DANwORD

Hyppighedsundersøgelser i moderne dansk

Bente Maegaard og Hanne Ruus

Formålet med projektet DANwORD er at undersøge et repræsentativt udsnit (på 1.25 million løbende ord) af moderne dansk med automatiske metoder, især med henblik på frekvens af ord. Udover de beregnede frekvenser har projektet interesse ved udviklingen af de automatiske metoder til bl.a. morfologisk analyse og entydiggørelse af homografer og ved anvendelse af databaseteknik til lagring af de fundne resultater.

Vi har arbejdet med projektet i ca. 1 1/2 år, og har i den forløbne periode indsamlet og registreret 250.000 løbende ord, dvs. en femtedel af vort materiale. Før vi kunne påbegynde denne indsamling, måtte vi bestemme, hvilke tekster der skulle indgå i vort materiale eller corpus. Vi skal her først gennemgå de overvejelser, man må gøre sig, når man skal sammensætte et corpus og de konklusioner vi er nået til, og derefter omtale, hvorledes vi vil behandle teksterne og hvilke oplysninger om teksterne, vi ønsker at finde og lagre.

I. Tekstgrundlag.

Vi har ved udvælgelsen af tekstmateriale anlagt et forbrugskriterium, således at vi har forsøgt at finde frem til det mest brugte sprog. Det mest brugte sprog består af de mest læste tekster og de mest hørte tekster; men af praktiske grunde har vi, ligesom alle andre, der har lavet større frekvensundersøgelser, måttet begrænse os til de mest læste tekster, altså til skriftsprog, selv om forbrugskriteriet peger både på ikke offentliggjorte skrevne tekster (breve, huskesedler, mv.) og på det talte sprog (der også findes i både offentliggjort (radio og fjernsyn) og ikke offentliggjort form).

DANwORD - Hyppighedsundersøgelser i moderne dansk
Bente Maegaard, Hanne Ruus
Proceedings of NODALIDA 1977, pages 65-74

Inden for det offentliggjorte skriftsprog, som altså er vores grundlagsmateriale, har vi som nævnt forsøgt at finde frem til det i nutiden mest anvendte, dvs. mest trykte og dermed mest læste, sprog. Ved nutiden har vi forstået 5-årsperioden 1970-1974, hvad vi senere skal vende tilbage til. Dette udvælgelsesprincip gør, at vore resultater med rette vil kunne opfattes som gældende for moderne dansk skriftsprog, idet vi dog igen skal understrege, at det sprog, vi undersøger, er det sprog, der læstes i begyndelsen af 1970'erne, ikke det sprog, der blev skrevet.

Det er væsentligt at være opmærksom på, at materialevalget har stor betydning for resultaternes generaliseringsværdi; det gælder nemlig om mange frekvensordbøger, at deres resultater opfattes som mere generelle end de egentlig er, idet kriterierne for udvælgelse af materiale ikke altid fremgår klart af titelbladet (en gennemgang af eksisterende frekvensordbøger fremkommer i Danske Studier 1978).

2. Valg af corpus.

Ved frekvensundersøgelser af andre sprog har man anvendt to hovedtyper af corpora: den ene type består udelukkende af avisprog, mens den anden er søgt sammensat således, at corpus afspejler det trykte sprogs sammensætning.

I begge tilfælde ønsker man, at resultaterne af undersøgelser i corpus skal gælde for det mest "almindelige" sprog inden for det undersøgte nationalsprog; men det er faktisk et spørgsmål, om der findes sådan noget som "det mest almindelige sprog", dvs. om man ud fra en undersøgelse af en blanding af tekster kan konkludere noget om sproget som helhed.

Ordhyppigheder i en bestemt tekst er nemlig for de fleste forskellige ords vedkommende specifikke for denne tekst og dens indhold, idet de afhænger af, hvad teksten handler om (jvf. Henning Spang-Hanssen, 1960). Hvis man tager en anden tekst, vil det være andre ord, der er hyppige. De eneste ord, hvis hyppighed er tekstuafhængig, er nogle få (et par hundrede) meget hyppige ord.

