Morfologische aspecten van het ideale woordenboek

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6006
fax: +31 30 253 6000
e-mail: lot@let.uu.nl
http://wwwlot.let.uu.nl/

Cover illustration: *Een lexicale kring*, door J.K. Loman (2005)

ISBN 90-76864-83-7
NUR 632

Morfologische aspecten van het ideale woordenboek

Een theoretische en empirische studie naar de lexicale samenhang van het Nederlands ten behoeve van een morfologische kennisbank

Morphological aspects of the ideal dictionary
A theoretical and empirical study to the lexical cohesion of the Dutch language for the purpose of a morphological data bank

(with a summary in English)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de Rector Magnificus, prof. dr. W.H. Gispen,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen
op woensdag 16 november 2005
des middags te 12.45 uur

door

Niels Oele Koornwinder

geboren op 13 april 1972 te Amsterdam

Promotor: prof. dr. H.J. Verkuyl - Uil OTS, Universiteit Utrecht
Copromotor: dr. J.J. Zuidema - Van Dale Lexicografie, Utrecht

Dankbetuiging

Als een onderzoeker zeven jaar aan een promotieproject werkt, kan hij heel veel mensen leren kennen (en weer bijna vergeten zijn). Dat ontdekte ik toen ik me aan het schrijven van dit dankwoord zette. Omdat ik de betrokkenen graag recht wil doen, is het dankwoord dus wat langer uitgevallen dan gebruikelijk. Maar met een proefschrift van ruim 300 pagina's is het schrijven van een extra bladzijde niet meer zo'n punt.

Eerst en vooral wil ik mijn beide promotoren danken. Omdat Henk en Johan vanaf het begin als tandem zijn opgetreden, lijkt het me aardig om ze ook allebei tegelijk in het zonnetje te zetten. Beiden hebben gedurende het hele project een grote, haast vaderlijke betrokkenheid getoond bij de opzet en uitvoering van mijn project, wat zowel in raad als in daad tot uitdrukking kwam. Als ik bijvoorbeeld eens een stuk inleverde, was het altijd binnen korte tijd gelezen en nauwkeurig gecorrigeerd. Henk en Johan zijn me ook altijd blijven steunen, ook toen de officiële einddatum van het project allang verstreken was. Hierbij hebben ze voor lief genomen dat ik vaak behoorlijk eigenwijs was en weinig respect toonde voor overeengekomen deadlines. Op dit punt heeft Johan mij wel iets steviger onder handen genomen dan Henk, met als resultaat dat ik mijn deadlines ook echt begon te halen. Voorwaar een prestatie! Hier kan ik Johan alleen maar dankbaar voor zijn.

Hoewel ik een tijdlang aan zijn gedrevenheid moest wennen, heb ik Johan geleidelijk aan leren kennen als een zeer betrokken en inspirerende begeleider. Voor Henk geldt dat ik een grenzeloze bewondering heb voor zijn vermogen om ingewikkelde dingen simpel voor te stellen, en alles stijlvol en met humor te benaderen. Elk mailtje van Henk leverde mij minstens twee dagen denkvoer op, vooral omdat Henk er een sport van maakte om (in antwoord op mijn epistels) zijn boodschap in zo weinig mogelijk woorden over te brengen. Het lijkt me de moeite waard om deze e-mailconversaties nog eens te bloemlezen: er zaten echt juweeltjes van formuleerkunst bij, maar op momenten van grote tijdsdruk konden de mailtjes ook zeer down-to-earth worden.

