

HET MENTALE LEXICON ALS EEN SEMANTISCH WEB

Simon De Deyne en Gert Storms

KU Leuven

Tot op heden is het onduidelijk hoe woordbetekenis in het mentale lexicon wordt gerepresenteerd. Onderzoek heeft zich in het verleden meestal toegespitst op concrete zelfstandige naamwoorden en heeft abstracte woorden, adjectieven en werkwoorden vaak genegeerd. Verder is het niet duidelijk welk soort van kenmerken de betekenis van woorden voldoende definieert. In dit onderzoek rapporteren we de bevindingen van een grootschalige Nederlandstalige studie die vertrekt van woordassociaties en die tot doel heeft een groot deel van ons lexicon in kaart te brengen en na te gaan welke types informatie in dit lexicon psychologisch centraal staan. Hierbij staan we stil bij de implicaties van een semantisch netwerk, opgebouwd op basis van deze associaties, voor het verwerven van woordbetekenis en bij het gebruik ervan in onder meer taalproductie en woordherkenning.

Inleiding

Een thesaurus laat ons toe om de betekenis van een woord gemakkelijk op te zoeken. Zo kunnen we er het Van Dale woordenboek van de Nederlandse taal op na slaan om de betekenis van een abstract woord als 'idee' op te zoeken en krijgen we:

Idee (het; de; o; v; zelfstandig naamwoord; meervoud: ideeën)

1. Het beeld dat de geest zich van iets heeft gevormd
2. In de geest gevormde voorstelling; denkbeeld
3. Mening
4. Inval, plan

Vanzelfsprekend vinden we er ook de betekenis van concrete woorden zoals paard in terug:

Paard (het; o; meervoud: paarden)

1. Zoogdier uit de familie van de eenhoevige,
2. Bep. Schaakstuk,
3. Bep. Gymnastiektoestel.

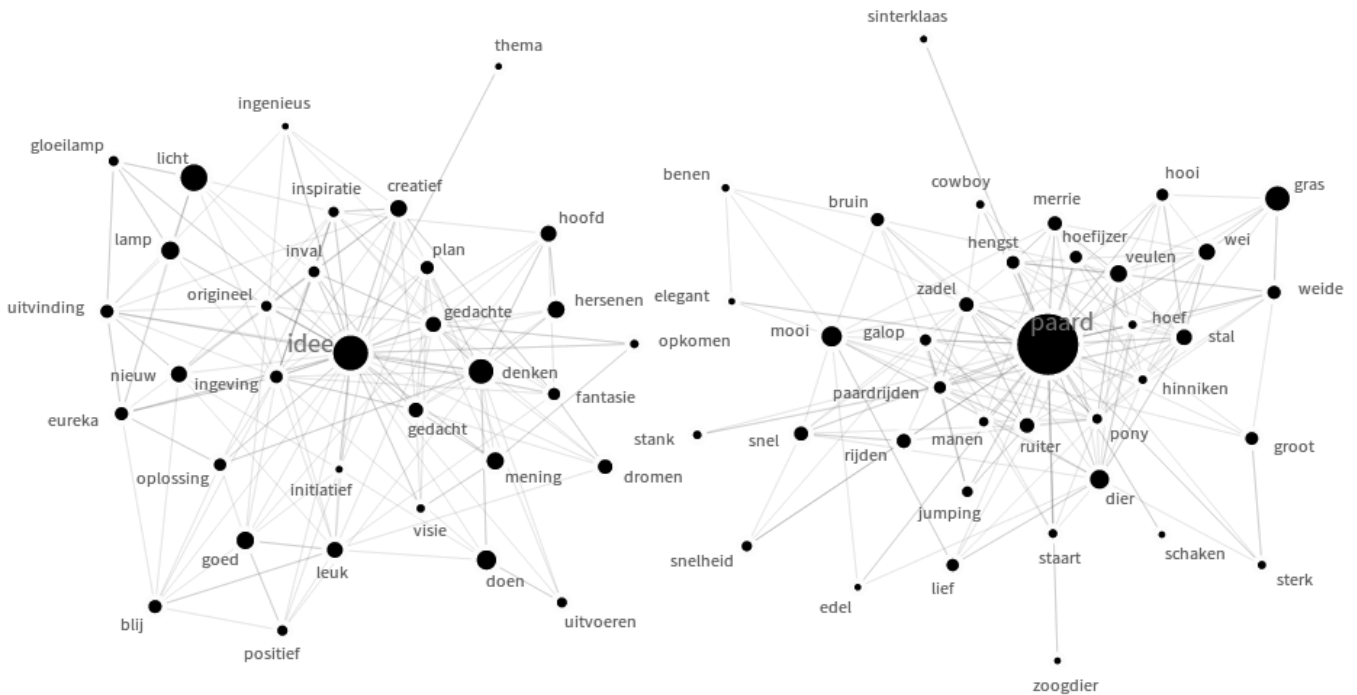
Hoewel iedereen het erover eens zal zijn dat een verklarend woordenboek een handig hulpmiddel is om de betekenis van een woord te achterhalen, heeft onderzoek over de mentale representatie van woordbetekenis aangetoond dat vanuit psychologisch oogpunt betekenis op een andere manier wordt verworven en gehanteerd. Ten eerste weten we dat de betekenis van veel woorden niet bepaald wordt door een kleine set noodzakelijke en voldoende kenmerken. Eerder gaat men er vanuit dat de betekenis van een woord vooral bepaald wordt door karakteristieke, eerder dan definiërende kenmerken, zoals het feit dat paarden lopen, snel kunnen lopen en dikwijls bruin zijn. Dit verklaart waarom Arabische volbloedpaarden typischer voorbeelden zijn van de categorie van paarden dan een Brabants trekpaard.

Wat een woordenboekdefinitie wel deelt met huidige opvattingen over de mentale representatie van woordbetekenis is het idee dat deze informatie niet willekeurig opgeslagen wordt in het lexicon, maar op een specifieke manier gestructureerd is. Voor dieren en planten wordt aangenomen dat we kennis hebben over de categorieën waartoe ze behoren en hoe deze categorieën samenhangen, waarbij een hiërarchische structurering centraal staat. Zo behoort een paard tot de familie van de eenhoevigen, die deel uitmaakt van een grotere groep zoogdieren.