Nu kan man ved ret begrænsede tællinger bestemme disse tekstuafhængige ord; men hvilke af sprogets øvrige ord man ellers får med, afhænger af valget af tekster og af, hvor stort et materiale man har. Hvis materialet er stort nok, og hvis man vælger tekstprøverne små nok og rimeligt varieret, er det muligt inden for en nærmere afgrænset, homogen, tekstart også at få gyldige resultater for andre end de meget hyppige ord. De enkelte tekstprøver bør være korte, fordi man derved mindsker risikoen for, at enkelte tekstspecifikke ord får en urimelig overvægt: tekstproverne kommer til at handle om forskellige ting.

Problemet bliver herefter at afgrænse en tekstart således, at den bliver homogen, dvs. således at den består af tekster af nogenlunde ensartet karakter. Det er ikke muligt på forhånd at angive, hvad der er en rimelig afgrænsning af en tekstart; man må skønsmæssigt opstille kriterier

for afgrænsning af hver enkelt tekstart, og kan først, når den faktiske sammensætning af tekstarten foreligger, undersøge, om den er homogen.

En homogenitetsprøve vil bestå i tilfældigt at udtage delmængder af hele tekstarten og undersøge, om tællinger i disse stemmer overens med tællinger i hele tekstarten.

3. Størrelse af tekstarter og tekstprøver.

Vi er ved fastlæggelsen af størrelsen af de enkelte tekstarter gået ud fra den alment bekræftede antagelse (Zipf's lov, omtales f. eks. i Charles Muller, 1968) at en tekstmasse på 1/4 million løbende ord indeholder ca. 5000 ord med en hyppighed på 5 eller mere. Da denne størrelse af en tekstart synes rimelig også med henblik på en sammenligning med andre undersøgelser, har vi besluttet at lade hver tekstart omfatte 250.000 løbende ord. For de enkelte prøver, der som allerede nævnt bør være forholdsvis korte, har vi valgt størrelsen 250 ord, dvs. ca. 1 side. Hver tekstart består altså af 1000 prøver. Prøverne er udvalgt i teksterne ved anvendelse af tilfældige tal. De tilfældige tal benyttes til at beregne den side og linie, hvor prøven starter. En prøve begynder altså altid ved det første hele ord på en linie, men af hensyn til entydiggørelse af homografer er ved indkodning af prøven også medtaget "forkonteksten", der strækker sig indtil den nærmeste periodegrænse før prøvens start, og "bagkonteksten", der strækker sig fra prøvens slutning til nærmeste periodegrænse.

En tekstprøve. Forkonteksten slutter ved //.

```
<< Anders Bodelsen
Tænk på et tal
Gyldendal 1970, s.47 l.26 >>
<<L18>>
                                                Midt i en
//mørk skov standsede han og stod ud af vognen.
  Det var bidende koldt. Han slukkede lygterne, øjnene vænnede sig til mørket og han hørte ikke nogen anden lyd end den kolde susen i grantræerne. Han gik et par skridt ind i skoven og borede med skosnuden i jorden, der allerede var fast af frost.
  En bil nærmede sig. Det lød som om den ville standse,
<< s.48 >>
men den satte atter farten op og lysvifften forsvandt bagude mellem træerne. Jorden var for hård, stedet for befærdet. Desuden havde han fået et nyt indfald.
  Han gik tilbage til bilen igen og kørte hjemad. Ideen han netop havde fået, forbandt han med de dybe, næsten berusende indåndinger af frostluft. Den var en inspiration. Han parkerede noget fra butikstorvet, gik roligt det sidste stykke med mappen i hånden, låste sig fra trappegangen ind i banklokalet og fandt lommelygten i bunden af skabet med de elektriske propper.
  Nøglerne til de ubenyttede boxe lå i Miriams skuffer. Han skrev en box-kvittering ud til F. Hjulmand, fandt frem til den tomme box via nummeret på nøglen, et hundrede og niogtyve, og låste ved hjælp af bankens universalnøgle til alle boxene og specialnøglen til et hundrede og niogtyve den blå madkasse ind i det snævre rum. Kartoteket med boxkvitteringerne var låst, denne del af manøvren måtte han udsætte. Han lagde lommelygten på plads i bunden af elektricitetsskabet, låste sig ud og sad lidt efter i sin vogn. Ingen var fulgt efter ham, ingen fulgte ham, da han startede vognen. Men han mødte vagtselskabets mand på hans cykel, idet han passerede butikstorvets udmunding.
```

