NICOLINE VAN DER SIJS

# Dijt onze woordenschat alsmaar uit?

## Over opbouw en geschiedenis van de Nederlandse woordenschat

RADBOUD UNIVERSITY PRESS

NEDERLANDS IN HET KLEIN

# DIJT ONZE WOORDENSCHAT ALSMAAR UIT?

*Nederlands in het klein* is een boekenreeks die is opgezet door de afdeling Nederlandse Taal en Cultuur aan de Radboud Universiteit in Nijmegen. In ieder deeltje werpt een expert licht op een specifiek onderdeel van de Nederlandse letterkunde, taalkunde of taalbeheersing. Het doel van de reeks is om lezers op een vernieuwende, aansprekende manier kennis te laten maken met recent neerlandistisch onderzoek. Om de boeken bruikbaar te maken in het middelbaar onderwijs, wordt op de website van de Werkgroep Onderzoek en Didactiek Nederlands (WODN) aanvullend lesmateriaal aangeboden. (in de kantlijn aangegeven met  )

Nicoline van der Sijs

# Dijt onze woordenschat alsmaar uit?

## Over opbouw en geschiedenis van de Nederlandse woordenschat

NEDERLANDS IN HET KLEIN

Dijt onze woordenschat alsmaar uit? Over opbouw en geschiedenis van de Nederlandse woordenschat

Uitgegeven door RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen
www.radbouduniversitypress.nl | radbouduniversitypress@ru.nl

Ontwerp: Andre Klijsen, VILLA Y
Print en distributie: Pumbo.nl

ISBN: 9789465151007
DOI: 10.54195/SFIA5632
Gratis te downloaden op: www.radbouduniversitypress.nl



RADBOUD
UNIVERSITY
PRESS

# Inhoud

# Voorwoord

Het meest opvallende aan taal is de woordenschat. Die krijgt vereweg de meeste aandacht in de media. 'Woorden doen ertoe', is de titel van een uitgave van de museumwereld, met als ondertitel 'een incomplete gids voor woordkeuze binnen de culturele sector'. Ook autoritaire wereldleiders zien het belang van woorden in: Poetin noemde de oorlog tegen Oekraïne eufemistisch een *speciale militaire operatie*, en Trump verbiedt ambtenaren om in overheidsstukken 'woke' woorden te gebruiken als *bias, diversity, gender, inclusivity, multicultural, lgbt, race* en *trans(gender)*. Schrijvers, rappers en zangers onderscheiden zich door hun originele en creatieve woordgebruik. En iedereen heeft wel eens de overtuigingskracht ervaren van de juiste woordkeuze of de pijn van een ongelukkig gekozen woord.

Woorden zijn veranderlijk en vaak inwisselbaar: je kunt hetzelfde op verschillende manieren zeggen. Nieuwigheden gaan bijna altijd vergezeld van een nieuw woord. Sommigen concluderen dan ook dat de Nederlandse woordenschat alsmaar groter wordt. Dat baseren ze op de jaarlijkse Woord van het jaar-verkiezingen, de traditionele eindejaarslijstjes in kranten met de nieuwe woorden van het afgelopen jaar, en het feit dat een uitgeverij als Van Dale bij iedere nieuwe druk van het bekende woordenboek een persbericht uitzendt met als lokkertje dat er weer duizenden nieuwe woorden zijn toegevoegd.

Anderen maken zich juist zorgen over een *afname* van de woordenschat: zij werpen zich op als redders van zogenaamde 'vergeetwoorden' – woorden die dreigen te verdwijnen – of verzamelen nostalgisch 'verdwijnwoorden' – woorden die al zijn verdwenen.[1] In 2017 richtte het radioprogramma *De Taalstaat* een heus Gezelschap van Geadopteerde Vergeetwoorden op, met Nelleke Noordervliet als beschermvrouwe, in de hoop het leven te rekken van woorden die in de vergetelheid dreigen te raken.

Hoe zit dat nu eigenlijk? Klopt het dat onze woordenschat alsmaar uitdijt? Kunnen we de woordenschat vergelijken met het heelal, en vond er na het eerste babygeluidje – of een knetterende oerknal – een steeds snellere uitdijing plaats? Of is er sprake van onregelmatige, fluctuerende bewegingen? Is de uitdijing van de woordenschat, net als die van het heelal, eindeloos, of bereiken we ooit een maximum aan mogelijke woorden? Op die vragen zal ik in dit boekje ingaan. In ieder hoofdstuk wordt de Nederlandse woordenschat vanuit een verschillend perspectief bekeken. Dan zal blijken dat het antwoord op de vraag in de titel niet een eenduidig ja of nee is.

Dit boekje gaat dus over Nederlandse woorden en de Nederlandse woordenschat. Die bewering levert echter direct al enkele existentiële vragen op. Wat is een 'woord'?[2] Wat is een 'Nederlands' woord? En wat is 'de' Nederlandse woordenschat? Voor de definitie van *woord* kunnen we natuurlijk kijken in het woordenboek. Dan vinden we iets als 'een reeks klanken of letters met een eigen betekenis'. Voor 99 procent van de gevallen dekt die definitie de lading. Iedereen is het erover eens dat *berg*, *gebergte*, *bergrug* en *verbergen* Nederlandse woorden zijn.

Maar als je wat verder doordenkt, kom je allerlei twijfelgevallen tegen. Is *p.m.* een woord? En hoe zit het met verbindingen als *ad hoc*, *ad rem*, *buiten kijf*, *geen snars*, *in petto*, *om de haverklap*, *op de bonnefooi*? Al deze verbindingen bevatten een of meer spaties, en lijken dus te bestaan uit meerdere reeksen letters, maar beide delen (*ad hoc*, *ad rem*) komen alleen samen als eenheid voor, of het tweede deel (*kijf*, *snars*, *petto*, *haverklap*, *bonnefooi*) komt uitsluitend voor in combinatie met het eerste en in geen enkele andere context. Moeten we deze verbindingen daarom als één woord rekenen, of vanwege de spatie toch als twee of drie? En hoe zit het met *blinde vink* tegenover *blindedarm* en *blindeman*? Maakt de willekeurige spellingregel die bepaalt dat we in het eerste geval een spatie spellen en in de andere gevallen niet, uit voor de beslissing of we met één of twee woorden te maken hebben?

Vergelijkbare vragen duiken op als we het hebben over een 'Nederlands' woord. Is dat een woord dat in het Standaardnederlands gebruikt wordt, of tellen we ook dialectwoorden als *tuinierskruipertje* ('winterko-

ninkje') mee? En hoe zit het met jargonwoorden of technische woorden (*catchy*, *riffelhamer*, *wilfen*), of met straattaalwoorden (*badsloef, gangsterlijk, faja, fittie*), die alleen gebruikt worden binnen een bepaalde groep? Zijn die geen Nederlands? Trouwens: hoe zit het met duidelijk herkenbare leenwoorden, zoals *sportfluencer* en *wokewashing*, die in 2023 meedongen naar de titel 'woord van het jaar', of *whopper, real deal* (met spatie) en de samentrekking *imho* (voor *in my humble opinion*), die niet in Van Dale staan maar frequent in de sociale media gebruikt worden? Is *2.0*, als aanduiding voor alles wat geheel vernieuwd is, een woord? En *huh*, om uit te drukken dat je iets niet begrijpt, of de verzuchting *hèhè*?[3] En hoe zit het met emoji's?[4] Die hebben wel een vaste vorm en eigen betekenis, maar geen klankvorm; desondanks verkoos het *Oxford English Dictionary* in 2015 een emoji tot woord (!) van het jaar.

Laat ik het maar verklappen: alle genoemde voorbeelden beschouw ik als Nederlandse woorden. Ze worden namelijk moeiteloos gebruikt in een Nederlandse zin en onderwerpen zich aan de Nederlandse grammatica: ze krijgen een Nederlandse verbuiging (*catchyer, catchyst*) of vervoeging (*hij wilfte, heeft gewilft*). Ze behoren dus allemaal tot de Nederlandse woordenschat, waarmee de vraag waaruit die bestaat, ook is beantwoord. De grenzen tussen woorden uit de standaardtaal, uit technische taal, uit de dialecten en uit jongerentaal zijn uiterst poreus. Regelmatig gebeurt het dat een woord dat bijvoorbeeld zijn leven in een dialect begon, eindigt als (informeel) standaardtaalwoord. Dat overkwam *flut, kanen, knakker, knoert, meuren* en *reuring*.

Ook woorden die slechts in een deel van het Nederlandse taalgebied gebruikt worden – bijvoorbeeld alleen in Nederland, of alleen in Vlaanderen, Suriname of de Nederlandse Antillen – tellen gewoon mee als Nederlandse woorden. Ook hier is de grens poreus: *kousenband, tayer, pom* en *schaafijs* komen oorspronkelijk uit het Caribische gebied, maar zijn inmiddels op alle Nederlandse markten te koop. Woorden als *living* ('woonkamer'), *plat water, prietpraat* en *uitbater*, en wielertermen als *afzien* ('lijden'), *met afstand* ('met een duidelijke voorsprong'), *demarrage, lossen, nipt, vals plat* en *vlammen* ('keihard fietsen') zijn alle afkomstig uit Nederlandstalig België maar inmiddels algemeen in gebruik.

Woorden hebben verschillende eigenschappen, afhankelijk van hoe je ernaar kijkt. En dat is wat ik in dit boekje ga doen: woorden en de woordenschat bekijken vanuit verschillende standpunten. Welke kenmerken hebben woorden in een woordenboek en woorden in een corpus, hoe zitten woorden in je hoofd, hoe heeft de woordenschat zich in de loop van de tijd ontwikkeld, hoe worden bestaande woorden uitgebreid tot samenstellingen of afleidingen, en hoe gedragen woorden zich in een netwerk: in een uitdrukking, in een tekstcorpus? Al die aspecten komen in de verschillende hoofdstukjes aan de orde. Tot slot zal ik antwoord geven op de beginvraag: dijt onze woordenschat alsmaar uit of stuit zo'n uitdijing op een grens?[5] En passant komt ook de vraag aan de orde of je kunt bepalen hoeveel leenwoorden en samengestelde woorden het Nederlands heeft, en of het Nederlands meer woorden heeft dan het Engels.

Woorden zijn mijn oude, zelfs mijn eerste wetenschappelijke liefde: als 20-jarige begon ik als redacteur van een Nederlands-Russisch woordenboek. Sindsdien ben ik altijd gefascineerd gebleven door de vele facetten van woorden. Ik hoop dat dit boekje iets van die liefde en fascinatie doorgeeft aan een volgende generatie, en verwondering oproept over de rijkdom van onze woordenschat, die we dagelijks zo achteloos gebruiken.

# 1 Woorden in het woordenboek

Een hardnekkig misverstand is dat een woord dat niet in het woordenboek staat, niet bestaat. Een ander hardnekkig misverstand is dat een woordenboekdefinitie uit de Dikke Van Dale uit de aard der zaak kan gelden als hard, juridisch bewijs in geschillen. Die misverstanden zijn te verklaren uit het feit dat mensen zekerheid zoeken en een objectieve scheidsrechter, en bovendien ten onrechte menen dat het woordenboek een officiële status heeft. Dat is echter niet het geval. Alleen de spelling is in de Lage Landen sinds 1804 officieel geregeld, en die officiële spelling is slechts verplicht voor het onderwijs en de overheid. Op schending ervan staat overigens geen sanctie, behalve als je een examen aflegt.

Het idee van het woordenboek als scheidsrechter gaat voorbij aan de functie van het woordenboek. Een woordenboek is bedoeld als hulpmiddel voor de gebruikers en beschrijft hoe een woord in de praktijk wordt gebruikt. Gebruikers kunnen er de spelling, woordsoort, grammaticale bijzonderheden, betekenis en gebruiksvoorbeelden van woorden in vinden. Vroeger was dat anders: tot in de eerste helft van de twintigste eeuw hadden woordenboeken een didactische, voorschrijvende functie: ze toonden het 'correcte' woordgebruik, ook als dat afweek van het dagelijkse gebruik. Inmiddels hebben lexicografen de wens om 'het volk te verheffen' – zoals dat vroeger heette – allang losgelaten. Onder taalgebruikers is het idee van het woordenboek als scheidsrechter echter blijven hangen.[6]

## Eentalige woordenboeken

Wat voor informatie nemen woordenboekschrijvers op? De meeste mensen denken bij een woordenboek aan een lijst alfabetisch geordende trefwoorden met een heldere betekenisomschrijving in het Nederlands. Dat type woordenboek bestaat echter pas anderhalve eeuw. Eerdere woordenboeken waren vertaalwoordenboeken. Vanaf 1799 publiceerde Petrus Weiland een woordenboek in elf delen, waarin hij als eerste de Nederlandse woordenschat *in* het Nederlands beschreef. Ondanks zijn elf delen

was het woordenboek in alle opzichten incompleet. Daarom zette Matthias de Vries het *Woordenboek der Nederlandsche Taal*, kortweg WNT, op. De samenstelling van dat woordenboek duurde maar liefst 137 jaar, maar toen had je ook iets: tussen 1864 en 2001 verschenen maar liefst 43 delen, waarin de Nederlandse woordenschat van 1500 tot 1976 is beschreven.[7]

Het WNT mikte op een wetenschappelijk publiek. Ondertussen zat het algemene publiek op zijn honger: dat had dringend behoefte aan een betrouwbaar, handzaam naslagwerk voor dagelijks gebruik. Voor dat publiek verschenen vanaf de tweede helft van de negentiende eeuw diverse eendelige handwoordenboeken. Die baseerden zich sterk op het WNT, maar zonder de historische invalshoek daarvan. De bekendste woordenboeken zijn naar hun makers genoemd, en ze zijn tot in deze eeuw bijgewerkt: Van Dale (1872-heden) en Koenen (1897-2006). In de twintigste eeuw volgde uitgeverij Prisma met een eveneens vele malen herzien woordenboek specifiek voor het onderwijs en gaf uitgeverij Van Goor een Nederlands woordenboek uit in een door J. Kramers in 1855 begonnen woordenboekenreeks.[8]

## Trefwoordkeuzestress

Welke informatie is er in zo'n handwoordenboek te vinden? De samenstellers van woordenboeken moeten allerlei keuzes maken. Om te beginnen moeten ze kiezen welke trefwoorden ze opnemen. Unaniem vermelden ze in hun inleidingen dat hun doel is 'de algemene woordenschat' te beschrijven. Dat houdt in dat technische termen, dialectwoorden en woorden van specifieke groepen als jongeren of ouderen, scholieren en militairen niet in aanmerking komen voor opname. Die horen thuis in gespecialiseerde woordenboeken. Persoons- en plaatsnamen krijgen alleen een plaatsje als ze een eigen betekenis hebben gekregen, dus soortnaam zijn geworden: dat geldt voor *casanova*, *guppy*, *jonassen*, *rugby* en *spa*. Bovendien moeten de woorden enige tijd in verschillende soorten teksten door diverse soorten taalgebruikers zijn gebezigd.

De criteria klinken stoer, maar in de praktijk zijn ze subjectief. Er bestaat namelijk een poreuze grens tussen algemene woorden enerzijds en technische termen en dialectwoorden en dergelijke anderzijds.

Zo komt tegenwoordig bijna iedereen onvermijdelijk in aanraking met massa's vaktermen uit de computerwereld en het bankwezen. Welke daarvan in het woordenboek terechtkomen is niet zozeer een inhoudelijke keuze als wel een kwestie van de beschikbare ruimte – plat gezegd: het aantal pagina's. En dat wordt bepaald door de woordenboekuitgevers. Die hebben andere belangen dan de woordenboekschrijvers. Zij kijken nauwlettend naar de verkoopcijfers en rekken de grenzen van de algemene woordenschat daarvoor graag wat op. Zo stimuleren zij de opname van woorden uit de vluchtige jongerentaal om nieuwe, jonge kopers te trekken, al behoren die woorden niet tot de algemene woordenschat en verdwijnen ze meestal snel weer. En ze haasten zich opvallende, nieuwe woorden toe te voegen lang voordat die hebben kunnen bewijzen dat ze blijvertjes zijn. Zo voegde de Dikke Van Dale in november 2024 *pieperaanval* aan het woordenboek toe, verwijzend naar een nieuw woord dat toen nog geen maand oud was.

Eén obstakel voor opname is in de loop van de tijd verdwenen, namelijk de herkomst van een woord. In 1851 betoonde Matthias de Vries zich nog streng puristisch en oordeelde dat tegen herkenbare leenwoorden 'met volle gestrengheid' moest worden opgetreden als tegen 'vijanden, waarbij een onophoudelijk waken en strijden vereischt wordt'. Dat is inmiddels veranderd: moderne woordenboekschrijvers werpen geen enkele barrière meer op voor leenwoorden en lijken zich soms zelfs op hen te verlekkeren.

Van oudsher gelden de zogenaamde 'onwelvoeglijke' of 'onkiese' woorden als problematisch. Ze worden veel gebruikt, maar behoren niet tot de standaardtaal en dus ook niet tot de algemene woordenschat. Horen ze dan thuis in het woordenboek of niet? Aanvankelijk wilde Matthias de Vries dergelijke woorden weren, maar in 1882 kwam hij daarop terug: 'De verrichtingen des lichaams, waarbij men geen getuigen toelaat en waarover men in gezelschap niet spreekt, zijn even natuurlijk als eten, drinken enz.'[9] In de twintigste eeuw, en met name in de losgeslagen jaren zestig en zeventig, lieten de lexicografen alle schroom varen. Ongeremd namen ze taboewoorden op: scheldwoorden als *klootzak*, *tyfushond*, vloeken en 'platte' woorden als *schijten*, *ruften*, *kotsen*, *neuken*, *kutzooi* en *klootjehannesen*. Dit tot ontzetting van sommige groepen taalgebruikers.

Om de kool en de geit te sparen voegden de woordenboekschrijvers aan deze woorden waarschuwingslabels toe als 'vulgair', 'informeel' en 'beledigend' – alsof de taalgebruikers niet zelf kunnen bedenken dat dergelijke woorden niet voor alle situaties geschikt zijn… Inmiddels vormen fijnmazige labels een belangrijke aanvulling op de tools die de lexicograaf ter beschikking staan.

De grootste keuze waar de lexicografen voor staan betreft echter de opname van samenstellingen. In het Nederlands kun je twee zelfstandige naamwoorden gemakkelijk samenvoegen tot één nieuw woord. Iedere taalgebruiker maakt dagelijks nieuwe samenstellingen en nooit leidt dat tot onbegrip bij de gesprekspartner. Hoewel *januariweer* niet in de Dikke Van Dale staat, begrijpt iedereen wat je bedoelt als je zegt: 'Het is vandaag echt januariweer.' En ook een reactie als 'Ja, echt beginjanuariweer' leidt tot geen enkel misverstand. Voor woordenboekschrijvers is het per definitie ondoenlijk alle denkbare samenstellingen te beschrijven, omdat hun aantal theoretisch onbeperkt lijkt (maar zie hoofdstuk 4).

Lexicografen hebben ook helemaal niet het *doel* om alle samenstellingen te beschrijven. Al in 1882 noemde Matthias de Vries het 'eene stellige dwaasheid' om alle samenstellingen op te nemen.[10] Begin twintigste eeuw introduceerden de woordenboekschrijvers de term *doorzichtige samenstelling* voor een samenstelling waarvan de betekenis moeiteloos kan worden afgeleid uit die van de afzonderlijke delen. Voor dergelijke samenstellingen – denk aan *hondenoor, kattenoor, koeienoor, leeuwenoor, mensenoor* – ruimen de woordenboeken geen plaats in. Een onvoorspelbare betekenis is vereist als entreebewijs; zo'n bewijs hebben *druiloor, kniesoor, ezelsoor* en *zeeoor* ('schelp').

Het al dan niet opnemen van een woord is kortom voor een belangrijk deel de persoonlijke keuze van de lexicograaf. Die heeft er in deze eeuw wel een hulpmiddel bij gekregen: nu er grote digitale tekstbestanden beschikbaar zijn, kan hij het gebruik en de frequentie van woorden voor het eerst objectief vaststellen. Dat neemt niet weg dat ieder woordenboek per definitie onvolledig is. Het zijn dan ook niet de woordenboeken die het bestaan van een woord bepalen, maar wij, de taalgebruikers.

## Betekenisomschrijvingen

De meest uitdagende taak voor de woordenboekschrijver is het maken van een begrijpelijke en waterdichte betekenisomschrijving. Dat is geen sinecure: probeer maar eens in enkele woorden te definiëren wat een *kat* is en waarin die verschilt van een *hond*, een *konijn* en een *cavia*. Lexicografen hebben van oudsher dan ook hun zware lot beklaagd. De beroemde achttiende-eeuwse Engelse schrijver Samuel Johnson omschreef een lexicograaf als 'a harmless drudge', dus een onschuldige zwoeger, en hij vergeleek een woordenboek met een klok: 'the worst is better than none, and the best cannot be expected to go quite true.' Johannes van Dale ging nog een stapje verder in het voorwoord van de eerste druk van zijn woordenboek in 1872:

> Verzekerde mij een mijner letterkundige vrienden, dat hij, die zijn vader en moeder vermoord heeft, nog te goed was om een Woordenboek te schrijven, ik heb myzelven vaak twijfelmoedig de vraag gedaan, of hij wel volkomen ongelijk had.

Misschien als compensatie lieten woordenboekschrijvers zich zo af en toe verleiden tot onverwachte frivoliteit. C. Kruyskamp, een van de bekendste bewerkers van de Dikke Van Dale, beschouwde het in 1976 'als een onbetwistbaar recht' dat de lexicograaf 'in zaken waarvan de waardering louter een kwestie van smaak is […] een persoonlijke noot mag laten horen'.[11] En die opvatting paste hij in de praktijk toe, zoals blijkt uit de onderstaande voorbeelden.

> *kosmonaut*, een ietwat hyperbolische benaming voor personen die een klein sprongetje in de kosmische ruimte doen, door zich b.v. naar de maan of een planeet van ons zonnestelsel te laten schieten.
> *popmuziek*, zekere, oorspr. op de rock-'n-roll gebaseerde, bij jeugdige en onrijpe personen in de smaak vallende hedendaagse amusementsmuziek.
> *volleybal*, gespeeld balspel waarmee sommige mensen zich vermaken, bestaande in het heen en weer slaan van een bal over een net.

Tegenwoordig is de omschrijving van *volleybal* in de Dikke Van Dale aanzienlijk objectiever, zij het wel érg gedetailleerd: 'balspel, gespeeld door twee ploegen van zes spelers op een veld dat door een net doormidden wordt gedeeld, waar de bal overheen moet worden geslagen (of gekopt) voor hij de grond van de eigen helft heeft geraakt.'

Hoe dan ook: betekenisomschrijvingen weerspiegelen de persoonlijke keuze en inbreng van de lexicograaf. Alleen al om die reden is het een misverstand om te denken dat een woordenboekdefinitie in alle gevallen juridische bewijskracht heeft. Als in een geschil de betekenis van een woord in het 'gewone, algemene taalgebruik' doorslaggevend is, kan Van Dale als scheidsrechter worden gebruikt – en in de praktijk gebeurt dat ook.[12] Als echter de exacte rechtskundige betekenis in het geding is, bijvoorbeeld in het verschil tussen *overtreding, misdaad, misdrijf, wanbedrijf* en *delict*, dan kun je beter een juridisch woordenboek of een ander terminologisch woordenboek raadplegen: dat heeft namelijk als doel om juridisch sluitende definities te geven, terwijl de omschrijvingen in een woordenboek als de Dikke Van Dale bedoeld zijn om gebruikers te informeren over de vele nuances en gebruiksmogelijkheden van een woord. Hoe ongeschikt zo'n algemene omschrijving is, blijkt uit het volgende waargebeurde verhaal uit begin vorige eeuw: een taalhistoricus, die als deskundige door de rechtbank was opgeroepen in een beledigingszaak, verklaarde dat een *smeerlap* een zeer nuttig werktuig is – Van Dale geeft als eerste betekenis 'lap om iets mee (in) te smeren' – en het woord daarom niet als een beledigend scheldwoord kon worden beschouwd.[13] De aangesprokene dacht daar heel anders over…

## De gebruikers

Hierboven stonden de keuzes van de woordenboekmakers centraal. Minstens zo belangrijk zijn natuurlijk de gebruikers. Waarvoor gebruiken zij een woordenboek? Uit onderzoek blijkt dat ze voornamelijk op zoek zijn naar de spelling of betekenis van een woord. Dat betekent automatisch dat een groot deel van de 'algemene' woordenschat nooit wordt opgezocht. Niemand zoekt naar de spelling of betekenis van een woord

als *kat*, *hond*, *de* of *en*, terwijl die woorden wel in alle handwoordenboeken – groot of klein – zijn opgenomen.

