

## FORESTILLINGEN OM "DANSK NORMALNUL" I FORBINDELSE MED ORDFORRÅD

Af Henrik Holmboe (Afdeling for Datalogvistik, HHÅ)

Dansk normalnul er det punkt, hvorudfra man måler, hvad der er højt og lavt i et landskab, hvad der er højvande og hvad der er lavvande ved vore kyster. Også uden dette normalnul som udgangspunkt kunne vi naturligvis have et indtryk af højvande og lavvande, af ebbe og flod, men kun et intuitivt, som ville overlade det til en individuel bedømmelse, om et aktuelt højvande er mere eller mindre højt end det, jeg bevidnede for et år siden, eller det, som samtidig iagttages 100 eller 200 km. væk. Et fastlagt nulpunkt er altså en nødvendig forudsætning for, at vi kan gøre rede for afvigelser fra - og størrelsen af afvigelser fra - det normale.

Tekstcorpora er typisk samlinger af tekster med indhold, der adskiller sig mere eller mindre fra hinanden: - mere, hvis der er tale om corpora, der er sammensat af tekster af forskellig genre eller hidrørende fra helt forskellige typer af publikationer; - mindre, hvis der er tale om corpora, der er sammensat af tekster inden for samme overordnede emne, men repræsenterende forskellige aspekter af dette emne. Hvad enten det ene eller det andet er tilfældet, vil man som læser forholdsvis let kunne gennemskue, at forskellige tekster handler om noget forskelligt, også uden en detaljeret forståelse af de enkelte tekster. Dette beror bl.a. på teksternes ordforråd, eller rettere på dele af ordforrådet, som af en eller anden grund er iøjnefaldende og karakteristisk for teksten. Det er fra netop denne del af ordforrådet man udtager keywords med henblik på indeksering og genfindning af teksterne. Foruden denne del af ordforrådet, den karakteristiske del, den del, der får teksten til at handle om det, den handler om, er der også en anden del af ordforrådet, som omvendt ikke får teksten til at handle om noget. Dette ordforråd vil jeg betegne som et normalnul ud fra en formodning om, at dette nulniveau er ganske generelt, og ud fra en idé om, at dette normalnul er et nødvendigt concept, hvis man vil udtale sig om teksters forskellighed på andet end en intuitiv måde. Normalnul vil utvivlsomt minde om hinanden fra sprog til sprog,

men en nøjere undersøgelse af ligheder og forskelle i så henseende, mangler vi endnu at foretage.

Paradigme-syntagmerelationen må være til stede på simpel og entydig måde, hvis man tager en tekst og over denne tekst laver en ordbog. Lad tekstens løbende ord være x-aksen; denne vil da være lige så lang som teksten, dvs. have lige så mange pladser, som der er løbende ord. Lad y-aksen være tekstens forskellige ord, som forlænger y-aksen efterhånden som de dukker op. x-aksen vil være forløbet, y-aksen systemet. I x-y-coordinatsystemet kan opsamles forskellig information om frekvens og rækkefølge, og ved en eller flere sorteringer af ordene på y-aksen vil man kunne producere et antal ordbøger over teksten på x-aksen. Ved hjælp af en y-akse-ordbog og en grammatik eller syntaks med angivelse af f.eks. frekvensmæssige rækkefølgerelationer må det være muligt at generere teksten på x-aksen, dvs. ud fra paradigmet og et regelsæt at komme frem til syntagmet.

Hvis man har en ordbog, der ikke har en så tæt relation til en bestemt tekst, vil der stadig i en vag forstand være en paradigme-syntagmerelation til stede; men meget vagere af en række grunde: der er ingen frekvensoplysninger med direkte relation til én eller flere bestemte tekster; der er ingen sikkerhed for, at alle ordbogens ord skal være brugt for at teksten kan påstås at være færdig eller korrekt; der vil på den anden side være mulighed for at producere ikke blot én men et antal korrekte tekster på grundlag af ordbogen. Et muligt postulat er, at én og kun én grammatik er tilstrækkelig til at generere de forskellige korrekte "x-akse-tekster" på grundlag af noget, som ikke er en rigtig y-akse-ordbog; altså et ét-til-mange forhold mellem system og forløb.