4. Valg af tekstarter.

Når man, som vi, udgår fra et forbrugskriterium, vil man være interesseret i, at udvalget af tekstarter tilsammen skal dække størstedelen af enkelttekster på det undersøgte sprog, her dansk. Flere forskellige undersøgelser (se f.eks. Socialforskningsinstituttets fritidsundersøgelser og Hans Hertel 1972) af voksne danskeres læsevaner viser, at 90 pct. dagligt læser avis, 75 pct. læser ugeblade og 50 pct. læser fiktionsprosa; på bibliotekerne låner børn 2-3 gange så meget som voksne og er således store forbrugere.

Hvis man kun vil undersøge én tekstart, er det altså klart, at den må bestå af avistekster; men da der også er andre tekstarter, der bruges af en ikke ubetydelig del af befolkningen, og da det vil være interessant at sammenligne forskellige tekstarter, har vi ikke ment at kunne nøjes med avisteksterne,og har valgt følgende 5 tekstarter: fiktionsprosa for voksne, fiktionsprosa for børn, aviser, ugeblade, faglitteratur.

Den første tekstart, vi behandler, er fiktionsprosa for voksne, og for denne er vi færdige med udtagelse og indkodning af prøverne, idet der dog mangler sidste korrekturlæsning på en del.

5. Udvælgelse af tekstprøver.

Også ved udvælgelse af de enkelte tekstprøver i en tekst anlægger vi et forbrugssynspunkt, dvs. at vi for fiktionsprosaen har forsøgt at finde frem til de mest læste danske forfattere. Det er imidlertid ikke så nemt, som man skulle tro, idet 1) oplagstal ikke kan bruges, fordi visse forlag ikke vil opgive det, 2) folkebibliotekerne ikke fører udlånsstatistik på titler, og 3) der ikke findes nyere undersøgelser af læsevaner, der når ned til enkelttitler. Vi har derfor valgt at bygge på udgivelsesfrekvens og har fundet frem til de mest udgivne forfattere i perioden 1970-1974 i Dansk Bogfortegnelse; vi har medtaget alle forfattere, der i denne periode har fået udgivet mindst 5 bøger, med én prøve fra hver af udgivelserne. Dette gav 976 tekstprøver, og de manglende 24 er derefter fundet ved forholdsmæssig fordeling på de allerede repræsenterede forfattere (en nærmere redegørelse for vores udvælgelsesprocedure fremkommer i Danske Studier 1978).

Grunden til, at den periode, vi undersøger, er forholdsvis lang og f.eks. længere end ét år, er, at udgivelsestallene for de enkelte forfattere viser ret store udsving fra år til år, men stabiliserer sig, når man betragter en længere periode.

Nedenfor angives de 20 mest udgivne forfattere i perioden 1970-1974, og antallet af tekstprøver for hver af dem.

nr.		antal
1	Cavling, Ib Henrik	58
2	Koch, Morten	43
3	Nielsen, Lars	38
4	Forsberg, Bodil	33
5	Panduro, Leif	28
6	Rifbjerg, Klaus	28
7	Hazel, Sven	27
8	Bodelsen, Anders	26
9	Nohr, Else-Marie	23
10	Edson, Ed	22
11	Andersen, H.C.	20
12	Søeborg, Finn	20
13	Kampmann, Christian	19
14	Poulsen, Erling	19
15	Ørum, Poul	19
16	Blicher, Steen Steensen	17
17	Hansen, Martin A.	16
18	Bang, Herman	15
19	Johansen, Orla	15
20	Scherfig, Hans	15