Uitgevers kunnen tegenwoordig gemakkelijk bijhouden welke woorden in digitale woordenboeken worden opgezocht. De woorden die gebruikers vergeefs opzoeken omdat ze nog ontbreken in het woordenboek, zijn een grote rol gaan spelen als bron van toevoegingen, terwijl de vaakst opgezochte woorden laten zien aan wat voor soort informatie gebruikers de meeste behoefte hebben. In Figuur 1.1 staan, in aflopende volgorde, de tien meest opgezochte woorden in de gratis Van Dale online, de Dikke Van Dale, het *Algemeen Nederlands Woordenboek* (ANW) en Woordenlijst.org (de officiële spellinglijst van het Nederlands, uitgegeven door de Taalunie).[14]

| Gratis Van Dale | Dikke Van Dale | ANW | Woordenlijst.org |
|---|---|---|---|
| braakliggend | kut | moshpit | willen |
| klaren | aardappel | ruwe bolster, blanke pit | updaten |
| initiatief | emoticon | n.t.b. | zijn |
| belangeloos | fuck | goedemiddag | creëren |
| conservatief | declareren | e.v. | kunnen |
| in | intern | zijn vruchten afwerpen | gebruikmaken |
| sex | entree | voor de kat z'n kut | hebben |
| imago | belminuut | goedemorgen | worden |
| traditiegetrouw | constructiefout | partir c'est mourir un peu | vinden |
| fonds | ad hoc | c'est le ton qui fait la musique | gaan |

Figuur 1.1 De meest opgezochte woorden in enkele moderne woordenboeken

De verschillen zijn opmerkelijk: in Van Dale worden woorden kennelijk opgezocht vanwege hun betekenis (*braakliggend*, *imago*), combinatiemogelijkheden (*in*), spelling (*ad hoc*, *emoticon*, *sex* – de officiële spelling is *seks*), en uit puberale opzoekwoede (*kut*, *fuck*). De informatie uit de Dikke Van Dale dateert uit 2003, en is een dagkoers. Volgens hoofdredacteur Ton den Boon is het aantal opgezochte schuttingwoorden in de top 10 van de Dikke Van Dale de afgelopen jaren sterk afgenomen ten gunste van zakelijke woorden als *paradigma*, *mitigatie*, *feedback*, *vermogensbeheerder* en

*interlocutoir*. Dat komt doordat de online Dikke Van Dale tegenwoordig vooral door het bedrijfsleven wordt geraadpleegd.

De top 10 van Woordenlijst.org bestaat uitsluitend uit werkwoorden: hier zoeken de gebruikers werkwoordsvervoegingen – waarvan Woordenlijst.org, anders dan gewone woordenboeken, een compleet overzicht geeft. In het ANW tot slot zoekt men vooral naar uitdrukkingen en afkortingen.

Het zijn ook de gebruikers die vragen om compleetheid van een woordenboek. Dat speelt met name in kringen van scrabbelaars. Zij willen discussies over de juistheid van een woord kunnen beslechten met een beroep op het woordenboek. In 2002 nam de Dikke Van Dale de handschoen op en publiceerde – ongetwijfeld mede ingegeven door commerciële motieven – de *Officiële woordenlijst voor Scrabble*. Ook die woordenlijst is, ondanks zijn 254.957 woorden van 2 t/m 9 letters, per definitie incompleet.

## Belangengroeperingen

Zeven eeuwen lang bepaalden deftige, witte mannen (en heel sporadisch een vrouw) de inhoud en toon van de woordenboeken, en schreven voor welke woorden je wel en niet kon of moest gebruiken. In de twintigste eeuw veranderde dat: voortaan lieten ze zich voor wat betreft hun opnamebeleid leiden door de taalgemeenschap en registreerden het woordgebruik zoals ze dat om zich heen hoorden en vooral lazen.

Dat bracht de taalgebruikers op een idee: als lexicografen het woordgebruik registreren, kun je beïnvloeden wat er in het woordenboek terechtkomt. Woordenboeken en hun schrijvers werden het mikpunt van actiegroepen. Die eisten dat betekenisomschrijvingen werden aangepast en woorden uit het woordenboek werden geschrapt, toegevoegd of de status van merknaam kregen.

Het begon in de jaren zestig van de vorige eeuw met een discussie rond de betekenisomschrijving van het woord *jood* als 'woekeraar, afzetter' en de opname van samenstellingen als *jodenstreek* en *jodenwoeker* in de Dikke Van Dale. Een deel van Joods Nederland raakte in rep en roer en eiste in een (verloren) kort geding aanpassing van het woordenboek. De

volgende decennia laaide de discussie als een veenbrand telkens opnieuw op. Taalkundigen vonden dat er sprake was van 'een restant van antisemitisme', terwijl de hoofdredactie als verdediging aanvoerde 'dat je met het schrappen van betekenissen en woorden de taalwerkelijkheid geweld aandoet'.[15] Onder druk van de media ging de uitgever toch overstag: er werden waarschuwingslabels toegevoegd en tientallen *joden*-samenstellingen werden geschrapt.

Producenten roken een kans: ook zij wilden een vinger in de woordenboekpap hebben. Zij beriepen zich op het merkenrecht en eisten, onder dreiging van een rechtszaak, dat aan bepaalde woorden de informatie werd toegevoegd dat het om een geregistreerde merknaam ® ging, ook al waren die merknamen in het dagelijkse spraakgebruik allang een soortnaam geworden. Iemand die een *luxaflex* bestelt, bedoelt zonwering binnenshuis, ongeacht van welk merk. De producenten wilden hun vaak tegen hoge kosten opgebouwd merk beschermen tegen het gebruik ervan door concurrenten en zagen het woordenboek bovendien als gratis reclame-uiting. En zo gelden in de Dikke Van Dale onder andere *biogarde, jiffypotje, leukoplast, maggi, muisjes, ranja* en *wokkels* als merknamen.[16] Om dat zo te houden, moeten de ondernemers wel werk verrichten: de overheid verlengt het merkenrecht na tien jaar namelijk alleen als ondernemers kunnen bewijzen dat ze hun merk beschermen tegen gebruik door anderen.[17]

In deze eeuw kwamen woorden als *negerzoen, moorkop, allochtoon, slaaf* en *Zwarte Piet* onder vuur te liggen. Het gold als politiek incorrect om deze woorden nog langer te gebruiken.[18] Dit leidde er niet toe dat ze werden geschrapt uit de woordenboeken, maar ze kregen wél een waarschuwingslabel. In de hoop het negatieve imago van bepaalde woorden af te schudden, werden telkens nieuwe eufemistische alternatieven in omloop gebracht. Vandaar de vele benamingen voor een niet-oorspronkelijke bewoner of werknemer: eerst *gastarbeider*, daarna *buitenlander, immigrant, allochtoon, medelander, nieuwe Nederlander, Marokkaanse/Turkse Nederlander* en tegenwoordig *iemand met een migratieachtergrond*. Al die namen staan in de woordenboeken, maar een nieuwe naam vijzelt helaas niet automatisch het imago op.

Ook op andere maatschappelijke veranderingen reageren de woordenboekschrijvers en -uitgevers. Zo hebben ze hun woordenboeken doorgelicht op vooroordelen en stereotypen. *Mensuren, menskracht* en *bemensing* werden als genderneutrale varianten opgenomen naast *manuren, mankracht* en *bemanning*. Voorbeeldzinnen als *zij is aan de afwas* en *hij is chirurg, evenals zijn vader* zijn gewijzigd in *hij is aan de afwas* en *zij is chirurg, evenals haar moeder*.

Aan persoonsaanduidingen zijn zoveel mogelijk vrouwelijke equivalenten toegevoegd, ongeacht of die in de praktijk daadwerkelijk gebruikt worden (*dominee – domina, lexicoloog – lexicologe, geograaf – geografe*). Nu bestaat er bij persoonsaanduidingen een verschil tussen het woordgeslacht en het natuurlijke geslacht of gender. *Dokter* en *burgemeester* zijn qua woordgeslacht mannelijke *de*-woorden, maar ze kunnen zowel naar een mannelijke als een vrouwelijke persoon verwijzen en, afhankelijk van dat natuurlijke geslacht, kun je zowel zeggen 'de dokter, *hij* liep daar' als 'de dokter, *zij* liep daar'. Hetzelfde verschil geldt voor *meisje*: het woordgeslacht is *het*, oftewel onzijdig, maar het natuurlijke geslacht is vrouwelijk (*zij*). In de digitale Dikke Van Dale (niet in de gedrukte) werd vanaf 2010 steeds vaker het natuurlijke geslacht toegevoegd aan het woordgeslacht: *burgemeester*, dat al sinds de dertiende eeuw het mannelijke woordgeslacht had, kreeg in de Dikke Van Dale de toevoeging m/v.

Ondertussen ontstond in de maatschappij een pleidooi voor het gebruik van genderneutrale voornaamwoorden als *hen* of *die* ('Morgen komt schrijver x. *Hen* of *die* geeft dan een lezing'), en voor een x-status als genderneutrale geslachtsaanduiding in officiële documenten als paspoorten. Dit pleidooi gaat over het natuurlijke geslacht en heeft niets te maken met het woordgeslacht. Daarom verraste het vriend en vijand dat de Dikke Van Dale in 2022 zo'n 15.000 persoonsaanduidingen voorzag van de aanduiding m/v/x. Hiermee is de vermenging van woordgeslacht en natuurlijk geslacht een feit. Het leidde onder taalkundigen tot discussies, en het blijft onduidelijk welk probleem hiermee is opgelost, want het woordgeslacht verhindert op geen enkele manier de keuze voor een genderneutraal voornaamwoord als *hen* of *die*. Bovendien ontbreekt de x bij woorden als *(het) bestuurslid, diensthoofd, jochie* en *meisje* of bij *(de) kraam-*

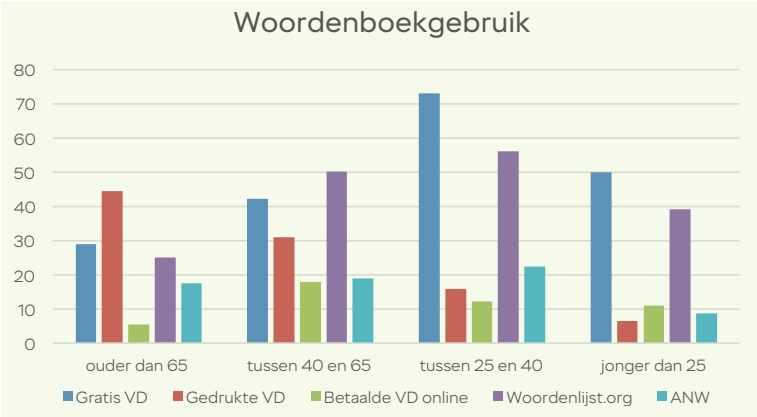
*vrouw, non, vader* en *vuilnisman*, suggererend dat naar deze woorden niet met een genderneutraal voornaamwoord kan of mag worden verwezen.[19] Misschien het meest opmerkelijke in deze kwestie is dat de uitgever niet het taalgebruik registreert, maar voor de troepen uit is gaan lopen.

## De digitale wereld

In deze eeuw worden de meeste woordenboeken niet langer gedrukt, maar digitaal aangeboden en geraadpleegd. Een gevolg daarvan is dat kwaadwilligen gemakkelijker informatie uit het woordenboek kunnen kopiëren. Om dat tegen te gaan voegen uitgevers spookwoorden met spookdefinities toe aan hun woordenboeken: woorden die niet bestaan en die, als ze ergens opduiken, afdoend bewijs zijn dat er sprake is van plagiaat. Welke die spookwoorden zijn is bedrijfsgeheim, maar van sommige woorden is dat inmiddels uitgelekt. Zo kende het Prisma woordenboek in het verleden de *lambadatol* ('tol die tijdens het draaien de lambada laat horen') en *gummelen* ('1. net op tijd de dans ontspringen, 2. wijn drinken'). De Dikke Van Dale voegde in 1999 *Detiger* toe, met de uitdrukking *stromen als de Rijn bij Detiger* ('zeer voortvarend en zonder scrupules te werk gaan'). Zowel *gummelen* als *Detiger* waren afgeleid van de achternaam van een medewerker van de uitgeverij.

Daarbij is het wel enigszins ironisch dat woordenboekschrijvers zich met spookwoorden tegen plagiaat beschermen, omdat juist zij eeuwenlang zwaar op voorgangers leunden en leunen. De Dikke Van Dale bevat nog steeds definities die op het WNT zijn gebaseerd. De bekende Franse schrijver en lexicograaf Charles Nodier wist het al in 1828: woordenboeken zijn 'plagiaat in alfabetische volgorde'.[20]

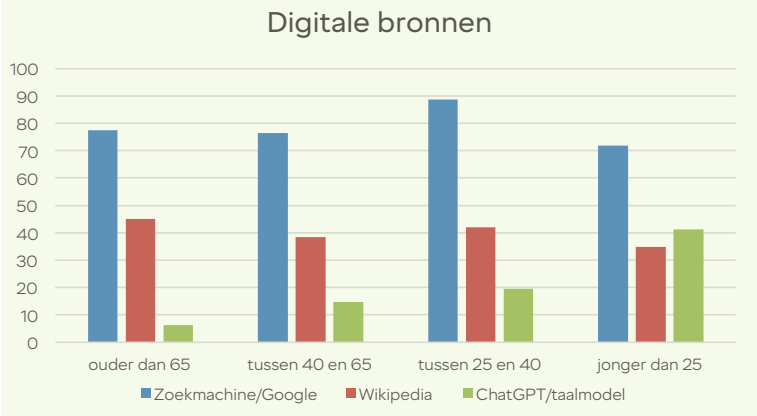
Ondertussen raadplegen gebruikers steeds vaker gratis online woordenboeken in plaats van gedrukte of betaalde online woordenboeken. Dat geldt voor de oudere generatie en nog veel meer voor de generaties daarna, zo blijkt uit Figuur 1.2, gebaseerd op een recente enquête naar woordenboekgebruik, ingevuld door 1424 personen.[21] De informanten konden meerdere bronnen aanvinken. De cijfers in de figuur zijn procentueel, om te compenseren voor het feit dat meer ouderen dan jongeren de enquête hebben ingevuld.




Figuur 1.2 Het gebruik van traditionele woordenboeken

In alle leeftijdscategorieën zijn de gratis Van Dale online (een uitgeklede versie van de Dikke Van Dale; donkerblauw) en Woordenlijst.org (de officiële spellinglijst; groen) het meest geraadpleegd – alleen mensen boven de 65 noemden vaker de gedrukte Dikke Van Dale (paars) en pas daarna deze twee bronnen. Het ANW (*Algemeen Nederlands Woordenboek*; lichtblauw) wordt het meest genoemd door de generatie tussen 25 en 40.

Maar nóg vaker gaan gebruikers te rade bij digitale bronnen, zo blijkt uit Figuur 1.3.




Figuur 1.3 Het gebruik van digitale bronnen

Onder alle leeftijdscategorieën is een zoekmachine als Google (blauw) het populairst. Op de tweede plaats staat Wikipedia/Wiktionary (rood), alleen onder de jongsten voorbijgestreefd door ChatGPT of een ander taalmodel (groen).

Dat woordenboeken tegenwoordig vooral online worden geraadpleegd, zal niemand verbazen. Maar uit de twee figuren blijkt dat gebruikers zich voor informatie niet langer direct wenden tot de bekende woordenboekuitgevers, maar vaker hun zoekopdracht intikken in een zoekmachine, een door vrijwilligers bijgehouden bron als Wikipedia, of een taalmodel. Dat geldt voor alle leeftijdscategorieën, al is het taalmodel onder ouderen het minst genoemd. De figuren laten zien dat de traditionele woordenboeken, geschreven door professionele lexicografen, op hun retour zijn en worden voorbijgestreefd door zoekmachines en taalmodellen.

Nu hebben zoekmachines en taalmodellen een belangrijke voorsprong op door lexicografen gemaakte woordenboeken: ze baseren zich op enorme hoeveelheden teksten en kunnen daardoor informatie – inclusief gebruiksvoorbeelden – geven over woorden die ontbreken in gedrukte of digitale woordenboeken. Toen ik ChatGPT vroeg wat de betekenis van *beginjanuariweer* is, gaf hij een foutloze en gedetailleerde definitie, die hij niet uit een bestaande bron heeft kunnen overschrijven, want dit woord komt in geen enkel woordenboek voor.

Via zoekmachines zoeken gebruikers losse woorden, en niet langer woorden in een lange alfabetische lijst – de manier waarop woorden zeven eeuwen lang waren gerangschikt. Het woord verliest zo zijn vertrouwde alfabetische omgeving. Een bijverschijnsel is dat veel jongeren de alfabetische volgorde niet meer kennen en moeite hebben een woord in een gedrukt woordenboek terug te vinden.

## Dijt de algemene woordenschat in woordenboeken uit?

Valt er iets te zeggen over de vraag of de algemene woordenschat die in woordenboeken is beschreven, in de loop van de tijd uitdijt? Eigenlijk niet, zolang er geen geaccepteerde en sluitende definitie bestaat van wat 'de algemene woordenschat' is. Wel bleek al dat woordenboekmakers in de loop van de tijd de definitie van algemene woordenschat hebben ver-

ruimd, doordat ze toeschietelijker werden in het opnemen van leenwoorden, vaktermen en 'onkiese' termen.

Een nóg recentere verruiming bestaat uit de systematische toevoeging van woorden die slechts in een deel van het Nederlandse taalgebied voorkomen. Belgisch-Nederlandse woorden als *aangroeipremie, bevallingsrust, croque-monsieur* en *gekend* ('bekend') werden al wat langer opgenomen en voorzien van het label BE. Uit het oogpunt van inclusiviteit zijn recent ook Surinaams-Nederlandse woorden als *buitenkind* ('buitenechtelijk kind'), *lawaaihemd* ('felgekleurd hemd') en *schaafijs* toegevoegd, met het label SR, en Antilliaans-Nederlandse woorden als *achterporch* ('veranda'), *basismand* ('eerste levensbehoeften, zoals melk, groente en fruit') en *speelschool* ('peuterspeelplaats'), met het label AN.

Sinds 2009 hebben dergelijke woorden bovendien een gelijkwaardige status gekregen. Voorheen golden de Nederlands-Nederlandse woorden stilzwijgend als 'normaal' of 'standaard', en werden alleen afwijkingen daarvan apart gelabeld. Tegenwoordig krijgen woorden die buiten Nederland onbekend zijn, het waarschuwingslabel NL. Van Dale vermeldt dat label bijvoorbeeld bij *bajes, gein, pinpas* en *reuring*.

Zo wordt de algemene woordenschat in het woordenboek dus uitgebreid met nieuwe soorten woorden. Als je daarbij bedenkt dat er wekelijks of zelfs dagelijks nieuwe woorden nodig zijn voor de onophoudelijke stroom aan maatschappelijke vernieuwingen, dan lijkt het vanzelfsprekend dat de woordenschat in het woordenboek groeit.

Er zijn echter tegenbewegingen. Ten eerste lukt het slechts een minderheid van de nieuwe woorden een felbegeerd plaatsje in het woordenboek te veroveren; de meeste neologismen zijn eendagsvliegen. Ten tweede verdringen nieuwe woorden regelmatig oudere woorden, die dan uit het woordenboek verdwijnen. De nieuwe en oude woorden zijn in principe communicerende vaten.

Maar het belangrijkste is het standpunt van de woordenboekuitgever; het kwam al ter sprake. Die bepaalt namelijk uiteindelijk hoe dik een woordenboek is en hoeveel woorden het bevat. Hij baseert dat op de doelgroep die hij voor ogen heeft en de kosten van het woordenboek. De omvang van een woordenboek zegt dus niets over de omvang van de alge-

mene woordenschat, maar is een doelbewuste keuze van de uitgever. Ter illustratie wordt in Figuur 1.4 het aantal trefwoorden in enkele woordenboeken vermeld.

| Dikke Van Dale | ANW[22] | Van Dale handwoordenboek Nederlands | Koenen woordenboek Nederlands | Prisma woordenboek Nederlands |
|---|---|---|---|---|
| 275.000 | 87.100 | 66.000 | 63.000 | 45.000 |

Figuur 1.4 Aantallen trefwoorden in enkele woordenboeken

De Dikke Van Dale is in dit verband een uitschieter: weliswaar is het woordenboek zijn leven begonnen als handwoordenboek, maar het is in de loop van anderhalve eeuw veranderd in een historisch woordenboek met als doel 'het hedendaagse Nederlands, met een terugblik op de voorbije 150 jaar' te beschrijven. Het gevolg is dat het van één deel is uitgedijd tot drie delen, met ca. 275.000 trefwoorden. Daaraan heeft het de bijnaam de *Dikke* te danken. De toename komt dus vooral doordat het woordenboek een steeds langere tijdsperiode beschrijft. Die omvang heeft het woordenboek ook autoriteit verleent, vandaar de scheidsrechterfunctie die het krijgt toebedeeld.

# 2 Woorden in een corpus

Geen enkel woordenboek bevat een compleet overzicht van al onze woorden, zo bleek in het vorige hoofdstuk. Tegenwoordig beschikken we over enorme tekstcorpora: digitale bestanden van literaire, wetenschappelijke en journalistieke werken. Bevatten die dan de complete Nederlandse woordenschat? Nee, dat is niet het geval, want woorden die specifiek gebruikt worden in de sociale media, op boodschappenbriefjes of sinterklaasgedichten zijn hierin bijvoorbeeld niet opgenomen. Wél vinden we in een corpus de woorden in hun natuurlijke omgeving en niet in de door mensen opgelegde alfabetische volgorde van het woordenboek. Welke en hoeveel woorden er in een corpus staan, kan de computer in een handomdraai uitrekenen. En dat leidt tot verrassende inzichten. Zo komen sommige woorden veel vaker voor dan andere, en dat blijkt geen toeval te zijn.

## Woordjes tellen

Het tellen van woorden lijkt een simpele exercitie, maar in de praktijk valt dat vies tegen. Want wát tel je eigenlijk: wat is een 'woord' in een tekstcorpus? Dat zijn niet alleen de trefwoorden uit een woordenboek, maar ook alle vervoegde en verbogen vormen, zoals meervoudsvormen en verkleinvormen van zelfstandige naamwoorden, en werkwoordvervoegingen. Daarnaast bevat een corpus veel elementen die ontbreken in een woordenboek omdat ze niet vallen onder wat een lexicograaf onder 'woord' verstaat, zoals getallen en leestekens (komma's, aanhalingstekens, gedachtestreepjes), de meeste namen van personen, plaatsen, bedrijven, instellingen en merken, veelgemaakte spelfouten (*zowiezo, produkt, asterix*) en jonge spelvarianten uit het chatverkeer (*w8, suc6, geweldiggg, goehoedd, leueueuk* en *leukkkk*).

De computer maakt zich niet druk om het onderscheid tussen een woordvorm, leesteken of cijfer. Hij telt mechanisch alles wat tussen twee spaties of een spatie en een leesteken staat en berekent hoe vaak elk zo'n

stukje tekst in het corpus voorkomt. Dat stukje tekst tussen twee spaties noemen we een *token*, omdat 'woord' of 'woordvorm' de lading niet dekt. De computer telt heel consequent: een afkorting als *p.s.* rekent hij als één token, volgens de regel dat alle symbolen, letters, cijfers en leestekens tussen twee spaties als token moeten worden opgevat, maar *Bakker* en *bakker* – met en zonder hoofdletter – vat hij als twee verschillende woorden op.

Dat mechanische tellen leidt, verrassend genoeg, niet tot een eenduidig resultaat. Zo blijken Microsoft Word, OpenOffice en Google Docs verschillende totalen op te geven als je ze voert met hetzelfde bestand. Het ene programma telt een leesteken als apart woord, het andere niet. Het ene beschouwt *secretaris-generaal* en *'s-Gravenhage* als één woord, het andere telt ze voor twee.

Ondertussen kan de computer wel razendsnel van ieder corpus berekenen welke tokens het bevat en hoe vaak die voorkomen. Die tellingen kunnen inhoudelijke verschillen tussen corpora aan het licht brengen: een roman bevat andere woorden, in andere verhoudingen, dan een kookboek of een medisch handboek. Maar wat blijkt als je de tokens sorteert op frequentie? De top 20, 30, 40 van de frequentste tokens is voor ieder corpus min of meer gelijk, alleen de onderlinge rangorde verschilt enigszins.

