Henrik Holmboe

Det endelige ordforråd

Abstract

Text analysis and text comprehension are problems closely connected with the question of vocabulary. The vocabulary comprises the units which text analysis cannot ignore as one of its goals, and text comprehension - irrespective of theoretical inclination - will always interact with word comprehension. You may claim that texts presuppose words and word semantics in so far as the words are the bricks in the building of texts, or you may claim that the meaning of words is an abstract concept presupposing the sentence, or the text, and not having its empirical status; but to the best of my knowledge nobody will claim that there is no interrelation between the comprehension of words and that of texts. Thus it is relevant to study the words and the vocabularies of languages considering i.a., whether the vocabulary is finite or infinite. This paper will argue the first case, even for a language like Danish with its wide range of word formation potentials which might present the latter case as the attractive alternative. The arguments will be based on facts about the structure of the words and on some considerations on functional and semantic load. The paper is a slightly revised and extended version of an oral contribution given at a conference “Danish Text Analysis on Computers” in November 1991 at the University of Copenhagen under the auspices of the Danish Research Council for the Humanities.

Man har i sprogvidenskaben ytret sig på forskellig måde om ordforråds størrelse. Vi ved bl.a., at Stephanus (Henri Éstienne) i sit store ordbogsværk Thesaurus Linguae Graecae (Genève 1572) satte sig for at registrere alle de overleverede græske ord, dvs. fra den klassisk græske litteratur kendte ord. Det lykkedes rigtig godt, og man har set et lignende initiativ gennemført for latinens vedkommende; men her meldte der sig et problem: klassisk græsk er et dødt sprog, en lukket klasse, et system, der kun i meget begrænset udstrækning, kun på en indirekte måde, innoveres alene af den grund, at der ingen modersmålstalende eller på anden måde produktive sprogbrugere er. Latin er aldrig på samme måde afgået ved døden; der kan komme

nydannelser til, og man kan aldrig være sikker på, at man har fået det hele med, al den stund vi endnu ikke har set den sidste tekst. Problemet blev med rette betragtet som accidentielt; rollerne kunne uden principielle vanskeligheder have været byttet om. Ingen blandt de to sprogs dyrkere var i tvivl om, at såvel det ene som det andet sprog ville være i stand til på tilstrækkelig og stringent måde at opfylde de kommunikationsbehov, der måtte opstå.

Da man i det nittende århundrede beskæftigede sig med såkaldt primitive sprog og med spørgsmål vedrørende sprogets oprindelse, tilkendegav man synspunkter, der i virkeligheden er nært beslægtede med det just anførte; nemlig at sproget i sin oprindelse og på de første primitive trin har bestået af ganske få ord med konkret betydning. Med kulturens udvikling vokser antallet af ting, sagsforhold og foreteelser, dvs. de relevante referenters antal ekspanderer, og dermed udvikler også sproget sig og herunder naturligvis ordforrådet. Ser vi på de tidlige tiltag inden for kvantitative undersøgelser af ordforråd som f.eks. Kädings og Zipfs kan vi konstatere, at de bl.a. beskæftiger sig med ordforråds fordeling på meget hyppige og mindre hyppige ord; men det ligger klart, at foruden de ord, de har taget i betragtning, er der endnu flere, som ville kunne behandles efter deres metoder, og som ville opføre sig på analog vis.

Den strukturalistiske og glossematiske sprogvidenskab har bl.a. beskæftiget sig med det sproglige tegns udtryksside og opstillet nogle taxonomier for udtrykssiden - eller rettere, anvist metoder til at opstille taxonomier for udtrykssiden - uden at drage slutninger med hensyn til ordforråds størrelse, idet det interessante var at fastslå, at klassen af lingvistiske tegn var åben alene derved, at et begrænset antal grundelementer altid kunne indgå nye kombinationer med hinanden og med kombinationer af hinanden.

Med Chomsky fik vi fastslået sprogets fundamentalt innovative og kreative karakter og dermed dogmet om det ubegrænsede ordforråd. Også her hviler argumenterne på, at vi kan antage, at mængden af mulige ting, sagsforhold og foreteelser, og vel dermed også mængden af mulige relevante referenter, er ubegrænset. Dette kan være rigtigt, - og jeg vil ikke anfægte synspunktets rigtighed - uden at det dermed er nødvendigt at antage, at ordforrådet er ubegrænset. En ubegrænset semantik fører os naturligvis til at

måtte antage et ubegrænset antal lingvistiske tegn, som selvfølgelig også skal have en udtryksside; men vi tvinges ikke til at antage, at ubegrænsetheden også skal gælde de tegn, der befinder sig på ordniveauet. Vi skal gennemgå nogle iagttagelser og argumenter, der tværtimod tyder på, at ordniveauets udtryksside ganske enkelt ikke kan bære ubegrænsetheden, og følgelig heller ikke gør det.