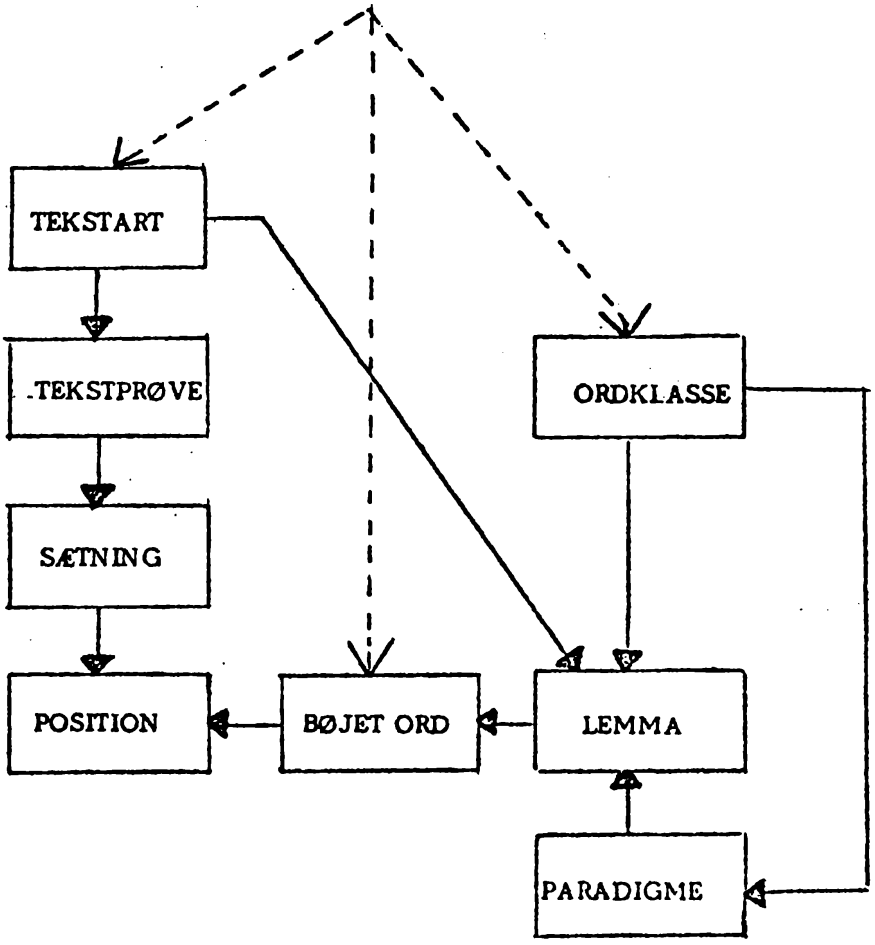
II. Den datamatiske behandling af teksterne.

Hovedformålet med projektet DANwORD er at lave hyppighedsundersøgelser. Derfor vil de primære resultater af vores bearbejdning af materialet være opgørelser over frekvenser af grafemer, lemmaer, ordklasser, paradigmer osv. Vor undersøgelsesmetode og den videre behandling af de analyserede tekstprøver vil imidlertid også bringe metodisk interessante resultater på andre områder: På det lingvistiske område giver arbejdet med at automatisere den sproglige analyse ny indsigt i form af mere strukturerede og udførlige grammatiske beskrivelser. På det datamatiske område vil lagringen af analyseresultaterne i en database bringe viden om samspillet mellem lagring af en større datamængde med en kompleks struktur og den effektive udnyttelse af de informationer, strukturen indeholder.

Første del af en automatisk sproglig analyse til brug for frekvensundersøgelser er lemmatiseringen. Denne kræver, at man opstiller en fuldstændig fleksionsgrammatik for dansk. Til brug for den morfologiske analyse disponerer vi over en udtømmende og sammenhængende grammatik over den regelmæssige bøjning i moderne dansk (se Hanne Ruus, 1977). Udover den morfologiske analyse omfatter den automatiske lemmatisering entydiggørelsesprocedurer for homografemerne. Ved entydiggørelsen af homografer fra lemmaer af forskellig ordklasse vil man kunne komme ganske langt automatisk, idet man kan bygge på ordklassernes forskellige syntaktiske funktion, ligesom homografer inden for samme lemma i stor

udstrækning vil kunne entydiggøres ved at søge efter kongruerende former i den nærmeste kontekst og ved også her at bygge på formernes forskellige syntaktiske funktion.