Wat zijn die frequentste tokens? Als voorbeeld dienen drie krantencorpora: de Nederlandse kranten uit het *Corpus Hedendaags Nederlands* (CHN), en de Vlaamse kranten *Het Nieuwsblad* en *Wablieft*. *Wablieft* is een wekelijkse krant die speciaal is geschreven in makkelijk te lezen Nederlands, terwijl *Het Nieuwsblad* een algemene, landelijke krant is. De kranten zijn dus geschreven voor verschillende doelgroepen en regio's en ze hanteren een verschillend taalgebruik. Maar dat blijkt niet uit de hoogstfrequente woorden, zo is te zien in Figuur 2.1; weliswaar zijn er verschillen in rangorde, maar die duiden niet op een wezenlijk verschil. Een enkel woord (*hebben, worden, ook*) komt slechts in de top 20 van één van de drie kranten voor, maar is er dan in de andere kranten nét buiten gevallen.

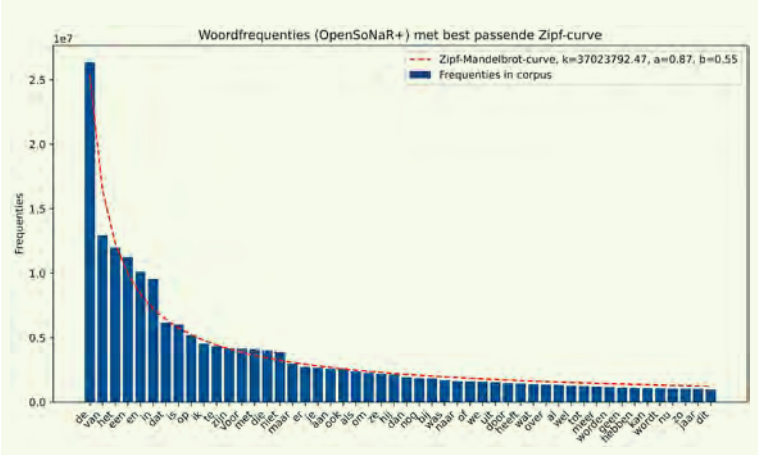
| Rangorde | Kranten in het CHN | Het Nieuwsblad | Wablieft |
|---|---|---|---|
| 1 | de | de | de |
| 2 | van | het | zijn |
| 3 | het | een | in |
| 4 | een | zijn | een |
| 5 | in | van | het |
| 6 | en | in | van |
| 7 | dat | en | dat |
| 8 | op | op | en |
| 9 | is | dat | op |
| 10 | te | te | veel |
| 11 | zijn | voor | voor |
| 12 | met | met | met |
| 13 | voor | we | niet |
| 14 | die | hebben | die |
| 15 | niet | worden | ik |
| 16 | ik | niet | je |
| 17 | maar | die | ze |
| 18 | aan | ik | er |
| 19 | er | er | ook |
| 20 | ook | om | hij |

Figuur 2.1 De 20 frequentste woorden in verschillende kranten[23]

## Woordjes samennemen

Als je wilt uitzoeken in welk opzicht corpora van elkaar verschillen, vormen de hoogfrequente woorden een hinderpaal. In plaats van alle woordvormen apart te tellen, telt men daarom identieke woorden slechts eenmaal: de vele woordvormen of *tokens* worden zo gereduceerd tot veel minder *types* (verschillende woordvormen). Een zin als *ik zag een kat die een andere kat had gezien* bevat tien tokens, maar twee daarvan (*een* en *kat*) komen tweemaal voor; het aantal types is dus slechts acht.

Al in 1949 heeft de Amerikaanse taalkundige George Kingsley Zipf geconstateerd dat de woordfrequentie van een taal – van iedere onderzochte taal! – een wetmatigheid vertoont: het meest voorkomende woord (token) komt ongeveer twee keer zo vaak voor als het tweede frequentste woord, ongeveer drie keer zo vaak als het derde frequentste, ongeveer vier keer zo vaak als het vierde frequentste, enzovoort.[24] Deze 'wet van Zipf' zegt dus dat de frequenties van woorden (types) in de tabel dalen naarmate hun rang in de tabel toeneemt, ofwel: de frequenties van woorden zijn omgekeerd evenredig aan hun rangorde. Alleen in de bovenste rangorde van de tabel zijn er behoorlijk wat uitzonderingen. Dat levert voor alle omvangrijke tekstbestanden een vaste distributie op zoals in Figuur 2.2, berekend volgens een verbeterde formule van de wiskundige Benoit Mandelbrot, met op de x-as de rangorde in de frequentielijst en op de y-as de relatieve frequentie van de woorden: in blauw de daadwerkelijke frequentieverdeling in het corpus, in rood de volgens Zipf verwachte relatieve frequentie.




Figuur 2.2 Top van de frequentielijst van het OpenSoNaR-corpus, met in rood de Zipf-curve (een deel van een zeer lange curve, berekend over meer dan 30 miljoen tokens)

Volgens de wet van Zipf vertegenwoordigen de tien hoogstfrequente types 25 procent van een (omvangrijk) corpus en honderd types representeren maar liefst 50 procent van het corpus. Daartegenover staat dat meer dan de helft van de types slechts eenmaal in het corpus voorkomt.[25] Een tweede wetmatigheid is dat de frequentste woorden in een taal het kortst zijn.[26] Zipf verklaart dit uit het feit dat sprekers en toehoorders zo efficiënt en economisch mogelijk willen communiceren: hoe korter de hoogfrequente woorden, hoe sneller de communicatie.

Nu valt er nóg iets op aan de tokens in Figuur 2.1 en 2.2, namelijk dat het vooral gaat om functiewoorden: woorden die geen eigen betekenis hebben, maar de betrekkingen tussen zinsdelen uitdrukken, zoals lidwoorden (*de, het, een*), voornaamwoorden (*ik, we, ze, zij*), voorzetsels (*in, van, op, voor, om*) en voegwoorden (*en, maar*). Verder zijn volgens de figuren de hulpwerkwoorden *zijn, hebben* en *worden* hoogfrequent (*zijn* is trouwens ook een voornaamwoord). In de top 20 staat geen enkel zelfstandig of bijvoeglijk naamwoord, en ook tussenwerpsels (zoals *ach, doeg, enfin, ja, of zo, zeg maar*)[27] en telwoorden ontbreken.

Opmerkelijk is nu dat het Nederlands – net als andere talen – slechts een heel beperkt arsenaal aan functiewoorden heeft in vergelijking met andere woordsoorten. Figuur 2.3 toont de percentuele verschillen tussen de woordsoorten in het Nederlands, gebaseerd op Molex (een digitale woordenlijst van 223.153 moderne trefwoorden, samengesteld door het Instituut voor de Nederlandse Taal), en het percentage in het OpenSoNaR-corpus met 554 miljoen tokens uit gedrukte media en nieuwe media van rond 2010.[28]

| Woordsoort | Percentage types in Molex | Percentage types in OpenSoNaR | Percentage tokens in OpenSoNaR |
|---|---|---|---|
| Zelfstandige naamwoorden | 82,43 | 40,98 | 16,48 |
| Bijvoeglijke naamwoorden | 8,51 | 6,24 | 6,40 |
| Werkwoorden | 6,33 | 7,65 | 13,56 |
| Bijwoorden | 0,99 | 0,42 | 5,82 |
| Eigennamen | 0,72 | 39,09 | 6,83 |
| Telwoorden | 0,45 | 5,29 | 2,49 |
| Tussenwerpsels | 0,32 | 0,05 | 0,31 |
| Voorzetsels | 0,12 | 0,11 | 11,10 |
| Lidwoorden en voornaamwoorden | 0,10 | 0,12 | 17,45 |
| Voegwoorden | 0,03 | 0,04 | 3,97 |
| Leestekens | -- | 0,01 | 15,59 |
| **Totaal** | **100%** | **100%** | **100%** |

Figuur 2.3 Percentages woordsoorten in Molex en het OpenSoNaR-corpus

De functiewoorden (lidwoorden, voornaamwoorden, voorzetsels en voegwoorden) vormen de kleinste woordsoortengroep in types, maar juist die woorden komen in een corpus het vaakst voor, zo blijkt uit het percentage van de tokens in het OpenSoNaR-corpus, en overtreffen in aantal tokens de naamwoorden en werkwoorden. Dat komt door de grammaticale rol van functiewoorden, die ze onmisbaar maakt in een Nederlandse zin. Opvallend is het grote aantal eigennamen in een corpus: een categorie die in woordenboeken systematisch wordt genegeerd, behalve als ze een eigen betekenis hebben gekregen, dus soortnaam zijn geworden, zoals *badminton*, *bintje* of *jack russell*.

## Type/token-verhouding

De verhouding tussen types en tokens verschilt per corpus, en dat verschil kan veelzeggend zijn. Figuur 2.4 toont bijvoorbeeld het aantal types, tokens en de verhouding daartussen in kranten die begin zeventiende eeuw in de Republiek (Couranten Corpus) en de Zuidelijke Nederlanden (Nieuwe Tijdinghen) zijn verschenen, en kranten uit het jaar 2023. De

type/token-verhouding is maximaal 1, namelijk als ieder token een apart type vormt; dat komt alleen in heel korte zinnen voor.

| Bron | Couranten Corpus 1620-1621 | Nieuwe Tijdinghen 1605-1629 | Kranten 2023 |
|---|---|---|---|
| Totaal aantal woorden (tokens) | 106.023 | 105.907 | 107.402 |
| Aantal verschillende woorden (types) | 11.469 | 14.544 | 19.740 |
| Type/token-verhouding | 0,11 | 0,14 | 0,18 |

Figuur 2.4 De type/token-verhouding in drie krantencorpora[29]

Wat opvalt in Figuur 2.4 is dat de type/token-verhouding in moderne kranten veel hoger ligt dan in de oude kranten. Nu bestond er in het verleden nog geen officiële spelling, waardoor er in oude teksten veel spelvarianten voorkomen (*paert, paerd, peert, paart, paard*), die allemaal als aparte types tellen. Daarom verwacht je dat het aantal types in de oude kranten hoger ligt dan in de moderne, terwijl het omgekeerde het geval is. Hoe dat komt, blijkt als je de kranten preciezer bekijkt. Dan komt naar voren dat in moderne kranten het nieuws draait om persoons-, plaats-, merk- en bedrijfsnamen. Daarvan staan er massa's in de krant en ze gelden allemaal als apart type. In de oude kranten spelen dergelijke namen een veel kleinere rol: zo komen daarin geen merk- en bedrijfsnamen voor. Daarnaast bevatten moderne kranten een groot aantal nieuwe termen als gevolg van technische en wetenschappelijke ontwikkelingen sinds met name de negentiende eeuw (zie Figuur 3.3).

De type/token-verhouding wordt gehanteerd als maat voor de woordenrijkdom van een tekst. Die lexicale diversiteit levert interessante gegevens op. Zo scoort een roman in lexicale diversiteit hoger dan een prentenboek of een kookboek: de roman is rijker aan woorden en synoniemen, maar daardoor ook moeilijker leesbaar. Hoevéél rijker hangt uiteraard af van de auteur. Uit onderzoek is gebleken dat de woordenschat van Nederlandse rappers – anders dan velen denken – even rijk is als die van literaire schrijvers. De meeste verschillende woorden worden

gebruikt door, in aflopende volgorde, Ilja Leonard Pfeijffer, Zo Moeilijk, Harry Mulisch, Louis Couperus en Opgezwolle.[30]

Er zijn sceptici die het tellen van woordjes maar triviaal vinden. Maar deze voorbeelden laten zien dat tellingen ertoe leiden dat je op een nieuwe manier naar gegevens kijkt en zo tot nieuwe inzichten kunt komen.

## Woordvormen samennemen

De meeste gebruikers die informatie in een corpus zoeken, zijn niet op zoek naar één specifieke woordvorm, maar juist naar alle mogelijke varianten. Ze willen met één zoekopdracht alle vervoegingen van *kunnen* vinden, in plaats van dat ze zelf moeten bedenken welke vormen allemaal mogelijk zijn (*kun, kan, kon, gekund,* ...). Als hulpmiddel hebben computerspecialisten daarom programma's geschreven die aan iedere woordvorm de basisvorm toevoegt, wat lexicografen de *trefwoorden* of *lemma's* noemen. De zin *ik zag een kat die een andere kat had gezien*, bevat zeven lemma's: *ik, zien, een, kat, die, ander, hebben*.

Voor dit lemmatiseren, zoals de technische term luidt, zijn de digitale woordenboeken die in hoofdstuk 1 aan de orde kwamen, een belangrijk hulpmiddel, maar ze zijn niet voldoende, omdat woordenboeken per definitie incompleet zijn. Het programma concentreert zich daarom op het laatste deel van een woord, en kan zo ook niet eerder aangetroffen samenstellingen of afleidingen analyseren. Een woord als *nijlpaardjongenpierenbadjes* ('pierenbadjes voor baby-nijlpaarden'), dat in geen enkel woordenboek voorkomt, lemmatiseert hij als *nijlpaardjongenpierenbad*. Scheidbare werkwoorden in een zin als *hij bood een cadeau aan* leveren nog wel problemen op, maar dankzij zelflerende systemen (machine-learning) wordt daarin vooruitgang geboekt.

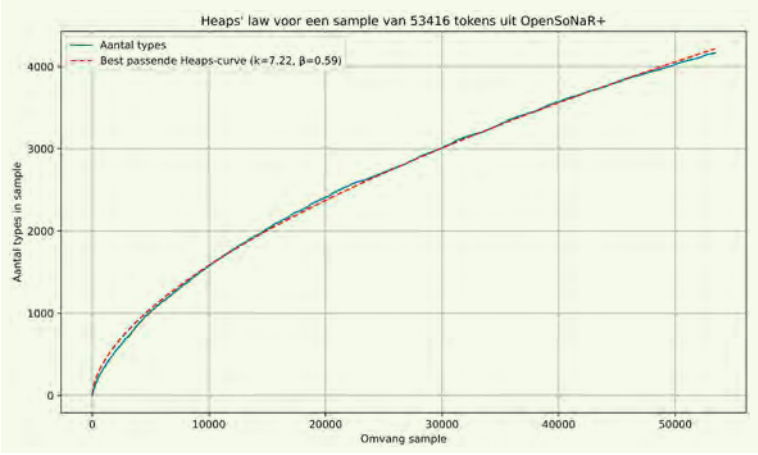
Bij het lemmatiseren kijkt het computerprogramma alleen naar de vorm van een woord, niet naar de betekenis. Het maakt geen verschil tussen het zelfstandige en het bijvoeglijke naamwoord *arm* of *rijk*. Als je dus op zoek bent naar informatie over de historische rijken of staten die in Europa in de loop van de tijd hebben bestaan, moet je eindeloze hoeveelheden voorbeelden overslaan waarin *rijk* de betekenis 'vermogend' heeft.

Ook voor dat probleem hebben de computerspecialisten een list bedacht. Ze hebben programma's ontworpen die automatisch aan iedere woordvorm een woordsoort toekennen, in het Engels *Part-of-Speech tagging* of kortweg *POS-tagging* genaamd. Daarvoor kijkt het programma over de woordgrenzen heen en maakt het gebruik van taalkundige regelmatigheden.[31] Zo'n regelmatigheid is: ken de woordsoort bijvoeglijk naamwoord toe aan *arm* als dat wordt voorafgegaan door een lidwoord en wordt gevolgd door een zelfstandig naamwoord (*de arme man, een arm mens*). Een andere regelmatigheid is: ken altijd de woordsoort zelfstandig naamwoord toe aan de vorm *armen*. In die laatste regel is echter niet verdisconteerd over welke betekenis het gaat: de woordsoort is in dit geval niet voldoende, vergelijk 'de armen bungelen aan zijn lichaam' met 'de armen gaan zonder eten naar bed'. Hier gaat voorlopig nog de eerste wet van de 'computer-taalkunde' op van de bekende taalkundige Hugo Brandt Corstius in zijn gelijknamige boek uit 1978: 'Wat je ook doet, de semantiek (betekenis) gooit roet.' Een computer kan geen chocola maken van de oude grap 'geld moet je onder de armen verdelen, daar houd je warme oksels bij'.

## Is het aantal woorden in een corpus eindig?

Uit tekstcorpora kunnen we veel informatie over woorden halen die niet in woordenboeken staat, zoals hoe vaak een woord voorkomt, met welke andere woorden het wordt gecombineerd, en hoe woordsoorten en woordlengte in teksten zijn verdeeld. Deze soorten informatie kun je gebruiken om teksten uit verschillende genres en tijden met elkaar te vergelijken, en zo te achterhalen welk woordgebruik kenmerkend is voor genres en tijdsperiodes. Daarvoor moeten de corpora wel groot genoeg zijn: de wet van Zipf toonde immers dat meer dan de helft van de woorden in het corpus maar eenmaal voorkomt, terwijl je, om zinnige conclusies te kunnen trekken, zoveel mogelijk woorden nodig hebt met een hogere frequentie dan één. Daarom is een corpus van 1 miljoen tokens eigenlijk al te klein; dat is ongeveer het aantal woorden dat in de Bijbel staat. Hoe groot een ideaal, betrouwbaar en representatief corpus moet zijn, hangt af van wat je erin wilt opzoeken, maar de meeste computerspecialisten zijn van oordeel dat het beste corpus een groter corpus is.

In het begin van dit hoofdstuk schreef ik dat er op dit ogenblik geen corpus bestaat van de complete Nederlandse woordenschat. Voor onderzoek naar veranderingen in taal, cultuur, maatschappij en wetenschap zou dat wel ideaal zijn. Maar hoewel er diverse belangrijke en omvangrijke Nederlandse corpora beschikbaar zijn (zoals CHN, OpenSoNaR en het Couranten Corpus), is er veel meer níét dan wél gedigitaliseerd. Stel nu echter dat de hele Nederlandse tekstproductie van de tiende eeuw tot heden, inclusief de inhoud van sociale media, zou zijn gedigitaliseerd, zouden we dan alle mogelijke Nederlandse woorden hebben gevangen? Volgens de wet van Heaps is het antwoord nee. Die wet van Harold Stanley Heaps uit 1960 stelt dat hoe groter het corpus is, hoe kleiner de type/token-verhouding wordt, anders gezegd: hoe groter het corpus wordt, hoe langer het duurt voordat je een nieuw type tegenkomt; zie Figuur 2.5.




Figuur 2.5 Een willekeurig sample van ruim 50.000 tokens uit OpenSoNaR, waaruit in overeenstemming met Heaps' wet blijkt dat het aantal nieuwe types afneemt naarmate het corpus groter wordt: de curve stijgt steeds minder snel (de rode stippelijn geeft de best passende theoretische curve weer)

Volgens diezelfde wet van Heaps wordt de type/token-verhouding nooit nul. Dat houdt in dat de woordenschat van een taal theoretisch oneindig is. Kernwoord is hier 'theoretisch': een dergelijke omvang gaat namelijk het menselijke voorstellingsvermogen verre te boven. En 'type' valt

in deze wet niet helemaal samen met wat lexicografen verstaan onder 'woord': een corpus bevat namelijk niet alleen woorden, maar ook getallen, leestekens, symbolen, tikfouten (*tto, boederij* in plaats van *tot* en *boerderij*), computerleesfouten en in het algemeen heel veel vervuiling en ruis, zoals blijkt uit deze willekeurige krantentekst uit 19-11-1650 op de online databank Delpher:

> 'jjr ff tCa talon (en Heeft men / Dat ficf) ïjet 1 |jlot -fiitr acn De Heeft S 7 ober-gegebe» / Daer npr 400 kranten nael)aec3£rmaDe/ Die niet ober 6000 tcseerbe enDe te boete dereb is getroeben 3i(n/ tiaec op Dé JBartftgraef ban|©o?ta;ia (jet 48** beden ooeit boetserom-gt-tcocctipeeet enDe lètffloet belegert Heeft/ toaer Doe? Co?tofa foo beeialsgeblotipiemts.

# 3 Woorden door de tijd

‘In den beginne was het Woord’, aldus het eerste hoofdstuk van het Bijbelse Evangelie volgens Johannes. We zouden natuurlijk dolgraag willen weten wat het eerste woordje van de mensheid was, maar daar zullen we nooit achter komen. Ook over hoe woorden en taal zijn ontstaan en of er ooit één oertaal was, tasten we in het duister. We kunnen het taalgebruik niet verder terug in de tijd reconstrueren dan tot ongeveer 5000 voor Christus. Toen sprak een groep mensen op de steppen ten noorden van de Zwarte Zee een taal die we nu Indo-Europees noemen. Die groep verspreidde zich over Europa en India. Een subgroep bevolkte Denemarken en de Duitse kust, en hun taal ontwikkelde zich daar tot het Germaans, de moedertaal van het Nederlands, Engels, Duits, Fries en de Scandinavische talen.[32]

De basiswoordenschat van het Nederlands is dus geërfd uit het Germaans. Dergelijke erfwoorden duiden de alledaagse belevingswereld van de mens aan. Het gaat bijvoorbeeld om namen voor familieleden, lichaamsdelen, dieren, planten, en natuur- en weersverschijnselen. Naarmate de maatschappij, techniek en wetenschap zich ontwikkelden, konden de taalgebruikers natuurlijk niet meer toe met de Germaanse erfwoorden, zelfs als ze de betekenis en het gebruik daarvan uitbreidden, bijvoorbeeld door *voet* voor ‘onderste deel’ te gebruiken (*aan de voet van de berg*), *hart* voor ‘midden’ (*in het hart van de winter*) of *kop(je)* voor ‘drinkgerei’. Er waren nieuwe woorden nodig voor nieuwe verschijnselen. Hoe kwamen ze aan nieuwe woorden? En kunnen we inschatten om hoeveel woorden het in de loop van de tijd gaat?

## Bronnen voor nieuwe woorden

Van oudsher wordt op twee manieren voldaan aan de behoefte aan nieuwe woorden: door nieuwvormingen op basis van erfwoorden en door ontleningen aan andere talen. Verreweg de grootste groep nieuwvormingen bestaat uit samenstellingen en afleidingen; daarover meer in het volgende hoofdstuk.

Theoretisch zouden taalgebruikers ook een nog niet gebruikte klankvorm kunnen kiezen om een nieuwe zaak mee te benoemen. Daarvan zijn er voldoende. Zo zijn *laam, lem, lim, liem, loem, lum* of *luum* nog beschikbaar, terwijl *lam* en *luim* al zijn gereserveerd voor 'jong schaap' en 'humeur'. De genoemde vormen druisen niet in tegen de klankregels van het Nederlands, anders dan vormen als *tlam, rlam* of *dlam*. Toch benoemen taalgebruikers een nieuwe zaak maar heel zelden met een nieuwe klankvorm, zoals de Belgische striptekenaar Peyo deed met *smurf*, Marten Toonder met *kukel* (*minkukel*) en de Amerikaanse schrijver Gelett Burgess met *blurb*. Alleen voor merknamen kiezen bedrijven juist heel bewust voor nieuwe klankvormen, zodat hun merknaam uniek is en in de digitale wereld gemakkelijk wordt gevonden.

De reden dat taalgebruikers geen compleet nieuwe klankvormen gebruiken, zal te maken hebben met de beperkingen van het geheugen: het begrijpen en onthouden van een nieuwe klankvorm belast het geheugen en vergt ook extra inspanning in de communicatie, want de spreker moet aan de hoorder duidelijk maken wat hij bedoelt. In principe geldt dat ook voor leenwoorden, maar die zijn vaak net als inheemse samenstellingen geconcentreerd in inzichtelijke woordfamilies. Bovendien komen ze als pakketje van vorm én betekenis een taal binnen, wat een andere situatie is dan het zoeken naar een geschikte vorm voor een nieuwe betekenis. Dat het op zich niet moeilijk is om nieuwe klankvormen te bedenken, blijkt uit de mooie beslisboom die Katinka Polderman in *de Volkskrant* van 25-10-2024 publiceerde, zie Figuur 3.1.

