). Man må altså klare sig med at søge på *ord*, hvad der betyder en alvorlig begrænsning for korpuslingvisten. Bare at opbygge en frekvensordliste er et sisyfosarbejde – længe inden man er færdig, har tekstmassen ændret sig signifikant, og så kan man begynde forfra. Men uden en pålidelig ordliste kan man ikke danne sig så meget som et første indtryk af korpus' distribution over tekstarter.

I denne artikel vil jeg foreslå en enkel og overkommelig metode til at få overblik over et kæmpekorpus – en opgave der kan minde om palæontologens, som ud fra nogle få knogler og nogle fornuftige slutninger skal fremmane billedet af en kæmpeøgle.

Artiklen har to dele: Først fremlægges selve målemetoden ved hjælp af et detaljeret eksempel. Derefter følger nogle afsnit med uddybende diskussioner (som dog kan springes over hvis det især er eksemplets konklusion der har læserens interesse).

ÉT OBJEKTKORPUS, TO STØTTEKORPORA

Metoden bygger på statistisk interpolation. Først udvælger man to små, kendte *støttekorpora* – de fungerer som endepunkter i en skala. Derefter tager man stikprøver af *objektkorpus* (www), og resultaterne bruger man til at vurdere hvor objektkorpus placerer sig i feltet mellem de to støttekorpora. Det er vigtigt at vælge de to støttekorpora med omhu, sådan at de afviger stærkest muligt fra hinanden på den parameter man vil undersøge.

PRÆMISSER

- De to støttekorpora kan have uens størrelse
- Om hvert støttekorpus skal man kende omfang og frekvensordliste (ikke andet)
- Der udtages fire sæt af *måleord* (20+20+20+20)
- Ethvert måleord skal være blandt de 100 hyppigste i et af de to støttekorpora
- Hvert støttekorpus kan være langt mindre end objektkorpus (ned til ca. 100.000 løbende ord)
- Objektkorpus' omfang og frekvensordliste indgår ikke i beregningen
- Der skal kun søges få gange i objektkorpus (<100 gange), og kun på fuldt specificerede ord
- Af søgeresultaterne skal kun bruges antal-fund, ikke fundene selv

Som det ses, er der fra starten taget hensyn til at metoden skal være håndtérlig, også på tidspunkter hvor www mest af alt står for world-wide-*wait*.

WWW – TALESPROG ELLER SKRIFTSPROG?

Den næste del af præsentationen tager form af et eksempel. Vi vil forsøge at besvare et påtrængende spørgsmål om teksten i det danske www-afsnit: Er stilen overvejende formelt-skriftsproglig (fx sagprosa), eller overvejende uformelt-mundtlig (fx chat)? Eller mere præcist: Hvor kan den søgbare del af www som helhed indplaceres i spektret mellem traditionel skriftlig og mundtlig stil?

Vi skal altså finde to støttekorpora som kan danne endepunkter i skalaen mellem skriftstil (S) og mundtlig stil (M); for at undgå alt for mange tabeller vælger vi korpora hvis frekvenslister er offentligt tilgængelige.

Som M-korpus anvender vi Projekt Bysociolingvistiks talesprogskorpus, siden 1998 tilgængeligt på http://www.cphling.dk/BySoc. Dette korpus, kaldet *BySoc*, tæller 1.277.822 leksikalske ord[2]; hvad angår tilgængelighed og størrelse er *BySoc* second to none i sin art.

Som modpol til *BySoc* vælger vi et af Maegaard et al.'s fire DANwORD korpora. Her står valget mellem *Ugeblade*, *Fagblade*, *Aviser* og *Børnebøger*. Hvert af disse fire korpora tæller 250.000 ord, og frekvensordlisterne har længe været i handelen (se ref.). I første omgang installerer vi *Fagblade* som S-korpus, men vi følger senere op med en sammenligning.

PROCEDURE

1. trin: Beregn 20 S_1-måleord
2. trin: Beregn 20 M_1-måleord
3. trin: Søg på de 40 måleord med AltaVista
4. trin: Placér www på skalaen mellem S-korpus og M-korpus

(Gentag proceduren med S_2 og M_2-måleord)

Første trin er at identificere de ordformer der er mest typisk for S-korpus. Hvorvidt et bestemt ord *w* er S-typisk, vurderer man ved at sammenligne *w*'s frekvens i S- og i M-korpus; er disse to tal næsten ens, eller ligefrem højest i M-korpus, så er *w* ikke S-typisk – er *w*'s frekvens derimod højest i S-korpus, er det S-typisk. De 20 mest S-typiske ordformer kaldes *S_1-måleord.*

S_1-måleord beregnes med formlerne herunder, efter ganske ukontroversielle principper. Den læser der ikke interesserer sig for formler, kan roligt springe henover – formlernes indhold bliver udlagt på dansk i det følgende.