Da vi vil opgøre frekvenser på forskellige lingvistiske niveauer, har vi brug for at have de analyserede tekstord med tilhørende oplysninger om forekomststeder, lemma osv. kædet sammen på forskellige måder. For at opnå dette vil vi lagre alle de fundne oplysninger om tekstordene i en database, således at det er muligt at uddrage ord eller sætninger med bestemte egneskaber. En database er en lagringsmetode, hvor man ved lagringen angiver, hvilke oplysninger der skal være kædet sammen, og derved strukturerer sine data. Vi vil f.eks. sørge for, at tekstordene er kædet sammen med deres lemma og deres ordklasse, herved bliver det nemt at opgøre frekvenser f.eks. for alle former, der hører til samme lemma og for alle ord fra samme ordklasse. I en database er det lettest at uddrage de oplysninger, der allerede er kædet sammen, men man kan selvfølgelig også finde frem til de delmængder af materialet, der ikke er nedlagt direkte i strukturen. Hvis man ønsker at kunne udtage mange forskelligartede delmængder, vil databasen imidlertid let få en ret indviklet struktur, så både førstegangslagring, lagerforbrug og søgninger vil blive tidskrævende og dermed dyre. Det drejer sig derfor om at begrænse databasen til den struktur, der er nødvendig for de hyppigste opslag. Vi forestiller os følgende databasestruktur:



De stiplede linjer angiver indgange til databasen, altså de kasser, man har direkte adgang til. Alle de andre kasser kan man finde frem til ved at benytte strukturen, dvs. følge pilene.

Kasserne i figuren kan opfattes som de steder, hvor man lagrer forekomsten af lemmaer, tekstord (bøjet ord) m.v. Lagringen er foretaget således, at et lemma f.eks. pen som angivet ved pilen fra LEMMA til BØJET ORD er forbundet med de forskellige, faktisk forekommende bøjningsformer af pen, altså f.eks. pen, penne, pennen, pennenes. Hver af de disse bøjningsformer peger via POSITION på de steder i teksterne, hvor ordformen er fundet. Et bøjet ords kontekst finder man i SÆTNING, hvor teksternes sætninger er lagret hver for sig. Pilen fra TEKSTPRØVE til SÆTNING angiver, at den enkelte tekstprøve peger på de sætninger, den indeholder.

2. Analyse og lagring.

Inden tekstprøverne kan lagres i databasen, skal de analyseres: bibliografiske oplysninger og for- og bagkontekst skal skilles fra, og selve prøven deles op i ord.

Når ordene fra en tekstprøve skal lagres i databasen foretages for hvert ord følgende:

I. Ordet slås op i basen. Findes ordet i basen, lagres en forekomst af det. Hvis det er en homograf, *) entydiggøres det først.

II. Hvis ordet ikke findes i basen endnu, skal det lemmatiseres. De allerhyppigste småord, hvoraf mange er ubøjelige, klares ved opslag i en liste på et par hundrede ord. I en forkortelsesliste eftersøges ord før punktum. De øvrige ord analyseres morfologisk, hvorved potentielle fleksiver fjernes og der opstilles et antal mulige lemmaer. De foreslåede lemmaer slås automatisk op i en let bearbejdet udgave af Nudansk Ordbog. Hvis flere af de foreslåede lemmaer findes i ordbogen, er det analyserede ord en homograf og sendes til entydiggørelse. Til sidst lagres tekstordet og analyseresultaterne i basen. De ord, der ikke kan findes i ordbogen, og de homografer, der ikke kan entydiggøres automatisk, skrives ud til håndbehandling.

Efterhånden som vi får lagret tekstprøverne i basen, vil stadig flere ord blive klaret under I, således at lemmatisering og ordbogsopslag begrænses mest muligt. En anden fordel ved den beskrevne fremgangsmåde er, at vi kan udtage frekvensoplysninger på forskellige niveauer på dele af materialet, inden vi har indsamlet prøver fra alle tekstarterne. Dele af den analyserede datamængde (f.eks. tekster af en forfatter) vil også kunne anvendes ved andre former for sprogvidenskabelige undersøgelser.

3. Fremtidig anvendelse af materialet.

Når vi har afsluttet frekvensundersøgelserne på de 1.25 million ord, vil vi bevare databasen som hjælpemiddel for sprog- og stilforskere. Ved at bruge 5000 sider gennemanalyseret tekst som basis for arbejdet med sprogbeskrivelse og tekstkarakteristik vil det være muligt kvantitativt at skelne mellem mere og mindre væsentlige regler i grammatikken, ligesom et gennemsnit af materialet for stilforskeren vil kunne være en tilnærmelse til den norm, der danner grundlaget for enhver tekstkarakteristik.