Wel worden van oudsher nieuwe woorden gevormd op basis van klanknabootsing. Bij klanknabootsingen heet een zaak naar het geluid dat het maakt: er bestaat dus een directe relatie tussen de vorm en de betekenis. Daarom zijn ze gemakkelijk te begrijpen en te onthouden, anders dan vormen als *laam* en *liem*. De oudste klanknabootsingen zijn tussenwerpsels; die worden zelfs wel als (een) bron van het ontstaan van taal beschouwd. Denk aan *haha, hopla, hup, plons, pats*. Klanknabootsende dierennamen zijn *grutto, koekoek, roek, tureluur*, en *krekel, mug, tor*. Daarnaast bestaan er veel klanknabootsende werkwoorden, zoals *blaffen, janken, kletsen, krijsen, loeien, piepen* en *zuchten*.[33]


Figuur 3.1 Beslisboom van Katinka Polderman in *de Volkskrant* van 25-10-2024 met nieuwe klankvormen

## Hoeveel nieuwe woorden?

In alle eeuwen zijn nieuwe woorden – nieuwvormingen, klanknabootsingen en leenwoorden – in het Nederlands geïntroduceerd. Om hoeveel woorden gaat het eigenlijk? Daarvoor kunnen we te rade gaan bij de (historische) woordenboeken die het Nederlands van de oudste tijd tot heden beschrijven. Als je het aantal trefwoorden in de verschillende periodewoordenboeken telt, krijg je een indruk van de omvang van de vroegere woordenschat; zie Figuur 3.2.

| Woordenboek | Periode | Trefwoorden | Betekenissen | Citaten |
|---|---|---|---|---|
| *Oudnederlands Woordenboek* (ONW) | 700-1200 | 9.000 | 12.619 | 30.025 |
| *Vroegmiddelnederlands Woordenboek* (VMNW) | 1200-1300 | 26.000 | 102.202 | 194.366 |
| *Middelnederlandsch Woordenboek* (MNW) | 1200-1500 | 75.000 | 144.714 | 400.619 |
| *Woordenboek der Nederlandsche Taal* (WNT) | 1500-1976 | 400.000[34] | 553.672 | 1.667.835 |
| *Algemeen Nederlands Woordenboek* (ANW) | 1976-heden | 87.100 | 46.696[35] | 275.233 |
| *Woordenboek van Nieuwe Woorden* (WNW) | 21ste eeuw | 4.380 | 4.710 | 17.034 |
| **Totaal** | | **601.480** | **864.613** | **2.585.112** |

Figuur 3.2 Aantallen trefwoorden, betekenissen en citaten in de historische woordenboeken van het Nederlands

Uit Figuur 3.2 zou je kunnen opmaken dat de woordenschat van na 1500 vier à vijf keer zo groot is als die in de middeleeuwen. Maar schijn bedriegt. Om te beginnen sluiten de woordenboeken in tijd niet netjes op elkaar aan, maar overlappen elkaar gedeeltelijk. Dat geldt bijvoorbeeld voor het *Vroegmiddelnederlands* en *Middelnederlandsch woordenboek*, en voor het *Algemeen Nederlands Woordenboek* en het *Woordenboek van Nieuwe Woorden*.

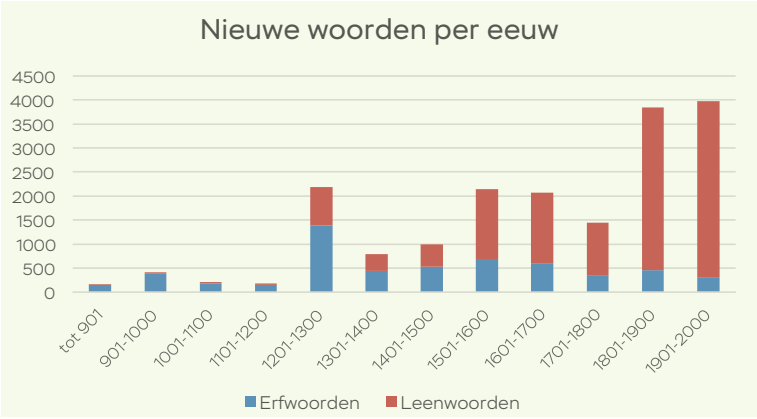
Er is nog een andere reden waarom de beschreven periodes niet goed met elkaar te vergelijken zijn. In het Oudnederlands zijn nauwelijks teksten geschreven: maar weinig mensen konden schrijven en áls ze schreven, was

dat in het Latijn. Veel woorden die toen wel bestonden, zijn dus niet genoteerd. Tussen 1200 en 1500 nam de tekstproductie wel toe, maar omdat de boekdrukkunst pas rond 1450 werd uitgevonden, werden de manuscripten slechts op kleine schaal gekopieerd en verspreid, en daardoor is veel verloren gegaan. Zo berekenden Mike Kestemont en Folgert Karsdorp onlangs dat er ongeveer 150 Middelnederlandse ridderromans moeten zijn geweest, waarvan slechts ongeveer de helft is overgebleven.[36] En dat zijn dan bijzondere, literaire teksten. Van alledaagse briefjes aan gemeentebesturen, buren of familieleden is sowieso nauwelijks iets bewaard gebleven.

Daarentegen is het WNT gebaseerd op gedrukte werken in grotere oplagen, waardoor veel minder verloren is gegaan. Bovendien is aan het WNT bijna anderhalve eeuw gewerkt, veel langer dan aan de andere woordenboeken, en dus kon er voor dit woordenboek ook veel meer materiaal worden verzameld: dat blijkt uit het grote aantal citaten en betekenissen.

En dan moet er nóg een voorbehoud worden gemaakt. Het is beslist niet zo dat álle trefwoorden die in een bepaald periodewoordenboek zijn genoemd, in die hele periode daadwerkelijk door alle taalgebruikers gebruikt werden of algemeen bekend waren. Veel woorden uit de periodewoordenboeken kwamen slechts in een korte periode of in een beperkt deel van het Nederlandse taalgebied voor. Zo gold *room* eind zestiende eeuw als 'Hollands en Vlaams', terwijl *zaan* in Brabant werd gezegd. Een woordenboek noemde eind zestiende eeuw *kaakappel* als verouderd woord voor 'wang'. Het woord verdween vrijwel direct. En met dit woord verdwenen er in de loop van de tijd nog veel meer, maar we hebben geen idee om hoeveel het gaat.

In plaats van naar het aantal trefwoorden in de periodewoordenboeken te kijken, kun je beter bekijken wanneer woorden hun intrede deden. Die informatie staat in het *Chronologisch woordenboek*, dat ik in 2001 publiceerde. Daarin is bij 18.540 trefwoorden (zowel geleende als niet-geleende) het jaartal van eerste voorkomen toegevoegd, gebaseerd op historische corpora en op de bovengenoemde woordenboeken. Tevens is bij ieder trefwoord genoteerd of het een erfwoord of afleiding daarvan is (inclusief klanknabootsingen), dan wel een leenwoord. Op Figuur 3.3. staat het aantal nieuwe woorden per eeuw.




Figuur 3.3 Aantal nieuwe woorden per eeuw uit het *Chronologisch woordenboek*

Uit Figuur 3.3 komt naar voren dat er in sommige eeuwen veel meer nieuwe woorden bijkwamen dan in andere. De woordenschat dijt in de loop van de tijd kennelijk niet exponentieel maar sprongsgewijs uit. De toppen kunnen rechtstreeks worden gerelateerd aan maatschappelijke ontwikkelingen. Zo werd in de dertiende eeuw voor het eerst op grote schaal in het Nederlands geschreven, en dat levert nieuwe woorden op als *avontuur*, *fabel*, *historie*, *klerk*, *kopie*, *poëet* en *rijm*. In de zestiende en zeventiende eeuw bloeiden kunst en wetenschap dankzij de renaissance. Dat bracht neologismen als *blijspel*, *opera*, *schouwburg*, *alchemie*, *algebra*, *tumor* en *tyfus*. In deze periode groeiden economie en handel dankzij overzeese reizen, waardoor veel nieuwe producten bekend werden als *ananas*, *maïs*, *tabak* en *tomaat*. In de negentiende en twintigste eeuw maakten techniek en wetenschap enorme sprongen voorwaarts, met nieuwe woorden als *accu*, *batterij*, *dieselmotor*, *electricien*, *fiets*, *monteur* en *psychiater*.

## Hoeveel leenwoorden?

Uit Figuur 3.3 blijkt ook dat leenwoorden belangrijk zijn voor de uitbreiding van de woordenschat: vanaf de zestiende eeuw overtreft het aantal leenwoorden dat van de erfwoorden. Hierbij moet wel bedacht worden dat het *Chronologisch woordenboek* is samengesteld voor etymologisch onderzoek, waardoor het relatief weinig samenstellingen en afleidingen

bevat. Waren die wel opgenomen, dan zou de verhouding erfwoord/leenwoord zeker anders liggen. Dat neemt niet weg dat leenwoorden een substantiële bron voor nieuwe woorden vormen.

Maar hoe groot is substantieel? De schattingen over hoeveel procent van de Nederlandse woordenschat is geleend, lopen sterk uiteen. Je hoort wel astronomische aantallen noemen van 50 à 75 procent, maar zonder dat daarbij gezegd wordt waarván dat percentage is. Voor een etymologisch woordenboek klopt dit wel min of meer, maar de trefwoorden in zo'n woordenboek zijn natuurlijk het resultaat van een aselecte steekproef. Van de trefwoorden in een algemeen Nederlands handwoordenboek als Van Dale of Koenen is naar schatting slechts 10 à 15 procent geleend.

Interessant genoeg komt dat percentage van 10 à 15 procent leenwoorden globaal overeen met wat je vindt in een groot corpus: zo blijkt dat in een krantencorpus van 1 miljoen woorden slechts ongeveer 10 procent van de tokens geleend is, waarbij de – zeer vele – namen van plaatsen, bedrijven, instellingen en personen niet zijn meegerekend.[37]

De meeste leenwoorden in zo'n corpus zijn ingeburgerde Franse of Latijnse woorden (*minister, politie, conferentie, kamer*), en niet, zoals vaak gedacht wordt, Engelse leenwoorden. Engelse leenwoorden vallen meer op omdat ze nog niet zijn ingeburgerd, maar in een corpus zijn ze in de minderheid. Wel neemt hun aantal toe: uit een kleine krantensteekproef bleek dat het percentage van Engelse leenwoorden (types) opliep van 2,3 procent in 1994 naar 6,9 procent in 2022, en ook het percentage tokens verdriedubbelde. In een onderzoek naar het taalgebruik van vloggers bleek verder dat in acht uur vlogs slechts 2,3 procent van het totale aantal woorden (tokens) van de vloggers een Engels woord was. En uit onderzoek onder Vlaamse kinderen van 7 tot 13 bleek tot slot dat slechts 3 procent van hun uitingen in algemene situaties Engelse woorden bevat; met vrienden kan dat oplopen naar 7 procent. Daarbij gaat het vooral om vluchtige modewoorden.[38] Zo bezien valt het met de Engelse invloed dus wel mee.

Uit de etymologische woordenboeken rijst een ander beeld van de Engelse invloed op. Ze tonen dat die invloed in deze eeuw groeit, wat overeenkomt met de intuïtie van taalgebruikers. Het omvangrijke *Etymo-*

*logisch woordenboek* van Van Dale uit 1997 bevatte 9.126 leenwoorden uit het Frans, 5.332 uit het Latijn en 'slechts' 2.326 uit het Engels (uit het Duits nog minder: 1.398). Inmiddels bevat een etymologische database uit 2025 maar liefst 10.356 Engelse leenwoorden. In deze eeuw overvleugelen de Engelse leenwoorden dus de Franse.[39]

Veel Engelse leenwoorden worden echter slechts door een specifieke groep (jongeren, computerspecialisten, economen) gebruikt, en vaak leiden ze in het Nederlands maar een kort leven: veertig procent van de Engelse leenwoorden die omstreeks 1990 in twee woordenboeken werden vermeld, kwam al niet meer voor in een groot corpus uit 2021.[40] De omloopsnelheid van moderne Engelse leenwoorden ligt hoog, terwijl oudere Franse, Latijnse en Duitse leenwoorden veelal blijvend zijn ingeburgerd.

## De woordsoorten van nieuwe woorden

Uit Figuur 2.3 bleek dat de meeste Nederlandse woorden (types) een naamwoord of werkwoord zijn; andere woordsoorten hebben veel minder vertegenwoordigers. Dezelfde verhouding vinden we – niet verrassend – bij neologismen, zo blijkt uit het *Chronologisch woordenboek*. Naamwoorden zijn immers de woordsoorten die het meest veranderlijk en modegevoelig zijn en de meeste synoniemen kennen, denk aan *gek, dwaas, dol, mal, zot, geschift, mesjogge, crazy, waus, wappie*. En werkwoorden zijn onmisbaar in een zin om een handeling of toestand uit te drukken.

Binnen de neologismen is de groep functiewoorden slechts klein. Vaak wordt beweerd dat het een gesloten klasse is die niet wordt uitgebreid. Dat is echter niet juist. Sterker nog: sommige categorieën functiewoorden, zoals lidwoorden en onderschikkende voegwoorden, bestonden nog helemaal niet in het Germaans; die zijn pas in het Middelnederlands ontwikkeld. Álle categorieën functiewoorden zijn bovendien in de loop van de tijd vermeerderd met nieuwe leden. Nieuwe voornaamwoorden zijn bijvoorbeeld *jullie, menigeen, verschillende, zoiets*, en nieuwe voegwoorden zijn *alsmede, echter, doordat, mits, sinds*. Het grootst is de groep voorzetsels: alleen al in de twintigste eeuw kwamen *middels, gezien, versus, gaandeweg, gaande, lopende* en *richting* in gebruik, en in de eenentwintigste eeuw *beyond* 'meer dan' (dat is beyond zielig).[41]

Het voorbeeld *beyond* verwijst een andere misvatting over functiewoorden naar de prullenbak, namelijk dat die niet uit andere talen worden overgenomen vanwege hun grammaticale betekenis, die specifiek voor het Nederlands is. Met name voorzetsels worden echter regelmatig geleend. Zo komen *contra, pro, per, circa, qua* uit het Latijn, *via, versus* uit het Engels (dat het weer uit het Latijn heeft), *conform* uit het Frans, en *namens* en *middels* uit het Duits. Uit het Duits hebben we verder de voornaamwoorden *ettelijke, ieder* en *zich* overgenomen.

Ook functiewoorden worden dus aangevuld, maar het aantal leden blijft uiterst gering vergeleken met dat van de andere woordsoorten. Het feit dat functiewoorden, die in de oudste tijd nauwelijks voorkwamen, inmiddels in iedere tekst het hoogst frequent zijn (zo bleek in hoofdstuk 2), bewijst dat de Nederlandse grammatica in de loop van de eeuwen fundamenteel gewijzigd is. Van een 'synthetische' taal, waarin grammaticale relaties werden uitgedrukt met behulp van naamvallen en verbuigingen, is het Nederlands veranderd in een 'analytische' taal, waarin die relaties worden omschreven met lidwoorden, voorzetsels en voornaamwoorden: zeiden onze voorouders *Crist gotes suno*, nu zeggen we *Christus, de zoon van God*.[42]

## Blijvertjes of passanten?

Figuur 3.3 toonde dat er iedere eeuw nieuwe woorden bij komen. Onzichtbaar blijft dat er ook iedere eeuw woorden verdwijnen. *Beanie, momjeans, hoodie, onesie* en *swikini* verdringen de oudere kledingstukken *boezeroen, hansop, kamizool* en *wambuis*. In het diachroon semantisch lexicon DiaMaNT staan veel synoniemen voor een algemeen begrip als 'vrolijk', zoals *aise, fruitig, galjaard, gemeed, hups, jolie, vreudelijk* en *wierig*, die inmiddels spoorloos zijn verdwenen.

Het is lastig te voorspellen of een nieuw woord een blijvertje is of niet. Om de overlevingskans van een nieuw woord te berekenen heeft Allan Metcalf in 2002 de zogenaamde FUDGE-test ontwikkeld.[43] De letters staan voor *Frequency, Unobtrusiveness, Diversity of users and situations, Generation of forms and meanings* en *Endurance of the concept*. Het komt er in het kort op neer dat de overlevingskans toeneemt naarmate een woord frequent wordt gebruikt, onopvallend is, door verschillende groepen in

verschillende situaties wordt gehanteerd, als het zelf nieuwe woorden voortbrengt en als het een duurzaam begrip beschrijft.

Daarbij ziet Metcalf volgens mij een belangrijke factor over het hoofd: de munter. Een woord gemaakt door een BN'er, BV'er of influencer raakt direct algemeen bekend en heeft een grote overlevingskans. Iedereen kent *doemdenken* en *regelneef* van Kees van Kooten en Wim de Bie, en *bovenbaas* en *denkraam* van Marten Toonder. Op termijn verdwijnt de munter uit het zicht, terwijl het woord overleeft: bijna niemand weet nog dat journalist Nico Scheepmaker achter *droste-effect* zit, Wim T. Schippers achter *gekte* en Multatuli achter *buitenissig*.

De FUDGE-test kan helaas niet gebruikt worden om met terugwerkende kracht te berekenen hoeveel woorden er in de loop van de tijd zijn verdwenen. Als alternatief heeft Freek Van de Velde een zogenaamde overlevingsanalyse gemaakt: van 500 woorden die meer dan eens in veertiende-eeuwse teksten vermeld zijn, bekeek hij of en hoe lang ze genoemd worden in de citaten van het WNT.[44] Het bleek dat de kans op overleven na 350 jaar nog ongeveer tachtig procent was. De hoogste overlevingskans hadden frequente, korte woorden. Bijvoeglijke naamwoorden verdwenen iets sneller dan zelfstandige naamwoorden en werkwoorden.

Dat zoveel woorden het eeuwenlang volhouden is een verrassend resultaat. Wel gaat het om een kleine steekproef, die uitgaat van veertiende-eeuwse oorkonden met een beperkte en formulaire woordenschat, terwijl eendagsvlinders werden uitgesloten, hoewel die invloed zouden hebben op de gemiddelde levensduur. Literaire werken uit die tijd hadden een veel rijkere en creatievere woordenschat, en de overlevingskans van dergelijke woorden zal kleiner zijn. Bovendien gaat het over de overlevingskans van middeleeuwse woorden: woorden die verwezen naar een agrarische samenleving die lange tijd voortduurde, tot de negentiende-eeuwse industriële revolutie. Sindsdien buitelen de technologische ontwikkelingen over elkaar heen, samen met nieuwe benamingen. Ik vermoed dan ook dat de omloopsnelheid van woorden sinds de negentiende eeuw enorm is toegenomen. Dat wordt bevestigd door het bovengemelde feit dat veertig procent van de Engelse leenwoorden uit 1990 na 30 jaar alweer is verdwenen.

Overigens is het meestal helemaal niet makkelijk het tijdstip van overlijden van een woord vast te stellen: vaak blijft het doorleven in bepaalde kringen of dialecten, of het houdt stand in een uitdrukking of spreekwoord. Zo kennen we *duit* ('koperen munt'), *duig* ('plank van de wand van een ton') en *kaak* ('schandpaal') nog dankzij de uitdrukkingen *een duit in het zakje doen*, *in duigen vallen* en *aan de kaak stellen*.

Sommige mensen betreuren het dat woorden verdwijnen of minder vaak gebruikt worden, en proberen hun leven te rekken door ze in een beschermd reservaat te plaatsen. Ze publiceren woordenboeken met vergeetwoorden en verdwijnwoorden, in de hoop dat dat taalgebruikers ertoe brengt de woorden te reanimeren. Maar woorden komen en gaan. Het zou interessant zijn te achterhalen of de nieuwkomers en de verschoppelingen elkaar in evenwicht houden, of dat er meer woorden bijkomen dan verdwijnen. Ik vermoed het laatste, maar het is momenteel niet mogelijk hiervoor harde bewijzen aan te voeren. Wellicht komen in de toekomst uitgebreide corpora met dwarsdoorsneden door de tijd beschikbaar. Aan de hand van die dwarsdoorsneden zou je dan kunnen berekenen of in de verschillende tijdsperiodes het aantal types toeneemt of afneemt. Iets om van te dromen!

## Wie heeft de grootste?

Hierboven bleek dat de historische woordenschat van het Nederlands in woordenboeken is vastgelegd. Vergelijkbare woordenboeken bestaan er ook voor andere talen, zoals het Frans, Duits en Engels. Die woordenboeken geven nogal eens aanleiding tot een wedstrijdje: taalgebruikers beroepen zich erop dat hún taal het grootste woordenboek of zelfs de grootste woordenschat heeft, waarbij vooral Nederlands en Engels met elkaar worden vergeleken. Wat is daarvan waar? In Figuur 3.4 staat het aantal trefwoorden en woordenboekdelen van het Nederlandse WNT, de Engelse *Oxford English Dictionary*, het *Deutsches Wörterbuch* van de gebroeders Grimm en het Franse *Trésor de la langue française* (TLF). Die woordenboeken beschrijven overigens niet dezelfde periodes.

| Woordenboek | Beschreven periode | Trefwoorden | Gedrukte delen |
|---|---|---|---|
| WNT (Nederlands) | 1500-1976 | 400.000 | 43 |
| OED (Engels) | 8e eeuw-heden | 500.000 | 20 |
| Grimm (Duits) | 1450-1961 | 330.000 | 17 |
| TLF (Frans) | 1800-2000 | 100.000 | 17 |

Figuur 3.4 De omvang van de historische woordenboeken van het Nederlands, Engels, Duits en Frans

Het WNT wordt vaak het grootste woordenboek ter wereld genoemd.[45] Dat klopt naar aantallen delen, maar niet naar aantallen trefwoorden: dan is de OED groter, al beschrijft die ook een veel langere periode. De OED wordt trouwens nog steeds online aangevuld en bevat inmiddels maar liefst ruim 600.000 trefwoorden. Toch moeten we dit alles met een korreltje zout nemen, want al eerder bleek dat de omvang van een woordenboek niet zoveel zegt over de werkelijke omvang van de woordenschat.

# 4 Het uitbreiden van bestaande woorden

In het vorige hoofdstuk bleek dat taalgebruikers een onverzadigbare honger hebben naar nieuwe woorden. De gemakkelijkste manier om die honger te stillen is door samenstellingen en afleidingen te maken van een of meer al bestaande woorden.[46] Die woorden hebben immers al een betekenis, zodat nieuwvormingen makkelijk te begrijpen zijn, zelfs als ze een overdrachtelijke betekenis krijgen, zoals *koevoet* ('hefboom in de vorm van de hoef van een koe') of *kraaienpootjes* ('rimpels in de ooghoeken'). Wat voor mogelijkheden biedt de taal, nemen die in de loop van de tijd toe? En klopt het, wat je vaak hoort beweren, dat het aantal samenstellingen en afleidingen (samen gelede woorden genoemd) eindeloos is?

## Samenstellingen

Samenstellingen uit twee of meer zelfstandige woorden bestaan sinds de oudste tijden. Zo dateren *koningsdochter, zwijnenvlees, vouwstoel, schildwacht, aankomen* en *geelblauw* al uit de dertiende eeuw of eerder. Aanvankelijk waren het twee aparte woorden (*des conincs dochter*), die samenklonterden tot één, met slechts één (hoofd)klemtoon. Moderne voorbeelden zijn *rodekool* en *sterkedrank*, die volgens de officiële spelling ook als *rode kool* en *sterke drank* mogen worden geschreven, terwijl *blinde vink* en *vlijtig liesje* altijd als twee woorden worden gespeld.

Het laatste deel in een samenstelling is bepalend voor de betekenis: een *waterput* is een soort put (voor water), *putwater* is een soort water (namelijk uit een put). Er zijn twee uitzonderingen. Voor samenstellingen waarin beide delen gelijkwaardig zijn, zoals *broekrok, puntkomma, zuurzoet* en *minister-president*, gaat die regel niet op. En het geldt evenmin voor samenstellingen waarbij de twee leden samen de bezitter of het bezit aanduiden van wat ze noemen: een *dikkop* is 'iemand met een dikke kop'. Andere voorbeelden van deze zogenaamde possessieve samenstellingen zijn *dikzak, domoor, driehoek, kletskous, roodborstje* en *volbloed*.

In de loop van de tijd is de manier waarop Nederlandse woorden worden samengesteld niet ingrijpend gewijzigd. Wel nieuw, twintigste-eeuws, zijn samenstellingen met als eerste deel een woordgroep, zoals *alles-moet-kunnenmentaliteit, blijf-van-mijn-lijfhuis, doe-het-zelfwinkel, je-weet-welkater* en *reken-je-rijkfilosofie*. Bovendien zijn twee soorten samenstellingen sinds eind negentiende eeuw populair geworden onder invloed van het Duits en Engels, namelijk de combinatie van een onverbogen bijvoeglijk naamwoord met een zelfstandig naamwoord (*groenvoer, groothandel, nieuwbouw, rauwkost, snelbinder*) en van een zelfstandig naamwoord met een voltooid deelwoord (*beursgenoteerd, drugsverslaafd, noodgedwongen* en *probleemgestuurd*).