Hvorledes nu i de situationer, hvor man etablerer en ordbog som en y-akse-ordbog i snæver forstand og så hævder, at den har gyldighed for mere end én x-akse? Dette må vel i det mindste stille skærpede betingelser til grammatikken. Endvidere: vil man kunne tage en del af en række således etablerede y-akse-ordbøger og hævde, at denne del er til stede i alle f.eks. danske tekster? Kan det indgå i en grammatik for dansk, at den rummer forskrifter for rækkefølgen af ordenes frekvensklasser? Vil dansk normalnul kunne være et sådant lille ordforråd, der ikke selv kan producere nogen tekst alene af den grund, at de højere rangklasser/lavere frekvensklasser ikke er til stede?

Når en tekst handler om noget, ligger denne omstændighed som anført ikke i hele tekstens ordforråd. Kan vi isolere den del af en teksts ordforråd, der får teksten til at handle om sit emne? Og kan vi evt. gøre det ved fra tekstens samlede ordforråd at subtrahere det ordforråd, den del af ordforrådet, der *ikke* handler om noget, eller bedre udtrykt, der ikke får teksten til at handle om noget? Man har forsøgt at udvikle algoritmer, der kunne ekstrahere keywords fra en tekst. Disse algoritmer udgår bl.a. fra den tankegang, at keywords, der er specifikke for en tekst, må have nogle egenskaber hvad angår frekvens og dispersion, som gør det muligt at fange dem uden en egentlig indholdsmæssig analyse og forståelse af teksten. I overensstemmelse hermed må vi som generelt udgangspunkt antage, at den omtalte ordmængde, der ikke får sin tekst til at handle om noget, ikke rummer kandidater til keywords.

En tekst har et ordforråd. Ligeledes en forfatter, en journalist eller for den sags skyld enhver af os. Et sprog har et ordforråd. Ligeledes en genre, et fag og en profession. Og hermed mener man også, at et bestemt ordforråd er det forråd, hvorudaf pågældende person eller profession bygger sine tekster. Ordforrådet leverer så at sige teksten. Når vi taler om ordforråd i forbindelse med normalnul, bruger vi ordet ordforråd lidt anderledes. Vi har her et ordforråd, om hvilket man ikke på samme måde kan sige, at en tekst eller en person har det, eller at det leverer bestemte tekster. Det specielle ved dette ordforråd - dette sprogets vandbærere - vil være, at formentlig alt og alle har det som en delmængde, at det er medleverandør ved alle tekster, at ingen kan klare sig uden det, men ej heller kan nogen nøjes med det.

Hvis man vil finde frem til dette ordforråd, kunne man gøre det ved at finde fællesmængden af de ordforråd, som en række tekster består af. Derved vil man ganske vist finde frem til en ordmængde, men det forekommer mig, at der ville mangle en række oplysninger om disse ord, f.eks. frekvensoplysninger, som det kunne være interessant at få med. Endnu et moment: uddrager man ordforrådet af en tekst, har dette ordforråd naturligvis altid afhængigt af tekstens omfang en ganske bestemt størrelse, og dets enkelte elementer kan beskrives absolut og i relation til hele mængden. Tager man blot fællesmængden af ord fra en række tekster, kommer man bl. a. til at mangle oplysninger om den vægt, hvormed de

enkelte elementer indgår i helheden. Jeg har derfor valgt en anden fremgangsmåde, der skulle sikre, at jeg får også disse oplysninger med mig:

Til undersøgelsen har jeg valgt 5 tekster på hver ca. 10.000 løbende ord fra teksttyperne (genrerne) litteratur, avis, samfundsrag, jura og sagprosa. Dette valg er foretaget ud fra en lidt løs fornemmelse af, at jeg hermed havde tekster med rimeligt forskelligt ordforråd, og også ud fra det praktiske hensyn, at det var tekster, som jeg havde i elektronisk manipulerbar form i forvejen. Jeg har derefter taget to af teksterne, målt deres omfang i løbende ord og set på fællesmængden i deres ordforråd. Ordene i fællesmængden optræder med en absolutfrekvens, der er summen af forekomsterne i den ene tekst og i den anden tekst. Disse forekomsttal har jeg normaliseret på følgende måde: lad os kalde de to tekster A og B; det normaliserede forekomsttal = (absolutfrekvensen af de enkelte ord i A) \* (antallet af løbende ord i B) / (antallet af løbende ord i A + antallet af løbende ord i B) adderet til (absolutfrekvensen af de enkelte ord i B) \* (antallet af løbende ord i A) / (antallet af løbende ord i A + antallet af løbende ord i B). Altså:

$$\begin{aligned} \text{Norm.forek.} = & \frac{(\text{abs.frq. af de enkelte ord i A}) * (\text{antal løb.ord i B})}{(\text{antal løb.ord i A}) + (\text{antal løb.ord i B})} \\ & + \frac{(\text{abs.frq. af de enkelte ord i B}) * (\text{antal løb.ord i A})}{(\text{antal løb.ord i A}) + (\text{antal løb.ord i B})} \end{aligned}$$

Herved får fællesmængdeordene en fiktiv absolutfrekvens; alle disse fiktive absolutfrekvenser kan vi addere og få et fiktivt omfang i absolutte tal, fiktivt, fordi der faktisk ikke umiddelbart ligger en skinbarlig tekst til grund for dette ordforråd. Dette ordforråd med de fiktive frekvensoplysninger sammenligner jeg nu - som før - med en reel tekst og itererer nu hele processen, indtil der ikke er flere tekster tilbage. Resultatet af denne proces udført på de fem tekster er et ordforråd, der ganske vist er fiktivt, men som har en række egenskaber, som om det var virkeligt: ordforrådet er på 4846 løbende ord, dvs. så stor er den fiktive eller kunstige, i alle fald non-eksistente tekst, der ville have resulteret i samme analyseresultater; i vor fiktive tekst er disse 4846 løbende ord fordelt på 206 forskellige ord, der altså har en gennemsnitsfrekvens på 23,52, de 14 hyppigste

ord dækker ca. 50% af ordforrådet, de 125 hyppigste ca. 95%, og på dette niveau er vi nede på en absolutfrekvens på 5, dvs. vi nærmer os grænsen for, hvad der er signifikant og hvad ikke. De sidste ca. 80 ord dækker kun ca. 5% af ordforrådet.

Til sammenligning anføres her nogle måleresultater fra en roman på 40.659 løbende ord; vi finder her 6.450 forskellige ord, der altså har en gennemsnitsfrekvens på 6,3. De hyppigste ca. 50 ord dækker ca. 50% af ordforrådet, de 815 hyppigste ca. 80%, hvilket bringer os til grænsen for det signifikante, hvilket igen betyder, at kun ca. 13% af de forskellige ord ligger over denne grænse.

Til slut en liste over de ord, der har udkrystalliseret sig som normalnøl på grundlag af de fem valgte tekster på i alt ca. 50.000 løbende ord efter at de har været behandlet som ovenfor anført. Listen er rangordnet og omfatter de 136 hyppigste ord, hvilket er til og med absolutfrekvens 5. Listens seks kolonner rummer:

løbenr-ord-rangtal-abs.frekvens-relativfrekvens-accumuleret relativfrekvens.  
De relative frekvenser er med henblik på sammenlignelighed med andre tilsvarende lister angivet i prodekamille. Der er ikke foretaget homografseparering.

|        |    |     |       |      |
|--------|----|-----|-------|------|
| i      | 1  | 301 | 621.1 | 621  |
| og     | 2  | 287 | 592.2 | 1213 |
| at     | 3  | 250 | 515.9 | 1729 |
| af     | 4  | 224 | 462.2 | 2191 |
| 5 det  | 5  | 174 | 359.1 | 2551 |
| en     | 6  | 169 | 348.7 | 2899 |
| er     | 7  | 164 | 338.4 | 3238 |
| til    | 7  | 164 | 338.4 | 3576 |
| for    | 9  | 142 | 293.0 | 3869 |
| 10 på  | 10 | 128 | 264.1 | 4133 |
| med    | 11 | 124 | 255.9 | 4389 |
| den    | 12 | 122 | 251.8 | 4641 |
| der    | 12 | 122 | 251.8 | 4893 |
| de     | 14 | 119 | 245.6 | 5138 |
| 15 som | 15 | 106 | 218.7 | 5357 |
| ikke   | 16 | 96  | 198.1 | 5555 |
| et     | 17 | 89  | 183.7 | 5739 |
| har    | 18 | 88  | 181.6 | 5920 |