Het chronologisch eerste psychologisch model om semantische informatie te vatten in een netwerk was het hiërarchisch model van Collins en Quillian (1969). Informatie wordt in dit model op een economische manier gerepresenteerd door uit te gaan van een boomstructuur, waarbij verschillende types van informatie gepresenteerd worden op het hoogst mogelijk niveau van de taxonomie. De eigenschap 'leeft' wordt in dit model opgeslagen op het niveau van dieren, omdat deze eigenschap van toepassing is op alle categorieën die resorteren onder deze categorie. Dezelfde redenering volgend zal dan de eigenschap 'heeft melkklieren' opgeslagen worden op niveau van zoogdieren etc. Dit model werd uitvoerig getoetst in een reeks experimenten, maar de basisassumpties bleken niet overeen te komen met empirische bevindingen over hoe we informatie uit dit netwerk ophalen. Om tegemoet te komen aan deze empirische discrepanties stelden Collins en Loftus (1975) een revisie van dit model voor waarin informatie niet langer in een taxonomische structuur gerepresenteerd werd, maar als een niet-hiërarchisch associatief netwerk werd voorgesteld. Dit model veronderstelde een mechanisme waarbij de betekenis wordt geactiveerd door middel van activatiespreiding (*spreading activation*).

Taal(ontwikkelings)stoornissen

Figuur 1. Voorbeeld van semantische netwerken voor 'idee' en 'paard'.



Hoewel niet zwart op wit is aangetoond dat de betekenis van woorden kan worden voorgesteld als een semantisch web waarbij elke knoop een woord representeert, heeft een dergelijke representatie een aantal voordelen die een verdere ontwikkeling van dit idee motiveren. Eén van de grootste voordelen is dat een dergelijke representatievorm erg flexibel is. Er zijn nauwelijks assumpties over het type informatie dat bijdraagt tot betekenis en het soort van concepten dat erin gerepresenteerd kan worden. Dit maakt het mogelijk om, naast concrete concepten zoals 'paard' of 'stoel', ook de betekenis van abstracte woorden zoals 'idee' of 'eerlijkheid', begrip-pen die minder ingebed zitten in een hiërarchisch systeem, te representeren. Een tweede kenmerk is de rol van de verbindingen tussen de knopen in zo'n netwerk. In tegenstelling tot een woordenboek zullen bepaalde aspecten van betekenis centraler zijn in ons lexicon dan andere. In het voorbeeld van 'paard' zal de zoogdier-betekenis primeren op die van schaakstuk of gymnas-tiektoestel en de verbinding tussen 'paard' en 'zoog-dier' zal daarom een hoger gewicht toegekend krijgen in het semantische netwerk. Verder wordt ook kennis gepresenteerd die we helemaal niet terugvinden in een woordenboek. Zo wordt het woord 'idee' geassocieerd met een licht dat gaat branden en is dit een belangrijke metafoor die dit abstracte begrip verankert in het se-mantische systeem.

De betekenis van de woorden zelf wordt in sterke mate bepaald door de context. Het netwerk bevat daarom ook

veel thematische links, waarbij de betekenis van 'paard' een context activeert die psychisch relevante informatie bevat voor courante situaties (frames of events) zoals 'zadel', 'ruiter', 'stal' of 'paardrijden'. De representatie van woordbetekenis in een associatief semantisch netwerk laat met andere woorden ook het gebruik van frame- of event-representaties toe. Ten slotte biedt deze representatie het voordeel van eenvoud in interpretaties, in tegenstelling tot complexe connectionistische modellen voor het semantisch geheugen waarbij woordbetekenis vervat zit in een activatiepatroon van een groot aantal knooppunten, eerder dan in één enkel knooppunt (zie bv. McClelland & Rogers, 2003).

Door de recente toename in rekenkracht wordt het mogelijk om dit model van woordbetekenis, dat is opgebouwd op basis van grote hoeveelheden woordassociaties, voor het eerst ten volle te ontwikkelen. Dergelijke schaalvergroting impliceert niet enkel een evolutie waarin we een grotere verscheidenheid aan woorden en hun betekenis kunnen representeren maar het leidt tot een revolutie in ons denken over hoe betekenis in het semantisch lexicon wordt gepresenteerd. Door het volledige lexicon in kaart te brengen kunnen we de macrostructuur blootleggen. Deze structuur vertoont immers een aantal unieke kenmerken. Zo speelt een klein aantal woorden een belangrijke rol in ons semantisch netwerk. De talrijke verbindingen met woorden als 'water' of 'geld' vormen een netwerk gekenmerkt door een klein aantal sterk verbonden knooppunten of zogenaamde hubs. De

aanwezigheid van dergelijke centrale en wijdvertakte knooppunten zorgen ervoor dat elk woord slechts een paar stappen verwijderd is van elk ander woord, waardoor informatie zich snel kan verspreiden in het netwerk. Net zoals andere informatienetwerken, zoals het internet en bepaalde neurale netwerken, zorgt de combinatie van een geclusterde structuur en korte afstanden voor een *small-world* structuur (De Deyne & Storms, 2008b, Steyvers & Tenenbaum, 2005).

Een tweede kenmerk van een semantisch netwerkmodel op ware schaal is dat het ons de mogelijkheid biedt na te gaan hoe gerelateerd de betekenis van twee woorden is door gebruik te maken van de *volledige* structuur die vervat zit in het netwerk. Een dergelijke zienswijze dat de betekenis van een woord bepaald wordt door de context waarin het gebruikt wordt is niet nieuw en vinden we terug bij onder meer Saussure (1915), Wittgenstein (1953) en Firth (1957). Concreet kunnen we stellen dat de betekenis van twee woorden bepaald wordt door de mate waarin ze gelijkaardige activatiepatronen in het semantische netwerk vertonen. Op die manier kunnen twee woorden die niet rechtstreeks met elkaar verbonden zijn indirect een bepaalde mate van gelijkenis vertonen (De Deyne, Peirsman, & Storms 2009). De manier waarop dit wordt berekend bespreken we uitvoeriger in de resultatensectie.