Blot for præciseringens skyld: det, jeg her forstår ved ord, er graford, som er strenge af bogstaver mellem to grænsesymboler i den trivielle betydning heraf. Mine eksempler i det følgende vil være hentet fra dansk; men i det omfang, mine overvejelser overhovedet er gyldige, vil de også være det for andre sprog.

Hvis man vil undersøge ord og ords opbygning fra udtrykssiden er det for en del formål for simpelt blot at konstatere, at ord består af en rækkefølge af bogstaver. Snarere vil vi påstå, at ord består af stavelser, og at stavelser består af vokaler og konsonanter efter en bestemt lovmæssighed, og at vokaler og konsonanter kan repræsenteres ved hjælp af bogstaver, hvoraf der på dansk er 29. Der er mig bekendt ingen restriktioner med hensyn til, hvorledes danske stavelser kan eller skal - og dermed evt. ikke må - efterfølge hinanden. I denne henseende er der altså en ubegrænset mulighed for kombination med henblik på at opbygge ord. Ser vi på opbygningen af stavelsen, er der en række restriktioner. En stavelse skal altid indeholde netop én vokal, hvis vi ser på den som model; de få eksempler, der er på stavelser med mere end én vokal, vil jeg betragte som enkeltstående og fossile derved, at de ikke er deriverede af en produktiv model. Forud for den obligatoriske vokal kan der være fra nul til tre konsonanter og efter vokalen fra nul til fire konsonanter. Dette medfører, at stavelsens omfang er fra ét til otte elementer, hvoraf ét er en vokal og de syv andre er konsonanter. Man kan nu gennemføre en beregning af de mulige strenge, der opfylder denne formel, og dermed finde frem til, at antallet af stavelser af længden én er $9*20^0 = 9$ og at f.eks. antallet af stavelser af længden 4 og 5 er henholdsvis $4*9*20^3$ (= 288.000) og $5*9*20^4$ (= 7.200.000), og dermed vil man få et helt forkert billede af antallet af stavelser, et billede, der måske ville være egnet til at fremkalde den opfattelse, at antallet var om end ikke ubegrænset så dog meget stort. Her kommer de grafotaktiske regler ind med en række sprogspecifikke begrænsninger: de tre og fire mulige konsonantpladser

henholdsvis før og efter vokalen har udfyldningsmuligheder, der indskrænkes mere og mere, jo flere konsonantpladser man tager i brug. Et enkelt eksempel kan illustrere dette: ser vi på stavelsesstrukturen KKKV i et system med 20 konsonanter og 9 vokaler, er det teoretiske maksimum for kombinationer 20^3*9=72.000; anskuer vi i stedet for problemet ud fra formelen $K_1K_2K_3V$ og ser på, hvor mange medlemmer hver af disse klasser kan have, bliver regnestykket et andet, idet K_1 har ét medlem, K_2 har 3, K_3 har 4 og V 9 medlemmer; herefter skal vi beregne 1*3*4*9=108. Vi kan nu undersøge, hvor mange af disse 108 muligheder, som systemet tilbyder os, vi faktisk udnytter; for overhovedet at få materiale nok har jeg måttet medtage alle ord med forlyden KKKV, altså også de ord, hvor stavelsen fortsætter med én til tre K-er efter vokalen; til dette brug har jeg benyttet mig af Nudansk Ordbog. Det næste skridt herefter bliver at undersøge, hvordan den faktiske sprogbrug udnytter eller snarere belaster de aktualiserede muligheder; til denne undersøgelse har jeg brugt Ruus & Maegaards frekvensordbøger omhandlende børnebøger og romaner, i det følgende HyDaBø og HyDaRo.

