Nederlandse samenvatting

Veilingmeesters en sportcommentatoren staan bekend om de enorme snelheid waarmee ze praten over biedingen en ballen. Om dit te kunnen, maken ze gebruik van een beperkte set van zinnen en zegswijzen die ze als kant-en-klare brokstukken aan elkaar kunnen knopen. Toch zijn het niet alleen de veilingmeesters en de sportcommentatoren die vaak in vaste formules praten — we maken er allemaal gebruik van in ons dagelijks leven.

De schattingen lopen uiteen, maar over het algemeen wordt aangenomen dat zeker de helft van onze gesproken en geschreven taal bestaat uit formules, standaardzinnen, en vaak voorkomende combinaties van woorden. Sommige van deze combinaties zijn ondoorzichtig: De betekenis van het geheel is niet af te leiden uit de som van de betekenissen van de losse woorden. *Daar komt de aap uit de mouw* gaat niet letterlijk over apen die uit kledingstukken klimmen. Er zijn echter ook veelvoorkomende combinaties waarvan de betekenis transparant is: Als je weet wat *in*, *de*, en *auto*, betekenen, dan weet je ook wat *in de auto* betekent. Deze transparante combinaties worden 'lexicale bundels' genoemd in de literatuur, en vormen het onderwerp van onderzoek van deze dissertatie.

Omdat lexicale bundels veel vaker gebruikt worden dan op basis van kans verwacht kan worden, rijst de vraag waarom juist die combinaties door taalgebruikers gebruikt worden. Binnen de taalwetenschap bestaat de gebruiksgebaseerde (*usage-based*) taalbenadering, die stelt dat de cognitieve representatie van taal voortkomt uit de manier waarop taal gebruikt wordt. Vanuit deze benadering is het aannemelijk dat vaakvoorkomende combinaties van woorden door gebruik samensmelten en ingesleten raken als brokstukken van taal die als eenheden verwerkt worden, zonder dat de gebruiker keer op keer de losse woorden moet samenvoegen op basis van grammaticale regels. Dit proces van samensmelten en inslijten staat bekend als *chunking*, en is een welbekend proces in andere cognitieve taken. Het zorgt ervoor dat deze taken snel, soepel en foutloos uitgevoerd kunnen worden. Het onthouden van een telefoonnummer, bijvoorbeeld, gaat makkelijker als de cijfers in brokstukken van meerdere losse getallen worden geleerd.

Deze gebruiksgebaseerde benadering voorspelt dus dat hoogfrequente combinaties van woorden als eenheden worden verwerkt. De afgelopen jaren zijn er steeds meer experimentele studies uitgevoerd waarvan de resultaten lijken te bevestigen dat vaakvoorkomende combinaties van woorden als eenheden in de verwerking gebruikt worden. Zo is gebleken dat de frequenties van gehele combinaties een grote rol spelen in het voorspellen van de snelheid waarmee mensen lexicale bundels lezen en uitspreken, los van de frequenties van de losse woorden waaruit die combinaties bestaan. Er waren echter nog erg weinig studies gedaan naar andere talen dan het Engels; er was nog nauwelijks onderzoek gedaan naar de verwerking van gesproken lexicale bundels; geavanceerde statistische modellen om goed te begrijpen welke factoren een rol spelen tijdens de verwerking van lexicale bundels werden nog weinig toegepast; en er is ook nog maar weinig gebruik gemaakt van computationale modellen om betere inzichten te krijgen in het verwerkingsproces van lexicale bundels. De onderzoeken in deze dissertatie pogen deze hiaten op te vullen.

De belangrijkste vraag die deze dissertatie tracht te beantwoorden is hoe lexicale bundels verwerkt worden door lezers, luisteraars, en sprekers van het Nederlands. Daarbij wordt gebruik gemaakt van geavanceerde statistische technieken en een computationeel model dat een cognitief plausibel model van leren biedt. Door gebruik te maken van grote corpora van het Nederlands, was het mogelijk om vast te stellen welke woordcombinaties van drie woorden zeer vaak voorkomen. Na checks door twee onafhankelijke codeerders kon vervolgens bepaald worden welke hoogfrequente combinaties een transparante betekenis hebben, en dus als lexicale bundel van het Nederland beschouwd zouden kunnen worden. De dissertatie is verdeeld in drie delen, waarbij het eerste deel zich richt op het lezen van lexicale bundels door zowel jongere als oudere lezers, het tweede deel op het luisteren naar lexicale bundels, en het derde deel op het lezen en het uitspreken van lexicale bundels, en welke extra inzichten in deze processen een computationeel model kan toevoegen.