Behalve mijn promotoren kende mijn project nog vele andere betrokkenen die van grote betekenis zijn geweest. Zo was er gedurende het eerste jaar van mijn project een werkgroep die elke twee à drie weken bijeen kwam om de koers van mijn project te bespreken. Behalve mijn promotoren bestond deze werkgroep uit twee hoogleraren met werkzaamheden binnen Van Dale, te weten Dirk Geeraerts en Franciska de Jong, en twee Van Dale medewerkers, namelijk de taaltechnologen Janneke Froon (die in die tijd ook aan een onderzoeksproject werkte) en Marc du Chatinier (die de centrale gegevensbank beheerde). Het was altijd erg plezierig om met hen te overleggen. Vooral Dirk Geeraerts heeft me inspirerende gedachten meegegeven (waaronder het idee om een netwerkmodel te gebruiken). Janneke en Marc hebben mij in de eerste fase van mijn project intensief ondersteund bij de aanmaak en bewerking van de Gegevensbank; hoewel ze het zelf ontzettend druk hadden, waren ze altijd snel beschikbaar als ik iets van ze nodig had. Dat geldt ook voor de systeembeheerders (van wie ik Bas, Anna en Menno noem). Hun allen, Cily en Rik Schutz komt dank toe.

In aanvulling op deze praktische begeleidingsgroep werd ook een wetenschappelijke begeleidingsgroep ingesteld, bestaande uit Martin Eveaert, Mieke Trommelen, Jan Don, en - op enige afstand - Wim Zonneveld. Zij hebben mijn project kritisch gevolgd en hierbij de nodige zorg uitgesproken over de grote lexicografische component in mijn werk. Voor mij was deze kritiek niet altijd makkelijk, maar ik vertrouwde op de visie van mijn promotoren en toen ik eenmaal aan het schrijven van mijn proefschrift begon, vond ik het een uitdaging om de geuite kritiek te weerleggen. De toekomst zal leren of ik hierin geslaagd ben.

Bij Van Dale heb ik verschillende medewerkers ontmoet die door hun persoonlijke belangstelling en interessante ideeën op een waardevolle manier hebben bijgedragen aan de tot standkoming van mijn lexicale theorie. Hierbij denk ik met name aan Peter Verhoeven, Josée Heemskerk en Rob Ermers; maar ook Rob Hekkers, Alex Kaiser, Michel Boekenstein, Marc Zuidema en Anneke Nunn sprak ik graag, om maar enkele personen te noemen. Als ik terugdenk aan de jaren dat ik op het kantoor van Van Dale werkte, roept dit veel warme herinneringen op. Hoewel het eerste jaar behoorlijk zwaar was (onder meer door het monomane werk aan mijn database), ging ik elke dag met plezier naar mijn werk; in de wandelgangen was er altijd wel tijd voor een interessant gesprekje en ik zat op een hele gezellige afdeling waar de taalgrappen je om de oren vlogen; in dit verband verdient Tjerk Hacquebord een buitenstzondere vermelding; ook de diverterende gesprekken met Anouk Monté zullen me bijblijven.

Zowel op het UiL OTS als aan andere unversiteiten zijn vele senioronderzoekers die mij (meestal in het kader van een LOT-school) geholpen hebben door interesse te tonen in mijn onderzoek, nuttige suggesties te doen of artikelen aan te reiken. Hierbij denk ik in de eerste plaats aan de volgende personen: Geert Booij (zijn LOT-cursus heeft grote invloed gehad op mijn verdere onderzoek), Harald Baayen (de eerste lezing die ik van hem meemaakte was een echte blikopener voor me), Michael Moortgat (van wie ik diverse cursussen heb gevolgd) en Jan Don (die deel uitmaakte van mijn begeleidingscommissie). Ook als aio-begeleider heeft Jan veel voor me gedaan, evenals Maaike Schoorlemmer; ik ben ze hier erkentelijk voor. In dit verband wil ik ook Keetje van den Heuvel en de secretariaatsmedewerkers van UiL OTS en LOT bedanken. Voorts wil ik de conciërges Jan en Jos van een pluim voorzien.

Andere onderzoekers met wie ik stimulerende gesprekken en e-mailwisselingen had zijn (o.a.) Jan van Eijck, Frank van Eynde, Antal van den Bosch, Dominiek Sandra, Anneke Neijt, Mirjam Ernestus, Nicoline van der Sijs, Mieke Trommelen, Jacqueline van Kampen, Arnold Evers, Jack Hoeksema, Crit Cremers, Ton van der Wouden, Joost Zwarts, Yoad Winter, Johan Kerstens, Frank Drijkoningen, Gertjan Postma, Arjen Versloot, Peter Ackema, Bart Hollebrandse, Arie Verhagen, Sieb Nooteboom, Tanya Reinhart en Eric Reuland. Enkele van deze onderzoekers zijn ook toegetreden tot de leescommissie van mijn proefschrift; deze commissie, bestaande uit Geert Booij, Franciska de Jong, Jack Hoeksema, Jan van Eijck en Martin Everaert, wil ik hartelijk danken voor haar medewerking.