In de volgende paragrafen wordt een grootschalige Nederlandstalige woordassociatiestudie beschreven waarin getracht wordt om zowel het bereik als de hoeveelheid informatie van het mentale lexicon voldoende te vatten om tot een beter inzicht te komen van de structuur en de werking ervan. Om alle aspecten van woordbetekenis aan bod te laten komen maken we gebruik van een vrije woordassociatietaak. Deze taak is vrij in de zin dat de deelnemers eender welke soort antwoord kunnen geven voor een bepaalde cue. Bovendien tracht deze taak de psychologisch relevante aspecten van woordbetekenis te vatten doordat de proefpersoon enkel woorden geeft die spontaan voor de geest komen.

Een beperking van de meest voorkomende versie van de associatietaak is dat voor sommige woorden het aantal verschillende antwoorden erg klein is door de aanwezigheid van een dominante associatie. Zo blijkt, bijvoorbeeld, dat meer dan 90% van de deelnemers de associatie 'regen' geeft na het lezen (of horen) van het cuewoord 'paraplu'. Dit fenomeen van dominante antwoorden kan leiden tot een vertekend beeld van het semantische netwerk en wordt als een belangrijke be-

perking beschouwd bij het gebruik van woordassociaties voor het onderzoeken van het mentale lexicon (Aitchison, 2003). Om dit tegen te gaan maken we gebruik van een meervoudige procedure waar iedere proefpersoon drie antwoorden geeft voor elk cuewoord, met andere woorden: de drie eerste woorden waaraan hij of zij denkt bij het lezen van een cuewoord. Uit de resultaten zal blijken dat deze procedure tot een dichter netwerk leidt, wat op zijn beurt beter toelaat om na te gaan hoe sterk twee woorden gerelateerd zijn.

Om ervoor te zorgen dat het lexicon een relevant deel van onze woordenschat beslaat maken we gebruik van een sneeuwbalprocedure, wat betekent dat nieuwe cues geleidelijk worden toegevoegd door op regelmatige intervallen de meest voorkomende antwoorden te tellen en deze op hun beurt te presenteren als cue. Naast het bespreken van de dichtheid en omvang van het netwerk zullen we ook dieper ingaan op de kwalitatieve kenmerken van woordassociaties, zoals het soort informatie dat vertegenwoordigd wordt en hoe die informatie verworven wordt.

Methode

Hieronder geven we een samenvatting van de eigenschappen van stimuli, deelnemers en de gebruikte procedure. Voor een gedetailleerd overzicht verwijzen we naar De Deyne, Navarro & Storms (2012). De data die hier gerapporteerd worden, hebben betrekking op de periode van 2003 tot november 2010.

Deelnemers. In totaal namen 71,380 personen deel aan deze studie (47,459 vrouwen, 22,966 mannen en 955 niet nader gespecificeerd). De gemiddelde leeftijd was 40 jaar en varieerde tussen 7 en 96 jaar. Naast het verzamelen van informatie over leeftijd en geslacht werd ook informatie bekomen over de regionale taalvariant. Slechts 6,875 deelnemers uit Nederland werden geregistreerd waardoor de meerderheid van de deelnemers Nederlandstalige Vlamingen (90%) waren.

Stimuli. De initiële set van woorden bestond uit exemplaren van de categorieën dieren, voorwerpen, fruit, groenten, sporten en beroepen (zie Ruts et al., 2008). Deze kleine set van 338 woorden werd uitgebreid aan de hand van de hierboven beschreven sneeuwbalprocedure waarbij de meest voorkomende antwoorden opnieuw aangeboden werden als cuewoorden. Bijkomend werden extra woorden ingesloten die van belang waren voor andere studies, zoals bijvoorbeeld prentbenoemingstaken (Severens et al., 2005).

De uiteindelijke set bestond uit 12,571 cues. Deze cues bestonden in hoofdzaak uit zelfstandige naamwoorden (64.5%), adjectieven (16.9%) en werkwoorden (15.7%).

Procedure. De afname van de woordassociatietaak gebeurde in hoofdzaak online aan de hand van een web-gebaseerde vragenlijst. De proefpersonen werden gevraagd om voor iedere cue drie associaties neer te schrijven die spontaan in hun opkwamen. Verder werden de proefpersonen erop gewezen geen volzinnen of strikt persoonlijke antwoorden te geven en enkel rekening te houden met het cuewoord zelf en niet met voorgaande antwoorden. Als een woord onbekend was, werd de proefpersoon gevraagd om een knop 'X' aan te klikken. Indien er slechts één of twee associaties spontaan bij hen opkwamen, dan werd hen gevraagd om dit aan te duiden door in de overige velden van het antwoordformulier 'geen antwoord' in te vullen.

Resultaten en Discussie

Aan de hand van een aantal criteria werd nagegaan of de proefpersonen de taak naar behoren uitgevoerd hadden. Dat gebeurde door de antwoorden te vergelijken met woordfrequentielijsten uit het CELEX woordfrequentie-corpus (Baayen, Piepenbrock, & van Rijn, 1993). Voor verdere details over het uitzuiveren van de data verwijzen we naar De Deyne, Navarro en Storms (2012). Per cue werden de antwoorden van ten minste 100 deelnemers weerhouden. Voor de 12,571 cues werd een totaal van 3.77 miljoen antwoorden verzameld. Dit kwam overeen met 201,356 verschillende woordvormen ongeacht de antwoordpositie (primaire, secundaire of tertiaire antwoord). De meeste woorden (97%) waren gekend bij minimum 90% van de deelnemers. Ook het genereren van een tweede en derde associatie vormde geen probleem: slechts 1.3% en 4.4% van respectievelijk de tweede antwoordassociaties en derde antwoordassociaties ontbrak.

