Dankzij een wiskundige wetmatigheid in taal kan Google in een fractie van een seconde een zoekvraag beantwoorden of een tekst automatisch vertalen.

Talen temmen door woordgebruik te tellen

Wat is het meest gebruikte woord in het Nederlands? Dat hangt af van wat voor soort taalgebruik je bekijkt. Gaat het om geschreven Nederlands in kranten en tijdschriften, dan is ‘de’ het meest gebruikte woord. Maar neem je het gesproken Nederlands, dan komt het woord ‘ja’ bovenaan te staan. En op Twitter is ‘ik’ de koploper.

Toch hebben geschreven Nederlands, gesproken Nederlands en Twitter-Nederlands één ding met elkaar gemeen: het meest voorkomende woord binnen een zo’n domein komt tweemaal zo vaak voor als nummer twee op de ranglijst, driemaal zo vaak als nummer drie, enzovoort. Wanneer we de frequentie van het meest voorkomende woord op 1 stellen, dan vormen de woordfrequenties de rij 1, 1/2, 1/3, 1/4…

Uitgedrukt in een wiskundige formule, heet dit patroon de wet van Zipf, naar de Amerikaanse taalwetenschapper George Zipf, die de wet in 1935 ontdekte. “Deze wet blijkt voor alle talen en voor alle verzamelingen teksten binnen een taal te gelden, of je nu kijkt in een Chinees wetboek, een Noorse bijbel of in Engelstalige e-mails van een groot bedrijf”, zegt Antal van den Bosch, hoogleraar aan de Radboud Universiteit Nijmegen en specialist in computationele taalkunde. “De wetboek van Zipf is een empirische wet, maar hij klopt vrij nauwkeurig. Alleen aan het begin, bij de top 10 van woorden, en aan de staart, bij de zeldzame woorden, wijkt de praktijk een beetje af van de wiskundige formule.”

De top-10 van geschreven Nederlands:

1. de
2. van
3. het
4. een
5. en
6. in
7. is
8. dat
9. op
10. te

De top-10 van gesproken Nederlands:

1. ja
2. dat
3. de
4. en
5. uh
6. ik
7. een
8. is
9. die
10. van

Efficiënt zoeken

Precies omdat de wet van Zipf universeel geldig is, kan Google's zoekmachine zo razendsnel antwoord geven. Van den Bosch: "De truc die Google gebruikt, is dat ze een woordindex van het Web hebben gemaakt en deze voortdurend bijwerken. De woordindex vertelt welk woord in welk document voorkomt. Met de wet van Zipf kun je nu laten zien dat die woordindex compact is. En dat betekent weer dat je deze compact op harde schijven kunt opslaan en gemakkelijk kunt distribueren naar datacentra over de hele wereld."

Waarom is die woordindex precies compact? Google heeft toegang tot tientallen miljarden webpagina's, maar het aantal woorden per taal loopt 'slechts' in de miljoenen, waarvan er trouwens meestal maar enkele honderdduizenden in een officieel woordenboek staan. De wet van Zipf leert ons nu dat de helft van het aantal woorden in een grote tekstverzameling maar eenmaal voorkomt. Dankzij Zipf weten we ook dat in de top 300 bijna alle functiewoorden staan (lidwoorden, voornaamwoorden, voorzetsels...) en de meest gebruikte inhoudswoorden (zelfstandige naamwoorden, werkwoorden, bijwoorden, bijvoeglijke naamwoorden). Deze beide eigenschappen maken de woordindex compact.

Van den Bosch: “Wanneer wij een zoekterm intikken, hoeft Google dus niet in tientallen miljarden documenten te zoeken, maar in de veel hanteerbaardere woordindex. En tikt iemand vier zoekwoorden in, dan neemt de zoekmachine de overlap van vier verzamelingen. Elke verzameling vertelt op welke webpagina het betreffende woord voorkomt. Die berekening is eenvoudig en dus razendsnel.”

Automatisch vertalen

Automatische vertaalmachines benutten een soort afgeleide eigenschap van de wet van Zipf, een eigenschap van het voorkomen van combinaties van woorden. Google Translate gebruikt een grote database met bestaande vertalingen, bijvoorbeeld officieel vertaalde teksten van het Europees parlement, of vertaalde ondertitels van films. Om een nieuwe tekst van bijvoorbeeld het Nederlands naar het Engels te vertalen, zoekt de vertaalmachine naar zo lang mogelijke woordcombinaties die in de bestaande vertalingen zo vaak mogelijk op dezelfde manier zijn vertaald.

De vertaalmachine ziet bijvoorbeeld dat het Shakespeare-citaat ‘Juliet is the sun’ altijd vertaald wordt als ‘Julia is de zon’. Dan zal dat wel de juiste vertaling zijn. Hoe vaak woordcombinaties in een bepaalde volgorde voorkomen, is ook weer verdeeld op een Zipf-achtige manier: slechts een beperkt aantal combinaties komt heel veel voor. En net zoals de woordindex compact is, zo is de index van woordcombinaties ook compact. Daarom doet een vertaalmachine zo snel zijn werk.

Deze statistische manier van vertalen werkt goed voor teksten die sterk lijken op al bestaande teksten. Maar hoe unieker en creatiever de tekst, hoe moeilijker de vertaalmachine het heeft. “Poëzie is notoir moeilijk”, zegt van den Bosch. “De heilige graal binnen mijn vakgebied is dan ook: hoe kunnen we ervoor zorgen dat machines taal ook echt begrijpen? Want dat doet Google Translate nog steeds niet, hoe handig hij vaak ook is.”