Udnyttede forlydsgrupper:	**skj-**	**skr-**	**skv-**	**spj-**	**spl-**	**spr-**	**stj-**	**str-**
Nudansk ordb.:								
antal ord	**18**	**104**	**10**	**3 1747**	**9117**			
HyDaBø:								
antal ord	**7**	**15**	**00**	**1 11 5**	**24**			
antal lemmata	**3**	**9**	**00**	**1 7 2**	**16**			
abs.frq.	**54**	**230**	**00**	**4 18332**	**287**			
gns.frq. - ord	**8**	**15**	**00**	**4 17 6**	**12**			
gns.frq. - lemma		**18**	**26**	**0 0 4**	**26**	**16**	**48**	
HyDaRo								
antal ord	**7**	**18**	**00**	**0 11 2**	**24**			
antal lemmata	**5**	**11**	**00**	**0 8 2**	**21**			
abs.frq.	**62**	**292**	**00**	**0 10210**	**236**			
gns.frq. - ord	**9**	**16**	**00**	**0 9 5**	**10**			**gns.frq. - lemma 12**
	27	**0**	**00**	**13 5 11**				

Skemaet viser, at to eller tre af de otte forlydsgrupper ikke har nogen belastning eller en meget lille, og at hovedparten af byrden bliver båret af ikke over halvdelen af den samlede styrke. Det samme billede vil vi få, hvis vi udvider undersøgelsen til at gælde alle enstavelsesord i en ordbog, og denne gang tager jeg Dansk Retrogradordbog som grundlag:

Enstavelsesord i Dansk Retrogradordbog fordelt efter stavelsesstruktur; forekomst angivet som abs.frq.:

				K
			K	K
			KK	K
		K	KK	K
	V	8	61	7414	5
	K V	82	639	615103	12
	K K V	64	501	38647	3
	K K K V	7	70	334	1

Enstavelsesord i Dansk Retrogradordbog fordelt efter stavelsesstruktur; forekomst angivet som rel.frq. i ‰:

				K
			K	K
			KK	K
		K	KK	K
	V	3	22	27 5	2
	K V	30	234	22538	4
	K K V	23	183	14117	1
	K K K V	3	26	12 1	0

(Skemaerne viser f.eks., at der er 501 forekomster af stavelsesstrukturen KKVK, og at disse udgør 183‰ af materialet)**.**

Stavelsesstrukturen i énstavelsesord i DR rangordnet efter rel.frq.i ‰:

	‰	ACCUM. ‰
KVK	234	234
KVKK	225	459
KKVK	183	642
KKVKK	141	783
KVKKK	38	821
KV	30	851
VKK	27	878
KKKVK	26	904
KKV	23	927
VK	22	949
KKVKKK	17	966
KKKVKK	12	978
VKKK	5	983
KVKKKK	4	987
KKKV	3	990
V	3	993
VKKKK	2	995
KKVKKKK	1	996
KKKVKKK	1	997
KKKKVKK	1	998

Tallene i den sidste kolonne viser, at de fire hyppigste stavelsesstrukturer dækker ca. 80% af materialet, og at den hyppigste halvdel af stavelsesstrukturerne dækker ca. 95% af materialet.

Ser vi på et corpus af løbende tekst, bekræftes billedet endnu en gang; grundlaget er denne gang corpus DK-87 på ca. 1 million løbende tekstord:

Enstavelsesord i DK-87, rangløbenumrene 1-6000 (_ 85% accumuleret relativfrekvens) fordelt efter stavelsesstruktur; forekomst angivet som rel.frq. i ‰:

			K	
			KK	
		K	KK	
	K	K	KK	
V	9	40	224	2
KV	61	272	22750	2
KKV	24	150	9915	0
KKKV	1	13	9 1	0

Stavelsesstrukturen i énstavelsesord i DK-87, rangløbenumrene 1-6000 (_ 85% accumuleret relativfrekvens) fordelt efter stavelsesstruktur; rangordnet efter rel.frq. i ‰:

	‰	ACCUM. ‰
KVK	272	272
KVKK	227	499
KKVK	150	649
KKVKK	99	748
KV	61	809
KVKKK	50	859
VK	40	899
KKVK	24	923
VKK	22	945
KKVKKK	15	960
KKKVK	13	973
KKKVKK	9	982
V	9	991
VKKK	4	995
VKKKK	2	997
KVKKKK	2	999
KKKV+ KKKVKKK	1	1000

Tallene i den sidste kolonne viser, at de fire hyppigste stavelsesstrukturer dækker ca. 75% af materialet, og at den hyppigste halvdel af stavelsesstrukturerne dækker ca. 95% af materialet.