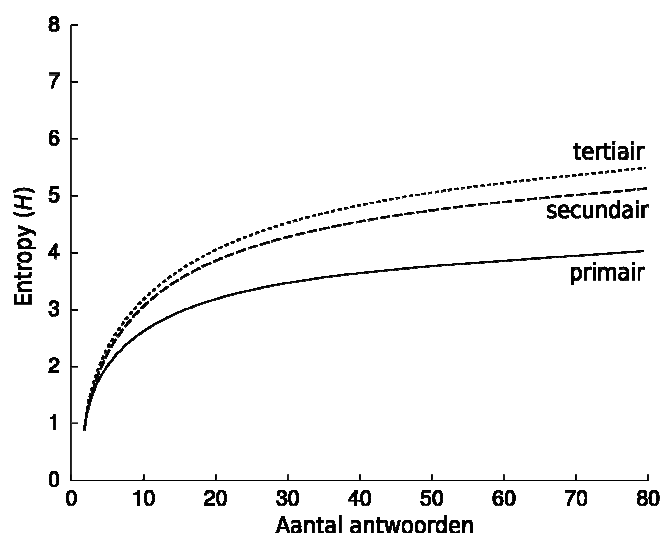
Als de antwoordpositie niet in acht genomen wordt zijn er per cue 300 associaties beschikbaar. Gemiddeld kwam dit overeen met 118 ($SD = 23.6$) verschillende antwoorden voor ieder cuewoord. Vooral de secundaire en tertiaire antwoorden zorgden voor extra heterogeniteit. Entropie is een maat die aangeeft in welke mate een verdeling heterogeen is. Entropie H wordt berekend als

$$H = - \sum_n p_i \log_2(p_i)$$

waarbij n het aantal verschillende antwoordtypes is en p_i de kans voor het antwoordtype i . H is de som van alle

kansen verschillende van nul. Als alle antwoorden identiek zijn krijgen we een $H = 0$. Als alle 300 antwoorden die verzameld werden per cue verschillen krijgen we een theoretisch maximum van $H = 8.22$ (zie y-as Figuur 2).

Figuur 2. Toename in entropie in functie van het aantal verzamelde antwoorden op basis van primaire, secundaire en tertiaire antwoorden.



Om na te gaan in welke mate nieuwe woorden opduiken naarmate er meer antwoorden verzameld worden werd voor iedere cue de entropie van de antwoordistributie berekend als een functie van het aantal getelde antwoorden (zie Figuur 2). Voor de latere antwoorden zien we hogere entropiecurves, ongeacht het aantal antwoorden dat verzameld werd. Dit geeft aan dat mensen meer verschillende antwoorden geven wanneer ook een tweede en derde associatie voor een cue gevraagd wordt. Verder toont Figuur 2 ook dat hoewel het aantal verschillende antwoorden afvlakt naarmate meer data verzameld worden, er geen sprake is van een beperkte set mogelijke responsen per cue.

Constructie van het semantisch netwerkmodel voor het mental lexicon

Een aantal interessante maten kan afgeleid worden wanneer de woordassociatiedata gerepresenteerd worden als een semantisch netwerk. Deze maten geven een indicatie van het belang van woorden als een functie van het aantal inkomende of uitgaande verbindingen. Deze informatie is enkel beschikbaar voor woorden die als cue en als antwoord voorkomen. Het netwerk wordt afgeleid op basis van de associatiedata door enkel antwoorden die ook als cue voorkomen op te nemen. Het uiteindelijke netwerk bevat 12,428 woorden waarbij ieder woord op zijn minst één uitgaande verbinding maakt met een an-

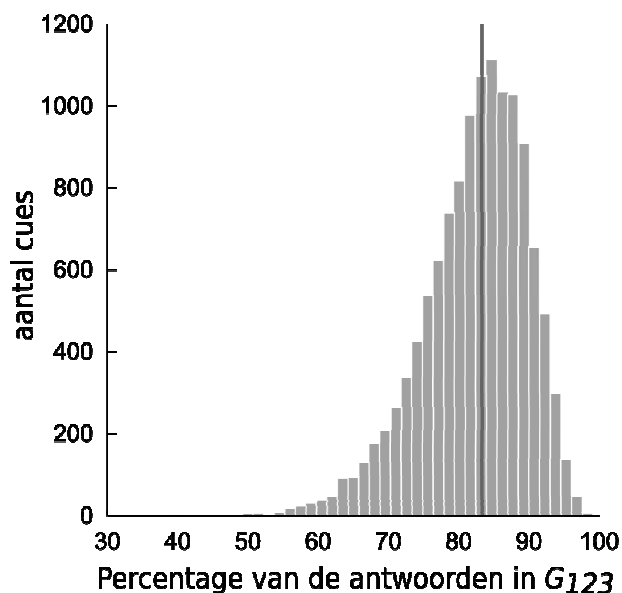
der woord. De verbindingen in dit netwerk geven aan hoe sterk twee woorden geassocieerd worden. De associatiesterkte tussen ieder paar van woorden wordt bepaald door het aantal proefpersonen dat een bepaalde associatie geeft voor een cue te delen door het totaal aantal associaties dat gegeven is voor een bepaalde cue. Het eindresultaat is een gewogen gericht (*directed*) netwerk. Het is gewogen op basis van associatiesterkte. Het feit dat dit netwerk gericht is, is typerend voor associatiedata, waar voor een cue als 'idee', 'licht' meer als antwoord gegeven wordt dan vice versa.

Via deze procedure kunnen netwerken geconstrueerd worden op basis van enkel primaire (G_1), enkel secundaire (G_2) en enkel tertiaire antwoorden (G_3). Tenzij anders vermeld zullen we in deze uiteenzetting echter gebruik maken van het netwerk waarin alle antwoorden gecombineerd worden (G_{123}). Voor meer details over de integratie van woordassociatiesterktes op basis van verschillende responseposities verwijzen we naar De Deyne, Navarro, & Storms (2013).

Hoe representatief is de semantische netwerkbenadering van het mentale lexicon?