Hoofdstuk 2 richt zich op de vraag hoe jongere en oudere volwassenen lexicale bundels lezen, en of er verschillen zijn tussen de leeftijdsgroepen. De vraag of er verschillen bestaan tussen jongere en oudere lezers, komt voort uit een *usage-based* standpunt dat aanneemt dat de representatie van taal in het brein bij ieder individu het gevolg is van de individuele ervaring die deze persoon met taal heeft gehad. Daardoor heeft eenieder ook een unieke taalrepresentatie, omdat iedereen weer andere ervaring met taal opdoet. Aangenomen wordt dat een grotere blootstelling aan een bepaalde combinatie van woorden tot het versmelten van deze combinatie leidt, waarbij deze lexicale bundel steeds meer als eenheid in taalverwerking gebruikt wordt. Omdat ouderen een veel grotere ervaring met taal hebben dan jongeren, en dus ook veel vaker frequente lexicale bundels tegen zijn gekomen, volgt uit een *usage-based* benadering dat bij ouderen lexicale bundels anders gerepresenteerd zijn dan bij jongeren. Het kan zijn dat ouderen een sterkere en uitgebreidere representatie van lexicale bundels hebben door de grotere blootstelling, of juist minder gebruik maken van kant-en-klare brokstukken, omdat ze meer oefening hebben dan jongeren in het

ophalen van losse woorden uit het lexicon en het samenvoegen daarvan volgens de regels van de grammatica. Als er een verschil bestaat in de representatie van lexicale bundels, dan heeft dat zeer waarschijnlijk ook gevolgen voor de manier waarop deze verwerkt worden.

Door gebruik te maken van eye-tracking is in een experiment in kaart gebracht hoe zowel 60-plussers als twintigers hoogfrequente Nederlandse lexicale bundels lezen. We hebben gebruik gemaakt van statistisch modelleren om in staat te zijn de effecten van verschillende linguïstische eenheden van verschillende groottes in kaart te brengen — zijn het vooral de losse woorden die bijdragen aan hoe snel een combinatie van woorden wordt gelezen, of spelen ook de frequenties van combinaties van twee of zelfs drie woorden mee?

Het gekozen statistisch model, een zogenaamd *generalized additive mixed-effects model* of GAMM, is een regressiemodel dat in staat is om niet-lineaire relaties te modelleren, en daarbij bovendien ook rekening houdt met de individuele verschillen tussen proefpersonen die losstaan van de kenmerken van de lexicale bundels zelf, en het tijdsverloop door het experiment heen. Deze modelleertechniek maakt het mogelijk om te zien welke linguïstische factoren een rol spelen in de duur van verschillende onderdelen van het lezen, en wat voor vorm deze relatie heeft. Er zijn duidelijke aanwijzingen dat representaties van gehele lexicale bundels een rol spelen in lezen, en al in een vroeg stadium, aangezien er frequentie-effecten van trigrammen zijn gevonden in modellen van de duraties van al de eerste fixaties gemaakt op de bundels. Deze frequentie-effecten spelen samen met verschillende oogmotorische kenmerken, zoals de positie van een fixatie, een rol in de duur en het aantal fixaties.

Opvallend is dat deze frequentie-effecten een andere richting hebben dan verwacht: Hoe frequenter een lexicale bundel, hoe langer de eerste fixatie op deze bundel duurt. Dit is het *Inverted Frequency Effect* genoemd, en zou verklaard kunnen worden door ofwel 1) lexicale competitie die hoger is zodra er sprake is van hoogfrequente combinaties, waarbij een grotere competitie leidt tot een vertraagde en dus langere verwerking; 2) een leesstrategie die proefpersonen (on)bewust inzetten bij lexicale bundels versus 'gewone' combinaties van woorden of 3) een andere verwerking van lexicale bundels dan losse woorden, omdat de verwerking van lexicale bundels een ander en trager proces is dan het verwerken van losse woorden.