FIGUR 1 Eksklusion_S, Kontrastfaktor_S (Formel_1)

$$T_M = \text{Løbende ord i M-korpus totalt} = 1.277.822$$
$$T_S = \text{Løbende ord i S-korpus totalt} = 250.000$$

$$C_S(w) = \text{Antal } w\text{-forekomster i S-korpus}$$
$$C_M(w) = \text{Antal } w\text{-forekomster i M-korpus}$$
$$F_S(w) = w\text{'s frekvens i S-korpus} = C_S(w)/T_S$$
$$F_M(w) = w\text{'s frekvens i M-korpus} = C_M(w)/T_M$$

$$\text{Eksklusion}_S(w) = \frac{F_S(w) - F_M(w)}{F_S(w) + F_M(w)} = \frac{C_S(w)\,T_M - C_M(w)\,T_S}{C_S(w)\,T_M + C_M(w)\,T_S}$$

$$C_S(w_1 \ldots w_{20}) = C_S(w_1) + C_S(w_2) + \ldots + C_S(w_{20}) = \sum_{i=1}^{20} C_S(w_i)$$

$$\text{Kontrastfaktor}_S = \frac{C_S(w_1 \ldots w_{20}) / T_S}{C_M(w_1 \ldots w_{20}) / T_M}$$

Beregningsproceduren er enkel nok: For hvert ord blandt de 100 hyppigste ord i S-korpus beregner man faktoren Eksklusion_S. Den er et mål for

ordets typikalitet som skriftsprogsord, eller som vi vil kalde det: Dets *eksklusion* i S-korpus. Derefter sorterer man de 100 ord efter deres Eksklusion$_S$, og de 20 højest rangerende ord udgør nu samlingen af S_1-måleord.

Denne beregningsprocedure vil vi referere til som *Formel*$_1$ – deraf det lille indeks-ettal i symbolet 'S_1'.

TABEL 1 S_1-måleord.

S_1-MÅLEORD	EKSKLUSION$_S$	RANG I S-KORPUS	RANG I M-KORPUS
samt	0.9964	#89	#13992
dette	0.9910	#41	#3570
disse	0.9848	#52	#3001
således	0.9765	#92	#3290
denne	0.8929	#37	#520
gennem	0.8487	#94	#796
mellem	0.7676	#66	#435
dansk	0.7673	#77	#476
nye	0.7291	#64	#377
større	0.6961	#91	#450
mod	0.6715	#87	#403
under	0.6637	#45	#262
af	0.5724	#4	#35
derfor	0.5252	#86	#298
uden	0.5189	#79	#272
store	0.5165	#70	#236
forskellige	0.4987	#100	#316
flere	0.4752	#97	#297
sin	0.4729	#84	#264
efter	0.4599	#35	#138

I tabel 1 ses resultatet af beregningen, nemlig de 20 S_1-måleord, som vi vil tolke som *de mest eksklusive danske skriftsprogsord.* Bemærk især at hvert måleord har stærkt afvigende rangnummer i S-korpus og M-korpus; de store rangforskelle er karakteristiske for måleord beregnet med Formel$_1$.

Ved hjælp af denne liste af S_1-måleord kan vi nu beregne Kontrastfaktor$_S$. Denne kontrastfaktor giver et mål for om S-korpus står effektivt i

kontrast til M-korpus eller ej. Den konkrete talværdi tolkes sådan at Kontrastfaktor$_S$= 1 betyder *ingen kontrast* (S_1-måleordene dækker den samme andel i hvert korpus), mens Kontrastfaktor$_S$>>1 betyder *stor kontrast* (S_1-måleordene dækker en langt større del af S-korpus end af M-korpus).

Vi har nu grundlag for en sammenligning. For hvert af de fire Maegaard-korpora kan vi beregne et sæt S_1-måleord og en Kontrastfaktor$_S$. Kontrastfaktoren er et mål for hvert enkelt korpus' egnethed som skalær modpol til *BySoc*.

TABEL 2 S-korpus.

S-KORPUS	S-DÆKNING	M-DÆKNING	KONTRAST-FAKTOR	FØRSTE FEM S_1-MÅLEORD
Fagblade	4,37%	0,98%	4,48	*samt, dette, disse, således, denne*
Aviser	6,64%	3,24%	2,05	*danske, denne, mod, nye, København*
Ugeblade	9,21%	5,15%	1,78	*denne, deres, sin, hans, uden*
Børnebøger	7,72%	2,84%	2,72	*sin, mod, lille, hans, store*

Det er tydeligvis *Fagblade* som er den stærkeste modpol til *BySoc* – de tyve mest mundfjendtlige ordformer dækker fire-en-halv gang bedre i *Fagblade* end i *BySoc*, et tal der ikke udfordres af de andre korpora. *Børnebøger* har dog en overraskende høj kontrastfaktor, men spørgsmålet er om ikke skalaen med *Børnebøger* ved den ene pol og *BySoc* ved den anden bør tolkes som et felt mellem børnestil og voksenstil, rettere end som et felt mellem formelt-skriftlig stil og uformelt-mundtlig stil (i sig selv ikke noget uinteressant felt – dog uden for denne artikels rammer).

MÅLEORD FRA M-KORPUS

Vi lader nu de to støttekorpora bytte plads i Formel$_1$-beregningen og får, mutatis mutandis, en liste af tyve M_1-måleord.

TABEL 3 M_1-måleord

M_1-MÅLEORD	EKSKLUSION$_M$	RANG I S-KORPUS	RANG I M-KORPUS
ja	0.9898	#547	#2
nej	0.9865	#1647	#26
nå	0.9717	#937	#33.
du	0.9582	#352	#16
altså	0.9221	#203	#20
jo	0.9026	#131	#18
bare	0.8959	#808	#71
sådan	0.8886	#144	#25
hun	0.8829	#272	#37
jeg	0.8733	#38	#3
tror	0.8330	#385	#63
sagde	0.7837	#464	#84
så	0.7684	#28	#6
noget	0.7619	#78	#32
min	0.7613	#265	#65
ham	0.7572	#331	#77
synes	0.7499	#309	#76
fordi	0.7487	#152	#48
kom	0.7358	#261	#72
han	0.7326	#48	#21

I tabel 3 ses de 20 M_1-måleord, som vi altså vil regne for *de mest eksklusive danske talesprogsord.* Hvor S_1-ordene især var præpositioner, adjektiver og upersonlige pronominer, er M_1-ordene interjektioner, personlige pronominer og verber. Konjunktioner finder man i begge ordlister – men med en klar stilforskel: 'samt', 'således' og 'derfor' er skriftsprogsord, hvor talesproget foretrækker 'altså' og 'fordi'.