De dichtheid of densiteit van het netwerk geeft aan hoeveel verbindingen er aanwezig zijn. Als elk woord met elk ander woord verbonden is bekomen we een densiteit van 1. De densiteit van het G_1 netwerk was erg laag: 0.22%. Het toevoegen van secundaire en tertiaire antwoorden uit G_2 en G_3 doet deze densiteit gevoelig stijgen tot respectievelijk 0.43% en 0.64%.

Figuur 3. Histogram voor het percentage van gerepresenteerde antwoorden per cue.



Omdat enkel antwoorden weerhouden worden bij het construeren van het netwerk die tevens in de set van cues voorkomen wordt het aantal antwoorden drastisch gereduceerd. In welke mate is het lexicon met ongeveer 12,000 woorden representatief voor de woordenschatgrootte van een gemiddelde volwassene? Hoewel meer dan 200,000 verschillende antwoorden bekomen werden, komen een groot deel van deze antwoorden slecht enkele malen voor. Daarom is het zinvol om deze vraag te beantwoorden door voor iedere cue te berekenen hoeveel antwoorden in het netwerk gepresenteerd worden. Dit aantal wordt bepaald door het aantal keer dat een antwoord in de lijst met cues voorkomt voor iedere cue afzonderlijk te berekenen. De resultaten van deze berekening worden weergegeven in het histogram in Figuur 3. Dit histogram duidt aan dat slechts een klein deel van de cues niet accuraat wordt gepresenteerd. Zo bedraagt de proportie cues die door minder dan 75% van zijn antwoorden wordt gerepresenteerd slechts 15.9%. Hierbij moet opgemerkt worden dat dit percentage een onderschatting is, aangezien spellingsfouten en lemmatizing de dekking verder kunnen verbeteren. Samenvattend kunnen we stellen dat het netwerk zowel een aanzienlijk deel van het lexicon representeert en de meervoudige versie van de woordassociatietask leidt tot een beduidend hogere dichtheid van het netwerk.

Afgeleide Maten

Uit het semantisch netwerk kan een aantal maten afgeleid worden dat aangeeft hoe belangrijk een woord is in het lexicon. Dergelijke centraliteitsmaten voorspellen de mate waarin bepaalde woorden onthouden kunnen worden (zie bv. Nelson et al., 1998) en de snelheid waarmee een woord geproduceerd en herkend wordt (bv. De Deyne, Navarro, & Storms, 2012; De Groot, 1989).

Een tweede type van maten maakt gebruik van de informatie over de verbindingen tussen meerdere knopen. Deze maten geven aan hoe sterk een woordpaar associatief of semantisch gerelateerd is. Evidentie voor dit onderscheid komt vooral uit associatief en semantisch primingonderzoek. In dit soort studies werd aangetoond dat een woord zoals paard sneller verwerkt wordt als het wordt voorafgegaan door een gerelateerd woord of *prime* (bv. ruiter). Een recent overzicht van de bevindingen in deze studies vindt men in Hutchison (2003) en Lucas (2000).

Netwerkcentraliteitsmaten

De centraliteit van een knoop in het netwerk is een indicatie voor het gemak waarmee een bepaald woord her-

kend of geproduceerd kan worden. Uit onderzoek blijkt dat deze maat een goed alternatief vormt voor andere centraliteitsmaten gebaseerd op woordfrequentie bij het verklaren van woordverwerking in de lexicale decisietaak (De Deyne, Navarro & Storms, 2012).

Met betrekking tot netwerkcentraliteit zijn de meest eenvoudige maten het aantal inkomende verbindingen (*in-degree*) en het aantal uitgaande verbindingen (*out-degree*). De eerste maat komt sterk overeen met het belang van een woord in het lexicon op basis van frequentie van voorkomen in gesproken en geschreven taal. De tweede maat geeft een indicatie van *set size* of setgrootte van betekenissen en laat toe om een onderscheid te maken tussen woorden met een diffuse set van betekenissen en woorden met een eerder compacte set van betekenissen. Op basis van *in-degree* vinden we dat 'water', 'auto', 'geld', 'pijn' erg centraal zijn in het lexicon, in tegenstelling tot 'paviljoen', 'pathogeen' en 'fuik' die een lage *in-degree* hebben. Voor *out-degree* hebben woorden als 'lidwoord', 'bestek' en 'zuivelproduct' hoge waarden. Voor deze woorden putten de proefpersonen uit een beperkt aantal mogelijkheden. Aan de andere kant van het spectrum vinden we woorden zoals 'telepathie', 'politieker' en 'manipulatie'. Voor deze woorden ontbreekt een eenduidig antwoord. Voor complexere maten, die rekening houden met de mate van verbondenheid binnen een set van woorden en de bereikbaarheid van willekeurige paren van knopen in het netwerk, verwijzen we o.a. naar De Deyne en Storms (2008b).

Semantische gerelateerdheid

De notie van similariteit of gerelateerdheid staat centraal in het onderzoek naar de representatie van concepten en woordbetekenis. Algemeen wordt verondersteld dat twee concepten gelijkend zijn als ze een groot aantal gemeenschappelijke kenmerken hebben (bv 'koe' en 'paard': *zijn zoogdieren, eten gras, hebben hoeven*, etc.). In de context van woordbetekenis kunnen we stellen dat twee woorden een gelijkende betekenis hebben naarmate ze in een gelijkaardige taalcontext voorkomen. In het geval van een semantisch netwerk wordt deze context gerepresenteerd door de knopen die verbonden zijn met een bepaald woord. Een maat van semantische gerelateerdheid die niet beïnvloed wordt door woordfrequenties van de woorden is de cosinus maat (zie Newman, 2009 voor de afleiding van deze maat in de context van netwerken). Deze maat drukt uit in welke mate twee knopen in het netwerk dezelfde buren hebben. Woorden die geen gemeenschappelijke buren hebben, zijn ongerelateerd en krijgen een waarde 0. Woorden die alle buren in dezelfde mate

delen zijn perfecte synoniemen en krijgen een waarde 1. Ter illustratie kunnen we de sterkst gerelateerde woorden vergelijken met de meest frequent gegeven associaties. Zo vinden we dat 'paard' sterk gerelateerd is met 'bizon', 'gazelle', 'hengst', 'draf', 'galop' en 'koe', terwijl we geen van deze gerelateerde woorden terugvinden bij de sterkste associaties ('dier', 'rijden', 'ruiter', 'groot', 'mooi', 'wei').