|          |    |    |       |      |
|----------|----|----|-------|------|
| var      | 19 | 75 | 154.8 | 6075 |
| 20 om    | 20 | 74 | 152.7 | 6228 |
| ved      | 21 | 60 | 123.8 | 6352 |
| han      | 22 | 56 | 115.6 | 6467 |
| fra      | 23 | 52 | 107.3 | 6574 |
| kan      | 24 | 49 | 101.1 | 6676 |
| 25 men   | 25 | 47 | 97.0  | 6773 |
| sig      | 26 | 46 | 94.9  | 6868 |
| så       | 27 | 39 | 80.5  | 6948 |
| blev     | 28 | 37 | 76.4  | 7024 |
| eller    | 28 | 37 | 76.4  | 7101 |
| 30 man   | 30 | 31 | 64.0  | 7165 |
| over     | 31 | 30 | 61.9  | 7227 |
| havde    | 32 | 29 | 59.8  | 7286 |
| kunne    | 33 | 24 | 49.5  | 7336 |
| også     | 33 | 24 | 49.5  | 7385 |
| 35 efter | 35 | 23 | 47.5  | 7433 |
| under    | 36 | 22 | 45.4  | 7478 |
| være     | 36 | 22 | 45.4  | 7524 |
| da       | 38 | 21 | 43.3  | 7567 |
| andre    | 39 | 20 | 41.3  | 7608 |
| 40 mod   | 40 | 19 | 39.2  | 7648 |
| skal     | 40 | 19 | 39.2  | 7687 |
| vil      | 42 | 18 | 37.1  | 7724 |
| end      | 43 | 17 | 35.1  | 7759 |
| kun      | 43 | 17 | 35.1  | 7794 |
| 45 år    | 43 | 17 | 35.1  | 7829 |
| alle     | 46 | 16 | 33.0  | 7862 |
| denne    | 46 | 16 | 33.0  | 7895 |
| fik      | 46 | 16 | 33.0  | 7928 |
| hvor     | 46 | 16 | 33.0  | 7961 |
| 50 noget | 46 | 16 | 33.0  | 7994 |
| s        | 46 | 16 | 33.0  | 8027 |
| sin      | 46 | 16 | 33.0  | 8060 |
| skulle   | 46 | 16 | 33.0  | 8093 |
| ud       | 46 | 16 | 33.0  | 8126 |
| 55 meget | 55 | 15 | 31.0  | 8157 |
| når      | 55 | 15 | 31.0  | 8188 |
| op       | 55 | 15 | 31.0  | 8219 |
| vi       | 55 | 15 | 31.0  | 8250 |
| været    | 55 | 15 | 31.0  | 8281 |
| 60 hvad  | 60 | 14 | 28.9  | 8310 |
| hvis     | 60 | 14 | 28.9  | 8339 |
| mere     | 60 | 14 | 28.9  | 8368 |
| selv     | 60 | 14 | 28.9  | 8397 |