In de 21ste eeuw is een nieuw soort samenstelling ontstaan, ook in het Duits en Engels, die bestaat uit reduplicatie (verdubbeling) van een woord, zoals *meisjemeisje*, *vriendvriend* en *omaoma*.[47] Die samenstellingen geven het prototype van het verdubbelde woord aan: een *meisjemeisje* is een heel meisjesachtig meisje, *thuisthuis* is het ouderlijk huis (dus niet de studentenkamer). Zo'n betekenis ontbreekt bij oudere reduplicaties als *blauwblauw* of *taaitaai*.

In samenstellingen zijn de aparte woorddelen gemakkelijk herkenbaar, zelfs bij heel oude samenstellingen. Dat is opmerkelijk, want samenstellingen hebben maar één klemtoon, meestal op de eerste lettergreep (*áppelboom, brúínbrood*), soms op de tweede (*blindedárm*). Taalgebruikers hebben de neiging om de klinkers in onbeklemtoonde lettergrepen gereduceerd uit te spreken. Je zou dus verwachten dat samenstellingen in de loop van de tijd korter en onherkenbaar worden. In de middeleeuwen was dat ook het geval: men spelde fonetisch *omoed, sprakonst, vrienscap* en *coman*. Een fonetische spelling leidt tot veel spelvarianten, naast *coman* bijvoorbeeld *comen, cooman, coopman, copeman*. Toen boeken in de renaissance in grotere oplagen werden gedrukt, werd die spellingvariatie voor uitgevers en lezers een struikelblok. Daarom ging men zich bezinnen op een uniforme spelling. Men koos voor het spellingbeginsel van de gelijkvormigheid, dat bepaalt dat *hond* eindigt op *d* vanwege het meervoud *honden*. Die gelijkvormige spelling paste men óók toe in samenstellingen. Men herstelde de oorspronkelijke samenstellende delen tot

*ootmoed, spraakkunst, vriendschap* en *koopman*, zodat de samenstellingen weer doorzichtig werden.

Natuurlijk zijn er uitzonderingen. Dat *laars, leidsel* en *vent* teruggaan op de samenstellingen *lederhose*, *leidzeel* (met *zeel* 'dik touw') en *vennoot* is niet langer te zien. Dergelijke verholen samenstellingen zijn echter zeldzaam, behalve onder functiewoorden en andere heel frequent gebruikte woorden. Deze woorden zijn wél ingekort, volgens de wetmatigheid genoemd in hoofdstuk 2: hoe frequenter een woord, hoe korter het is. Zo gaat *maar* terug op *neware* ('ware het niet'), *niet* op *niewiht* ('niet wicht, geen ding'), *iets* op *iowiht* ('een of ander ding'), *welk* op *wie* + *lijk* ('wat voor een vorm hebbend') en *elk* op *een* + *lijk* ('iedere persoon afzonderlijk').[48]

## Afleidingen

Anders dan samenstellingen bevatten afleidingen een woorddeel dat niet zelfstandig voorkomt. Dat woorddeel kan als voorvoegsel vóór het grondwoord staan, of als achtervoegel erna, en kan niet met íéder grondwoord worden verbonden: daarvoor zijn vaste regels, waarbij ook de woordsoort een rol speelt. Zo wordt het voorvoegsel *aarts-* gecombineerd met een persoonsnaam of bijvoeglijk naamwoord en versterkt de betekenis daarvan, vergelijk *aartsdeugniet* en *aartsdom*. Het achtervoegsel *-s* leidt een bijwoord af van verschillende woordsoorten, bijvoorbeeld in *anders, eens, ergens, langs* en *zelfs*. Het voorvoegsel *ge-* plus het achtervoegsel *-te* duiden samen een verzamelnaam aan: *gebeente, gevogelte*.

Van enkele voor- en achtervoegsels weten we dat ze in het Germaans een zelfstandig woord waren. In de loop van de tijd vervaagde de betekenis ervan, verzwakte de klemtoon en verloor het zijn zelfstandigheid. Dat geldt bijvoorbeeld voor de voorvoegsels *be-* en *ver-* in werkwoorden als *bedenken* en *verzorgen*, die teruggaan op de bijwoorden *bij* en *voor*. De achtervoegsels *-dom, -heid* en *-schap* (in *rijkdom, gelegenheid* en *boodschap*) waren oorspronkelijk zelfstandige naamwoorden die iets vaags als 'stand, toestand' betekenden, en waarmee men daarna nieuwe zelfstandige naamwoorden ging afleiden. Volgens sommigen zijn álle voor- en achtervoegsels ontstaan uit zelfstandige woorden, maar dat is niet meer met zekerheid na te gaan.

De taalgebruikers breidden iedere eeuw het aantal voor- en achtervoegsels uit, met ieder een eigen functie en betekenis. Zelfstandige naamwoorden gingen ze bijvoorbeeld afleiden met *-nis*, *-ing*, *-te*, bijvoeglijke naamwoorden met *-achtig*, *-baar*, *-lijk*, *-loos*, *-zaam*, bijwoorden met *-halve* en *-weg*, en werkwoorden met *-igen*, *-elen*, *-eren*. Daar staat tegenover dat enkele achtervoegsels hun productiviteit verloren: er werden dus geen nieuwe woorden meer mee gemaakt. Zo vormen we verkleinwoorden tegenwoordig met *-je* (*stipje*) en niet langer met *-el* (*stippel*). Maar daarmee verdwijnen oude afleidingen met dat achtervoegsel niet: we spreken nog steeds over *eikel* (afgeleid van *eik*), *ijzel* (van *ijs*), *navel* (van *naaf*) en *tepel* (van *tip*).

Nieuwe voor- en achtervoegsels zijn ook geleend uit andere talen, en wel uit het Latijn, Frans, Duits en Engels. Als er uit een taal veel woorden met een bepaald achtervoegsel zijn overgenomen, herkennen de taalgebruikers dit als achtervoegsel en gaan ze het gebruiken in nieuwvormingen. Eerst wordt zo'n achtervoegsel meestal gecombineerd met een leenwoord, maar dat hoeft niet per se de taal te zijn waaruit het achtervoegsel is geleend. Sommige voor- en achtervoegsels blijven in dit stadium steken: ze zijn alleen productief met vreemde woorden. Dat geldt bijvoorbeeld voor de Franse achtervoegsels *-ant* (*muzikant*), *-aris* (*jubilaris*), *-atie* (*diplomatie*) en *-eur/-euse* (*chauffeur, chauffeuse*), en voor de Latijnse achtervoegsels *-icus* (*neerlandicus*), *-(a)tor/-trix/-trice/* (*conservator, conservatrix, coördinator, coördinatrice*).[49] Nieuwe wetenschappelijke termen worden dikwijls gevormd van Latijnse of Griekse woorden, zoals *biologie* en *automobiel*; soms krijgen die woorden een nieuwe wetenschappelijke betekenis, zoals *lymfe* ('weefselvocht') in *lymfocyt*, van Latijn *lympha* ('helder water').[50]

Andere geleende achtervoegsels worden moeiteloos gecombineerd met inheemse woorden. Figuur 4.1 geeft daarvan voorbeelden; sommige achtervoegsels zijn inmiddels zo ingeburgerd dat de vreemde herkomst onherkenbaar is geworden.

| Brontaal | Achtervoegsel | Voorbeelden |
|---|---|---|
| Latijn | *-aar, -er* | leraar, minnaar, bakker |
| | *-ster* | naaister, werkster |
| Frans | *-aard, -erd* | gierigaard, grijsaard, leukerd, goeierd |
| | *-age* | bagage, lekkage, pluimage, vrijage |
| | *-es, -esse* | prinses, lerares, secretaresse |
| | *-ette* | wasserette, affairette, boerderette |
| | *-ier* | herbergier, scholier, tuinier |
| | *-ij* | abdij, heerschappij |
| | *-(i)teit* | flauwiteit, stommiteit, puberteit |
| | *-eren* | kleineren, schakeren |
| Frans of Latijn | *-ement* | dreigement, gruzelement |
| Duits | *-isch* | esthetisch, evangelisch, kritisch |
| | *-haftig* | heldhaftig, manhaftig |
| | *-matig* | rechtmatig, regelmatig |
| | *-vol* | eervol, stijlvol |
| | *-waardig* | begerenswaardig, bewonderenswaardig |
| Engels | *-burger* | kaasburger, kipburger |
| | *-freak* | milieufreak, filmfreak |
| | *-gate* | Margaritagate, dieselgate |
| | *-minded* | sportminded, kunstminded |
| | *-proof* | crisisproof, generatieproof |

Figuur 4.1 Voorbeelden van geleende achtervoegsels

Heel bijzonder is dat er in deze eeuw dankzij Engelse invloed op kleine schaal een totaal nieuw type afleiding is geïntroduceerd in het Nederlands, waarbij het onzelfstandige element niet voor of achter het grondwoord staat, maar er midden in. Zo'n infix vinden we alleen in tussenwerpsels, het dient als versterking en bestaat vooralsnog alleen uit de Engelse leenwoorden *bloody* en *fucking*. Direct uit het Engels geleend zijn *abso-bloody-lutely, out-bloody-rageous, abso-fucking-lutely* en *im-fucking-possible*. De tussenwerpsels worden echter steeds meer aan het Nederlands aangepast,

vergelijk *fan-bloody-tastisch, onge-bloody-looflijk, absofuckingluut, fan-fucking-tastisch, knetter-fucking-gek, onfokkinggelooflijk, onge-fucking-zellig, per ongefuckingluk* en *super-fucking-snel*. De voorbeelden laten zien dat er variatie is in spelling – al dan niet met verbindingsstreepjes – en in de plaats van het infix, vergelijk *onfuckinggelooflijk* naast *onge-fucking-looflijk*.

## Samenstellende afleidingen

Na de middeleeuwen werd een nieuw woordvormingsprocedé in het Nederlands populair, namelijk dat van de samenstellende afleiding: een combinatie van samenstelling en afleiding. Zo'n samenstellende afleiding, zoals *blauwogig*, bestaat uit twee grondwoorden die niet samen voorkomen als samenstelling of als afleiding: *blauwoog* noch *ogig* zijn bestaande woorden. Samenstellende afleidingen zijn gevormd met de achtervoegsels *-ig* (*vijfledig, kleinschalig, loslippig*), *-s* (*onderhuids, wijdbeens*) en *-er* (*bevelhebber, gezaghebber, dijenkletser, driewieler*). Het laatste deel kan ook bestaan uit een deelwoord, vergelijk *adembenemend, toonaangevend, breedgeschouderd* en *hooggehakt*.

Ook werkwoorden als *afzwakken, indikken, opfrissen, overschaduwen* en *uitzieken* zijn samenstellende afleidingen. Een bijzondere groep vormen werkwoorden met als tweede deel een lichaamsdeel, resulterend in beeldende omschrijvingen als *klappertanden, knikkebollen, likkebaarden, schuimbekken, stampvoeten* en *tandenknarsen*.

## Verkortingen en blends

In de loop van de eeuwen werden steeds weer nieuwe samenstellingen en afleidingen gevormd. De bronwoorden werden zo steeds langer. *Denken* werd verlengd tot *bedenken, nadenken, marktdenken, zwart-witdenken*. Aan het eind van de twintigste eeuw kwam er een tegenbeweging op gang: taalgebruikers gingen zich steeds korter en efficiënter uitdrukken. Dat deden ze onder andere door in nieuwvormingen het laatste deel van een bestaand woord weg te laten, vergelijk *dino(saurus), noncha(lant), poli(kliniek), promo(filmpje)* en *sax(ofoon)*. Sommige van die woorden gaan uit op *-o* (*aso, pedo*) en dat kreeg de status van achtervoegsel in bijvoorbeeld *brabo, lesbo, limbo, lullo*.[51]

Bovendien gingen taalgebruikers nieuwe woorden vormen uit de beginletters van bij elkaar horende woorden. Zo wordt een *buitengewoon opsporingsambtenaar* kortweg *boa* genoemd. Het procedé ontstond in het Engels, maar werd al gauw door Nederlandse taalgebruikers geïmiteerd. Voorbeelden van initiaalwoorden (die letter voor letter worden uitgesproken) zijn *apk, cd, cv, pc, wc*; voorbeelden van letterwoorden (die als gewoon woord worden uitgesproken) zijn *ahob, beha, havo, vip, vut*, en voorbeelden van lettergreepwoorden (waarbij de eerste lettergrepen samen een nieuw woord vormen) zijn *horeca* (hotel, restaurant, café), *holebi* (homo, lesbienne of biseksueel) en *vrijmibo* (vrijdagmiddagborrel).

Een laatste nieuw woordvormingsprocedé bestaat uit de samentrekking van twee woorden tot één nieuw woord, zoals *bromance* (brother + romance), *brunch* (breakfast + lunch), *conculega* (concurrent + collega), *cronut* (croissant + donut) en *stagflatie* (stagnatie + inflatie). Dit procedé wordt *blend* genoemd of *portmanteau-woord* (een vondst van de Engelse schrijver Lewis Carrol). Het is geen toeval dat het een Engelse naam heeft, want ook dit procedé hebben we overgenomen uit het Engels, al passen we het zelf inmiddels ook enthousiast toe.

## Woordfamilies

Dankzij de eeuwenlange woorduitbreidingen zijn er hele woordfamilies ontstaan rond een grondwoord. Denk aan *kopen, koop(je), koper, koopster, verkopen, verkoopster, bekopen, opkopen, opkoper, doorverkopen, vrijkopen, koopziek, koopwaardig, koopzucht*. Eenzelfde soort woordfamilie klontert rond leenwoorden, bijvoorbeeld *expliceren, explicateur, explicatie, expliciet, expliciteit*. De achtervoegsels bij inheemse en geleende grondwoorden verschillen van elkaar, maar taalgebruikers weten feilloos welke achtervoegsels bij welke grondwoorden horen.

Bij een woordfamilie die uit leenwoorden bestaat, speelt de taal van herkomst geen rol voor de taalgebruikers. Ze herkennen het grondwoord en kunnen gemakkelijk de betekenis van bijbehorende afleidingen herleiden. Zo is de woordfamilie rond *organ-* opgebouwd uit leenwoorden uit verschillende talen: *orgaan, organisator* (Latijn), *organiek, organisatie, organiseren, organisme, organist* (Frans), *organisch* (Duits) *organizer* (Engels),

en *organigram* (Latijn + Grieks). Dergelijke woordfamilies komen ook in andere West-Europese talen voor, wat de onderlinge verstaanbaarheid vergemakkelijkt.

## Hoe productief zijn de woordvormingsprocedés?

Welk van de genoemde woordvormingsprocedés levert nu de meeste nieuwe woorden op in het Nederlands? Over hoe dat in het verleden was, valt helaas niets te zeggen, maar voor de huidige tijd kunnen we profiteren van twee digitale woordenboeken: het ANW (*Algemeen Nederlands Woordenboek*), dat het Nederlands van 1970 tot heden beschrijft en dat 87.100 trefwoorden bevat, en het WNW (*Woordenboek van Nieuwe Woorden*), waarin 4.380 woorden zijn beschreven die vanaf het jaar 2000 zijn ontstaan. In die woordenboeken is bij ieder trefwoord aangegeven hoe het is gevormd. In Figuur 4.2 heb ik het van ieder woordvormingsprocedé genoteerd hoeveel procent het uitmaakt van het totale aantal trefwoorden in de twee woordenboeken. Bij het ANW heb ik alleen gekeken naar de trefwoorden die van vóór 2000 dateren en dus niet als neologisme worden beschouwd, 78.741 in totaal. Dat maakt het mogelijk om een vergelijking te trekken tussen de manieren waarop nieuwe woorden in de twintigste en de eenentwintigste eeuw zijn gevormd.

| Woordenboek | Percentage van ANW-trefwoorden | Percentage van WNW-trefwoorden |
|---|---|---|
| Afleidingen | 24,4 | 9,3 |
| Blends | 0,03 | 4,3 |
| Letter(greep)woorden, initiaalwoorden | 0,6 | 2,3 |
| Samenstellingen en -koppelingen | 62,8 | 63,3 |
| Verkortingen en afkortingen | 0,8 | 2,0 |
| Woordgroepen | 0,9 | 1,3 |
| Leenwoorden | (geschat 6 à 8) | 17,5 |
| **Totaal** | **89,53%** | **100%** |

Figuur 4.2 Het percentage afleidingen, blends, letter(greep)woorden, samenstellingen, verkortingen en woordgroepen in twee moderne woordenboeken

De ruim 10 procent die in het ANW niet is ingevuld, bestaat uit een restcategorie van ongelede (dus niet-samengestelde of afgeleide) woorden die deels zijn geleend (*chick*, *mitella*, *muffin*) en deels erfwoorden zijn met recente betekenissen (*ader* 'verkeersader, goudader'). De twee categorieën zijn in het woordenboek samengenomen en daarom kan ik niet het exacte percentage leenwoorden invullen maar slechts een benadering. Zeker is in ieder geval dat het percentage leenwoorden in het WNW hoger ligt dan in het ANW, wat betekent dat het aantal leenwoorden in deze eeuw toeneemt – vooral uit het Engels. Een onderzoekje naar Engelse leenwoorden in kranten bevestigt dit beeld.[52]

Uit Figuur 4.2 blijkt dat samenstellingen het meest voorkomen, in het ANW gevolgd door afleidingen en in het WNW door leenwoorden. Er zijn echter opmerkelijke verschillen tussen ANW en WNW: het percentage afleidingen ligt in het WNW ver onder dat van het ANW, terwijl de blends, letter(greep)woorden, verkortingen en woordgroepen in het WNW een veel hoger percentage hebben dan in het ANW. Kennelijk zijn die woordcategorieën in deze eeuw in opmars en komen ze in de plaats van de oude categorie afleidingen. Het is natuurlijk slechts een tendens, maar het lijkt erop dat de vorming van neologismen in deze eeuw een innovatie doormaakt.

## Oneindig uitdijend?

In de loop van de tijd zijn telkens nieuwe woorden gevormd op telkens nieuwe manieren. Van het ene woord kwam letterlijk het andere. Zo breidt de woordenschat zich uit (al is niet ieder woord een blijvertje, zo bleek hiervoor). Samenstellingen en afleidingen zijn productief: er kunnen telkens nieuwe woorden mee worden gemaakt. Ze zijn ook 'recursief', herhaalbaar: je kunt telkens nieuwe woorden maken door het tweede deel te herhalen. Een *pizzadoos* ('doos voor een pizza') past in een *pizzadoosdoos* ('doos voor een pizzadoos'), die weer past in een *pizzadoosdoosdoos*, die weer gaat in een *pizzadoosdoosdoosdoos*, en ga zo maar door. Hetzelfde geldt voor *raket – antiraket – antiraketraket*, of *fotoalbum – fotoalbumalbum – fotoalbumalbumalbum*, etc.

Vanwege deze eigenschappen wordt vaak beweerd dat de Nederlandse woordenschat oneindig is: ieder woord kan immers worden gecombineerd met een willekeurig ander woord. Maar dat is een misvatting. Je loopt namelijk op tegen twee beperkingen: die van de betekenis en die van je geheugen. Zo kan *doos* alleen gecombineerd worden met iets wat past in een doos (*brillendoos, naaidoos, verbanddoos*) of de vorm heeft van een doos (*doosschakelaar, doosvrucht*). Je kunt niet zeggen *watervaldoos, vloeddoos* of *liefdedoos*. Nu hoor ik je al tegenwerpen: maar die woorden staan hier toch, dan bestaan ze toch? Dat klopt natuurlijk voor wat betreft de vorm, maar die is niet gekoppeld aan een betekenis, aan iets wat in de realiteit of de geest bestaat. Om dezelfde reden gold *liem* in hoofdstuk 3 niet als Nederlands woord.

Woorden waarvan de betekenissen met elkaar botsen, kunnen dus niet met elkaar worden gecombineerd. En ook het geheugen en de behoefte om helder te communiceren perken de mogelijkheden in: *pizzadoosdoosdoosdoos* etc. is inhoudelijk niet meer te bevatten, en *kindercarnavalsoptochtvoorbereidingswerkzaamhedenbesprekingsvergadering* is door zijn lengte nauwelijks te doorzien, zelfs niet als je van rechts naar links leest, en komt alleen voor als rariteit in een langstewoordenwedstrijdje. In het wild kom je zo'n woord niet tegen.

De vorming van afleidingen is in het Nederlands nog sterker beperkt dan die van samenstellingen: niet alleen door de betekenis maar ook door de woordvormingsregels van het Nederlands. Weliswaar kun je een mannelijke persoonsnaam afleiden met het achtervoegsel *-erik*, maar dat kan alleen van bijvoeglijke naamwoorden (*bangerik, gemenerik, stommerik*), en niet eens van álle bijvoeglijke naamwoorden; zo komen *armerik, beroerderik, dementerik, driesterik* niet voor.

Het aantal samenstellingen en afleidingen is dus in de praktijk begrensd. Waar de precieze grenzen liggen, dus hoeveel potentiële samenstellingen en afleidingen er bestaan, valt momenteel niet te berekenen. Misschien kan dat in de toekomst, als we meer inzicht hebben in de regels waaraan woordbetekenissen en woordvormingen zijn gebonden. Zeker is wel dat het aantal Nederlandse samenstellingen en afleidingen heel groot is. Maar oneindig is het niet.

## Hoeveel ongelede woorden?

En hoe zit het met het tegenovergestelde: de ongelede, dus niet-samengestelde of afgeleide woorden? Hoeveel hebben we er daarvan, en wat is de verhouding tussen het aantal ongelede en gelede woorden? Worden beide types in de loop van de tijd in gelijke mate uitgebreid, en is hun verhouding dus min of meer constant, of vindt uitbreiding van de woordenschat voornamelijk plaats door de vorming van nieuwe samenstellingen en afleidingen? In dat laatste geval wordt de verhouding ongeleed/geleed steeds kleiner naarmate de woordenschat groeit; dat is vergelijkbaar met de eerder genoemde wet van Heaps, die stelde dat de type/token-verhouding afneemt naarmate een corpus groter wordt.

De ongeleed/geleed-verhouding is momenteel niet eenvoudig en eenduidig te berekenen. Slechts drie Nederlandse woordenlijsten geven informatie over de woordgeleding. Het gaat om het ANW, CELEX (gebaseerd op de inhoud van de *Woordenlijst van de Nederlandse taal* uit 1990) en de Dikke Van Dale uit 1999, waaraan Oele Koornwinder informatie over de woordvorming heeft toegevoegd.[53] Een probleem is dat de drie woordenlijsten verschillende definities hanteren van wat een ongeleed woord is. Het pijnpunt zit hem vooral in de leenwoorden. Over woorden als *productie, producent, produceren* en *productief* bestaat wel overeenstemming: die zijn geleed. Maar hoe zit het met *product*? ANW beschouwt dit als ongeleed, Koornwinder als geleed, en in CELEX ontbreekt een analyse.[54] *Cluster* is volgens ANW en Koornwinder ongeleed en volgens CELEX geleed.

Eenzelfde verdeeldheid treedt op bij wetenschappelijke termen die zijn gebaseerd op twee of meer Latijnse of Griekse woorden. *Autocraat* en *parallellogram* noemt ANW 'in de brontaal geleed', Koornwinder 'geleed' en CELEX analyseert het niet. Kenners van de klassieke talen herkennen de verschillende woorddelen, maar voor de gemiddelde taalgebruiker zullen dit toch wellicht eerder ongelede woorden zijn. En hoe moet je afleidingen analyseren waarvan een basiswoord ontbreekt (*meisje, gedeisd*), of jonge verschijnselen zoals verkortingen (*afko, lesbo*), letterwoorden (*vip, wc*) en blends (*conculega*)?