Det samlede indtryk er, at det er forholdsvis få og forholdsvis simple stavelsesstrukturer, der virkelig bruges i sproget.

Vi kan endelig supplere med et skema over fordelingen af ord, hvor ordlængde sammenholdes med vokalantal. Materialet er igen hentet fra DK-87, rangløbenumrene 1-6000 (_ 85% accumuleret relativfrekvens), angivet i absolutte tal.

Tallene 1-19 (lodret) angiver ordlængde; tallene 0-8 (vandret) angiver vokalantal, hvilket med ubetydeligt få undtagelser er lig med stavelsesantal.

	0	1	2	3	4	5	6	7	8
1:	31	10	0	0	0	0	0	0	0
2:	29	116	1	0	0	0	0	0	0
3:	13	366	36	2	0	0	0	0	0
4:	1	440	290	2	0	0	0	0	0
5:	0	189	851	42	0	0	0	0	0
6:	0	29	932	169	4	0	0	0	0
7:	0	1	432	427	21	0	0	0	0
8:	0	1	87	425	62	0	0	0	0
9:	0	0	9	242	127	7	0	0	0
10:		0	0	2	95	155	21	1	0 0
11:		0	0	0	32	87	24	5	0 0
12:		0	0	0	7	49	40	2	0 0
13:		0	0	0	0	12	20	9	0 0
14:		0	0	0	0	4	8	5	2 0
15:		0	0	0	0	3	5	3	0 0
16:		0	0	0	0	0	2	3	0 0
17:		0	0	0	0	0	3	1	0 1
18:		0	0	0	0	0	1	1	0 3
19:		0	0	0	0	0	0	0	0 0

Samme skema med angivelser i ‰:

	0	1	2	3	4	5	6	7	8	
1:	5	2	0	0	0	0	0	0	0	
2:	5	19	0	0	0	0	0	0	0	
3:	2	61	6	0	0	0	0	0	0	
4:	0	73	48	0	0	0	0	0	0	
5:	0	32	142	7	0	0	0	0	0	
6:	0	5	155	28	1	0	0	0	0	
7:	0	0	72	71	4	0	0	0	0	
8:	0	0	15	71	10	0	0	0	0	
9:	0	0	2	40	21	1	0	0	0	
10:		0	0	0	16	26	4	0	0	0
11:		0	0	0	5	15	4	1	0	0
12:		0	0	0	1	8	7	0	0	0
13:		0	0	0	0	2	3	2	0	0
14:		0	0	0	0	1	1	1	0	0
15:		0	0	0	0	1	1	1	0	0
16:		0	0	0	0	0	0	1	0	0
17:		0	0	0	0	0	1	0	0	0
18:		0	0	0	0	0	0	0	0	1
19:		0	0	0	0	0	0	0	0	0

- og endelig kombinationerne af ordlængde og vokalantal eller stavelsesantal rangordnet efter frekvens med angivelse i ‰ og med angivelse af accumulering:

Ordlængde	x	vokalantal	‰ACCUM.	‰
6	**x**	**2** **155**	**155**	
5	**x**	**2** **142**	**297**	
4	**x**	**1** **73**	**370**	
7	**x**	**2** **72**	**442**	
7	**x**	**3** **71**	**513**	
8	**x**	**3** **71**	**584**	
3	**x**	**1** **61**	**645**	
4	**x**	**2** **48**	**693**	
9	**x**	**3** **40**	**733**	
5	**x**	**1** **32**	**765**	
6	**x**	**3** **28**	**793**	
10	**x**	**4** **26**	**819**	
9	**x**	**4** **21**	**840**	
2	**x**	**1** **19**	**859**	
10	**x**	**3** **16**	**875**	
8	**x**	**2** **15**	**890**	
11	**x**	**4** **15**	**905**	
8	**x**	**4** **10**	**915**	
12	**x**	**4** **8**	**923**	
5	**x**	**3** **7**	**930**	
12	**x**	**5** **7**	**937**	
3	**x**	**2** **6**	**943**	
11	**x**	**3** **5**	**948**	

Tallene i den sidste kolonne viser, at de fem hyppigste kombinationer dækker ca. 50% af materialet. Der er i alt 62 forskellige kombinationer, hvoraf vi kun har bragt de første 23, som til gengæld dækker ca. 95% af materialet; de sidste ca. 5% er altså fordelt på 39 forskellige kombinationer.