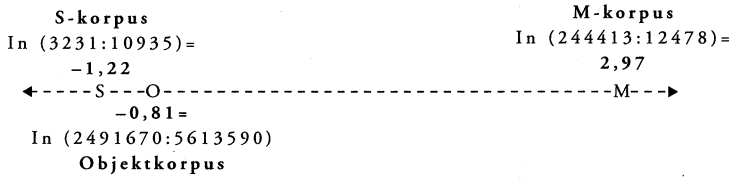
STILEN I WWW: FØRSTE RESULTAT

Det tredje trin i proceduren er at søge på de 20+20 måleord fundet ved hjælp af Formel$_1$; søgeresultaterne med AltaVista er fremlagt i slutningen af artiklen. I tabellen herunder ses de 40 måleords andele i hver af de tre korpora. Og så er vi parat til trin 4, til at indplacere www-tekstmassen på skalaen udspændt mellem formel-skriftlig stil og uformel-mundtlig stil.

TABEL 4 Måleordenes dækning i S- og M-korpus.

FOREKOMSTER ...	... I S-KORPUS	... I M-KORPUS	... I OBJEKTKORPUS
... af de 20 S-måleord	10.935	12.478	5.613.590
... af de 20 M-måleord	3.231	268.835	2.491.670

FIGUR 2 S → M placering (modulo Formel$_1$)



Den foreløbige konklusion er at objektkorpus placerer sig i området mellem S-korpus og M-korpus, langt tættest på S. Hvis skalaen fra S til M regnes fra 0-100%, befinder Objektkorpus sig ved 9,7%.

Www-tekstmassen er altså, efter denne bestemmelse, nogenlunde lige så formelt-skriftsprogligt som fagbladet, og i alt fald langt mindre mundtligsprogligt end den uformelle samtale i den danske dagligstue.

Bemærk at intet forhindrer et Objektkorpus i at placere sig udenfor området mellem de to støttekorpora, hvis det fx har et endnu mere skriftsprogligt forhold mellem S-måleord og M-måleord end *Fagblade* og dermed, så at sige, slår S-korpus på hjemmebane. Det gives der flere eksempler på herunder.

STILEN I WWW: ANDET RESULTAT

En god måde at kontrollere det første resultat på er at gentage eksperimentet, men med dets mest teoriladede del udskiftet. Vi erstatter derfor Formel$_1$ med en ny beregningsprocedure, Formel$_2$, der måler typikalitet på en anden måde end Formel$_1$ og afleverer et andet sæt måleord. Hvor Formel$_1$-måleord først og fremmest er *eksklusive*, er Formel$_2$-måleord først og fremmest *dominerende* i måleteksten. Formel$_2$ erstatter Eksklusion$_S(w)$ med Dominans$_S(w)$ i beregningen af måleord, men følger ellers samme opskrift.

FIGUR 3 Dominans$_S$, Kontrastmasse$_S$ (Formel$_2$)

$$\text{Eksklusion}_S(w) = \frac{F_S(w) - F_M(w)}{F_S(w) + F_M(w)} = \frac{C_S(w)\,T_M - C_M(w)\,T_S}{C_S(w)\,T_M + C_M(w)\,T_S}$$

$$\text{Dominans}_S(w) = F_S(w) - F_M(w)$$

$$\text{Kontrastfaktor}_S = \frac{C_S(w_1 \ldots w_{20})/T_S}{C_M(w_1 \ldots w_{20})/T_M}$$

$$\text{Kontrastmasse}_S = C_S(w_1 \ldots w_{20})/T_S - C_M(w_1 \ldots w_{20})/T_M$$

(formlerne for Eksklusion$_S(w)$ og Kontrastfaktor$_S$ er gentaget herover for sammenligningens skyld). Betragt til eksempel det første S-måleord i hhv. Formel$_1$- og Formel$_2$-listen, nemlig 'samt' og 'af'.

TABEL 5 Eksklusion og Dominans for de to første S-måleord.

S-MÅLEORD	EKSKLUSION$_S$ (FORMEL$_1$)	DOMINANS$_S$ (FORMEL$_2$)
samt (først i S_1-listen)	0,9964	0,0086%
af (først i S_2-listen)	0,5724	1,51%

TABEL 6 Eksklusion og Dominans for de to første M-måleord.

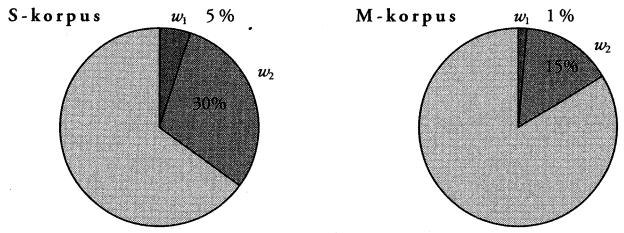
M-MÅLEORD	EKSKLUSION$_M$ (FORMEL$_1$)	DOMINANS$_M$ (FORMEL$_2$)
ja (først i M_1-listen)	0,9898	3,25%
det (først i M_2-listen)	0,5674	4,07%

Måleordet 'samt' er et stærkt *eksklusivt* skriftsprogsord, fordi det dækker en langt større del af S-korpus end af M-korpus; i denne egenskab er det topscorer på S_1-måleordslisten, sådan som det fremgår af tabel 1. I sammenligning med 'samt' er måleordet 'af' langt mindre eksklusivt, men derimod

stærkt *dominerende*, fordi det breder sig over 1,51% mere af den samlede tekstmasse i S-korpus end i M-korpus, hvad der giver det en førsteplads i S_2-måleordslisten.