Vragen, vragen, vragen... Er is geen 'goed' of 'fout' antwoord, maar zolang we niet allemaal hetzelfde antwoord geven, is het lastig iets zin-

nigs te zeggen over de ongeleed/geleed-verhouding. Desondanks komt uit een vergelijking van de drie woordenlijsten toch een algemeen beeld naar voren, zie Figuur 4.3.

| Woordenboek | Aantal trefwoorden | Percentage ongeleed | Percentage geleed |
|---|---|---|---|
| ANW | 87.100 | 7% | 93% |
| CELEX | 102.581[55] | 9,6% | 90,4% |
| Dikke Van Dale 1999 | 245.000 | 8% | 92% |

Figuur 4.3 Een (globale) indicatie van de verhouding ongelede/gelede woorden in drie woordenlijsten

Uit Figuur 4.3 blijkt dat het percentage ongelede woorden in alle drie de woordenlijsten tussen de 7 en 10 procent ligt, ondanks het feit dat de lijsten enigszins verschillende definities voor 'geleedheid' hanteren, het totale aantal trefwoorden per woordenlijst sterk verschilt en het doel van de woordenlijsten verschillend is: CELEX als spellinglijst, ANW als beperkt modern woordenboek en de Dikke Van Dale als historisch verklarend woordenboek met relatief veel leenwoorden en wetenschappelijke termen.

Het lijkt er dus op dat de verhouding ongeleed/geleed min of meer constant is, althans voor het moderne Nederlands. Hoe het zit met het oudere Nederlands, staat nog open.

# 5 Woorden in het hoofd

Het eerste woord dat een baby uitspreekt, is een belangrijk moment voor jonge ouders – zeker als dat woord, althans in hun perceptie, *mama* of *papa* luidt. Vanaf dat moment neemt de woordenschat van een kind toe. Hoe werkt dat eigenlijk? Hoe slaan we woorden op in ons langetermijngeheugen? En hoe omvangrijk is ons mentale lexicon, het interne woordenboek in ons geheugen?

## Woordinformatie in de hersenen

Hoe informatie over woorden in de hersenen is opgeslagen, kunnen we alleen indirect onderzoeken. Dat gebeurt op verschillende manieren. Een daarvan is om te bekijken wat er nodig is om in het dagelijkse taalgebruik woorden te herkennen en te produceren.[56] Die taalverwerking wordt mogelijk doordat het mentale lexicon bij ieder woord informatie bewaart over de klank en de uitspraak, de interne opbouw (hoe het is samengesteld of afgeleid) en, nadat we hebben leren schrijven, over de spelling. Om een woord in een zin te kunnen gebruiken is bovendien informatie beschikbaar over de woordsoort, de verbuiging en vervoeging, de betekenis, en de context waarin een woord gebruikt kan worden, bijvoorbeeld dat *guur* alleen gecombineerd wordt met weersomstandigheden.

Al deze informatie wordt opgeslagen bij de kleinste betekeniseenheid. Dat kán een woord zijn, maar zo'n eenheid kan ook uit meerdere woorden bestaan, zoals vaste verbindingen met een eigen betekenis, denk aan *nieuwsgierig Aagje, in petto, een blauwtje lopen, een doekje voor het bloeden* of *om de haverklap*. Deze vormen samen de kleinste betekeniseenheid.

Daarnaast bevat het geheugen regels over de afleiding van regelmatige taalvormen, zoals de manier waarop een regelmatige meervoudsvorm of verkleinvorm wordt gemaakt van een zelfstandig naamwoord. Die vormen hoeven dan niet apart in het geheugen te worden opgeslagen, anders dan onregelmatige vormen als *kalveren, kinderen, gelederen*.

## Snelheid van woordherkenning

Woorden worden al herkend voordat iemand ze helemaal heeft gehoord of gelezen, en sommige woorden worden sneller herkend dan andere. Kennelijk bevat het mentale lexicon informatie die invloed heeft op de snelheid van woordherkenning. Wat voor informatie is dat? Onderzoekers hebben vijf factoren gevonden die de herkenningssnelheid blijken te beïnvloeden.[57]

De eerste factor is vertrouwdheid met een woord: hoe vaker iemand een woord is tegengekomen, dus hoe hoger de woordfrequentie, hoe sneller het wordt herkend. De tweede is woordlengte: hoe korter het woord, hoe sneller we het herkennen. Deze twee factoren hangen hoogstwaarschijnlijk samen, want in hoofdstuk 2 bleek al dat de hoogstfrequente woorden het kortst zijn. Uit deze twee factoren kunnen we in ieder geval afleiden dat frequente en korte woorden in het geheugen hoger in de hiërarchie staan dan laagfrequente en lange woorden.

De derde factor is de betekenisrelatie van een woord met voorafgaande woorden: het woord *jongen* wordt sneller herkend in een context waarin sprake is van *meisje* dan van bijvoorbeeld *meimaand*.

De vorm en klank blijken een vierde factor: als er van een woord meerdere vorm- of klankovereenkomsten bestaan, herkennen we het sneller. Dat bewijst dat verbuigingen, vervoegingen, samenstellingen en afleidingen rond een woord samen worden opgeslagen in een zogenaamde woordfamilie. Zo'n woordfamilie (zoals *helpen, helpt, geholpen, helper, helpfunctie, behelpen, meehelpen*) maakt herkenning kennelijk makkelijker. Hetzelfde geldt voor woorden die passen binnen een veelvoorkomend klanktype: *bakken* wordt vanwege vergelijkbare werkwoorden als *lakken, bukken, ballen* sneller herkend dan een 'eenling' als *blurb* of *jungle*. Uit deze voorbeelden blijkt tevens dat woorden niet alfabetisch in het geheugen worden opgeslagen, zoals in een woordenboek. Wel worden gelijkluidende woorden met elkaar in verband gebracht, anders zouden taalgebruikers geen woorden met dezelfde beginklanken door elkaar halen (*flamingo* en *flamenco, organisme* en *orgasme*) of woorden verhaspelen (*inleveringsvermogen* in plaats van *inlevingsvermogen*, *miniscuul* voor *minuscuul*, *polshoogte* voor *poolshoogte*).

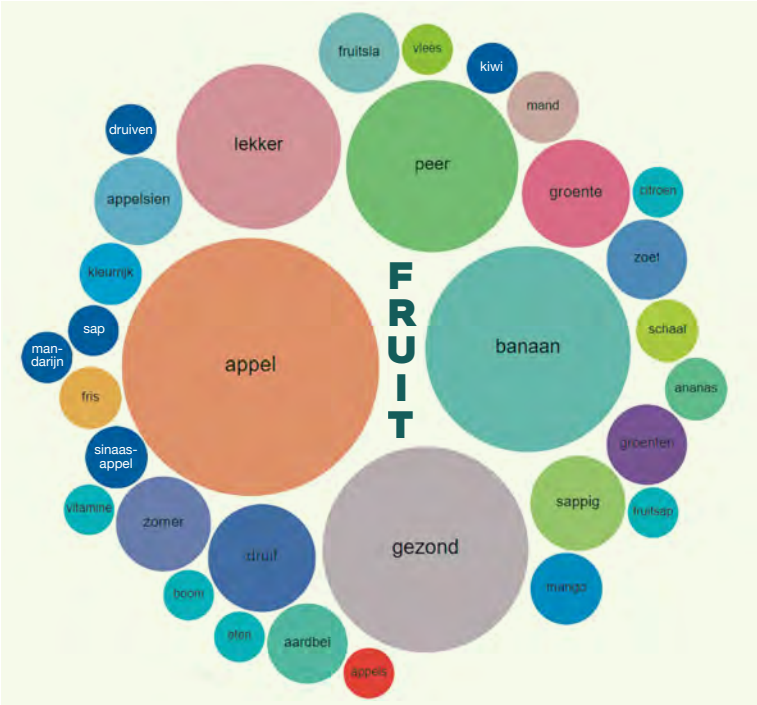
De laatste factor is de verwervingsleeftijd: een woord dat op jonge leeftijd is geleerd, wordt sneller herkend dan een woord waarmee we op latere leeftijd kennismaakten. Dat lijkt logisch, want met die woorden zijn we het langst in aanraking geweest en ze zijn dus het meest vertrouwd. Maar het blijkt óók te gelden voor woorden die we na onze kindertijd nauwelijks nog tegenkomen, zoals *draak*, *eenhoorn* of *elfje*. Dergelijke jonggeleerde woorden blijven bovendien het langst bewaard bij mensen met een hersenziekte als dementie.

## Woordassociaties

Woorden waarvan de betekenissen aan elkaar zijn gerelateerd (*meisje – jongen*), bleken sneller te worden herkend dan woorden waarbij dat niet het geval is. Dat roept de interessante vraag op hoe woorden in het mentale lexicon eigenlijk met elkaar zijn verbonden. Om dat te achterhalen hebben Vlaamse taalpsychologen een grootschalig online-onderzoek uitgevoerd naar de associaties die taalgebruikers bij een bepaald woord hebben.[58] Aan informanten werd gevraagd om drie woorden op te schrijven die een bepaald begrip, zoals *kat*, bij ze opriep. Ze noteerden bijvoorbeeld 'huisdier', 'spinnen', 'miauw'.

Aan het onderzoek deden ruim 100.000 Nederlanders en Vlamingen mee, die bij 17.000 woorden hun woordassociaties noteerden. Het bleek dat sommige associaties algemeen zijn: de meeste informanten associëren *hamer* met 'spijker' of 'nagel', *bank* met 'geld' en *bloed* met 'rood'. Onverwachter is dat de meerderheid *idee* associeert met 'lamp': ze zien een idee als een lichtje dat gaat branden.

De associaties kunnen in een zogenaamd semantisch netwerk worden afgebeeld. Figuur 5.1. toont het semantische netwerk voor *fruit*.




Figuur 5.1 De woordassociaties van 'fruit' in een semantisch netwerk[59]

Uit Figuur 5.1 blijkt dat 'appel', 'banaan' en 'peer', met de grootste cirkels, gelden als het prototype voor *fruit* (die woorden werden het vaakst genoemd), en dat 'gezond' en 'lekker' de kenmerkende eigenschappen zijn. Dat is heel andere informatie dan die we vinden in woordenboeken of corpora. Zo definieert de Dikke Van Dale *fruit* als 'vruchten die ontoebereid plegen te worden gegeten, als voedsel of als gewas'. En in het *Corpus Hedendaags Nederlands* blijken de meestgenoemde eigenschappen van *fruit* niet 'gezond' en 'lekker' te zijn, maar, in aflopende volgorde, 'vers', 'Turks', 'rood', 'laaghangend', 'gedroogd', 'exotisch' en 'rijp'. In het mentale lexicon roept *fruit* een vrij sterke associatie met 'groente(n)' op, terwijl die twee woorden elkaar volgens de woordenboekdefinities uitsluiten: fruit is geen groente en andersom. De woordrelaties in het mentale lexicon

komen dus niet overeen met het gebruik en de frequentie van een woord in het dagelijks leven en ook niet met een woordenboekdefinitie.

De woorden in zo'n semantisch netwerk zijn altijd slechts een paar stappen (een paar associaties) van elkaar verwijderd, en sommige woorden blijken een centrale positie te bekleden: ze worden vaak als associatie bij een ander woord genoemd en vormen daardoor een knooppunt in het netwerk. Dat geldt bijvoorbeeld voor woorden die eigenschappen en kleuren uitdrukken, maar ook voor woorden als *eten, water, dier, geld, auto* en *pijn*. Dit zijn vaak woorden die we op jonge leeftijd hebben geleerd. De betekenis van woorden die we later leren wordt opgehangen aan de betekenis van woorden die er al zijn, en zo komen de knooppunten steeds sterker, centraler in het netwerk te staan.

## Tweetaligheid

Een vraag die zowel onderzoekers als ervaringsdeskundige taalgebruikers al lange tijd bezighoudt, is hoe het brein in staat is twee (of meer) talen te verwerken, en of tweetaligheid leidt tot leerachterstand. Tot voor kort dacht men dat tweetaligen twee onafhankelijke taalverwerkingssystemen hebben, eentje voor de moedertaal en eentje voor de tweede taal. Dankzij allerlei onderzoeken is echter gebleken dat dat niet klopt.[60] Zo blijken tweetaligen Nederlands-Engels het woord *meisje* niet alleen sneller te herkennen als ze kort voordien het Nederlandse woord *jongen* hebben gehoord, maar óók als ze zijn geconfronteerd met het Engelse *boy*. Dat bewijst dat *jongen* en *boy* in het mentale lexicon gekoppeld zijn aan een gedeelde betekenis; de betekenis is dus niet apart bij beide woorden opgeslagen. Bovendien blijkt dat bij vloeiend tweetaligen dezelfde hersengebieden actief zijn, ongeacht of ze de moedertaal of de tweede taal hanteren.

Daarom gaat men er tegenwoordig van uit dat tweetaligen één gemeenschappelijk lexicon hebben dat alle bekende woorden in beide talen bevat. Controlemechanismen zorgen ervoor dat de taalgebruiker de talen niet – of slechts zelden – met elkaar verwart. Dat verklaart dan óók waarom het makkelijker is om een nieuwe taal te leren die lijkt op een taal die je reeds kent: daarbij kun je dan voortbouwen op al opgeslagen ken-

nis. De optelsom van beide talen zorgt ervoor dat woordfamilies uitgebreid worden, en leden ervan makkelijker herkend worden. Kennis van één taal blijkt dus de kennis van een tweede taal te versterken, en twee- of meertaligheid brengt op de lange duur vooral voordelen met zich mee.[61]

## Hoeveel woorden bevat het mentale lexicon?

Tot slot de hamvraag: hoe omvangrijk is ons mentale lexicon? Een eenduidig antwoord op die vraag bestaat niet. Om te beginnen bestaat er een verschil tussen de passieve en de actieve woordenschat: woorden die we herkennen en woorden die we gebruiken. Verder kennen kinderen natuurlijk minder woorden dan volwassenen. Naar de woordenschat van kinderen is veel onderzoek gedaan, omdat zo leerachterstanden aan het licht kunnen komen. Onderzoekers vragen kinderen bijvoorbeeld te benoemen wat er op een plaatje staat. Uit dat onderzoek blijkt dat de passieve en actieve woordenschat in de kindertijd snel toenemen, vooral na het vijfde jaar, als kinderen naar school gaan en leren schrijven (zie Figuur 5.2). Het gaat uiteraard om gemiddelden en om indicaties.[62]

| Leeftijd | Gemiddelde passieve woordenschat | Gemiddelde actieve woordenschat |
|---|---|---|
| 3 jaar | 1.250 | 1.000 |
| 4 jaar | 2.500 | 2.000 |
| 5 jaar | 3.500 | 3.000 |
| 6 jaar | 14.000 | 6.000 |
| 12 jaar | 26.500 | 17.000 |

Figuur 5.2 De gemiddelde passieve en actieve woordenschat van kinderen

En hoe zit het met de woordenschat van volwassenen? Daarover bestaan vele berekeningen en schattingen met zeer uiteenlopende aantallen. Het grootste en recentste onderzoek naar de omvang van de passieve woordenschat is in 2013 uitgevoerd door de Vlaamse taalpsycholoog Marc Brysbaert. Aan dit online-onderzoek deden 400.000 Nederlanders en Vlamingen van 12 jaar en ouder mee. De deelnemers gaven van 53.000 woorden aan of ze het kenden of niet. De lijst bevatte ook nepwoorden (vergelijkbaar met de spookwoorden uit hoofdstuk 1), zoals *beslopping*,

*lopsig*, *roog*, *tippelspin* en *soeptang*.[63] Er deden relatief gezien meer oudere Nederlanders, tussen 30 en 65, mee dan oudere Belgen.

Uit het onderzoek blijkt dat de deelnemers gemiddeld ongeveer 40.000 woorden kennen, maar ook dat de leeftijd sterke invloed heeft. Het blijkt namelijk dat de woordenschat tot op hoge leeftijd constant blijft groeien: 12-jarigen kennen gemiddeld 26.500 woorden en 80-jarigen gemiddeld 42.500. Dit is een verschil van bijna 16.000 woorden! Na je twintigste leer je gemiddeld iedere twee dagen een nieuw woord, terwijl je de oude woorden behoudt.

Verder hangt, niet heel verrassend, de omvang van de woordenschat af van iemands opleidingsniveau: hoe hoger dat is, hoe groter de woordenschat. Een laatste interessante conclusie is dat deelnemers die naast het Nederlands als moedertaal meerdere talen spreken, een groter aantal Nederlandse woorden kennen, en hoe meer talen men spreekt, hoe groter de woordenschat van het Nederlands is. Dit heeft waarschijnlijk te maken met het opleidingsniveau, maar misschien ook met een vergroting van de woordfamilies dankzij de verschillende talen. Opnieuw een bewijs dat meertaligheid loont.

Er bestaan interessante verschillen tussen groepen taalgebruikers. Zo zijn er woorden die de meeste Vlamingen wél en de meeste Nederlanders niet kennen, en die dus typisch voor het Belgisch-Nederlands zijn, zoals *aprilvis*, *bissen*, *denkpiste*, *inwijkeling*, *kattin*, *mattentaart*, *nefast*, *resem*, *tweewoonst* en *wegdeemsteren*. Andersom, woorden die de meeste Nederlanders kennen en Vlamingen niet, zijn *atjar*, *eigenheimer*, *gajes*, *kassiewijle*, *katenspek*, *kliko*, *omkukelen*, *vernachelen* en *vlaflip*.

Ook tussen mannen en vrouwen bleken verschillen te bestaan. Woorden als *apparatsjik*, *debuggen*, *infotainment*, *kevlar* en *konterfeitsel* zijn vooral bekend bij mannen, terwijl vrouwen meer vertrouwd zijn met *boothals*, *kooikerhondje*, *sleehak*, *smokwerk* en *zielenroerselen*.

Onderzoek onder Engelstalige taalgebruikers levert vergelijkbare gegevens en extra informatie op. Zo kwam onderzoeker Paul Nation voor wat betreft de omvang van de passieve woordenschat van volwassen taalgebruikers met een middelbareschoolopleiding uit op rond de 50.000 verschillende woorden.[64] Als we deze gegevens extrapoleren naar verbo-

gen en vervoegde woordvormen voor het Nederlands, komen we op minstens 100.000 woordvormen. De woorden waren te herleiden tot 20.000 woordfamilies. Zoals boven bleek, zijn woorden die tot één woordfamilie behoren, makkelijker te herkennen. Nation berekende dat kennis van 8.000 à 9.000 woordfamilies voldoende is om een geschreven Engelse tekst (roman, krant) te kunnen lezen, en kennis van 6.000 à 7.000 woordfamilies volstaat om gesproken tekst te kunnen volgen.[65] Er is geen reden te veronderstellen dat die aantallen voor het Nederlands heel anders zijn.

De actieve woordenschat, de woorden die een moedertaalspreker daadwerkelijk gebruikt, is veel kleiner dan de passieve: een volwassene met een redelijk opleidingsniveau beschikt actief over zo'n 30.000 woorden. Zowel de passieve als de actieve woordenschat van een individu is dus een stuk kleiner dan het aantal trefwoorden dat is opgenomen in een algemeen woordenboek, en komt niet in de buurt van het aantal types in een corpus. Die individuele woordenschat is echter prima geschikt voor de dagelijkse communicatie.

# 6 Woorden in een netwerk

Al in 1957 constateerde de Engelse taalkundige J.R. Firth: 'You shall know a word by the company it keeps'. Daarmee bedoelde hij dat de betekenis van een woord blijkt uit de context. Als *been* voorkomt met woorden als *hond, kluiven* of *soep*, dan weet je dat het gaat om *been* in de betekenis 'bot'. Komt het voor in de context van *armen, kousen, lopen, staan* of *breken*, dan is er sprake van de betekenis 'onderste ledemaat'. Ook toont de context dat *been* vaak voorkomt in bepaalde woordgroepen, zoals *op eigen benen staan, goed ter been zijn, een blok aan het been zijn, ergens geen been zien in*. Die woordgroepen hebben een eigen betekenis. In dit laatste hoofdstuk kijken we over de woordgrens heen naar dit soort woordgroepen. Neemt hun aantal toe? Zit er systeem in de manier waarop woorden samenklonteren? Wat kunnen we over woorden leren via de netwerken die ze vormen, en hoe helpt de computer daarbij?

## Voorspelbaarheid

Ieder woord kan theoretisch met ieder ander woord in een woordgroep of zin worden gecombineerd, maar in de praktijk blijkt dat bepaalde woordcombinaties veel vaker voorkomen dan andere. Dat zie je in een corpus en het ligt ook vast in ons geheugen. Als je leest of hoort: *jongens en…* vul je al snel aan *meisjes* of *mannen*, zeker als je de eerstvolgende klank (*ma-* of *mei-*) hoort. Hetzelfde geldt voor *man en…* (*paard, vrouw, macht*), *appels en… (peren)*.

Van veel woordcombinaties is dus voorspelbaar wat het volgende woord zal zijn. Op die voorspelbaarheid zijn de algoritmes voor woordsuggestie en autocorrectie op telefoons en computers gebaseerd: ze geven het meest waarschijnlijke vervolg aan, en hoe meer letters je intikt, hoe preciezer het programma kan voorspellen wat het volgende woord gaat worden. Erg origineel blijkt ons taalgebruik dus niet te zijn…

## Van vrije verbinding naar vaste verbinding

Sommige woordcombinaties zijn niet alleen voorspelbaar, maar ze liggen ook min of meer vast: de variatie in gebruiksmogelijkheid is beperkt en ze hebben een betekenis die niet direct is af te leiden uit die van de afzonderlijke woorden. Daarom noemen we ze vaste verbindingen.[66] Denk aan *de sterke arm* voor 'de politie'. Als je de vorm wijzigt (*een sterke arm*), verandert ook de betekenis: dan gaat het opeens over *arm* als lichaamsdeel. Een beperkte vormvariatie is vaak wel mogelijk: *ik ben in mijn sas* of *op mijn hoede*, maar *wij zijn in onze sas of op onze hoede*. Die vaste verbindingen worden beschreven in woordenboeken en grammatica's, en we slaan ze als betekeniseenheid op in ons mentale lexicon.

In zijn boek *Taal op Drift* uit 2013 beschrijft Joop van der Horst dat het aantal vaste verbindingen in het Nederlands in de loop van de tijd toeneemt. Hij wijst er bovendien op dat vaste verbindingen zoals *aan de drank, een gepasseerd station, in de gaten houden, in de war, onder vier ogen, op de hoogte, stuk voor stuk, zacht gezegd* en *zijn best doen* een relatief jong verschijnsel zijn in het Nederlands. Ze zijn rond de zeventiende eeuw ontstaan, maar de meeste zijn pas in de laatste honderd of tweehonderd jaar gevormd.[67]

Soms gaan ze terug op oudere vrije verbindingen, die in de loop van de tijd hun variatie verliezen: de vaste verbindingen worden steeds vaster. De varianten *voor mijn ogen komen, onder mijn ogen komen* en *onder de ogen komen* verdwenen ten gunste van de vaste verbinding *onder ogen komen*, waarin het lidwoord en bezittelijk voornaamwoord zijn verdwenen. We zeggen *afscheid nemen*, en niet meer, zoals vroeger, *een/mijn/ons afscheid nemen*. In deze zelfde periode kregen ook steeds meer werkwoorden en naamwoorden een vast voorzetsel, waar vroeger een naamvalsvorm werd gebruikt of variatie in voorzetsels bestond, vergelijk *lachen om* (vroeger ook: *lachen in/met/van*), *geloven/geloof in, kijken naar, trots op, opzien tegen, zich verbazen over, wachten op*.

Het aantal vaste verbindingen in het Nederlands is heel groot, veel groter dan de meeste mensen zich realiseren. Sommige bestaan uit een uitgebreid zelfstandig naamwoord, zoals *een open deur, een koekje van eigen deeg, één pot nat, een steuntje in de rug* en *slappe hap*. Andere vormen bij-

woordelijke uitdrukkingen: *bij nacht en ontij, heen en weer, hoog en droog, naar verluidt, op goed geluk, op de valreep, zacht gezegd*. Weer andere treden op als werkwoord: *aan de haal gaan, afscheid nemen, beet hebben, belangstelling hebben voor, hemel en aarde bewegen, het ijs breken, het paard achter de wagen spannen, in iemands voetsporen treden, 'm smeren, naar iets kunnen fluiten, op z'n beloop laten*.[68]

Dan zijn er nog uitroepen als *tot ziens, helaas pindakaas, hiep hiep hoera, jemig de pemig*, stopwoordjes als *zeg maar* en *of zo*, en routineformules of taalclichés als *je tante!, handen thuis, laat maar waaien, morgen brengen* of *dat hoor je mij niet zeggen*.[69] Een populaire subcategorie bestaat uit vergelijkingen: *dronken als een tol, lopen als een trein, trots als een pauw* en *trillen als een juffershondje*.