Er is geen enkel verschil gevonden in de manier waarop jongeren en ouderen lexicale bundels lezen. Dit is onverwacht vanuit een *usage-based* perspectief, waar de voorspelling zou zijn dat een verschil van dertig tot veertig jaar aan taalervaring een groot verschil in representaties in het lexicon tot gevolg zou moeten hebben, en dus ook op online taalverwerking. Het zou kunnen zijn dat taalrepresentaties bij jongvolwassenen al gestabiliseerd zijn en nog maar weinig veranderen in de jaren daarna. Het is ook mogelijk dat de stimuli gebruikt voor dit experiment niet optimaal waren, of dat er door toeval geen effect is gevonden — een 'false negative'. Hoe het ook zij, het zal interessant zijn om in toekomstige experimenten vast te stellen of een grotere taalervaring daadwerkelijk geen effect heeft op de verwerking van lexicale bundels, of dat

de *usage-based* benadering deels herzien zal moeten worden.

Hoe mensen gesproken lexicale bundels verwerken, is het thema van hoofdstuk 3. Er bestond nog vrijwel geen onderzoek naar de online verwerking van gesproken lexicale bundels, en dit hoofdstuk bespreekt een experiment waarin proefpersonen moesten luisteren naar allerlei frequente lexicale bundels zoals *aan de beurt* en een laagfrequente tegenhanger zoals *aan de prins*. Hoewel beide combinaties bestaan uit losse woorden die gematcht zijn op hun frequentie, en met dezelfde twee woorden beginnen, zijn er grote verschillen in de frasale frequenties: *Aan de beurt* komt veel vaker voor dan *aan de prins*. Wanneer een proefpersoon deze lexicale bundels hoort, zal zij pas aan het einde van het tweede woord doorhebben wat het laatste woord zou kunnen zijn — een verwachte continuatie die het einde van een hoogfrequente lexicale bundel vormt (*aan de beurt*) of juist een onverwacht, maar even frequent woord, *prins*, dat in combinatie met *aan de* weinig voorkomt. Deze twee condities zijn vervolgens met elkaar vergeleken door met machine learning de ERP-data te analyseren.

De analysetechniek die voor deze dataset is gebruikt, is een *conditional inference random forest* (CForest). CForests zijn een krachtig machine learning algoritme waarbij een grote groep van verschillende beslisbomen worden gegenereerd. Iedere afzonderlijke beslisboom is gebaseerd op een willekeurige subset van de data, en voor iedere splitsing in de beslisboom wordt steeds uit een willekeurige subset van predictoren bepaald welke predictor de beste tweedeling in de data maakt. Dit zorgt voor een grote variatie in de afzonderlijke beslisbomen, die samen een bos of 'forest' vormen. Random forests zijn in staat om niet-lineaire relaties in de data vast te leggen, en staan bekend om hun grote nauwkeurigheid en stabiele voorspellingen.

Naast deze voordelen van random forests, is een belangrijke reden om bij deze EEG-studie voor CForests te kiezen, dat CForests het mogelijk maken om de effecten te beoordelen van sterk aan elkaar gecorreleerde predictoren. De frequentie van de gehele lexicale bundel is vaak sterk gecorreleerd met de frequenties van de bigrammen en unigrammen waaruit deze is opgebouwd. In regressie-analyses is het daarom niet mogelijk om al deze predictoren tegelijkertijd in één model mee te nemen — terwijl het heel goed mogelijk is dat al deze eenheden in parallel een effect hebben op de verwerking van lexicale bundels. Bovendien is EEG-data afkomstig van verschillende electrodes niet onafhankelijk van elkaar — het signaal gemeten door een willekeurige electrode is sterk gecorreleerd met het signaal van aangrenzende electrodes.

In de EEG-data is een duidelijk verschil te zien tussen het signaal gegenereerd door frequente lexicale bundels, en het signaal gegenereerd door de controle-items. Er is een continu en vroegbeginnend negatief signaal dat bij de controle-items een nog negatievere voltage had. In het random forest model is de vorm van het verloop van de voltages gemodeleerd door het effect van de lengte van de stimuli, de kans dat een woord op de derde plek van een stimulus zou staan, de frequenties van de losse woorden, de bigrammen, en de trigrammen mee te nemen, rekening te houden met de status van een item (een lexicale bundel of controle-item), het tijdsverloop, en de electrode waar het sig-

naal gemeten is. Door naar een representatieve boom uit de random forest te kijken, is het mogelijk om een deel van het random forest model te doorgronden en hypotheses te vormen over hoe auditief gepresenteerde lexicale bundels verwerkt worden.