Hvor $Formel_1$ optimerer kontrast*faktoren*, er det kontrast*massen* der optimeres med $Formel_2$. Forskellen ses tydeligt i et lagkagediagram.

FIGUR 4 Cirkler og cirkeludsnit



Måleordet w_1 dækker 5 gange mere af S-korpus end af M-korpus i modellen herover; w_1 er altså meget *eksklusivt* og et godt S_1-måleord. Ordet w_2 breder sig over 15% mere af S-korpus end af M-korpus; w_2 er altså meget *dominerende* og dermed et godt S_2-måleord.

TABEL 7 Kontrastfaktor og Kontrastmasse.

S-MÅLEORD	$KONTRASTFAKTOR_S$	$KONTRASTMASSE_S$
S_1-måleord	4,48 [14,8]	3,40% [17,8%]
S_2-måleord	1,94 [4,52]	6,29% [26,5%]

De 20 S_2-måleord (beregnet med $Formel_2$) ses her, sorteret efter faldende Dominans:

> *af* [#4], *i* [#1], *at* [#3], *til* [#7], *for* [#9], *en* [#6], *som* [#15], *den* [#12], *med* [#13], *på* [#11], *et* [#17], *fra* [#23], *denne* [#37], *sig* [#27], *dette* [#41], *om* [#19], *efter* [#35], *disse* [#52], *under* [#45], *vil* [#25], *ved* [#22], *mellem* [#66], *nye* [#64], *dansk* [#77].

Tallene i klammer angiver rangnummeret i S-korpus. De tilsvarende M_2-måleord (med deres rang i M-korpus) er:

det [#1], *ja* [#2], *jeg* [#3], *så* [#6], *var* [#9], *ikke* [#8], *du* [#16], *der* [#7], *jo* [#18], *altså* [#20], *er* [#4], *sådan* [#25], *nej* [#26], *han* [#21], *men* [#17], *nå* [#33], *vi* [#22], *noget* [#32], *har* [#11], *også* [#29], *hun* [#37].

Man ser en klar forskel mellem $Formel_1$- og $Formel_2$-listerne, i fordelingen af rangnumre. Hvor $Formel_1$-måleord fordelte sig nogenlunde jævnt over rangnumrene i det tilladte område 1-100, er de to lister af $Formel_2$-måleord klart korrelleret med frekvensordlisterne for henholdsvis S-korpus og M-korpus. Følgelig dækker $Formel_2$-måleordene en langt større del af de respektive korpora end $Formel_1$-måleordene. Til gengæld er der betydeligt mindre kontrast mellem dækningen af S-måleord og M-måleord – sammenlign de logaritmiske forholdstal herunder med tallene i fig. 2.

FIGUR 5 S → M placering (modulo $Formel_2$)

TABEL 8 S → M placering for diverse objektkorpora.

OBJEKTKORPUS	S → M, $FORMEL_1$	S → M, $FORMEL_2$
Fagblade	0	0
www_{Danish} (AltaVista)	9,7%	+0,7%
Aviser	17,0%	+3,6%
Ugeblade	38,9%	+20,9%
Børnebøger	62,2%	+41,4%
VOKSENTALE	105,9%	+88,3%
BØRNETALE	97,4%	+116,1%
BabySoc (kontrol)	99,97%	+99,63%
BySoc	100%	100%

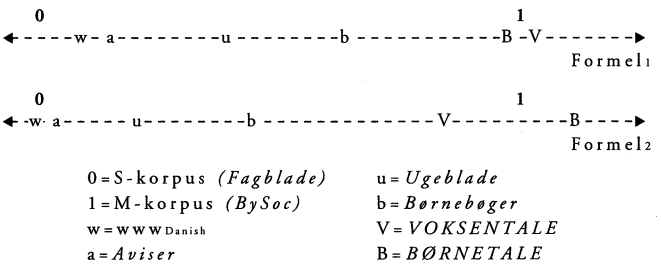
$Formel_2$ placerer Objektkorpus 0,7 procentpoint til venstre for S-korpus, det vil sige cirka 10 procentpoint fra resultatet i fig. 2. Er der tale om en systematisk forskel her, eller skal man regne forskydningen for almindelig måleusikkerhed?

For at undersøge forholdet nærmere prøver vi at gentage beregningerne med nogle uafhængige objektkorpora.

Resultaterne er her angivet som procenttal, sådan at 0 svarer til en placering ved S-korpus og 100% til en placering ved M-korpus (alle tal i beregningerne kan findes enten i teksten eller via referencerne). Korpus *BabySoc* er en lille del af *BySoc*, ca. 100.000 ord, udvalgt tilfældigt, men repræsentativt (jf. Henrichsen (1998)).

De fem mindste korpora placerer sig uden de store overraskelser, efter faldende skriftsprogskarakter: *Aviser* → *Ugeblade* → *Børnebøger* → *VOKSENTALE&BØRNETALE*[3]. Bemærk at det lille korpus *BØRNETALE*, der består af transskriberet børnesprog, rækker et godt stykke ud i det hypermundtlige overdrev. Det virker absolut ikke urimeligt at en børnesprogstransskription flygter så langt væk fra tør sagprosa som vel muligt – dets reelle placering er endda ca. 20 procentpoint længere til højre, når man kompenserer for den specielle deiktiske brug af pronominet 'den', som diskuteret herunder i afsnittet om dansk-danske homografer.