Eerder bleek dat functiewoorden behoren tot een kleine woordklasse die in de loop van de tijd maar beperkt wordt aangevuld. Vanaf de achttiende eeuw kreeg het aantal voorzetsels en voegwoorden echter een forse 'boost', doordat woordgroepen de functie van voorzetsel of voegwoord op zich namen. Zo ontstonden nieuwe voorzetseluitdrukkingen als *aan het adres van, aan de hand van, bij gelegenheid van, in plaats van, in tegenstelling tot, met behulp van, op grond van* en *ten behoeve van*, en nieuwe voegwoordelijke verbindingen als *elke keer dat, (enkel en) alleen als, in het besef dat, in het geval dat* en *in de veronderstelling dat*.

Doordat de vorm van vaste verbindingen grotendeels vastligt, bevatten ze regelmatig nog naamvalsvormen die kenmerkend zijn voor formele geschreven teksten en die allang uit het dagelijkse taalgebruik zijn verdwenen – scholieren moeten ze dan ook in het hoofd stampen. Voorbeelden zijn *bij monde van, in koelen bloede, te allen tijde, te elfder ure, te goeder trouw, ten laste van, ten tijde van, ter plaatse, van ganser harte* en *van goeden huize*. Ook bevatten ze soms woorden die verder niet meer voorkomen, vergelijk *in arren moede, te berde brengen, ergens de brui aan geven, van heinde en verre, de hort op zijn, iemand een loer draaien, op het nippertje*.

## Spreekwoorden en uitdrukkingen

Uit de gegeven voorbeelden blijkt dat vaste verbindingen kunnen bestaan uit een variabel aantal woorden. Sterker nog: ze kunnen zelfs een

hele zin beslaan. Dat geldt bijvoorbeeld voor spreekwoorden als *wie een kuil graaft voor een ander, valt er zelf in* of *een kat in het nauw maakt rare sprongen*. Spreekwoorden gelden vaak als oubollig. Men spreekt ze dan liever niet helemaal uit, maar beperkt zich tot een verwijzing: *de situatie van de spreekwoordelijke kat*.

Een andere manier om het clichématige karakter van spreekwoorden en uitdrukkingen te doorbreken is door ze te verkorten. Dat past binnen de algemene hang naar efficiënt en kort taalgebruik die, zo bleek, ook leidt tot woordverkortingen, maar het geeft de uitdrukking bovendien een ironische lading. Iedereen snapt wat je bedoelt met *dat is een gevalletje klok-klepel* of nóg korter *klok-klepel*, dus er is geen noodzaak om voluit *de klok horen luiden, maar niet weten waar de klepel hangt* te zeggen. En zo kun je ook volstaan met *(gevalletje) spijker-kop, zwaluw-lente, mosterd-maaltijd* en *boer-kiespijn*.[70]

Anders dan velen denken komen er, ondanks hun oubollige imago, nog steeds nieuwe spreekwoorden bij, recent bijvoorbeeld *een slimme meid is op haar toekomst voorbereid; doe maar gewoon, dan doe je al gek genoeg; kennis is macht; nood breekt wet* en *op een oude fiets moet je het leren*. In de moderne tijd zijn ook enkele citaten spreekwoordelijk geworden, zoals *de leugen regeert* van (toen nog) koningin Beatrix en *alles van waarde is weerloos* van Lucebert.[71]

## Vaste constructies

Naast vaste verbindingen bestaan er ook vaste constructies of formules, waarin verschillende (maar niet willekeurige) woorden kunnen worden ingevuld, zoals:

- *van x tot x* (*van dag tot dag, van minuut tot minuut, van deur tot deur*);
- *x op, x af* (*trap op, trap af; berg op, berg af*);
- *maar wat aan* (*rommelen, rotzooien, klooien*);
- *een x van een y* (*een schat van een kind, een loeder van een mens, een boom van een kerel, een kast van een huis*).

Enkele van die vaste constructies zijn in de loop van de tijd sterk uitgebreid, zo laat Jack Hoeksema zien.[72] Zo vertonen uitdrukkingen met *geen*

*x* voor 'niets' een grote groei: van 28 in de zeventiende eeuw naar 189 in het begin van deze eeuw; vergelijk *geen bal, geen barst, geen biet, geen donder, geen fluit, geen greintje, geen klap, geen lor, geen mallemoer, geen moer, geen sikkepit, geen spat, geen steek, geen snars* en *geen zier*.

Eenzelfde toename zien we bij de zogenoemde '*wemel*-constructie': *het wemelt van de muizen*. In oudere teksten komen maar vier varianten van deze constructie voor, in de eerste helft van de twintigste eeuw elf en tegenwoordig maar liefst 38, waaronder *barsten van, bol staan van, krioelen van, ritselen van, sterven van, stikken van, vergeven zijn van* en *zwart zien van.*

## Onzichtbare netwerken

Vaste verbindingen bestaan uit woorden die in elkaars directe nabijheid staan. Er bestaan echter ook netwerken tussen verder van elkaar gelegen woorden: netwerken die voor ons taalgebruikers onzichtbaar zijn maar die computers kunnen blootleggen. Zo rekenen de gratis Voyant Tools bijvoorbeeld in enkele seconden uit wat de belangrijkste verschillen in woordgebruik en woordfrequentie zijn tussen twee teksten. Ik heb die tool gevoerd met een essay uit 1959 van Hella Haasse (*Dat weet ik zelf niet*) en de verzameling gedichten over *Het schaap Veronica* van Annie M.G. Schmidt uit 1960. In Figuur 6.1. staan de gegevens die Voyant Tools ophoest; vergelijk dit ook met Figuur 2.4.

| Bron | Hella Haasse 1959 | Annie M.G. Schmidt 1960 |
|---|---|---|
| Totaal aantal woorden (tokens) | 22.911 | 17.916 |
| Aantal verschillende woorden (types) | 5.199 | 2.982 |
| Type/token-verhouding | 0,23 | 0,17 |
| Aantal woorden per zin | 22,4 | 10,6 |
| Typerende woorden | *kind, kinderen, vader, jeugd, mens* | *dominee, Veronica, zeiden, kijk, hè* |

Figuur 6.1 Informatie die Voyant Tools haalt uit teksten van Hella Haasse en Annie M.G. Schmidt

Uit de gegevens van Figuur 6.1 blijkt dat Schmidt veel kortere zinnen schrijft dan Haasse en minder verschillende woorden gebruikt (de type/token-verhouding is kleiner). Informatief is het lijstje van typerende woorden, woorden waarin de twee teksten zich van elkaar onderscheiden. Voor Schmidt blijken de spreektalige vormen *kijk* en *hè* kenmerkend te zijn. Kortere zinnen, minder woordvariatie en spreektalige woorden duiden op een eenvoudigere tekst, en dat past natuurlijk helemaal bij het feit dat de gedichten van Schmidt bestemd waren voor kinderen, terwijl de tekst van Haasse een essay voor volwassenen was. Computers blijken dus in staat om dit verschil in genres en doelgroepen op te sporen. Iets ingewikkeldere programma's kunnen bovendien helpen om de auteur te achterhalen van een onder pseudoniem gepubliceerde literaire tekst, of de schrijver van een anonieme dreigbrief te ontmaskeren.

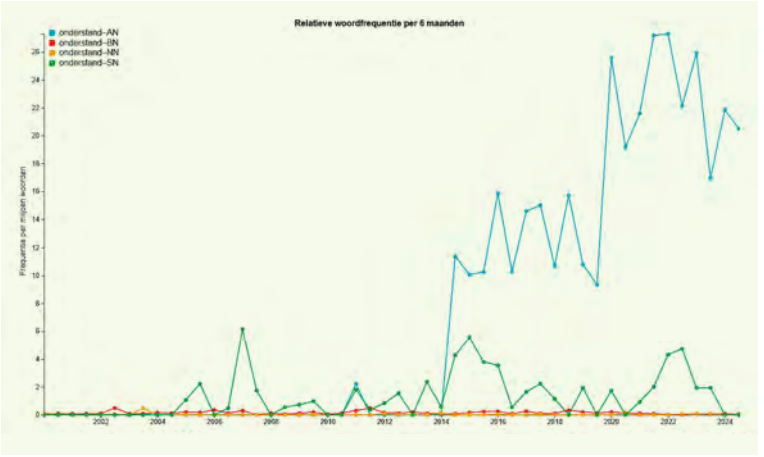
Een verdere toepassing bestaat eruit dat de computer berekent hoe makkelijk of moeilijk leesbaar een tekst is. Er bestaan veel verschillende programma's die leesbaarheid meten. Daarbij baseren ze zich op verschillende factoren, zoals gemiddelde woordlengte, aantal woorden per zin, aantal lettergrepen per woord, type/token-verhouding, woordfrequenties en concreetheid van woorden.[73] Voyant Tools geeft de tekst van Haasse een leesbaarheidsscore van 14,6, dat is ongeveer het niveau van een derdejaars student, en die van Schmidt een score van 13,2.

Dergelijke programma's kunnen ook stijlverschillen tussen auteurs aantonen.[74] Erg aardig en leerzaam is het in Antwerpen ontworpen programma Stylene, waarmee je je eigen schrijfstijl kunt analyseren. Van een ingevoerde tekst voorspelt Stylene het geslacht, de leeftijd, het opleidingsniveau en de persoonlijkheid van de auteur, tot welk genre de tekst behoort en bij welke literaire schrijver de tekst stilistisch het meest aansluit. Foutloos zijn dergelijke voorspellingen uiteraard niet: volgens Stylene is *Het schaap Veronica* geschreven door een hoogopgeleide man van boven de 50, de tekst sluit het meest aan bij Brusselmans, en het leesbaarheidsniveau is voor een leeftijd van 8 à 9 jaar. Dat laatste is in ieder geval correct.

Sommige programma's gaan nog een stapje verder en doen suggesties voor tekstverbeteringen. Voor Nederlandstaligen is LanguageTool

erg interessant: de tool kan je tekst herschrijven in standaardtaal, formele taal of eenvoudige taal.[75] Voor iedere schrijver erg leerzaam!

Een laatste nuttige toepassing vormt de Woordpeiler van het Instituut voor de Nederlandse Taal. Daarin kun je vinden hoe vaak een woord genoemd wordt in kranten tussen 2000 en nu, en of een woord in het hele Nederlandstalige taalgebied voorkomt of slechts in een deel ervan. Figuur 6.2 laat het resultaat voor het woord *onderstand* zien, uitgesplitst in het Antilliaans-, Belgisch-, Nederlands- en Surinaams-Nederlands.



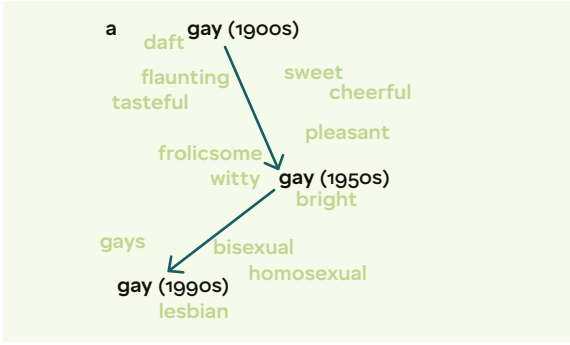

Figuur 6.2 *Onderstand* in Antilliaans-, Belgisch-, Nederlands- en Surinaams-Nederlandse kranten vanaf het jaar 2000

*Onderstand* is in het Antilliaans-Nederlands (blauw) en het Surinaams-Nederlands (groen) het normale woord voor 'bijstand', en heeft in die varieteiten een relatief hoge frequentie. In het Belgisch-Nederlands (rood) komt het een enkele maal voor, meestal in historische zin, als verkorting van *openbare onderstand* – een benaming die in 1976 officieel werd vervangen door *sociale bijstand*. In het Nederlands-Nederlands (oranje) was *onderstand* vroeger normaal, maar in 1963, toen de Bijstandswet inging, werd het vervangen door *bijstand*; de enkele keer dat *onderstand* in de eenentwintigste eeuw voorkomt in Nederlandse kranten, verwijst het naar de situatie van de Nederlandse Antillen of Suriname.

## Betekenisverschuivingen

De computer kan zelfs betekenisverbanden tussen woorden leggen. Daarvoor zetten speciale computerprogramma's die gebruik maken van zogenaamde taalmodellen, de woorden in een corpus om in rijtjes getallen ('woordinbeddingen' genaamd) en daaraan voegen ze informatie toe over de contexten waarin die woorden voorkomen, opnieuw in de vorm van rijtjes getallen ('vectoren'). De vectoren van woorden die in betekenis verwant zijn, liggen dicht bij elkaar. Zo kan de computer woorden vinden die min of meer synoniem zijn, zoals *stoel, zetel, fauteuil*, of een andere betekenisrelatie hebben, zoals *kat* en *hond* (twee huisdieren) of *koning, koningin, prins* en *prinses* (koninklijke figuren die verschillen in leeftijd en gender).

Nu kunnen betekenissen in de loop van de tijd verschuiven, en het blijkt dat computerprogramma's kunnen helpen bij het opsporen daarvan. Daarvoor hebben Amerikaanse onderzoekers de computer gevoed met corpora uit verschillende periodes en een programma geschreven dat de betekenisrelaties in de verschillende periodes met elkaar vergelijkt.[76] In Figuur 6.3 staat het resultaat voor het Engelse woord *gay*. Hieruit blijkt dat *gay* rond 1900 in dezelfde contexten voorkwam als woorden met de betekenis 'nonchalant, opgewekt'; 50 jaar later kwam *gay* meer in de buurt van woorden voor 'vrolijk, grappig', en in de moderne tijd verschoof de betekenis naar 'homo, lesbienne'.




Figuur 6.3 De betekenisverschuiving van *gay* in het Engels, volgens een computervergelijking van historische corpora

De onderzoekers vonden bovendien enkele statistische wetmatigheden in dit soort betekenisveranderingen. Ten eerste bleek dat de betekenis van frequente woorden langzamer verandert dan die van infrequente woorden. Die wetmatigheid is overigens ook bekend van sterke werkwoorden: de verleden tijd van frequente sterke werkwoorden, zoals *vroeg*, verandert niet, terwijl die van infrequente werkwoorden vaak zwak wordt (*joeg* wordt vervangen door *jaagde*). Ten tweede bleek dat de betekenis van een woord met veel verschillende betekenissen sneller verandert dan die van een woord met slechts een enkele betekenis.

## ChatGPT

De meest recente computerontwikkeling – het kan niemand zijn ontgaan – is de lancering van grote taalmodellen als ChatGPT: tekstrobots of chatbots die woorden aaneenrijgen tot vloeiende zinnen en teksten, en antwoorden geven op vragen die de gebruikers stellen.[77] De systemen zijn zelflerend en gebaseerd op een enorme hoeveelheid teksten, waaruit ze hun antwoorden destilleren. Hoe het onder de motorkap werkt, is onduidelijk.

Terwijl de meeste computerprogramma's grote moeite hebben het verschil te zien tussen letterlijk en figuurlijk taalgebruik, draait Chattie (zoals ik haar noem) daar haar hand niet voor om. Op mijn vraag wat *ik reed hem in de wielen* betekent, meldde de gratis versie 3.5 in januari 2025: 'Het betekent dat je iemand dwarsboomt of hindert, vaak op een manier die hun plannen of vooruitgang belemmert.' *Hun* zal wel de genderneutrale keuze zijn. Chattie lijkt dus betekenis te 'snappen' en daarmee de eerste wet van Hugo Brandt Corstius te ontkrachten, die inhield dat semantiek in de computer-taalkunde altijd roet in het eten gooit. Daarna raakt Chattie echter het spoor bijster, want met grote stelligheid beweert hun: 'Het komt oorspronkelijk uit de wielersport, waar het letterlijk betekent dat je met je fiets de wielen van een andere fietser raakt, waardoor die persoon gehinderd wordt.' Hier is nog enige ruimte voor verbetering. De uitdrukking dateert namelijk al uit 1797, lang voor de uitvinding van de fiets, en verwijst naar de wielen van een rijtuig.

Het bijzondere van Chattie en andere recente zelflerende computerprogramma's is dat woorden, woordcombinaties en zinnen – tot nu toe het exclusieve terrein van de mens – voor het eerst ook door computers worden 'begrepen'. Om tot dat 'begrip' te komen, moeten de woorden weliswaar omgezet worden in getallen, maar daarmee kunnen de computerprogramma's dan voor de mens onzichtbare verbanden tussen woorden leggen. Bovendien is hun woordenschat onbeperkt uitbreidbaar met nieuwe woorden: computers lopen, anders dan mensen, niet aan tegen de grenzen van hun geheugen, tenminste zolang de elektriciteit werkt.

# 7 En, dijt onze woordenschat alsmaar uit?

Dit boekje gaf in kort bestek een inkijkje in wat we weten over woorden en hun eigenschappen. Daarmee is natuurlijk nog lang niet alles gezegd over de combinatiemogelijkheden van woorden, hun betekenisschakeringen en hun betekenisrelaties. Wél werd duidelijk dat het Nederlands prat kan gaan op een rijkgeschakeerde en omvangrijke woordenschat.

In de verschillende hoofdstukken bleek dat er geen eenduidig antwoord bestaat op de vraag of die woordenschat uitdijt of niet. Het hangt er maar net van af wat je verstaat onder woord en welke woordenschat je bedoelt. 'Woord' blijkt namelijk een dubbelzinnige en vage term te zijn. Lexicografen zien een 'woord' meestal als een trefwoord of lemma in een woordenboek, in een corpus spelen types en tokens de hoofdrol, in veel computerprogramma's draait het om getallen als representatie van woorden, en in het mentale lexicon gaat het om betekeniseenheden, en dat kunnen ook vaste verbindingen zijn als *om de haverklap*.

Er bestaan belangrijke verschillen tussen de woordenschat in een woordenboek, in een corpus en in het hoofd. In algemene woordenboeken blijft de omvang van de woordenschat min of meer in evenwicht of groeit maar beperkt: enerzijds voegen lexicografen nieuwe woorden toe voor maatschappelijke, technische en wetenschappelijke vernieuwingen, anderzijds schrappen uitgevers trefwoorden, zodat de woordenboeken niet te dik en te duur worden. Dat laatste gebeurt niet in historische woordenboeken: die dijen alsmaar uit, omdat er in de loop van de tijd telkens nieuwe woorden bijkomen, terwijl de woorden die niet langer door de taalgebruikers gehanteerd worden, hun plaatsje in het woordenboek behouden. Daarom geven ze geen goed tijdsbeeld van hoeveel woorden er op een bepaald moment in omloop zijn.

Als je kijkt naar de woordenschat in ons hoofd, het mentale lexicon, dan blijkt dat de woordenschat zich uitbreidt naarmate we ouder worden en een hogere opleiding volgen, maar daaraan zit een grens, want we kun-

nen niet alles onthouden. Het aantal woorden in een corpus is daarentegen oneindig: het aantal blijft toenemen zolang het corpus groeit. Zo'n enorm corpus gaat het bevattingsvermogen van de mens verre te boven.

Het vormen van nieuwe samenstellingen en afleidingen op basis van bestaande woorden blijkt dan weer tegen een grens te lopen. De combinatiemogelijkheden zijn weliswaar groot, maar in de praktijk niet onbeperkt: niet ieder woord kan qua betekenis of woordvorm met ieder ander woord of woorddeel worden gecombineerd.

Een belangrijke eigenschap van woorden bleek de woordsoort te zijn waartoe ze behoren: sommige woordsoorten – met name naamwoorden en werkwoorden – hebben veel leden en worden gemakkelijk met nieuwe leden uitgebreid, terwijl functiewoorden (lidwoorden, voorzetsels en voegwoorden) nauwelijks aangroeien. In een corpus zie je precies het tegenovergestelde: het kleine groepje functiewoorden heeft de hoogste frequentie en komt dus het vaakst voor. Dat komt door hun functie in de zin. Van woorden bouwen we zinnen door veranderlijke naamwoorden en werkwoorden aan elkaar te plakken. Daarbij dient het kleine groepje functiewoorden als lijm. In de loop van de tijd is daar de vaste verbinding als extra bouwsteen bij gekomen als tussenstap tussen woord en zin. Het aantal vaste verbindingen groeit nog steeds, en de combinatiemogelijkheden van woorden en vaste verbindingen is eindeloos: dat betekent dat de hoeveelheid zinnen in een taal onbeperkt is.

## De toekomst

We weten nog lang niet alles over woorden en woordrelaties. Maar onze kennis neemt de laatste tijd wel snel toe. Slimme computeralgoritmes leggen voor de mens onzichtbare betekenispatronen en grammaticale relaties in teksten bloot. Oude teksten leveren daarbij echter voor de computer nog problemen op, door de grote spellingvariatie, verouderde constructies en veranderde betekenissen. Voor die oude teksten moeten nóg slimmere computeralgoritmes worden geschreven. Of we dat kunnen overlaten aan een taalmodel als ChatGPT is de vraag. Chattie kan verrassend veel en is zelflerend, maar op oudere teksten bijt ze haar tandjes nog stuk, niet alleen vanwege hun complexiteit, maar ook vanwege hun relatieve schaarste.

Mensen moeten bovendien de juiste vragen stellen – van Chattie kun je die niet verwachten – en alleen mensen kunnen beoordelen of de antwoorden die ze geeft aan de verwachtingen voldoen. Een 'historische' ChatGPT, die gevoed is met teksten van de oudste tijden tot nu en nieuwe inzichten kan geven in de veranderende opbouw van de woordenschat en woordrelaties, lijkt voorlopig nog een utopie.

Er is dus nog veel fascinerend en uitdagend werk voor jonge onderzoekers. Veel theoretische vragen zijn nog onbeantwoord: waarom vormen sommige woorden grote netwerken of families, en blijven andere eenzaam? Hoe en waardoor veranderen betekenissen? Wat is de gemiddelde levensduur van een woord en welke factoren verlengen of verkorten die? Daarnaast ligt er nog veel praktisch werk in het uitdenken van computertoepassingen voor bijvoorbeeld leesbaarheidsscores, forensische taalkunde en stijlonderzoek, en in het waarborgen dat taalmodellen transparant, objectief en inclusief zijn, of liever: worden. Al dat werk draait om het woord en zijn eigenschappen. Ik zou zeggen: geen woorden, maar daden!

# Dankwoord

Roland de Bonth, Paul Hulsenboom, Marc van Oostendorp, Ewoud Sanders en Rik Schutz lazen de hele tekst mee en deden nuttige suggesties. Jesse de Does voorzag de hoofdstukken 2 en 6 van commentaar en vervaardigde de figuren 2.2 en 2.5, terwijl Gerard Kempen nuttige kritiek had op de hoofdstukken 2, 5 en 6. Mijn collega's Katrien Depuydt, Kris Heylen, Carole Tiberius, Vincent Vandeghinste, Boukje Verheij en Vivien Waszink van het Instituut voor de Nederlandse Taal leverden data voor figuren die gebaseerd zijn op woordenboeken of databanken van het instituut, en deden literatuursuggesties. Aan Oele Koornwinder dank ik de gegevens voor Figuur 4.3 met een analyse van de (on)gelede woorden in de Dikke Van Dale 1999. Ik ben hen allen heel dankbaar voor hun waardevolle opmerkingen.

# Literatuur en verder lezen


- Booij, Geert en Ariane van Santen (1998), *Morfologie – de woordstructuur van het Nederlands*, Amsterdam.
- Booij, Geert (2002), *Morphology of Dutch*, Oxford.
- Boon, Ton den (2010), *Van Dale Modern verdwijnwoordenboek. Van aamborstigheid tot zwijmelgeest*, Utrecht/Antwerpen.
- Boon, Ton den (2015), *Het nieuwe verdwijnwoordenboek*, Varik.
- Brandt Corstius, Hugo (1978), *Computer-taalkunde*, Muiderberg.
- Brysbaert, Marc (2003), 'Hoe werkt tweetaligheid?', in: *Neuron* 8: 16-21.
- Brysbaert, Marc en Ton Dijkstra (2006), 'Changing views on word recognition in bilinguals', in: *Proceedings of the Belgian Academy of Psychology*, red. J. Morais en G. d'Ydewalle, pp. 25-37.
- Brysbaert, Marc, Emmanuel Keuleers, Pawel Mandera en Michaël Stevens (2013), *Woordenkennis van Nederlanders en Vlamingen anno 2013: Resultaten van het Groot Nationaal Onderzoek Taal*, Gent.
- Brysbaert, Marc, Michaël Stevens, Pawel Mandera en Emmanuel Keuleers (2016), 'How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age', in: *Frontiers in Psychology* 7, 1116.
- Brysbaert, Marc (2021), 'De invloed van verwervingsleeftijd op woordherkenning', in: *Wat gebeurt er in het Nederlands?! Over taal, frequentie en variatie*, red. N. van der Sijs, L. Fonteyn en M. van der Meulen, Gorredijk, pp. 233-238.
- Butterfield, Jeremy (2008), *Damp squid. The English language laid bare*, Oxford.
- Cornelisse, Paulien (2025), *Hèhè. Over wat we zeggen zonder dat we het doorhebben*, Amsterdam.
- Dalen-Oskam, Karina van (2021), *Het raadsel literatuur. Is literaire kwaliteit meetbaar?*, Amsterdam.