We stellen voor dat bij de verwerking van gesproken lexicale bundels drie stadia te onderscheiden zijn. Als eerste worden er voorspellingen gemaakt over wat er zou kunnen komen, terwijl er tegelijkertijd volop *bottom-up* verwerking is. Na deze eerste stappen worden mogelijke concurrende vormen actief onderdrukt, terwijl er een competitie ontstaat tussen andere mogelijke lexicale kandidaten. Deze competitie zorgt ervoor dat de verwerking van lexicale bundels met vele lexicale concurrenten moeilijker is voor het cognitieve systeem. In de derde en laatste fase vindt de lexicale integratie plaats van alle vrijgekomen informatie. In al deze drie fasen is duidelijk dat de frequenties van zowel enkele woorden, bigrammen, en de gehele trigram een rol spelen, vaak parallel aan elkaar.

In hoofdstuk 4, ten slotte, worden de eerste stappen gezet naar een beter begrip van lexicale toegang tot lexicale bundels. Hoewel er een groeiend aantal studies is dat frequentie-effecten voor vaakvoorkomende combinaties van woorden vindt, wat veel onderzoekers doet vermoeden dat deze bundels een cognitieve realiteit hebben, is het niet duidelijk hoe het brein toegang krijgt tot deze bundels. Om beter te kunnen begrijpen wat een lexicale bundel is, helpt het om expliciet in een computermodel vast te leggen hoe lexicale toegang zou kunnen verlopen, en dan te testen of predictoren uit een dergelijk model even goed of zelfs beter in staat zijn om experimentele data te beschrijven dan traditionele predictoren zoals frequenties.

We hebben twee experimenten uitgevoerd, een leesexperiment waarbij we data van oogbewegingen registreerden met behulp van eye-tracking, en een productie-experiment, waarbij we registreerden hoe snel proefpersonen begonnen met hardop voorlezen van lexicale bundels, en hoe lang ze erover deden om deze bundels helemaal uit te spreken. De data van beide experimenten zijn gemodelleerd met zowel traditionele predictoren zoals de frequentie van de lexicale bundels, als predictoren uit een computationeel model van lexicale toegang, de Naive Discriminative Learner (NDL), waarbij lexicale bundels expliciet in het model zijn opgenomen.

NDL is een eenvoudig neuraal netwerk dat uit slechts twee lagen bestaat, een inputlaag waar fonemen, letters, of losse woorden de *cues* vormen die verbonden zijn met de outputlaag, een set van *outcomes* of uitkomsten, in dit geval symbolische eenheden, *lexomes*. Deze lexomen wijzen naar de locatie van lexicale bundels in een semantische ruimte, en zijn stabiele eenheden die een connectie vormen tussen immer veranderende taalvormen en betekenissen. In een NDL netwerk zijn alle *cues* met alle uitkomsten verbonden, en worden connecties gevormd via de Rescorla-Wagner leerregels. Deze leerregels zijn erg succesvol gebleken in het modelleren van uiteenlopende gedragingen van dieren, en vormen daarmee een cognitief plausibel algoritme dat gebruikt kan worden om te modelleren hoe mensen hun linguïstische kennis in de loop der jaren

opbouwen.

Volgens het NDL-model vormt een taalgebruiker op basis van woord-*cues* verwachtingen over welke lexicale bundel hij kan verwachten. Behalve verwachtingen op basis van de input, heeft een taalgebruiker ook verwachtingen op basis van eerdere ervaringen, zodat een lexicale bundel die vaker is gebruikt, ook eerder verwacht wordt. Door verwachtingen op basis van zowel de input als eerdere ervaringen te combineren, en dat te vergelijken met de daadwerkelijke combinaties van woorden in het signaal, leert het systeem van zowel correcte als incorrecte voorspellingen. De kracht van NDL zit niet alleen in het feit dat het rekening houdt met woorden die vaak samen voorkomen, maar dat het ook inzichtelijk maakt hoe onderscheidend een *cue* is: het woordje *een* kan door vele andere woorden gevolgd worden, en is dus een slechte *cue*, terwijl *paarse* een sterke *cue* vormt voor *krokodil*.