Algemene Bespreking

Om de context te creëren waarbinnen het gebruik van het op associaties gebaseerde semantisch netwerk binnen de logopediepraktijk mogelijk te maken, onderscheiden we taalrepresentaties en mentale representaties. Vervolgens gaan we dieper in op de inhoudelijke kenmerken van de verbanden in het netwerk. Hiervoor wordt een semantische codering toegepast waarin zowel syntactische (paradigmatische en syntagmatische verbanden), psychologische (conceptgebonden kenmerken, verbanden die connotatie of thematische informatie uitdrukken) als linguïstische verbanden (antoniemen, hyponiemen, hyperoniemen, meroniemen, etc.) opgenomen worden.

Taalrepresentaties en mentale representaties

In welke mate kan men aannemen dat het mentale netwerk gebaseerd op woordassociaties, afkomstig van een grote proefgroep, een goed model vormt voor het mentale netwerk van een bepaald individu? Hoewel deze empirische vraag tot op heden nog niet volledig beantwoord is, kan het onderscheid tussen taalrepresentaties en mentale representaties hier enige houvast bieden.

Een recente strekking tracht semantische relaties in kaart te brengen aan de hand van tekstcorpora. Een voorbeeld van zo'n model is het latente semantische analyse model van Landauer en Dumais (1997). Dergelijke modellen verklaren tal van taalgerelateerd gedrag zoals blijkt uit hun prestaties op woordenschatmeerkeuzevragen die ook bij leerlingen gebruikt worden (Landauer & Dumais, 1997). Theoretisch baseren deze modellen zich op het idee dat woorden die frequent samen voorkomen ook in het lexicon aan elkaar gekoppeld worden (Spence & Owen, 1990).

Evenals de corpus-gebaseerde semantische modellen is het semantische netwerk op basis van woordassociaties eerder opgevat als een gedeeld sociolinguïstisch artefact dat de talige kennis binnen een gemeenschap samenvat dan het mentale netwerk op individueel niveau. Dit netwerk is voor individuen in verschillende mate toegankelijk, afhankelijk van communicatie met leden van

de talige gemeenschap. Aangezien communicatie één van de belangrijkste functies van taal is en efficiënte communicatie afhangt van convergentie met dit gedeelte van de lexicon verwachten we dat dit netwerk in belangrijke mate informatief is voor de mentale representaties binnen het individu. In tegenstelling tot andere talige representatievormen zoals bv. tekstcorpora verwachten we dat een mentaal lexicon op basis van woordassociaties een aantal eigenschappen van de mentale representaties dichter benadert. Hiervoor kan een aantal redenen worden aangehaald. De belangrijkste redenen zijn de verschillen in de pragmatische vereisten in de woordassociatietaken enerzijds, en het schrijven van een artikel of het voeren van een discussie anderzijds. In de associatietaken worden de pragmatische vereisten geminimaliseerd, waardoor saliente algemene kennis (bv. 'banaan' – 'geel') expliciet gegenereerd wordt. Dit kan nagegaan worden aan de hand van tekstcorpora door de frequentie waarmee twee woorden naast elkaar in een zin voorkomen (i.e. een bigram) te vergelijken met de associaties terkte tussen de twee woorden van het bigram. Hieruit blijkt dat er slechts een geringe overeenkomst is tussen woorden die samen voorkomen in de taal en woorden die in de associatietaken gegeven worden (zie bv. Hahn & Sivley, 2011, Wettler, Rap & Sedlmeier, 2005). Een tweede bron van verschillen tussen mentale en taalgebaseerde representaties is het belang van woordconnotatie. Uit een omvangrijke studie over adjectieven blijkt dat de belangrijkste bron van structuur in het lexicon bepaald wordt door de mate waarin woorden zoals 'probleem' of 'vakantie' een positieve of negatieve connotatie of valentie hebben (Deese, 1965).

Op basis van deze argumenten en evidentie uit recente overzichtsartikelen kunnen we besluiten dat de woordassociatietaken niet enkel gebruik maakt van talige informatie zoals bigramfrequenties, maar in belangrijke mate de semantische verbanden in het lexicon weerspiegelt (McRae, Khalkhali, & Hare, 2011; Mollin, 2009).

Informatie gecodeerd in het semantisch netwerk

Als woordassociaties in belangrijke mate semantische relaties weerspiegelen, stelt de vraag zich over welk soort relaties dit gaat en wat hun relatieve aandeel is in het lexicon. Bij de analyse van antwoorden in de vrije woordassociatietaken wordt een onderscheid gemaakt tussen *syntagmatische* en *paradigmatische* antwoorden. Syntagmatische relaties komen voor bij paren waar cue en antwoord een verschillende grammaticale functie in de zin vervullen (bv. 'paard' – 'lopen'). Paradigmatische relaties geven aan dat zowel cue als antwoord een in-