- Deyne, Simon De en Gert Storms (2013), 'Het mentale lexicon als een semantisch web', in: *Logopedie* juli/augustus, pp. 1-11.
- Dijkstra, Ton en Gerard Kempen (1984), *Taal in uitvoering. Inleiding tot de psycholinguïstiek*, Groningen.
- Dingemanse, Mark e.a. (2013), 'Is "Huh?" a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items', in: *PLoS ONE* 9(4): e94620.
- Dingemanse, Mark (2015), 'Emoji's: waarom we taal en schrift niet moeten verwarren', in *Neerlandistiek*.
- Engelsman, Jaap (2004), *Bekende citaten uit het dagelijks taalgebruik*, Den Haag.
- Everaert, Martin (1993), 'Vaste verbindingen (in woordenboeken)', in: *Spektator* 22 (1): 3-27.
- Fritschy, Yannick (2019), *De stam van het woord. Over taalevolutie en de eerste taal ter wereld*, Utrecht.
- Groot, Hans de (2006), *Van Dale Groot Uitdrukkingenwoordenboek. Verklaring en herkomst van moderne uitdrukkingen*, Utrecht.
- Haas, Wim de en Mieke Trommelen (1993), *Morfologisch handboek van het Nederlands. Een overzicht van de woordvorming*, 's-Gravenhage.
- Hamans, Camiel (2021), *Border cases in morphology*, Amsterdam.
- Hamilton, William L., Jure Leskovec en Dan Jurafsky (2016), 'Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change', in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1489-1501.
- Haspelmath, Martin (2023), 'Defining the word', in: *Word* 69 (3): 283-297.
- Heyvaert, Frans e.a. (red.) (1998), *Het grootste woordenboek ter wereld. Een kijkje achter de kolommen van het Woordenboek der Nederlandsche Taal (WNT)*, Den Haag.
- Hoeksema, Jack (2005), 'Rijkdom en weelde van het Nederlands', in: *Tabu* 34: 1-12.
- Horst, Joop en Kees van der (1999), *Geschiedenis van het Nederlands in de twintigste eeuw*, Den Haag.

- Horst, Joop van der (2013), *Taal op drift. Lange-termijnontwikkelingen in taal en samenleving*, Amsterdam.
- Jansen, Mathilde, Nicoline van der Sijs, Fieke van der Gucht en Johan De Caluwe (2017), *Atlas van de Nederlandse taal*, Tielt.
- Jansen, Mathilde (2023), 'Meertaligheid: geen nadeel, maar voordeel', in: *NEMO Kennislink*.
- Kestemont, Mike en Folgert Karsdorp (2024), *Het mysterie van de verdwenen ridderromans*, Borgerhout.
- Koornwinder, Oele (2005), *Morfologische aspecten van het ideale woordenboek*, Utrecht.
- Meesters, Gert (2002), *Marginale Morfologie in het Nederlands: paradigmatische samenstellingen, neoklassieke composita en splintercomposita*, Leuven.
- Melchers, Fleur (2019), 'Engels in the language van Nederlandse vloggers. Een onderzoek naar de invloed van de Engelse taal op het taalgebruik van Nederlandstalige vloggers', masterscriptie Utrecht.
- Metcalf, Allan (2002), *Predicting New Words. The secrets of their success*, Boston.
- Nation, Paul (2006), 'How large a vocabulary is needed for reading and listening?', in: *Canadian Modern Language Review* 63: 59–82.
- Nation, Paul en Averil Coxhead (2021), *Measuring native-speaker vocabulary size*, Amsterdam/Philadelphia.
- Noordervliet, Nelleke (2015), *1000 vergeetwoorden om te koesteren*, Utrecht/Antwerpen.
- Oostendorp, Marc (2015), 'Nu komt de uit de', in: *Neerlandistiek*.
- Oostendorp, Marc (2022), 'De m/v/x', in: *Neerlandistiek*.
- Pander Maat, Henk en Nick Dekker (2016), 'Tekstgenres analyseren op lexicale complexiteit met TScan', in: *Tijdschrift voor Taalbeheersing* 38 (3): 263-304.
- Permentier, Ludo (2022), 'De non-binaire non', in: *Woorden weten alles*.
- Reuneker, Alex, Vivien Waszink en Ton van der Wouden (2017), 'Sanskriet op de beat', in: *Neerlandistiek*.
- Samuel, Mounir (2023), *Je mag ook niets meer zeggen. Een nieuwe taal voor een nieuwe tijd*, Uitgeverij Nieuw Amsterdam.

- Sanders, Ewoud (2024), *Jood, de vergeten geschiedenis van een beladen woord*, Zutphen.
- Schuring, Melissa en Eline Zenner (2022), 'English from scratch: Preadolescents' developing use of English lexical resources in Belgian Dutch', in: *Frontiers in Communication*, vol. 6.
- Sijs, Nicoline van der (1998), *Geleend en uitgeleend. Nederlandse woorden in andere talen & andersom*, Amsterdam.
- Sijs, Nicoline van der (2001), *Chronologisch woordenboek van het Nederlands. De ouderdom en herkomst van onze woorden en betekenissen*, Amsterdam.
- Sijs, Nicoline van der (2005), *Van Dale Groot Leenwoordenboek. De invloed van andere talen op het Nederlands*, Utrecht/Antwerpen.
- Sijs, Nicoline van der (2019), *15 eeuwen Nederlandse taal*, Gorredijk.
- Sijs, Nicoline van der (2021), *Taalwetten maken en vinden: het ontstaan van het Standaardnederlands*, Gorredijk.
- Sijs, Nicoline van der, Lauren Fonteyn en Marten van der Meulen (red.) (2021), *Wat gebeurt er in het Nederlands?! Over taal, frequentie en variatie*, Gorredijk.
- Sijs, Nicoline van der (2022), 'Chattie', in: *Neerlandistiek*.
- Sijs, Nicoline van der (2023), 'Een "boom" aan Engelse leenwoorden', in: *Onze Taal* 1: 20-23.
- Sijs, Nicoline van der (2023a), 'De overlevingsgraad van Engelse leenwoorden', in: *In termen van taal. Liber amicorum Frieda Steurs*, red. P. van Sterkenburg, R.de Bonth en K. Heylen, pp. 286-296.
- Sijs, Nicoline van der (2024), 'Van "zpel" tot "sich". Zeventiende-eeuwse kranten in de Republiek en de Spaanse Nederlanden', in: *Neerlandistiek*.
- Sterkenburg, Piet van (2011), *Van woordenlijst tot woordenboek. Een geschiedenis van woordenboeken van het Nederlands*, Schiedam.
- Stroop, Jan (2014), 'Ofzo: het landingsgestel van de Nederlandse zin', in *Neerlandistiek*.
- Taylor, John R. (red.) (2015), *Oxford Handbook of the Word*, Oxford.

- Toorn, Maarten van den, Wil Pijnenburg, Arjan Van Leuvensteijn en Joop van der Horst (red.) (1997), *Geschiedenis van de Nederlandse taal*, Amsterdam.
- Vandeghinste, Vincent en Bram Bulté (2019), 'Linguistic proxies of readability: Comparing easy-to-read and regular newspaper Dutch', in: *Computational Linguistics in the Netherlands Journal* 9: 81-100.
- Veen, Pieter van en Nicoline van der Sijs (1997), *Etymologisch woordenboek. De herkomst van onze woorden*, Utrecht/Antwerpen.
- Velde, Freek Van de (2020), 'Zwindende woorden', lezing op symposium 'Op je woorden letten', 5 november.
- Velde, Freek Van de & Alek Keersmaekers (2020), 'What are the determinants of survival curves of words? An evolutionary linguistics approach', in: *Evolutionary Linguistic Theory* 2(2): 127-137.
- Velde, Freek Van de (2022), 'Het uitdijende bestand van woordgeslacht in Van Dale', in: *Neerlandistiek*.
- Vooys, Cornelis de (1957), *Nederlandse spraakkunst*, 4de druk herzien door M. Schönfeld, Groningen/Djakarta.
- Vries, Matthias de (1851), *Ontwerp van een Nederlandsch woordenboek*, Brussel.
- Walch, Jan (1928), *Uit de levensgeschiedenis van woorden*, Zutphen.
- Waszink, Vivien, Alex Reuneker en Ton van der Wouden (2018), 'Als ik praat, dan praat ik money', in: *Neerlandistiek*.
- Waszink, Vivien (2022), *Dat mag je óók (al niet meer) zeggen. Welke woorden kunnen en welke juist niet?*, Onze Taal.
- Wingerden, Wouter van en Pepijn Hendriks (2015), *'Dat hoor je mij niet zeggen!' De allerbeste taalclichés*, Amsterdam.
- Wolf, Henk (2019), 'Het meisjemeisje was thuisthuis', in: *Neerlandistiek*.


## Digitale woordenboeken en corpora

Op de website van het Instituut voor de Nederlandse Taal is een groot aantal corpora, lexica, woordenboeken, grammatica's en tools van het Nederlands te vinden. Op de volgende pagina een overzicht van de belangrijkste daarvan, aangevuld met andere relevante bronnen.

- ANW (Algemeen Nederlands Woordenboek).
- CHN (Corpus Hedendaags Nederlands; achter log-in).
- Couranten Corpus (zeventiende-eeuwse kranten).
- DAGENTA (Database Geschiedenis Nederlandse Taalkunde).
- DBNL (Digitale Bibliotheek voor de Nederlandse Letteren).
- Delpher (online databank met gedigitaliseerde boeken, kranten en tijdschriften).
- DiaMaNT (Diachroon seMAntisch lexicon van de Nederlandse Taal).
- Van Dale (gratis, verkorte versie).
- e-ANS (elektronische Algemene Nederlandse Spraakkunst).
- Etymologiebank, samengesteld door Nicoline van der Sijs.
- eWND (elektronische Woordenbank van de Nederlandse Dialecten, samengesteld door Nicoline van der Sijs).
- FUDGE-test (test naar de houdbaarheid van nieuwe woorden).
- GLAD (Global Anglicism Database).
- Historische woordenboeken van het Nederlands (WNT, MNW, VMNW, ONW).
- LanguageTool (meertalige grammatica- en spellingscontrole).
- OpenSoNaR (corpus gedrukte media en nieuwe media van rond 2010; achter log-in)
- Stylene (tool die de schrijfstijl van een auteur analyseert).
- Textalyzer (voor het tellen van het aantal woorden, lettergrepen, zinnen in een tekst).
- Uitleenwoordenbank (Nederlandse woorden geleend door andere talen, samengesteld door Nicoline van der Sijs).
- Voyant Tools (tool die teksten analyseert).
- WNW (Woordenboek van Nieuwe Woorden).
- Woordassociaties (uitgebreide databank van woordassociaties).
- Woordcombinaties (tool die toont welke woorden vaak met elkaar gecombineerd worden).
- Woordenlijst Nederlandse Taal (de lijst met de officiële spelling van het Nederlands).
- Woordpeiler (tool die woordtrends van 2000 tot nu toont).

# Noten

1 Den Boon 2010, 2015; Noordervliet 2015.
2 Zie de 'Introduction' in Taylor 2015, en Haspelmath 2023.
3 Uit onderzoek van Mark Dingemanse en anderen (2013) is naar voren gekomen dat *huh* een universeel woord is dat overal ter wereld op min of meer dezelfde manier wordt gebruikt als mensen iets niet begrijpen. Voor het gebruik van *hèhè* zie Cornelisse 2025.
4 Zie Dingemanse 2015.
5 Zie ook hoofdstuk 2 'How Many Words Are There?' door Kilgarri in Taylor (red.) 2015.
6 Over de geschiedenis van woordenboeken, spelling en grammatica zie Van der Sijs 2021. Zie ook hoofdstuk 3 'Words and Dictionaries' door Alexander in Taylor (red.) 2015.
7 Voor een beschrijving van het WNT zie Heyvaert e.a. 1998.
8 Van Sterkenburg 2011 beschrijft de geschiedenis van woordenboeken van het Nederlands.
9 De Vries 1851 en 'Inleiding' van het eerste deel van het WNT, 1882: xlv-xlvi.
10 'Inleiding' van het eerste deel van het WNT, 1882: xlviii.
11 'Inleiding' van de tiende druk van Van Dale, 1976: xv.
12 Zoek maar eens naar 'Van Dale' in de gerechtelijke uitspraken op de site rechtspraak.nl: https://tinyurl.com/hdj2abma
13 Walch 1928: 110.
14 Zie https://www.vandale.nl/gratis-woordenboek/meest-opgezochte-woorden; en https://onzetaal.nl/schatkamer/lezen/varia/taalrecords/het-meest-opgezochte-woord; de informatie van de Dikke Van Dale betreft één dag in 2003. De gegevens van het ANW (op https://anw.ivdnt.org/search) en Woordenlijst.org zijn afkomstig van het Instituut voor de Nederlandse Taal en betreffen de woorden die in het jaar 2024 het meest zijn opgezocht.
15 Sanders 2024: 163-174.
16 https://www.vandale.nl/merknamen
17 Zie https://ondernemersplein.kvk.nl/merkenrecht/
18 Zie voor meer voorbeelden Waszink 2022 en Samuel 2023.
19 Zie voor de discussie Van Oostendorp 2022, Permentier 2022, Van de Velde 2022.
20 'Les dictionnaires sont en général des plagiats par ordre alphabétique', aldus Charles Nodier, *Questions de littérature légale: du plagiat, de la supposition d'auteurs, des supercheries qui ont rapport aux livres*, 1828: 37-38.

21 https://neerlandistiek.nl/2025/02/woordenboekgebruik/
22 Het ANW omvat vrij veel trefwoorden die (nog) geen betekenisomschrijving hebben, waaronder een groot aantal doorzichtige samenstellingen.
23 De gegevens voor CHN kreeg ik van het Instituut voor de Nederlandse Taal. De overige gegevens zijn afkomstig uit Vandeghinste en Bulté 2019; onder de frequentste tokens vielen ook enkele leestekens, maar die heb ik in de figuur weggelaten.
24 Meer in hoofdstuk 5 'Word Frequencies', door Sorell in Taylor (red.) 2015.
25 Butterfield 2008: 16-20.
26 Meer in hoofdstuk 6 'Word Length' door Grzybek in Taylor (red.) 2015.
27 Een overzicht van tussenwerpsels staat in de e-ANS: https://e-ans.ivdnt.org/topics/pid/topic-13371852971604587; over *of zo* en *zeg maar*: Stroop 2014.
28 Zie https://taalmaterialen.ivdnt.org/download/tstc-gigant-molex2-0-c/ en Jansen e.a. 2017: 60-63.
29 Zie Van der Sijs 2024.
30 Reuneker e.a. 2017, Waszink e.a. 2018.
31 Tegenwoordig wordt vooral gewerkt met machinelearning: een algoritme dat getraind wordt aan de hand van een dataset en op basis daarvan nieuwe data kan analyseren.
32 Zie Fritschy 2019 en Van der Sijs 2019.
33 Een overzicht van klanknabootsingen geeft Van der Sijs 2001: 190:203.
34 Inclusief trefwoorden zonder behandeling maar met alleen een verwijzing komt dit aantal op 467.288; zie voor meer precieze gegevens 'Het WNT in cijfers' in Heyvaert e.a. 1998: 319-338.
35 Nog niet alle woorden in het ANW en het WNW zijn beschreven, dus deze aantallen zullen nog hoger worden.
36 Kestemont en Karsdorp 2024.
37 Van der Sijs 2019: 235-237.
38 Van der Sijs 2023; Melchers 2019; Schuring en Zenner 2022; overzichten van concrete leenwoorden in Van der Sijs 2005.
39 Van Veen en Van der Sijs 1997; Van der Sijs 1998: 158-188; zie voor de recentste Engelse leenwoorden in het Nederlands de Global Anglicism Database (GLAD). Een deel van die woorden (*boss, gin, cookie, skate*) had het Engels eerder ontleend aan het Nederlands, want Nederlandse woorden zijn veelvuldig uitgeleend, zie de Uitleenwoordenbank.
40 Van der Sijs 2023a.
41 Van der Sijs 2001: 501-518; zie verder de passage over nieuwe voorzetseluitdrukkingen en voegwoordelijke verbindingen in hoofdstuk 6 van dit boek.
42 Meer hierover in Van der Sijs 2019.
43 Zie Metcalf 2002; Jansen e.a. 2017, hoofdstuk 16 en 17.

44 Zie Van de Velde 2020, Van de Velde & Keersmaekers 2020; https://diamant.ivdnt.org/diamant-ui/.
45 In 1998 verscheen een boek getiteld *Het grootste woordenboek ter wereld. Een kijkje achter de kolommen van het Woordenboek der Nederlandsche Taal (WNT)*, onder redactie van Heyvaert e.a. Zie ook https://ivdnt.org/woordenboeken/historische-woordenboeken/woordenboek-der-nederlandsche-taal/ De aantallen trefwoorden zijn opgegeven door de redacties van de verschillende woordenboeken; onduidelijk is daarbij of ze op dezelfde manier hebben geteld: tellen ze bijvoorbeeld opnoemers (woorden zonder betekenisomschrijving die slechts in het voorbijgaan worden genoemd) en verwijstrefwoorden (x zie y) als apart trefwoord?
46 Booij en Van Santen 1998; Booij 2002; De Haas en Trommelen 1993, en grammatica's van het Nederlands zoals de e-ANS: https://e-ans.ivdnt.org/topics/pid/topic-16402164761856575 geven beschrijvingen van de manieren waarop samenstellingen en afleidingen in het moderne Nederlands worden gevormd. De ontwikkelingen in het verleden zijn te vinden in De Vooys 1957, Van der Sijs 2001: 147-184 en Van der Sijs 2019. Vanwege ruimtegebrek geef ik hier slechts de grote lijn.
47 Zie Wolf 2019.
48 Zie etymologiebank.nl
49 Overzichten van geleende voor- en achtervoegsels staan in Van der Sijs 2005.
50 Voorbeelden in Meesters 2002.
51 Zie voor verkortingen en blends Hamans 2021.
52 Van der Sijs 2023.
53 Koornwinder heeft de 245.000 trefwoorden van de Dikke Van Dale uit 1999, aangevuld met woorden en woordkenmerken uit de derde druk van het *Groot Handwoordenboek Hedendaags Nederlands*, omgezet naar een Morfologische Gegevensbank met circa 84.000 geanalyseerde basiswoorden (voor details zie Koornwinder 2005). Omdat de analysemethode voor een specifiek doel is gehanteerd, wijkt ze wat af van die in CELEX en ANW; dat hebben we zoveel mogelijk proberen recht te trekken: zo beschouwen we – anders dan in de Morfologische Gegevensbank – werkwoorden als *filmen, raden, spreken* als ongeleed (tegenover gelede werkwoorden als *verfilmen, aanraden, inspreken*). Hoe dan ook blijft de telling een globale schatting.
54 Ongeveer 14 procent van de woorden in CELEX is helaas niet geanalyseerd.
55 Alleen de geanalyseerde trefwoorden zijn meegeteld.
56 Dijkstra en Kempen 1984.
57 Brysbaert 2021; hoofdstuk 27 'Accessing Words from the Mental Lexicon' door Schiller en Verdonschot in Taylor 2015.

58 De Deyne en Storms 2013; hoofdstuk 26 'Word Associations' door De Deyne en Storms in Taylor (red.) 2015; resultaten van het project op https://smallworldofwords.org/nl/project/home.
59 Figuur opgevraagd via https://smallworldofwords.org/nl/project/visualize
60 Brysbaert 2003; Brysbaert en Dijkstra 2006; hoofdstuk 28 'The Bilingual Lexicon', door Williams in Taylor (red.) 2015.
61 Jansen 2023.
62 Hoofdstuk 30 'First words' door Clark in Taylor (red.) 2015; https://wij-leren.nl/woordenschat.php; https://cedinonderwijs.nl/artikel/het-belang-van-woordenschat/; https://bruuttaal.nl/pdfs/woordenschat.pdf
63 Brysbaert e.a. 2013, 2016.
64 Nation en Coxhead 2021. Het blijven beredeneerde schattingen; Brysbaert e.a. 2016 komen voor een gemiddelde 20-jarige Amerikaanse student op 42.000 woorden en 4.200 verbindingen, afgeleid van 11.100 woordfamilies.
65 Nation 2006.
66 Ze worden ook wel idioom, meerwoordige uitdrukking, constructie en dergelijke genoemd, maar ik veeg alles gemakshalve samen onder de term 'vaste verbinding'. Zie voor voorbeelden en een overzicht Everaert 1993, De Groot 2006, Van der Horst 2013, de e-ANS, hoofdstuk 7 'Multi-word Items, door Moon in Taylor (red.) 2015.
67 Van der Horst 2013: 15, 154-157.
68 Voor combinatiemogelijkheden van werkwoorden en zelfstandige naamwoorden in het algemeen zie https://woordcombinaties.ivdnt.org/
69 Van Wingerden en Hendriks 2015.
70 Van Oostendorp 2015.
71 Zie Engelsman 2004.
72 Hoeksema 2005.
73 Pander Maat en Dekker 2016.
74 Zie bijvoorbeeld Van Dalen-Oskam 2021.
75 Voor het tellen van woorden is ook bijvoorbeeld de Engelstalige tool https://seoscout.com/tools/text-analyzer geschikt.
76 Hamilton e.a. 2016.
77 Zie ook de voorbeelden in Van der Sijs 2022.

# Over de auteur

Nicoline van der Sijs (1955) is opgeleid als slavist. Ze is gastonderzoeker bij het Instituut voor de Nederlandse Taal in Leiden en emerita hoogleraar historische taalkunde van het Nederlands aan de Radboud Universiteit in Nijmegen. Eerder werkte ze aan Van Dale-woordenboeken en was ze senior onderzoeker op het Meertens Instituut. In 2021-2022 bekleedde ze de Koning Willem-Alexander leerstoel voor Lage Landen-studies te Luik. Ze publiceerde een groot aantal boeken en artikelen over etymologie en de geschiedenis van de Nederlandse taal. Ze is oprichter van etymologiebank.nl, vaste medewerker van Onze Taal en winnaar van o.a. de Taalboekenprijs en de Prins Bernhard Cultuurfonds Prijs. In 2024 ontving ze de Matthias de Vriespenning voor lexicologische verdiensten en in 2025 de Everwinus Wassenbergh Penning.

Op de website van de Werkgroep Onderzoek en Didactiek Nederlands (WODN) is aanvullend lesmateriaal te vinden, dat aansluit bij verschillende onderwerpen in dit boek (in de kantlijn aangegeven met ).

Het materiaal is te bereiken via:

Weinig dingen zijn zo alledaags als woorden, maar tegelijkertijd blijken weinig dingen ook zo wonderlijk: wat verstaan we eigenlijk onder een woord? Hoeveel woorden kent een gemiddelde spreker van het Nederlands? En hoe verhoudt zich dat tot de totale woordenschat van de taal?

Nicoline van der Sijs neemt je in *Dijt onze woordenschat alsmaar uit?* mee op reis door de geschiedenis en opbouw van de Nederlandse woordenschat. Ze onderzoekt vanuit verschillende wetenschappelijke perspectieven hoe woorden ontstaan, veranderen en verdwijnen, en stelt daarbij prikkelende vragen: groeit de Nederlandse woordenschat bijvoorbeeld eindeloos, of is er een grens aan het aantal woorden dat we kunnen maken?

*Dijt onze woordenschat alsmaar uit?* is het eerste deel van een reeks beknopte, toegankelijke boeken over actuele ontwikkelingen in de neerlandistiek, getiteld *Nederlands in het klein*. Ieder jaar zullen een paar deeltjes verschijnen, waarbij telkens los daarvan didactisch materiaal wordt ontwikkeld voor behandeling in het voortgezet onderwijs.

ISBN 978-94 6515 058 1

9 789465 150581

Radboud University

www.radbouduniversitypress.nl