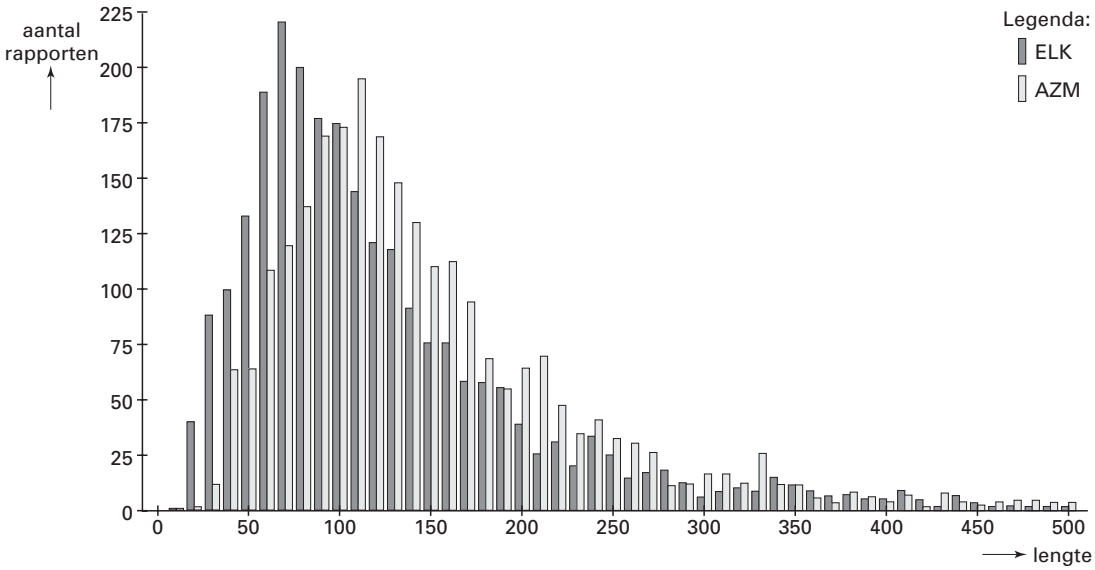
## Teksten vergelijken

In ziekenhuizen worden vaak medische rapporten geschreven. Bij een onderzoek naar de inhoud van dergelijke rapporten zijn 2500 rapporten van het Elkerliek Ziekenhuis (ELK) te Deurne vergeleken met 2500 rapporten van het Academisch Ziekenhuis Maastricht (AZM). Van elk rapport is de lengte bepaald; de lengte van een rapport is het aantal woorden dat het bevat. In figuur 3 zijn de gegevens weergegeven in een gecombineerd staafdiagram met klassenbreedte 10.

figuur 3




Voor de lengte van de rapporten van de ene verzameling geldt:
I *1e kwartiel is 68, mediaan is 100 en 3e kwartiel is 149*
Voor de lengte van de rapporten van de andere verzameling geldt:
II *1e kwartiel is 92, mediaan is 127 en 3e kwartiel is 184*

3p **10** □ Welke van deze series gegevens, I of II, hoort bij de rapporten van het ELK? Licht je antwoord toe.

Uit het onderzoek bleek dat de mediaan en het gemiddelde die horen bij de rapporten van het AZM niet even groot zijn.

4p **11** □ Geef met een redenering, dus zonder een berekening, aan of de mediaan groter of kleiner is dan het gemiddelde.

De rapporten van beide ziekenhuizen 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:

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

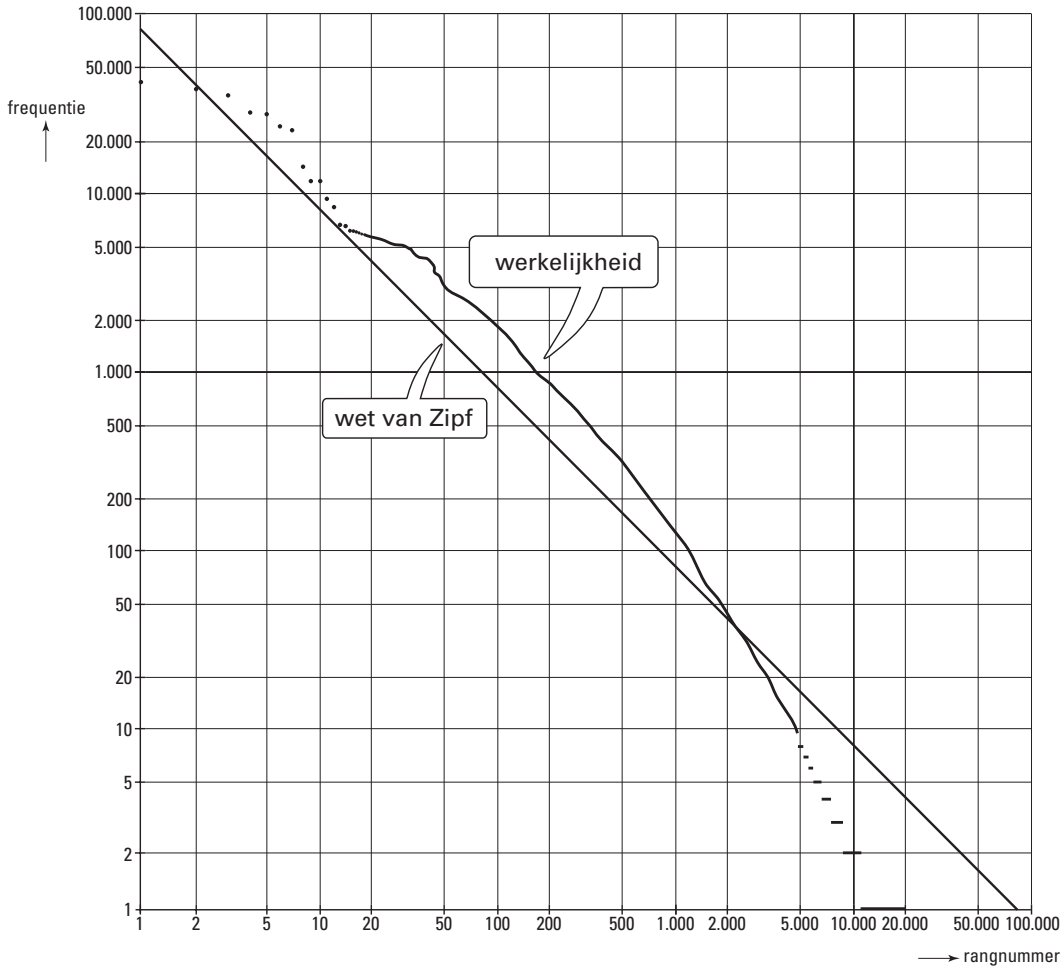
De rapporten van het AZM bevatten samen 495 378 woorden.

3p **12** □ Bereken de waarde van $C$ die bij de rapporten van het AZM hoort. Rond af op duizendtallen.

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

In figuur 4 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 4 terug als de bovenste 10 punten.
Om de wet van Zipf en de werkelijkheid met elkaar te kunnen vergelijken, is in figuur 4 ook de grafiek van $f_r = \frac{88000}{r}$ getekend. Figuur 4 is ook afgedrukt op de uitwerkbijlage.

figuur 4




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

wiskunde A1 (nieuwe stijl)

**Uitwerkbijlage bij vraag 13**

**Examen VWO 2004**

Tijdvak 2
Woensdag 23 juni
13.30 – 16.30 uur

**Examennummer**

.....

**Naam**

.....

**Vraag 13**