Op het UiL OTS kon ik mij in een warme kring van collega's verheugen. Omdat ik hier ook tijdens mijn studiejaren had rondgelopen, kende ik de meeste collega's al, zowel onder de staf als onder de aio's. Helaas had ik niet veel gelegenheid om met ze op te trekken, want ik verbleef het grootste deel van mijn tijd bij Van Dale. Maar tijdens de cursusperiodes van de LOT onderzoekschool (die ik altijd enorm inspirerend vond) was er alle tijd om mijn collega's beter te leren kennen; dat was nog sterker het geval bij de zomerschool in Potsdam (waar vrijwel alle taalkunde-aio's uit Nederland vertegenwoordigd waren) en de ESSLLI-zomerschool in Birmingham (waar ik met vier andere Utrechters aan deelnam). Die beide buitenlandse reizen waren onvergetelijk, zowel in wetenschappelijk opzicht als in sociaal opzicht.

Op het UiL OTS zijn eveneens collega's die een speciale vermelding verdienen. Dan denk ik in de eerste plaats aan mijn kamergenoot, Mike Huiskes: Mike, ik zal onze urenlange conversaties over taalkundige, maatschappelijke en andere vraagstukken nooit vergeten. Je gerichte belangstelling en kritische blik hebben heel veel bijgedragen aan mijn zelfvertrouwen en de ontwikkeling van mijn taalkundige ideeën. Helaas zijn onze wegen wat uiteen gaan lopen. Maar voor mij was je echt de ideale gesprekspartner (en dat ben je nog steeds!). Behalve met Mike heb ik ook vele aangename uren doorgebracht met de andere 0.52-ers, namelijk met Ninke Stukker, Jacqueline Evers-Vermeul en Judith Kamalski. Hoewel ik de

laatste jaren veel thuiswerkte, was het altijd een feest (wilde thee!) om weer een dagje bij jullie op de werkkamer te zijn.

Van de vele (LOT-)promovendi met wie ik interessante gesprekken heb gevoerd, zijn er enkele die ik speciaal wil noemen, omdat ze veel betekend hebben voor mijn onderzoek. Het gaat om Rick Nouwen, Anna Młynarczyk, Østein Nilsen, Willemijn Vermaat, Iris Mulders, Inge Zwitserlood, Fabien Reniers en Corrien Blom. Er zijn natuurlijk tal van andere promovendi en postdocs die me hebben geïnspireerd, met wie ik in leesgroepjes heb gezeten of waarmee ik gewoon een gezellige tijd heb gehad. Hen wil ik bedanken door middel van een alfabetische opsomming van hun namen: Annemarie Mineur, Annemarie Kerkhof, Benjamin Spektor, Christine Versluis, Christophe Costa Florêncio, Dimitra Papangeli, Elise de Bree, Elma Blom, Esther Janse, Gerben Mulder, Hedde Zeijlstra, Heleen Hoekstra, Ingeborg van Gijn, Joost Kremers, Judith van Wijk, Juliette Waals, Kakhi Sakhltkhutsishvili, Maarten Janssen, Marijana Marelj, Mirna Pit, Nada Vasić, Nadezhda Vinokurova, Olaf Koeneman, Olga Borik, Patrick Brandt, Paz Gonzáles, Raffaela Bernardi, Richard Moot, Sarah Kennelly, Sergio Baauw, Sharon Unsworth, Silke Hamann, Suzanne Aalbers en Wouter Kusters.