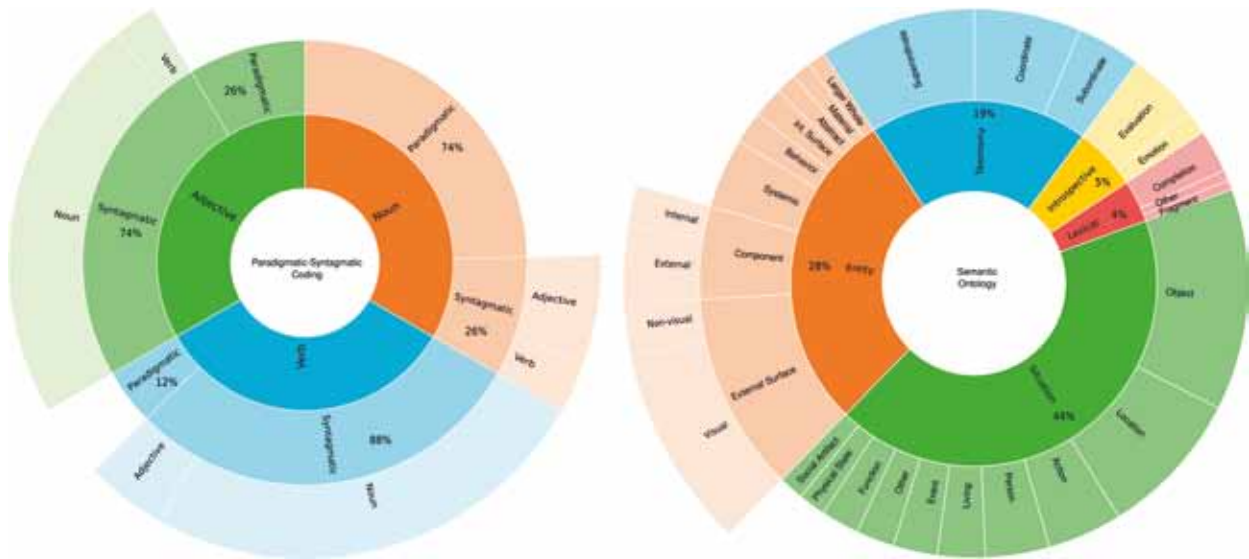
wisselbare rol spelen in de context van een zin (bv. 'idee' – 'gedachte'). Deze typering is vooral van belang binnen de taalontwikkeling aangezien kinderen tot 9 jaar vooral paradigmatische antwoorden geven en het antwoordpatroon pas later syntagmatisch wordt (K. Nelson, 1977). Deze shift kan verklaard worden door herstructurering tot een meer formele structuur in het lexicon met een toename aan abstracte kennis (zoals de categorieën waartoe een woord behoort) en anderzijds door een uitbreiding van het lexicon waardoor synoniemen en coördinaten toegevoegd worden aan het netwerk. Het voorkomen van paradigmatische antwoorden kan daarom beschouwd worden als een marker die aangeeft in welke mate de mentale representatie de structuur van de talige omgeving benadert.

In het volwassen lexicon komen zowel syntagmatische als paradigmatische relaties voor. Zoals Figuur 4 toont varieert dit aandeel naargelang het om adjectieven, zelfstandige naamwoorden of werkwoorden gaat. Zo kunnen we aan de hand van Figuur 4 opmaken dat voor zelfstandige naamwoordcues 74% van de antwoorden paradigmatisch zijn, terwijl dit voor werkwoorden veel minder voorkomt (in ongeveer 12% van de gevallen).

Dit onderscheid kan verder verfijnd worden door het gebruik van een semantische ontologie waarbij verschillende types van informatie die het lexicon vorm geven worden opgenomen (zie De Deyne & Storms, 2008b). In Figuur 4 wordt ook een overzicht gegeven van de semantische ontologie toegepast op 128,929 cue-associatieparen uit De Deyne en Storms (2008b). De meeste paren (44%, aangeduid als *situation*) zijn situatie-gebonden, thematisch gerelateerde paren zoals 'dokter' – 'ziekenhuis'. Ook conceptkenmerken waaronder meroniemen zoals 'olifant' – 'slurf' of visuele kenmerken zoals 'melk' – 'wit' komen frequent voor (28% van de paren, aangeduid door *entity*) en geven aan dat het lexicon ook kennis bevat over karakteristieke kenmerken van concepten. Deze informatie staat centraal in exemplaar en prototype-theorieën over het semantisch geheugen (Storms, De Boek, & Ruts, 2000). Taxonomische relaties, waaronder coördinaten zoals 'appel' – 'peer', superordinaten of hyperniemen zoals 'meeuw' – 'vogel' en subordinaten of hyponiemen zoals 'hond' – 'poedel' komen eveneens voor (19% van de paren, gelabeld als *taxonomy*). Connotatie-gerelateerde en andere introspectieve antwoorden (*introspective* in Figuur 4) zoals persoonlijk verwijzingen komen voor in de gecodeerde paren, maar slechts in mindere mate³. Een laatste type wordt gevormd door lexicale relaties (gelabeld *lexical* in Figuur 4) waarbij er louter een fonologisch

Taal(ontwikkelings)stoornissen

Figuur 4. Paradigmatisch-Syntagmatische codering (links) en Semantische codering (rechts).



verband is tussen de paren. Een voorbeeld van dit type antwoord zijn *clang*-responsen zoals 'wafel'-'tafel' en zijn uitgebreid onderzocht in ontwikkelingsstudies, patiëntenstudies en de tweedetaalverwerving waar ze duiden op een lexicon dat nog niet volledig ontwikkeld is, of af-takelt (zie bv Fitzpatrick, 2006). De semantische ontologie in Figuur 4 toont naast situatie-gebonden, entiteiten, taxonomische, lexicale en introspectieve klassen nog een fijnere opdeling. Meer details hierover worden gegeven in De Deyne en Storms (2008b).

Beknopt kunnen we stellen dat aan de hand van deze semantische analyse duidelijk het belang van situatie-gebonden of thematisch informatie naar voor komt. Dit onderstreept één van de kenmerkende aspecten van de woordassociatieprocedure: in tegenstelling tot taalkundig gemotiveerde thesauri of klassieke benaderingen waarbij betekenis beperkt wordt tot kenmerken van de entiteit zelf, wordt woordbetekenis bepaald door een amalgaam van verbanden, waarbij contextgebonden informatie een belangrijk rol toebedeeld krijgt.