---

*) Man siger, at to ord er homografer, hvis de har samme grafiske form - staves ens -, men enten tilhører forskellige lemmaer (<u>helt</u> substantiv, <u>helt</u> adjektiv) eller er forskellige bøjningsformer af samme lemma (<u>kunne</u> infinitiv, <u>kunne</u> præteritum). At entydiggøre en homograf betyder at bestemme hvilket lemma, den tilhører, og hvilken bøjningsform, den er, i den aktuelle kontekst.

Nedenfor anføres de 75 hyppigste graford i vore skønlitterære prøver, og til sammenligning de 75 hyppigste graford fra Noesgaards undersøgelser af skønlitteratur (Aksel Noesgaard: Hyppigheds Undersøgelser II, 1937).

DAN-ORD				NOESGAARD			
1	OG	43	OVER	1	OG	43	MAN
2	DET	44	NOGET	2	I	44	SIN
3	I	45	HVAD	3	DET	45	IND
4	AT	46	KUNNE	4	HAN	46	HANS
5	HAN	47	MAN	5	AT	47	HENDE
6	VAR	48	SIN	6	VAR	48	NAAR
7	JEG	49	EFTER	7	EN	49	HVOR
8	EN	50	IND	8	DEN	50	MIG
9	PÅ	51	BLEV	9	TIL	51	SKULDE
10	IKKE	52	SKAL	10	PAA	52	DEM
11	TIL	53	VIL	11	SAA	53	EFTER
12	ER	54	KOM	12	DER	54	KAN
13	DE	55	VÆRE	13	MED	55	NOGET
14	MED	56	JO	14	DE	56	NED
15	HUN	57	GIK	15	AF	57	STOD
16	DEN	58	SELV	16	IKKE	58	JO
17	DER	59	HVOR	17	FOR	59	HVAD
18	SÅ	60	JA	18	SOM	60	OGSAA
19	AF	61	VILLE	19	HUN	61	VILDE
20	FOR	62	DIG	20	SIG	62	ALLE
21	SIG	63	HENDES	21	JEG	63	VÆRE
22	MEN	64	OGSÅ	22	HAVDE	64	LILLE
23	HAVDE	65	ELLER	23	ER	65	SELV
24	SOM	66	NED	24	MEN	66	SKAL
25	ET	67	HER	25	ET	67	HER
26	OM	68	MEGET	26	OM	68	HEN
27	DU	69	NÅR	27	HAM	69	JA
28	HAR	70	MIN	28	VED	70	HENDES
29	HAM	71	HAVE	29	OVER	71	HAVE
30	SAGDE	72	MÅ	30	DA	72	GAMLE
31	MIG	73	IGEN	31	FRA	73	FIK
32	NU	74	SKULLE	32	NU	74	MOD
33	VED	75	LIDT	33	SAGDE	75	ELLER
34	VI			34	UD		
35	UD			35	HAR		
36	KAN			36	KUNDE		
37	DA			37	OP		
38	OP			38	DU		
39	FRA			39	VI		
40	HANS			40	BLEV		
41	DEM			41	KOM		
42	HENDE			42	GIK		

Liste over værker, der er henvist til i foredraget:

Hans Hertel: Det litterære system i Danmark (i Robert Escarpit: Bogen og læseren, Reitzel 1972).

Bente Maegaard og Hanne Ruus: DANwORD, Baggrund og materiale (i Danske Studier 1978 (i Trykken)).

Charles Muller: Initiation à la statistique linguistique, Dunod 1968.

Hanne Ruus: Ordmekanik (i SAML III, 1977, s. R1 - R28).

Socialforskningsinstituttets fritidsundersøgelser:
P.-H. Kühl, Inger Koch-Nielsen, Kaj Westergaard: Fritidsvaner i Danmark, 1966.
P.-H. Kühl og Inger Koch-Nielsen: Fritid 1975, 1976.

Henning Spang-Hanssen: Aksel Noesgaards ordstatistiske pionerarbejde (i Danske Studier 1960, s. 81-90).