Ik ben vast nog vele collega's vergeten, maar in de voorbije jaren heb ik zoveel mensen leren kennen dat het niet lukt om iedereen recht te doen. Iemand die ik op de valreep nog een eervolle vermelding wil geven is Zhang Shuyuan xiâojie. Deze Chinese dame kwam eind 2004 naar Utrecht voor een verblijf als gastonderzoeker. Omdat ze mijn kamergenote werd, kregen we al gauw een bijzondere band. Shuyuan heeft ook een speciale luister gegeven aan mijn laatste maanden als promovendus, en heel veel bijgedragen aan mijn geluk door me kennis te laten maken met mijn geliefde Min.

Het laatste deel van dit dankwoord wil ik wijden aan mijn vrienden en kennissen buiten de taalkundige wereld en natuurlijk aan mijn familie. Voor hen allen geldt dat ik de afgelopen drie jaar heel veel steun heb gehad aan alle belangstelling die ik heb mogen ondervinden voor mijn onderzoek. Met sommige vrienden heb ik veel over mijn onderzoek gepraat en mijn zorgen gedeeld, namelijk Valentijn en Aafke, Arthur, Maarten, Peter, Martijn, Tuvit, Henk en Christa, Eelco, Catherina, zeilzwerver Ruben en Beer (die zich op alle fronten een zeer behulpzaam huisgenoot heeft getoond). Bedankt! Jullie luisterend oor en adviezen waren van grote waarde. In mindere mate geldt dit ook voor al mijn andere vrienden en kennissen, maar dit dankwoord zou overladen worden als ik hen allemaal zou noemen. Wie ik wel wil noemen is mijn pianoleraar, Gert, die op de achtergrond van grote betekenis is geweest voor mijn hele ontwikkeling tot "vrijdenker". In zijn kielzog noem ik ook Fred de Bree, die een bijzondere dienst voor me verrichtte. Voorts is een speciale vermelding op zijn plaats voor Tigran (mijn nieuwe werkgever: een droombaas!), en voor de (hele) familie Bletterman, bij wie ik de afgelopen jaren heb ingewoond: jullie gastvrijheid is werkelijk hartverwarmend...

Iemand die ik niet meer kan bedanken, maar die ik wel wil memoreren is Jan Loman. Hij is een jaar terug totaal onverwacht en veel te jong overleden. Zijn ouders hebben zijn gemis enigszins opgevangen door - in plaats van Jan - contact met mij te onderhouden en naar mijn voortgang te informeren. Omdat zijn vader (J.K. Loman) een bekwaam schilder is, lag het voor de hand om deze te vragen de afbeelding voor de voorkant te maken. Het ontwerp hiervoor is in samenwerking met Valentijn Visch ontstaan. Hiermee heeft deze afbeelding naast een filosofische en decoratieve ook een hele persoonlijke betekenis gekregen.

Tot slot wil ik mijn familie bedanken. Hinde en Bas, jullie zorgden met jullie huwelijk en de geboorte van Lukas voor welkome afleiding. En ik heb jullie belangstelling en vertrouwen in een goede afloop enorm gewaardeerd! Dit geldt nog sterker voor mijn ouders, die zelfs materieel zijn bijgesprongen toen het water me tot aan de lippen stond. Tom en Marita, dit proefschrift zie ik als de bekroning van de opvoeding die jullie me gaven.

Inhoudsopgave

1 Het onderzoekskader

1.1 Doelstelling en aanpak

1.1.1 Introductie

In dit boek doe ik verslag van een theoretisch en empirisch onderzoek naar de morfologische structuur van de Nederlandse woordenschat.[1] Dit onderzoek had als doel om een bijdrage te leveren aan de systematisering van de woordkenmerken in de gegevensbank die ten grondslag ligt aan de Nederlandse woordenboeken van uitgever Van Dale Lexicografie (VDL). Hiertoe is een Morfologische Gegevensbank voor het Nederlands (MGBN) ontwikkeld, een project dat centraal staat in dit proefschrift. Bij de opzet en analyse van deze gegevensbank heb ik een brug proberen te slaan tussen lexicografisch (systematisch inventariserend) en linguïstisch (cognitief verklarend) onderzoek naar de Nederlandse woordstructuur. Deze brug bestaat uit een lexicontheorie met een dynamische wisselwerking tussen individuele en collectieve taalkennis.[2] Hierbij is morfologische structuur een bijproduct van lexicale kenniscompressie. Dankzij deze theorie kan een structureel verband worden gelegd tussen de morfologische structuur van het mentale lexicon en die van het woordenboek.