FIGUR 6 De seks objektkorporas indbyrdes placering



De store korpora er placeret længere til højre på $Formel_1$-aksen end på $Formel_2$-aksen (årsagen diskuteres herunder); men hvad der er mere interessant i denne sammenhæng, er at ordningen af de fem store korpora er den samme over begge skalaer, og endda også deres indbyrdes afstand i god tilnærmelse. Jeg vover derfor det ene øje og fremsætter to påstande.

To påstande om tekstmassen i den danske del af Internet (som indekseret af AltaVista, maj 1999)

1. Stilen i www_{Danish} er altovervejende formelt-skriftsproglig
2. Stilen i www_{Danish} er nærmere beslægtet med fagbladet og avisen end med ugebladet

Med disse to antagelser kan vi estimere den samlede ordmasse i det (søgbare) danske www-afsnit.

Ordmasse = (2.491.670+5.613.590)/(3.231+10.935)x250.000 ≈
143 mio. løbende ord

Dette tal dækker den *indekserede* ordmasse, der – som diskuteret herunder – er lidt mindre end den *søgbare* ordmasse. Tallet Ordmasse er altså et lavt estimat.

DISKUSSIONER

Indtil nu har forfatteren givet sig selv en easy case ved hverken at berøre metodens problemer eller de deraf følgende ad hoc løsninger. De omtales i det følgende.

HVILKEN FORMEL ER DEN MEST ATTRAKTIVE?

Metodisk set er $Formel_2$ den elegante. $Formel_2$-måleord har lavt rangnummer (høj hyppighed), i praksis langt lavere end rang-100, som er den grænse der er defineret i præmisserne i begyndelsen af artiklen. Det giver altså ingen forskel i måleordslisten hvor man end flytter ranggrænsen hen, så længe den blot er sat højt nok. Den kan fx sættes ved rang-∞, og dermed har man skåret en arbitrær konstant ud af beregningen.

Men der klæber imidlertid tre praktiske problemer til $Formel_2$, som alle har at gøre med AltaVistas idiosynkrasier. Det første problem handler om blokering. Med $Formel_2$ høster man stærkt *frekvente* måleord. Frekvente ord er typisk *korte* ord. Mange korte danske ord har stærkt frekvente homografer i de vestlige hovedsprog, og mange sådanne er blokeret for søgning. Kun 35 ud af de 40 oprindelige $Formel_2$-måleord kan søges på med AltaVista, mens de sidste fem er blokerede, nemlig 'der', 'i', 'at', 'for' og 'en'. Der er derimod ingen blokeringer på de 20+20 $Formel_1$-måleord.

Da måleusikkerheden kan forventes at være mindst ved de højest placerede måleord – der hvor kontrasterne er størst – giver AltaVistas blokeringer $Formel_1$ et (lille) fortrin frem for $Formel_2$.

Det andet problem med $Formel_2$ er beslægtet med det første, men er mere kritisk. Somme tider returnerer AltaVista et resultat med et modificerende *about* knyttet til sig; sammenlign disse to søgninger på 'fra' og 'du':

"fra": AltaVista found about 1,035,029 Web pages for you.
"du": AltaVista found 773270 Web pages for you.

Output som disse kunne tyde på et skift af algoritme omkring 1 million hits, sådan at søgeresultater derover kun er estimerede. Langt de fleste AltaVista-søgninger bekræfter den hypotese; men der er besynderlige undtagelser ind i mellem, fx som her:

"med": AltaVista found 1,773,270 Web pages for you.
"han": AltaVista found about 130752 Web pages for you.

Hvad 'about' skal betyde, og hvornår det bruges, er ikke dokumenteret i AltaVista's hjemmeside, og når man henvender sig personligt til AltaVista, er svaret naturligvis: Forretningshemmelighed! Man har altså ikke, som bruger, begreb om måleusikkerheden. Ved at søge igen og igen på de samme højfrekvente søgeord kan man sætte usikkerheden til ikke under 3%. Men den *øvre* grænse for måleusikkerheden kan man selvfølgelig ikke finde ad denne vej – den indgår som en ubekendt fejlkilde i alle beregninger, rimeligvis især med stærkt frekvente søgeord, altså især med $Formel_2$.

Det tredje problem ved $Formel_2$ skyldes en begrænsning som AltaVista har til fælles med alle de store internationale www-søgetjenester: Hvad de returnerer som søgeresultat er ikke antal-forekomster, men antal-dokumenter. Eller sagt på en anden måde: Kun den *første* forekomst i hvert www-dokument tæller i output. Dette skævtrækker naturligvis målingerne på de ord der ofte har mere end 1 forekomst i hvert dokument, dvs. især de stærkt frekvente ord, hvad der igen tjener til argument imod $Formel_2$.[4]

Summa summarum: der er skavanker ved såvel $Formel_1$ som $Formel_2$, men heldigvis ikke af samme slags. Snarere står svaghederne i komplementært forhold. Det kan man tage som argument for ikke at *vælge* mellem $Formel_1$ og $Formel_2$, men tværtimod gennemføre begge beregninger og lade resultaterne komplettere hinanden, sådan som det er gjort herover. At

man ad to så forskellige veje når til stort set samme resultater, giver grund til en vis optimisme. Det tyder på noget bedre end ren tilfældighed.

HVORFOR NETOP ALTAVISTA?