Vi kan nu vende tilbage til spørgsmålet om ordforrådets ubegrænsethed. Det er umiddelbart klart, at der ikke er ubegrænset mange ord af en hvilken som helst ordlængde. Kun hvis vi anskuer det samlede ordforråd som summen af ord med længden ét plus ord af længden 2 plus ord af længden 3 osv. og samtidig er villige til at mene, at ord kan være ubegrænset lange, kan vi opretholde forestillingen om det ubegrænsede ordforråd. Bortset fra, at tanken om ubegrænset lange ord kan virke contraintuitiv, må vi også sige, at den mangler teoretisk sandsynlighed af et par grunde.

Lad os undersøge to antagelser:

Den ene er, at vort ordforråd ganske vist er stort, men ikke ubegrænset. Dette vil betyde, at ordene før eller siden alle sammen er brugt i den forstand, at en forøgelse af vor tekst i antallet af løbende ord ikke vil fremkalde nye frekvens-ét ord, men derimod vise et genbrug af de allerede anvendte ord, som altså vil få en stigende frekvens; vi kan også udtrykke det på den måde, at der ikke uden for teksten er flere frekvens-nul ord, som kan hentes ind. Hvis vi forøger en tekst med stadig flere løbende ord, må vi altså imødese en ændring af den meget lavfrekvente del af tekstens profil, fordi der, efterhånden som teksten vokser, bliver færre og færre frekvens-nul ord, som kan hentes ind.

Den anden antagelse er, at vort ordforråd er ubegrænset. Dette burde betyde, at tekstens profil ved en forøgelse ikke ændrede sig; der vil hele tiden være en lang tynd hale af frekvens-ét ord, al den stund der hele tiden vil være ubegrænset mange frekvens-nul ord, som vil og kan flyde ind i teksten, efterhånden som den vokser. Lad os se, om vi kan finde empiriske holdepunkter for det ene eller det andet synspunkt.

Mit materiale denne gang er fire danske tekstcorpora à 1 million løbende ord, DK-87, -88, -89 og -90. Vi skal se på udviklingen af frekvens-ét ord og på disses fordeling på ordlængder:

Antal frq.-ét ord:		abs.	0/000
DK	-90	44369	443
DK	-90		
+	-89	76319	382
DK	-90		
+	-89		
+	-88	100045	333
DK	-90		
+	-89		
+	-88		
+	-87	119353	298

Tallene viser, at når teksten vokser fra 1 til 4 mill. løbende ord, falder den relative andel af frekvens-ét ordene. En undersøgelse af frekvens-to ordenes andel viser til gengæld en stigning. (Ved frekvens-to (og -tre) ord forstår jeg hér de ord, der er frekvens-ét ord i DK-90 og DK-89 (og DK-88) og som overgår til at være frekvens-to (og-tre) ord ved sammenlægning af de fire corpora):

Antal frq.-to ord:		abs.	0/000
DK	-90		
+	-89	5621	27,34
DK	-90		
+	-89		
+	-88	12594	40,76
DK	-90		
+	-89		
+	-88		
+	-87	19695	48,18

Da vi savner holdepunkter for at antage, at denne tendens er karakteristisk netop for intervallet 1-4 mill. ord, vil vi antage, at det generelt gælder, at når en løbende teksts længde går mod uendeligt, så går antallet af frekvens-ét ord mod nul. Dette fører os til at antage, at også antallet af frekvens-nul ord går mod nul.

Frekvens-ét ordenes fordeling på ordlængder kan vi iagttage i følgende skema, der viser den procentvise andel af ordene fordelt efter længde og den gennemsnitlige ordlængde i de forskellige grupper af corpora:

Ordlængde:	**1 - 7**	**8 -14**	**15-21**	**22->**	**gns.ln**
DK-90	**25,845**	**60,060**	**13,118**	**0,978**	**10,24**
DK-90 + -89	**24,782**	**59,742**	**14,326**	**1,151**	**10,42**
DK-90 + -89 + -88	**23,645**	**59,692**	**15,450**	**1,213**	**10,6**
DK-90 + -89 + -88 + -87	**22,830**	**59,688**	**16,117**	**1,366**	**10,73**