|           |     |    |      |      |
|-----------|-----|----|------|------|
| ville     | 60  | 14 | 28.9 | 8426 |
| 65 gennem | 65  | 13 | 26.8 | 8452 |
| mellem    | 65  | 13 | 26.8 | 8479 |
| dem       | 67  | 12 | 24.8 | 8504 |
| første    | 67  | 12 | 24.8 | 8529 |
| få        | 67  | 12 | 24.8 | 8553 |
| 70 må     | 67  | 12 | 24.8 | 8578 |
| to        | 67  | 12 | 24.8 | 8603 |
| andet     | 72  | 11 | 22.7 | 8626 |
| deres     | 72  | 11 | 22.7 | 8648 |
| have      | 72  | 11 | 22.7 | 8671 |
| 75 helt   | 72  | 11 | 22.7 | 8694 |
| nu        | 72  | 11 | 22.7 | 8716 |
| blevet    | 77  | 10 | 20.6 | 8737 |
| derfor    | 77  | 10 | 20.6 | 8758 |
| findes    | 77  | 10 | 20.6 | 8778 |
| 80 fordi  | 77  | 10 | 20.6 | 8799 |
| her       | 77  | 10 | 20.6 | 8820 |
| siden     | 77  | 10 | 20.6 | 8840 |
| anden     | 83  | 9  | 18.6 | 8859 |
| samme     | 83  | 9  | 18.6 | 8877 |
| 85 sit    | 83  | 9  | 18.6 | 8896 |
| store     | 83  | 9  | 18.6 | 8915 |
| uden      | 83  | 9  | 18.6 | 8933 |
| allerede  | 88  | 8  | 16.5 | 8950 |
| dag       | 88  | 8  | 16.5 | 8966 |
| 90 flere  | 88  | 8  | 16.5 | 8983 |
| forhold   | 88  | 8  | 16.5 | 8999 |
| først     | 88  | 8  | 16.5 | 9016 |
| gjort     | 88  | 8  | 16.5 | 9032 |
| godt      | 88  | 8  | 16.5 | 9049 |
| 95 grund  | 88  | 8  | 16.5 | 9065 |
| hans      | 88  | 8  | 16.5 | 9082 |
| hver      | 88  | 8  | 16.5 | 9098 |
| mens      | 88  | 8  | 16.5 | 9115 |
| nogle     | 88  | 8  | 16.5 | 9131 |
| 100 nye   | 88  | 8  | 16.5 | 9148 |
| række     | 88  | 8  | 16.5 | 9164 |
| sammen    | 88  | 8  | 16.5 | 9181 |
| samtidig  | 88  | 8  | 16.5 | 9197 |
| senere    | 88  | 8  | 16.5 | 9214 |
| 105 blive | 105 | 7  | 14.4 | 9228 |
| del       | 105 | 7  | 14.4 | 9243 |
| før       | 105 | 7  | 14.4 | 9257 |
| gang      | 105 | 7  | 14.4 | 9272 |

|           |     |   |      |      |
|-----------|-----|---|------|------|
| hvert     | 105 | 7 | 14.4 | 9286 |
| 110 lige  | 105 | 7 | 14.4 | 9300 |
| mindre    | 105 | 7 | 14.4 | 9315 |
| nok       | 105 | 7 | 14.4 | 9329 |
| sine      | 105 | 7 | 14.4 | 9344 |
| stor      | 105 | 7 | 14.4 | 9358 |
| 115 tid   | 105 | 7 | 14.4 | 9373 |
| arbejde   | 116 | 6 | 12.4 | 9385 |
| både      | 116 | 6 | 12.4 | 9397 |
| endnu     | 116 | 6 | 12.4 | 9410 |
| finde     | 116 | 6 | 12.4 | 9422 |
| 120 hele  | 116 | 6 | 12.4 | 9435 |
| hvilket   | 116 | 6 | 12.4 | 9447 |
| ingen     | 116 | 6 | 12.4 | 9459 |
| kom       | 116 | 6 | 12.4 | 9472 |
| ny        | 116 | 6 | 12.4 | 9484 |
| 125 bedre | 125 | 5 | 10.3 | 9494 |
| fået      | 125 | 5 | 10.3 | 9505 |
| give      | 125 | 5 | 10.3 | 9515 |
| går       | 125 | 5 | 10.3 | 9525 |
| hos       | 125 | 5 | 10.3 | 9536 |
| 130 især  | 125 | 5 | 10.3 | 9546 |
| kommer    | 125 | 5 | 10.3 | 9556 |
| lade      | 125 | 5 | 10.3 | 9567 |
| se        | 125 | 5 | 10.3 | 9577 |
| sådan     | 125 | 5 | 10.3 | 9587 |
| 135 taget | 125 | 5 | 10.3 | 9598 |
| tale      | 125 | 5 | 10.3 | 9608 |