Man behøver ikke megen erfaring med de internationale søgemaskiner før AltaVista skiller sig ud som den eneste mulighed, hvis formålet er korpuslingvistik. AltaVista har, ved et forsigtigt skøn, indekseret ti gange så stor en del af det danske www-afsnit som søgetjenesten Infoseek, som af tekniske grunde er det eneste alternativ. Infoseek giver iøvrigt upålidelige søgeresultater – sammenlign for eksempel søgninger på *i*, *forskellige*, *det* og *af*.

SØGERESULTATER MED INFOSEEK

i	419 fund
forskellige	4.063 fund
det	5.327 fund
af	14.911 fund

Alle de øvrige offentligt tilgængelige søgemaskiner dækker lige så småt som Infoseek i det danske område, og hver af dem har i tilgift ét eller flere af disse problemer:

- Forvalg af dansk søgeområde ikke mulig (fx Excite, Hotbot)
- Søgeblokeringer på stærkt frekvente danske ord, såsom 'og' og 'jeg' (fx $\text{Yahoo}_{\text{Danish}}$, Jubii)
- Rapporterer ikke antallet af fund (fx Lycos, LookSmart)

Valget af søgemaskinen AltaVista kalder på et forbehold. Det er ikke www's tekstmasse som sådan man har mulighed for at undersøge, men de måske 20% som AltaVista, i følge almindeligt internationalt omdømme, har indekseret. Dette spolerer ikke www's værdi som *korpuslingvistisk* ressource, når blot der er mulighed for at undersøge den indekserede tekstmasses fordeling – sådan som jeg har argumenteret for i denne artikel. På den anden side er det ikke klart hvor meget den fremlagte metode har at sige om www's værdi som *sociologisk* ressource. Hvordan den samlede danske www-tekstmasse er distribueret, hvordan den ændrer sig over tid, og hvad dét betyder – det er forskningsopgaver som den nye IT-højskole i hovedstaden bør sætte højt på dagsordenen.

HVORFOR NETOP 20 MÅLEORD?

Når måleproceduren foreskriver 20 måleord, er der tale om et kompromis mellem praktiske og teoretiske hensyn. Søgningerne citeret i slutningen af artiklen tog mindre end to timer at gennemføre i Internettets myldretid (dvs. indenfor det amerikanske forretningslivs åbningstider), så måske er det ikke uoverkommeligt at udvide måleordssættet; men man skal dog være sikker på at kunne gennemføre hele søgningen inden for en enkelt dag – AltaVista indekserer 6 millioner nye Internet-sider i døgnet!

Set i retrospekt er der måske grund til at forlænge visse af måleordslisterne, hvis det er praktisk muligt. Dermed kunne man undgå den højreforskydning som ses i de seks objektkorporas placering på $Formel_1$-aksen i fig. 2. En foreløbig analyse viser at forskydningen skyldes en asymmetri i S_1- og M_1-måleordslisterne. Prøv at sammenligne det sidste S_1-måleord i tabel 1 med det sidste M_1-måleord i tabel 3. I S_1-måleordslisten er Eksklusion faldet til 0,46 omkring det 20. måleord, mens det tilsvarende tal i M_1-måleordslisten er hele 0,73; det giver uensartede værdier af $Kontrastfaktor_S$ og $Kontrastfaktor_M$, som det ses i tabellen herunder. Til sammenligning er S_2- og M_2-listerne mere symmetriske: I begge lister er Dominansen ved det 20. måleord faldet til ca. en tiendedel af det første måleords. Mine foreløbige undersøgelser viser at $Formel_1$ og $Formel_2$-beregningerne giver sammenlignelige resultater, hvis man blot – i det aktuelle eksempel – forlænger M_1-listen (eller forkorter S_1-listen), sådan at de to måleord der står sidst i de to lister har nogenlunde samme Eksklusion.

Der er dog også to andre, nok så attraktive, muligheder: Enten at vælge to støttekorpora hvis ordlister har omtrent samme frekvensfordeling (dét har *Fagblade* og *BySoc* ikke!). Eller også at give sig tilfreds med den standende procedure, for ret beset er resultaterne ikke dårlige som de står.

Sammenhængen mellem antal-måleord og Kontrastfaktor ses i tabellen:

TABEL 9 $Kontrastfaktor_S$ og $Kontrastfaktor_M$ som funktion af antal-måleord

ANTAL-MÅLEORD PR. LISTE	$KONTRASTFAKTOR_S$	$KONTRASTFAKTOR_M$
40	1,92	7,13
30	2,46	10,95
20	4,48	14,80
15	4,93	17,24
10	14,27	27,76

I *Fagblade* forekommer 'jo' formentlig altid som adverbium og 'nå' som verbum, mens 'jo' og 'nå' i *BySoc* har mange forekomster som interjektion. 'Kom' er formentlig især en præteritum i *Fagblade*, mens 'gå' er en infinitiv og 'løb' et substantiv; i *BySoc* ses 'kom', 'gå' og 'løb' også som imperativer, fx i konteksterne "kom så herover", "gå ned med dig", "løb fjorten gange rundt om huset".

Hvad sker der med målingerne, når der optræder dansk-danske homografer i måleordslisterne? I det værst tænkelige tilfælde er den ene læsning fremherskende i de to støttekorpora, mens den anden læsning dominerer i Objektkorpus. Lad os antage at en bestemt homograf *h*, med læsningerne *h1* og *h2*, er stærkt frekvent i S-korpus, men infrekvent i M-korpus (og dermed et kvalificeret S-måleord); lad os videre antage at *h* i begge støttekorpora har læsningen *h1* på alle forekomster, samt at *h* er stærkt frekvent i Objektkorpus, men her altid med læsningen *h2*. I denne situation vil Objektkorpus, tilfældigt, blive trukket mod S-polen og altså give støj i målingen.