I den første kolonne falder tallene samtidig med, at tekstmængden stiger; vi vil tolke dette som et tegn på, at ordene af disse ordlængder er ved at blive brugt op; dette harmonerer med, at der er en stigning i denne ordlængdegruppe, hvis vi ser på frekvens-to og -tre ordene. I anden kolonne er tallene næsten stabile; vi kan sammenholde dette med, at den gennemsnitlige ordlængde bevæger sig en smule opad, men inden for netop det interval, som anden kolonne dækker. I tredie kolonne stiger tallene med næsten samme intervaller, som de falder med i første kolonne; i denne gruppe af ordlængder er der altså stadigvæk et ubrugt ordforråd, som kan

inddrages. I fjerde kolonne med de lange ordlængder stiger tallet også, men kun svagt. Min tolkning heraf er, at det skyldes, at området med de høje ordlængder er meget tyndt befolket; der er simpelthen ikke ord at føre ind i teksten. Den omstændighed, at frekvens-nul ordene på de kortere ordlængder gradvis vil blive opbrugt sammenholdt med, at teksten næsten ikke opviser et øget forbrug af lange ordlængder, er for mig en indikation om, at ordforrådet er endeligt. Jeg formoder altså, at nye millioner løbende ord vil følge den samme tendens, som vi har set her, således at ordforrådet bliver stedse mindre innovativt og stedse mere repetitivt.

Frekvens-to ordenes fordeling på ordlængder:

Ordlængde:	**1 - 7**	**8 -1415-21**	**22->**	**gns.ln**
DK -90				
+ -89	**32,38**	**60,636,67**	**0,3**	**9,23**
DK -90				
+ -89				
+ -88	**31,38**	**61,177,09**	**0,36**	**9,34**
DK -90				
+ -89				
+ -88				
+ -87	**31,71**	**60,017,87**	**0,42**	**9,38**

Frekvens-tre ordenes fordeling på ordlængder:

Ordlængde:	**1 - 7**	**8 -1415-21**	**22->**	**gns.ln**
DK -90				
+ -89				
+ -88	**30,98**	**63,1 5,85**	**0,08**	**9,15**
DK -90				
+ -89				

+	**-88**				
+	**-87**	**31,62**	**62,697,11**	**0,08**	**9,15**

I det foregående så vi, at frekvens-ét ordene på 1 mill. ord udgjorde 443 0/000 og på 4 mill. ord 298 0/000. I DAJUR, et corpus på 1 mill. løbende ord juridisk tekst, udgør frekvens-ét ordene 226 0/000. Dette sidste tal kan forklares med, at DAJUR i sammenligning med DK-corporaene kun udgør én genre. Vi har set, at når man lægger flere corpora med flere genrer sammen, falder den relative andel af frekvens-ét ord. Dette vil jeg tolke på den måde, at når den løbende tekstmængde stiger, konvergerer genrerne mod én, hvilket er i overensstemmelse med det just anførte synspunkt, at ordforrådet bliver stedse mere repetetivt.

Hvad betyder nu dette for tekstfortolkningen, for den semantiske mangfoldighed? Trues disse ikke af visnedøden, hvis jeg har ret i min profeti om, at ordforrådets kilder før eller siden tørrer ud? Det tror jeg slet ikke. Jeg vil vove den påstand, at mangfoldigheden er at finde i mangetydigheden. De allerhyppigste ord, de små og korte, er måske nok mangetydige, men ikke på en spændende måde. Den næste snes tusinde derimod er de ord, der kan præstere tilsyneladende uudtømmelige kombinationer af mangetydighed. De sjældnere ord, dvs. de længere og længere ord, bliver mere og mere éntydige, ikke nødvendigvis lette at tyde og forstå, men dog éntydige og dermed uspændende i det perspektiv, der handler om fortolkningsmæssig og semantisk mangfoldighed, så om der er lidt flere eller lidt færre af dem, behøver ikke at bekymre os. - Et klaver har 85 tangenter, en fløjte har 9 huller og en violin har 4 strenge; har det været en mærkbar begrænsning for musikken? Vi kan sagtens forestille os en uendelighed af tekster, selv om de er sammensat af det endelige ordforråd.

Litteratur

Holmboe, Henrik (1978): *Dansk Retrogradordbog*. København: Akademisk Forlag.

Maegaard, Bente og Hanne Ruus (1981): *Hyppige Ord i Danske Børnebøger*. København: Gyldendal.

Maegaard, Bente og Hanne Ruus (1981): *Hyppige Ord i Danske Romaner*. København: Gyldendal.