De MGBN is tot stand gekomen door alle woorden uit Van Dale's Groot Woordenboek van de Nederlandse Taal (c.q. Grote Van Dale), dat in totaal 250.000 woorden telt,[3] in basislexemen op te delen en deze ca. 80.000 basislexemen van morfeemstructuur te voorzien. Deze informatie is niet alleen nuttig met het oog op lexicografische toepassingen, maar kan ook worden ingezet voor linguïstisch onderzoek naar de morfologische eigenschappen van de Nederlandse woordenschat en de onderliggende structuurprincipes. In het kader van deze studie heb ik de MGBN aan een reeks statistische analyses onderworpen en de zo verkregen informatie aan de bestaande morfologische kennis getoetst door deze systematisch te vergelijken met het Morfologisch Handboek van het Nederlands (MHB) van De Haas en Trommelen (1993). Hieruit blijkt dat de MGBN tot een aanzienlijke uitbreiding van de affixkennis leidt. Bovendien biedt de MGBN een zeer uitgebreide inventarisatie van wortels, sequenties en affixparadigma's, iets waarover het MHB weinig te melden heeft. Op dit terrein overtreft de MGBN ook automatisch geannoteerde tekstcorpora als CELEX en het Corpus Gesproken Nederlands (CGN).

Bij de opzet van de MGBN heb ik me laten leiden door het uitgangspunt dat de hierin vastgelegde kennis aan de eisen van een Ideaal Woordenboek (IW) moet voldoen. Deze eisen zijn vastgelegd in het IW-model van Verkuyl & al. (1998). Dit model berust op het idee dat een lexicografisch informatiesysteem bruikbaarder wordt naarmate het beter in staat is om voor nader te specificeren taken (zoals spellingadvies en betekenisduiding) de rol van een menselijke taalexpert over te nemen. Hiertoe dient het onderliggende informatiesysteem dezelfde kennisstructuur te krijgen als het mentale lexicon, terwijl de hierin opgeslagen kennis aan hoge eisen moet voldoen ten aanzien van de zogeheten c-criteria, te weten *consistentie*, *compleetheid* en *correctheid*. Verder dient het informatiesysteem over een gebruiksvriendelijke zoekcomponent te beschikken; deze moet de gebruiker helpen om te achterhalen of het opgegeven woord deel uitmaakt van de bekende of mogelijke woordenschat van de bevraagde taal, en contextgevoelige informatie verstrekken over woordkenmerken als spelling, uitspraak, betekenis, vervoeging, interne structuur en syntactische eigenschappen. In deze studie wordt

[1] In aanvulling op dit boek zal een website worden ingericht met allerhande aanvullingen, waaronder detailinformatie over de gegevensbank, complete datarapporten, aanvullingen op mijn theorie en nieuwe publicaties. Deze website wordt ondergebracht bij het UiL OTS; nadere informatie is verkrijgbaar via UiL-OTS@let.uu.nl.
[2] Everaert (2004) spreekt in dit verband van *binnentaal* en *buitentaal*. Zie ook voetnoot 53.
[3] Als men ook inflectievormen meerekent, bevat dit woordenboek meer dan een miljoen Nederlandse woorden.

dit lexicografische ideaal stap voor stap omgezet in een concreet structuurmodel voor een lexicografisch kennissysteem; dit model dient tevens als leidraad voor de MGBN.