Det er netop hvad der sker i analysen af korpus *BØRNETALE* (se tabel 8). I dette korpus er formen 'den' ekstremt frekvent (#5; 4,17%), men næsten altid i forbindelse med deiksis – dette 'den' danner en homograf med det determinative og anaforiske 'den', som det forekommer i *Fagblade*. Formen 'den' er langt hyppigere i *Fagblade* end i *BySoc* (#12 vs. #28), kort sagt: Situationen er tæt på den 'værst tænkelige'. Man kan naturligvis kompensere ved at gå ind og blokere for netop måleordet 'den': Derved flytter *BØRNETALE* position mellem S-polen og M-polen fra 97,4%/116,1% til 112,1%/138,2% (se fig. 6), hvad der synes en rimeligere placering. Men en sådan punktvis blokering er naturligvis en ad hoc løsning, og bedre er det at gardere sig med mere generelle midler, især ved at udvælge støttekorpora med henblik på helt at undgå den værst tænkelige situation – altså undgå forsøgsopstillinger som den nævnte, hvor Objektkorpus adskiller sig markant fra begge de to støttekorpora.

Iøvrigt er problemet med homografer ikke enestående for metoden her, det optræder, med forskellige ansigter, i alle undersøgelser af korpora uden grammatisk annotation. Man kan godt ærgre sig over at dansk ortografi ikke noterer de mest elementære forskelle i realiseringen af stød og tryk – det ville eliminere en god del af problemet med homografi. På dette punkt vildleder den danske skriftkode i en grad så at selv den trænede oplæser gør prosodiske fejl, når han realiserer en tekst prima vista (ikke mindst omkring verbaler med enhedstryk).

DANSK-UDENLANDSKE HOMOGRAFER

En mulig fejlkilde i forbindelse med www-undersøgelser er udenlandsk tekst. Både i .dk domænet og i de andre danske områder finder man en ikke uvæsentlig tekstmængde på engelsk, fransk og tysk – man kan derfor konstatere en spuriøs overfrekvens af ord som er sjældne i dansk, såsom 'did', 'das' og 'le'. Så længe www kun spiller rollen som Objektkorpus, vil disse overfrekvenser være usynlige – de pågældende homografer vil naturligvis ikke blive valgt som måleord, hvis de er infrekvente i de to støttekorpora. Men hvis man som støttekorpora installerer tekster hentet direkte fra www – hvad der i sig selv er en fornuftig idé – bør man holde øje med procentdelen af ikke-dansk tekst.

Det omvendte problem – spuriøs underfrekvens af ord som er hyppige i dansk – spiller næppe nogen rolle i praksis, for af tekstmassen i det danske www-afsnit indekseret af AltaVista fylder udenlandske tekster mindre end 10%, bedømt ud fra stikprøver. Endnu mindre udgøres af computerkode – højst nogle få promille, bedømt ud fra søgning på reserverede ord i HTML, Java, C, Unix og Perl.

HETEROGRAFER

De hyppige skriftformer 'kr.', 'ca.', 'kl.' og 'nr.' er problematiske som måleord, af to grunde.

For det første fordi BySoc notationen ikke tillader forkortelser (på nær 'hr.', akademiske titler og egennavne). Det er ærgerligt, fordi især leksemet 'cirka' er et velegnet måleord; det ville, hvis stavet ens i *BySoc* og *Fagblade*, stå som nummer syv i tabel 1 herover. Da *BySoc* hviler på en notationsformalisme uden forkortelser, som fx www-dokumenter ikke følger, kunne man nok, uden teoretisk blusel, harmonisere ortografien og erstatte 'cirka' med 'ca.' i *BySoc* ordlisten. Ændringen er næsten neutral i forhold til *Fagblade*, hvor formen 'ca.' er mere end 25 gange hyppigere end 'cirka'. Gevinsten er slående: Alene harmoniseringen af 'ca.' øger Kontrastfaktor_S fra 4,48 til 4,71.

Men her kommer *for det andet*: Netop forkortelser er udelukkede som måleord af en rent praktisk grund, når formålet med hele øvelsen er at lave målinger på www-tekstmassen. Som nævnt har man ikke mange frihedsgrader, når man søger med AltaVista og de andre søgemaskiner. Der søges grundlæggende på *ord*, og endda har man ikke indflydelse på orddefinitionen. AltaVista normaliserer den søgeprofil man indgiver.

- Lower-case erstattes med ubestemt case ("anna" matcher 'Anna', men ikke omvendt)
- Søgning på hyperhyppige former standses: "of", "la", "is"...
- Redundante blanktegn fjernes (" mifunes sidste sang " finder det samme som "mifunes sidste sang")
- Ikke-alfabetiske tegn inden i ord gøres ubestemte ("o!s#v{" matcher 'o.s.v.', men ikke 'osv')
- Efterstillede ikke-alfabetiske tegn ignoreres ("kr!"/"kr."/"kr" matcher såvel 'kr' som 'kr.')

I vores sammenhæng udgør det sidste punkt et problem. Søgestrengen "ca." matcher en mængde uønskede forekomster som 'Ca'Luna', 'KMD-CA' og – værst – 'CA'. I et akronymtilplastret korpus som www er det altså på forhånd håbløst at søge efter konventionelle forkortelser.

