Woorden tellen

In ziekenhuizen verschijnen veel rapporten die over de behandeling van patiënten gaan. In dergelijke rapporten komen, naast het gewone taalgebruik, ook veel medische termen voor. Bij twee ziekenhuizen heeft men onderzoek gedaan naar het woordgebruik in deze rapporten. Hiervoor heeft men van 5000 rapporten geteld hoe vaak ieder woord in totaal voorkwam.
Deze rapporten bevatten samen 996 734 woorden. Toch waren er in totaal slechts ongeveer 20 000 verschillende woorden. Dit komt omdat er woorden zijn die heel vaak gebruikt worden. Om je hiervan een idee te geven zie je in tabel 2 de tien woorden die het meest frequent in de rapporten werden gebruikt.

tabel 2

woord	een	de	van	met	en	het	in	is	ik	geen
frequentie	40 361	36 485	34 231	27 667	26 869	22 965	22 082	13 681	11 416	11 363
rangnummer	1	2	3	4	5	6	7	8	9	10

Je ziet dat in de tabel de woorden op *rangnummer*, in volgorde van hun frequentie, zijn genoemd. Zo kun je bijvoorbeeld aflezen dat het woord ‘met’ in totaal 27 667 keer is geteld en dat dit woord rangnummer 4 heeft.

De onderzoekers J. B. Estoup en G. K. Zipf hebben geprobeerd in allerlei teksten een verband te vinden tussen het rangnummer r van een woord en de bijbehorende frequentie f_r. In 1949 vond Zipf de formule:

$$f_r = \frac{C}{r}$$

Deze formule wordt ook wel de ‘wet van Zipf’ genoemd.
De waarde van C hangt af van het totale aantal woorden in de tekst. Volgens Zipf is C de oplossing van de vergelijking:

$$2{,}3 \cdot C \cdot \log C = \textit{aantal woorden in de tekst}$$

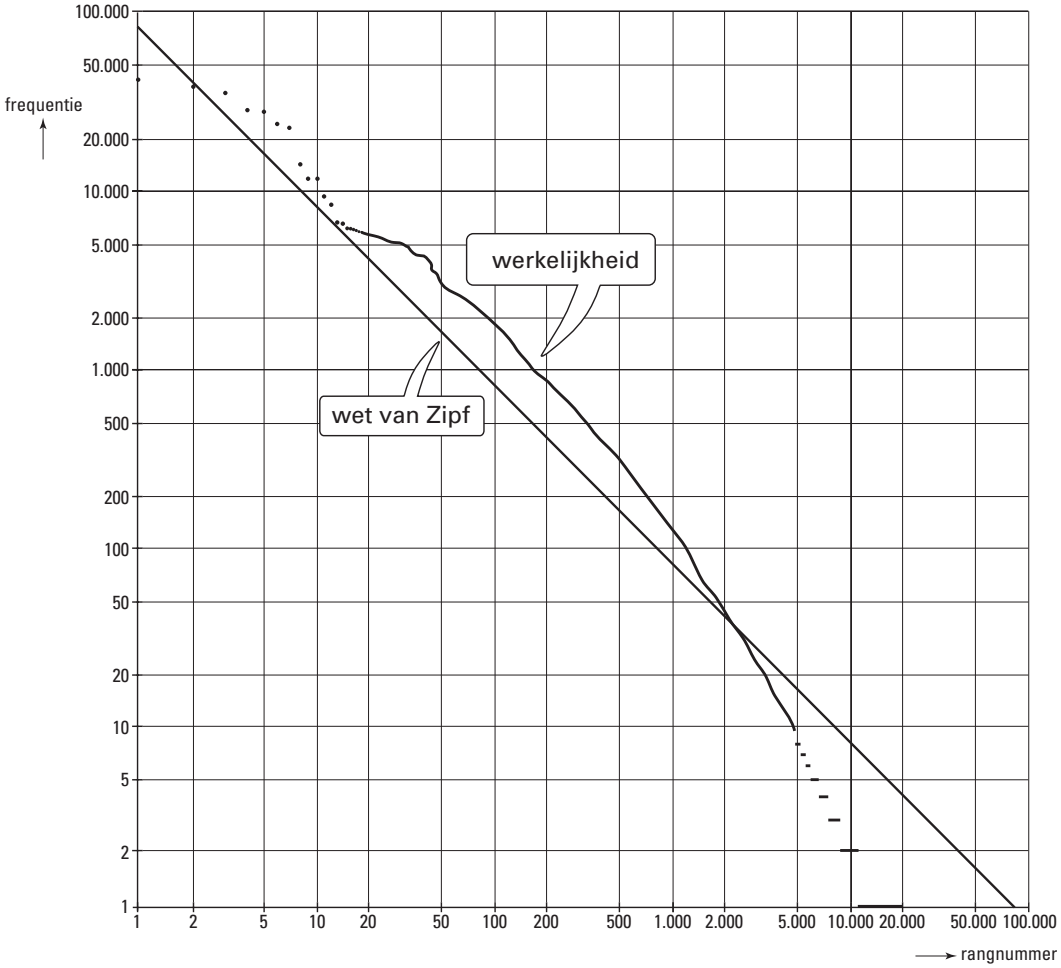
De rapporten van één van de ziekenhuizen bevatten samen 495 378 woorden.

3p **9** ☐ Bereken de waarde van C die bij de rapporten van dit ziekenhuis hoort. Rond af op duizendtallen.

Voor de 996 734 woorden in de rapporten van beide ziekenhuizen *samen* geldt $C = 88\,000$.

In figuur 3 zijn van alle gebruikte woorden de frequenties uitgezet tegen de rangnummers. Op beide assen is gekozen voor een logaritmische schaalverdeling. De woorden uit tabel 2 vind je in figuur 3 terug als de bovenste 10 punten.
Om de wet van Zipf en de werkelijkheid met elkaar te kunnen vergelijken, is in figuur 3 ook de grafiek van $f_r = \frac{88000}{r}$ getekend. Figuur 3 is ook afgedrukt op de uitwerkbijlage.

figuur 3



De wet van Zipf geldt voor algemene teksten zoals krantenartikelen en dergelijke. Omdat medische rapporten niet 'algemeen' zijn, vertonen de grafieken opmerkelijke verschillen.

Tussen de rangnummers 2 en (ongeveer) 2200 zijn de werkelijke frequenties groter dan de frequenties volgens de wet van Zipf.

4p **10** □ Onderzoek of dit verschil bij $r = 100$ groter is dan bij $r = 500$. Licht je antwoord toe.

Iemand trekt uit figuur 3 de volgende twee conclusies:

1. In deze medische rapporten heeft een meerderheid van de gebruikte woorden een hogere frequentie dan de wet van Zipf voorspelt voor teksten met deze omvang.
2. Deze medische rapporten bevatten minder verschillende woorden dan de wet van Zipf voorspelt voor teksten met deze omvang.

4p **11** □ Geef over elk van deze conclusies een gemotiveerd oordeel.

In figuur 3 zie je dat er in de medische rapporten woorden voorkomen die dezelfde frequentie hebben. Volgens de wet van Zipf zou dit niet kunnen. Deze wet, $f_r = \frac{88000}{r}$, zegt dat f_r steeds minder snel afneemt naarmate r toeneemt.

4p **12** □ Stel de afgeleide van f_r op en toon met deze afgeleide aan dat voor de wet van Zipf inderdaad geldt dat f_r steeds minder snel afneemt als r toeneemt.

Uitwerkbijlage bij vraag 10

Vraag 10