Unieke woorden

Teksten bestaan uit woorden (en leestekens, maar die laten we in deze opgave buiten beschouwing). Deze woorden zijn niet allemaal verschillend. Dat wil zeggen dat ze niet allemaal uniek zijn. Hoe meer unieke woorden je naar verhouding tegenkomt, hoe moeilijker de tekst is.

In deze opgave kijken we naar het percentage unieke woorden in een tekst. Dit percentage wordt bepaald aan de hand van twee grootheden:

U: het aantal unieke woorden in een stuk tekst;
T: het totaal aantal woorden in dat stuk tekst.

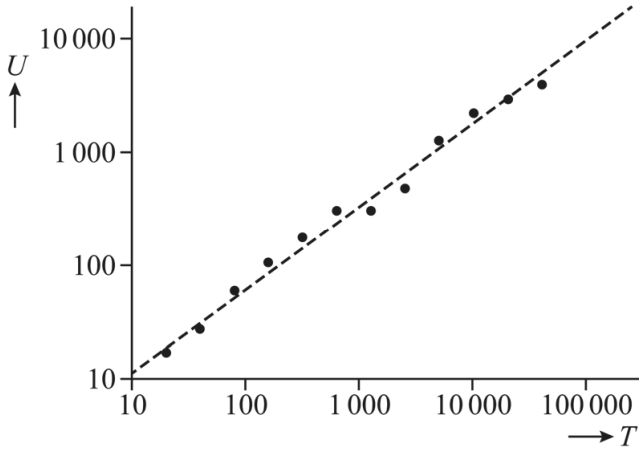
We bekijken de eerste twee zinnen van deze opgave:

> Teksten bestaan uit woorden (en leestekens, maar die laten we in deze opgave buiten beschouwing). Deze woorden zijn niet allemaal verschillend.

2p **15** Bepaal het percentage unieke woorden in de eerste twee zinnen van deze opgave samen. Geef je antwoord als geheel getal.

Van het boek *On The Origin of Species* van Charles Darwin is het verband tussen U en T bepaald. Zie figuur 1.

figuur 1



In figuur 1 is op beide assen een logaritmische schaal gebruikt. De gestippelde lijn geeft een benadering van het verband tussen U en T. Figuur 1 staat ook vergroot op de uitwerkbijlage.

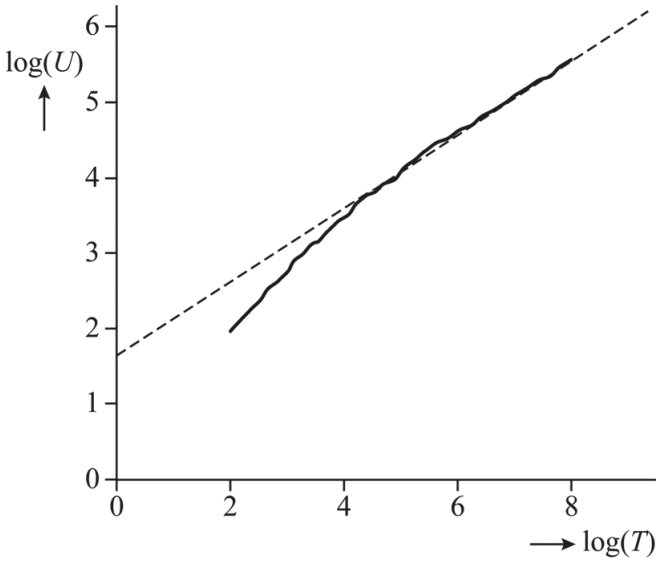
On The Origin of Species bevat in totaal 191 740 woorden en er komen 8842 unieke woorden in voor. Naarmate je verder leest, kom je steeds minder nieuwe unieke woorden tegen. Als je een kwart van dit boek hebt gelezen, ben je al meer dan de helft van het totaal aantal unieke woorden tegengekomen.

5p **16** Bereken met behulp van de gestippelde lijn in de figuur op de uitwerkbijlage hoeveel procent van het totaal aantal unieke woorden je dan al bent tegengekomen. Geef je antwoord als geheel getal.

De taalkundige Gustav Herdan ontdekte een algemeen verband tussen U en T voor grotere teksten. Dit verband werd door Harold Stanley Heap bekendgemaakt en wordt de **wet van Herdan-Heap** genoemd.

De internationale nieuwsdienst Reuters heeft een database – de zogeheten **RCV1** – beschikbaar gesteld ten behoeve van taalonderzoek. Onderzoekers hebben voor RCV1 het verband tussen U en T bepaald. Zie figuur 2, waarin $\log(U)$ tegen $\log(T)$ is uitgezet.

figuur 2



De grafiek in figuur 2 geeft het werkelijke verband tussen U en T in RCV1 en de gestippelde lijn geeft een benadering volgens de wet van Herdan-Heap.

Iemand leest een tekst die bestaat uit de eerste 7432 woorden uit RCV1.

2p **17** Ga met behulp van figuur 2 na of deze tekst voldoet aan de wet van Herdan-Heap.

Een formule voor de gestippelde lijn in figuur 2 is

$$\log(U) = 0,49\log(T) + 1,64$$

3p **18** Benader met behulp van deze formule het aantal unieke woorden in de eerste 1 000 000 woorden in RCV1. Geef je antwoord in duizenden.

De formule $\log(U) = 0,49\log(T) + 1,64$ kan geschreven worden als $U = 43,65 \cdot T^{0,49}$.

Stel nu dat je RCV1 in zijn geheel gaat lezen. Als je dan drie keer zo ver bent gekomen, wil dat niet zeggen dat je ook drie keer zo veel unieke woorden bent tegengekomen. Met behulp van de formule $U = 43,65 \cdot T^{0,49}$ kun je berekenen hoeveel procent meer unieke woorden je dan wel bent tegengekomen.

4p **19** Bereken dit percentage. Geef je antwoord als geheel getal.