FRA EKSEMPEL TIL VIDENSKAB

I denne artikel har vi, af hensyn til reproducerbarheden, ladet de to støttekorpora udgøre af danske standardkilder af ældre dato. Ingen af dem rummer tekst som er hentet fra www selv, og derfor kan det diskuteres om de er egnede som skalære modpoler ved www-undersøgelser. Bedre er det naturligvis at installere korpora bestående af tekst i særligt udvalgte www-genrer. Dermed får man resultater som kan tolkes mere entydigt (men man vil i tilsvarende grad være nødt til at træffe teoriladede tekstvalg).

Download fx en stor portion chat (ikke under 100.000 ord). Download en tilsvarende portion kursusbeskrivelser fra de højere læreanstalter. Generér de to frekvensordlister, og beregn S- og M-måleord som beskrevet herover. Med de forsøgsbetingelser kunne nås en endegyldig vurdering af det danske www-afsnits placering mellem skæg og snot.

Det Perl-script der er brugt til beregningerne, kan frit hentes på www.cphling.dk/~pjuel

SØGERESULTATER (ALTAVISTA, SØGEOMRÅDE 'DANISH', MAJ 1999)

SØGEORD	ANTAL FUND	SØGEORD	ANTAL FUND
af	2.480.296	*kom*	81.880
altså	24.319	*man*	328.630
at	*(blokeret)*	*med*	1.771.900
bare	58.400	*mellem*	87.023
da	278.340	*men*	391.980
dansk	264.323	*min*	173.520
den	1.400.512	*mod*	127.320
denne	411.330	*nej*	14.808
der	*(blokeret)*	*noget*	67.430
derfor	77.441	*nye*	260.300
det	1479.392	*nå*	13.620
dette	373.540	*og*	2.338.448
disse	170.810	*også*	337.780
du	772.900	*om*	1.260.626
efter	337.410	*på*	1.809.697
en	*(blokeret)*	*sagde*	11.547
er	2.417.540	*samt*	236.860
et	1.147.840	*sig*	332.110
flere	158.640	*sin*	131.790
for	*(blokeret)*	*som*	1.143.170
fordi	37.556	*store*	142.560
forskellige	69.724	*større*	45.911
fra	1.032.860	*synes*	19.575
gennem	56.470	*så*	452.440
ham	28.650	*sådan*	40.310
han	130.390	*således*	48.925
har	1.220.490	*til*	1.749.136
hun	47.340	*tror*	14.827
hvad	268.580	*uden*	76.636
i	*(blokeret)*	*under*	241.850
ikke	669.480	*var*	260.360
ja	56.960	*ved*	626.800
jeg	341.810	*vi*	646.720
jo	44.720	*vil*	461.530

NOTER

1. Skønnet ud fra stikprøver ligger 80-90% af den danske www-tekst i domænet .dk, mens resten er fordelt over .com, .org og andre. Den førende internationale søgetjeneste, AltaVista, dækker en dansk tekstmasse på mindst 140 mio. ord, sandsynligvis mere (maj 1999, beregning følger). AltaVistas internationale dækning passerede 100 mio. www-steder i 1998; ingen indeksinformation er mere end 1 måned gammel (data fra AltaVistas hjemmeside). DDO's tekstarkiv (se http://coco.ihi.ku.dk/~ddo/ddokorpd.htm) har ca. 40 mio. ord.

2. Med 'leksikalske ord' menes ord transskriberet hen til en RO96-normal form, dvs. fraregnet stammen, selvafbrydelse, o.lign.; leksikalske ord fanges med søgeprofilen " +[\'\.]?\s" (nærmere detaljer om søgesproget regular expressions: se ref.); de 1.277.822 fund er distribueret over 32.215 former. Søgningen er foretaget i transskriptionsversion a (*BySoc*'s hovedkorpus). Som søgealgoritme er brugt: *SØG-HURTIGT*, som udskrifttype er valgt: Liste:Rangorden

3. *BØRNETALE* og *VOKSENTALE* er uddrag af korpus *CHILDES*, et stort, internationalt transskriptionskorpus af overvejende voksen-børnesamtaler (se ref.); *BØRNETALE* er et ekstrakt af den danske del, nemlig samtlige ord udtalt af børn (alder 1-3 år), ialt 43.101 registrerede ord (1.035 former). *VOKSENTALE* er det tilsvarende ekstrakt af de voksne speakere (152.303 ord, 4.019 former).

4. Man kan let vise at skævtrækningen ikke influerer på indplaceringen af Objektkorpus mellem de to støttekorpora, hvis blot den er nogenlunde ligeligt fordelt mellem S-måleord og M-måleord. Derimod er skævtrækningen synlig i estimatet Ordmasse, som jo benytter *summen* af S- og M-måleordenes dækning. Beregnet ud fra $Formel_2$ bliver Ordmasse på 124 mio. ord.

LITTERATUR

Henrichsen, P. J. (1997): Talesprog med Ansigtsløftning; Kbh. Univ.: Instrumentalis 10

Henrichsen, P. J. (1998): Talesprog med Netstrømper; Kbh. Univ.: Instrumentalis 12

MacWhinney, B. (1995): The CHILDES Project: Tools for Analyzing Talk; Hillsdale, N.J.: L. Erlbaum; *se også* http://childes.psy.cmu.edu/

Maegaard, B. et al. (1981): Hyppige Ord i Danske Børnebøger; Gyldendal

Maegaard, B. et al. (1986): Hyppige Ord i Danske Aviser, Ugeblade og Fagblade; 2 bd.; Gyldendal

Nørretranders, T. (1997): Stedet Som Ikke Er – Fremtidens Nærvær, Netværk og Internet; Aschehoug.