Een getraind NDL-netwerk vormt een mathematische karakterisatie van de toestand van het lexicon. Uit een dergelijk netwerk kunnen allerlei predictoren worden gehaald, zoals hoe sterk de verwachting van een bepaalde lexicale bundel op voorhand al is (een predictor die sterk lijkt op een traditionele frequentiemaat), hoe sterk bepaalde losse woorden de verwachting opwekken van bepaalde lexicale bundels, en hoe makkelijk bepaalde lexicale bundels van elkaar te onderscheiden zijn. Uit hoofstuk 4 is gebleken dat deze NDL-predictoren beter in staat zijn om de experimentele data te beschrijven dan traditionele frequentiematen alleen, en bovendien meer inzichten verschaffen.

Lexicale toegang tot lexicale bundels vindt plaats vanuit zowel een *top-down* als een *bottom-up* proces, waarbij trigram-frequenties een grote rol spelen, en een grotere co-activatie van vergelijkbare items het uitspreken van lexicale bundels versnelt. Als alleen gebruik gemaakt wordt van frequentiematen, zouden *bottom-up* en *top-down* processen niet los van elkaar beschouwd kunnen worden. Daarnaast blijkt uit de eye-trackingdata dat lezers sneller lexicale bundels lezen als ze meer tijd besteden aan de *first pass*, de eerste keer dat ze een stuk tekst van links naar rechts lezen.

Door de hele dissertatie heen komt keer op keer naar voren dat eenheden groter dan het woord een rol spelen in lezen, luisteren en spreken in het Nederlands. Uit de data blijkt dat het tijdsverloop en de processen betrokken bij taalverwerking vrijwel hetzelfde zijn bij losse woorden en bij lexicale bundels, wat suggereert dat hoogfrequente lexicale bundels op een zelfde manier functioneren en gerepresenteerd zijn. Het lijkt er bovendien op dat vaakvoorkomende lexicale bundels niet alleen vanwege hun frequentie als eenheid in het lexicon functioneren — semantische eigenschappen van deze bundels spelen vermoedelijk ook een rol. Kijkende naar de items gebruikt in deze dissertatie, valt op dat dit voornamelijk discoursmarkeringen zijn zoals *ik denk dat*, *affordances* zoals *op de tafel* en complexe tijd- op ruimtemarkeringen zoals *op de dag* of *in het midden*. Hoewel deze items puur op frequentie van de trigrammen geselecteerd zijn, blijken ze in het algemeen een soort functionele eenheden te zijn. Deze items worden toevallig door meerdere woorden uitgedrukt in het Nederlands, maar worden in (sommige) morfologisch rijke talen uitgedrukt als

een enkel woord. Deze dissertatie laat hiermee zien dat eenheden van vorm en betekenis niet altijd overeen hoeven te komen wat wij wegens orthografische redenen als meerdere losse woorden beschouwen.

Toch betekent dit niet dat lexicale bundels ondoorzichtige brokstukken zijn: Ook de kleinere eenheden waaruit de lexicale bundels zijn opgebouwd, de bigrammen en unigrammen, spelen een rol in de verwerking, parallel aan de gehele lexicale bundels zelf. Dit laat zien dat, hoewel vaak voorkomende combinaties als brokstukken verwerkt worden, het taalsysteem deze brokstukken ook nog steeds opbreekt in kleinere delen, die ieder op zich ook van belang zijn in de verwerking. Dit is aanvullend bewijs voor een model van taalverwerking waarin meerdere eenheden van verschillende groottes parallel worden verwerkt.

De belangrijkste toevoegingen aan bestaand onderzoek zijn a) de focus op geavanceerde statistische modellen die subtiele patronen van verwerking aan het licht kunnen brengen; b) de eerste data die laten zien hoe oudere volwassenen lexicale bundels verwerken; c) een uitgebreidere analyse van de manier waarop gesproken lexicale bundels online verwerkt worden, en d) het inzetten van een computationeel model van lexicale toegang, waarin lexicale bundels als eenheden zijn opgenomen — dit maakt het mogelijk om lexicale toegang tot lexicale bundels in subprocessen op te delen en daardoor beter te begrijpen, en daarmee ook de status van lexicale bundels in het mentale lexicon te beschrijven.