Gebruik van woordassociaties in de praktijk

Aan de hand van de voorgaande typering hopen we een duidelijker beeld te scheppen over welke kennis in het lexicon geactiveerd wordt door de woordassociatietask. Het gebruik van een gedetailleerde semantische ontologie kan een alternatief bieden voor het rudimentair onderscheid tussen paradigmatische en syntagmatische antwoorden, wat ons zou toelaten om uitspraak te doen over de structuur van het lexicon tijdens verschillende taalontwikkelingsfasen. Verder onderzoek over het mentale lexicon in ontwikkeling is nodig om hierover definitief uitspraak te doen.

De belangrijkste bijdrage situeert zich in de omvang en diepgang van het lexicon dat in kaart gebracht wordt. Dit laat toe om zowel voor abstracte woorden als concrete woorden, werkwoorden, adjectieven als zelfstandige naamwoorden gedetailleerde informatie te raadplegen over hun belang in het lexicon en de manier waarop ze gehanteerd worden in context. Voor het eerste aspect hebben we een aantal netwerkgebaseerde centraliteitsmaten beschreven die indicatief zijn voor het belang van een woord.

In eerste instantie levert het lexicon een bron van informatie over welke woorden belangrijk zijn vanuit een perspectief dat dichter aansluit bij een mentale benadering van het lexicon dan een taalgebaseerde benadering aan de hand van woordfrequentienormen of streeflijsten (zie bv Kohnstamm, & Schaerlaekens, 1981).

Het tweede aspect, de context waarbinnen een woord gebruikt wordt, kan bestudeerd worden door enerzijds na te gaan met welke andere woorden een cuewoord verbonden is en anderzijds door abstractie te maken van deze context en na te gaan in welke mate het in gelijkaardige contexten gebruikt wordt. Op die manier kan een semantisch veld in kaart gebracht worden. In beide gevallen kunnen deze data bijdragen tot een praktisch instrument in tal van contexten zoals linguïstische geïnspireerde therapievormen bij kinderen, afasiepatiënten of tweede-taalonderwijs. Verder kunnen deze gegevens ook aanleiding geven tot nieuwe diagnostische testen zoals de Semantische Associatie Test (Visch-Brink, Stronks, Denes, 2005) die aantrekkelijk blijven omwille van hun gunstige psychometrische eigenschappen voor het identificeren van semantische stoornissen, de eenvoud van de taak en korte afnameduur.

Voetnoten

1. Op dit moment loopt de studie nog steeds verder. Deelnemen kan op <http://www.smallworldofwords.com/nl>
2. Informatie en data voor beide maten is beschikbaar op <http://www.smallworldofwords.com/visualize>
3. Connotatie is latent wel aanwezig en komt onder meer naar boven in gerelateerdheidsnormen. Deze dimensie komt bij adjectieven als de belangrijkste dimensie tevoorschijn na multidimensionale schalering [Deese, 1965; De Deyne et al., 2013].

Referenties

- Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon*. Wiley-Blackwell.
- Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database [CD-ROM]*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.
- Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 9, 240–247.
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6), 407–428.
- Deese, J. (1965). *The structure of associations in language and thought*. Baltimore: Johns Hopkins Press.
- De Deyne, S., Navarro, D.J., & Storms, G. (2012). Better explanations of lexical and semantic cognition using networks derived from continuous rather than single word associations. *Behavior Research Methods*.
- De Deyne, S., Peirsman Y., & Storms, G. (2009). Sources of Semantic Proximity. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 1834–1839). Austin, TX: Cognitive Science Society.
- De Deyne, S., & Storms, G. (2008a). Word Associations: Network and Semantic properties. *Behavior Research Methods*, 40, 213–231.
- De Deyne, S., & Storms, G. (2008b). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, 40, 198–205.
- de Groot, A. M. (1989). Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(5), 824–845.
- Firth, J. R. (1957). *Studies in linguistic analysis*. Oxford, England: Basil Blackwell.
- Fitzpatrick, T. (2006) Habits and rabbits: word associations and the L2 lexicon. *EUROSLA Yearbook 6*: 121–46.
- Kohnstamm, G. A., & Schaerlaekens, A. M. (1981). *Nieuwe streeflijst woordenschat: voor 6-jarigen; gebaseerd op onderzoek in Nederland en België*. Swets & Zeitlinger.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211–240.
- McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature reviews. Neuroscience*, 4(4), 310–22.
- McRae, K., Khalkhali, S., & Hare, M. (2012). Semantic and associative relations: Examining a tenuous dichotomy. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (Eds.), *The Adolescent Brain: Learning, Reasoning, and Decision Making* (pp. 39–66). Washington, DC: APA.
- Mollin, S. (2009). Combining corpus linguistics and psychological data on word co-occurrence: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory*, 5, 175–200.
- Nelson, D.L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review*, 105, 299–324.
- Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: A review of research and theory. *Psychological Bulletin*, 84, 93.
- Newman, M. (2009). *Networks: an introduction*. OUP Oxford.
- Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, 36, 506–515.
- Severens, E., Van Lommel, S., Ratinckx, E., & Hartsuiker, R. J. (2005). Timed picture naming norms for 590 pictures in Dutch. *Acta Psychologica*, 119, 159–187.

Taal(ontwikkelings)stoornissen

Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, 19, 317- 330.

Saussure, F. (1959). *Course in general linguistics*. (W. Baskin, Trans.) New York: McGraw-Hill.

Steyvers, M., & Tenenbaum, J. B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29, 41–78.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar based information in natural language categories. *Journal of Memory and Language*, 42, 51-73.

Visch-Brink, E. G., Stronks, D. L., & Denes, G. (2005). *De semantische associatie test*. Harcourt Test Publishers.

Wettler, M., Rapp, R., & Sedlmeier, P. (2005). Free Word Associations Correspond to Contiguities Between Words in Texts. *Journal of Quantitative Linguistics*, 12(2-3), 111–122.

Wittgenstein, L. J. J. (1953). *Philosophical investigations*. Oxford, England: Basil Blackwell.

Correspondentieadres

Simon De Deyne (simon.dedeyne@ppw.kuleuven.be)

Gert Storms (gert.storms@ppw.kuleuven.be)

Experimentele Psychologie

Tiensestraat 102 – bus 3711

3000 Leuven