In de voorgaande decennia werd het vakgebied van de morfologie sterk gedomineerd door onderzoek naar algemeen geldige regels voor de woordvorming. In die visie dient een morfologische grammatica uitsluitend uit *productieve* woordvormingsregels te bestaan, d.w.z. regels die aangeven hoe men bekende woorden kan uitbreiden tot nieuwe. Daarbij bepaalt de taalkundige intuïtie wat mogelijke en wat onmogelijke woorden zijn, en derhalve wat geldige en niet-geldige regels zijn. Door de focus op productieve woordvorming is relatief weinig bekend over de vraag in hoeverre bekende woorden morfologische structuur bezitten en hoe men deze structuur kan achterhalen. In mijn eigen visie op lexicale kennisrepresentatie is het niet wenselijk om uit te gaan van een statische grammatica met morfologische structuurregels, maar dienen deze regels langs inductieve weg uit de lexicale structuur van bekende woorden te worden afgeleid. Hierbij dienen de woordrepresentaties zoveel mogelijk uit gemeenschappelijke bouwstenen te worden opgebouwd, maar zonder dat er informatie verloren gaat. Dit compressieprincipe vormt het centrale uitgangspunt van de in dit proefschrift voorgestelde theorie: de L-KRING-theorie (Lexicale KennisRepresentatie door Inductieve NaamGeving).

De L-KRING-theorie stelt dat het mentale lexicon met een hiërarchisch netwerk van lexicale indexen correspondeert. Hierbij staat de term *index* voor een arbitraire naam (bijv. een getal), terwijl de lexicale inhoud (c.q. *denotatie*) van deze naam met een geheugenlocatie correspondeert waar een unieke kenmerkenbundel wordt gedefinieerd (analoog aan de relatie tussen letter en bijbehorend foneem). Indien men zich op de morfologische structuurdimensie richt, corresponderen de indexen per definitie met vaste vorm-functie-eenheden[4], namelijk woorden en woordinterne eenheden zoals lexemen (in het geval van samenstellingen) en morfemen; verder kan elke index zelf ook weer interne structuur bezitten. Indien zo'n lexicale eenheid (c.q. index) door meerdere woorden wordt gedeeld, ontstaat morfologische structuur. De hierbij aangemaakte bouwstenen kunnen ook worden benut voor de constructie of interpretatie van nieuwe woorden. In dat geval is sprake van productief gebruik van indexen. De L-KRING-theorie biedt dus perspectief op een zelflerend lexiconmodel, d.w.z. op een lexicaal representatiesysteem dat in staat is om zijn eigen woordvormingsregels te formuleren.

1.1.2 Onderzoeksdoelen

Het in deze studie beschreven onderzoek heeft als centraal doel om een psycholinguïstisch gemotiveerde bijdrage te leveren aan de systematisering van de woordkenmerken in VDL's WoordKenmerkenBank Nederlands (WKB-Ned), d.w.z. de lexicografische kennisbank die ten grondslag ligt aan de formele woordkenmerken in de Nederlandstalige woordenboeken van VDL. Deze centrale doelstelling bestaat uit vier subdoelen, te weten een lexicologische, een lexicografische, een analytische en een evaluatieve doelstelling:

Lexicologische doelstelling: de ontwikkeling van een lexicale representatietheorie die een structureel verband legt tussen het mentale lexicon en een lexicografisch kennissysteem, en die cognitieve criteria verschaft voor morfologische structuurtoekenning.

Lexicografische doelstelling: de ontwikkeling van een morfologische gegevensbank door alle lexemen uit de WKB-Ned langs semi-automatische weg van morfologische structuurinformatie te voorzien.

[4] Hierbij kan de 'functie' zowel met een syntactische als met een morfosyntactische eigenschap corresponderen.

Analytische doelstelling: de systematische beschrijving van de Nederlandse woordbouw door integrale analyse van de MGBN:

- inventarisatie van prefixen en hun combinatiemogelijkheden
- inventarisatie van suffixen en hun combinatiemogelijkheden
- inventarisatie van prefix-suffix-interacties
- analyse van de onderliggende structuurcriteria

Evaluatieve doelstelling: de evaluatie van de MGBN door
a) externe evaluatie: toetsing van de MGBN door de hierin aanwezige affixkenmerken met de kennis in het Morfologisch Handboek (MHB) te vergelijken en vice versa;
b) interne evaluatie: toetsing van de consistentie van de aan de MGBN ontleende analyserapporten door het bijbehorende functieverloop te beoordelen.

1.1.3 Aanpak

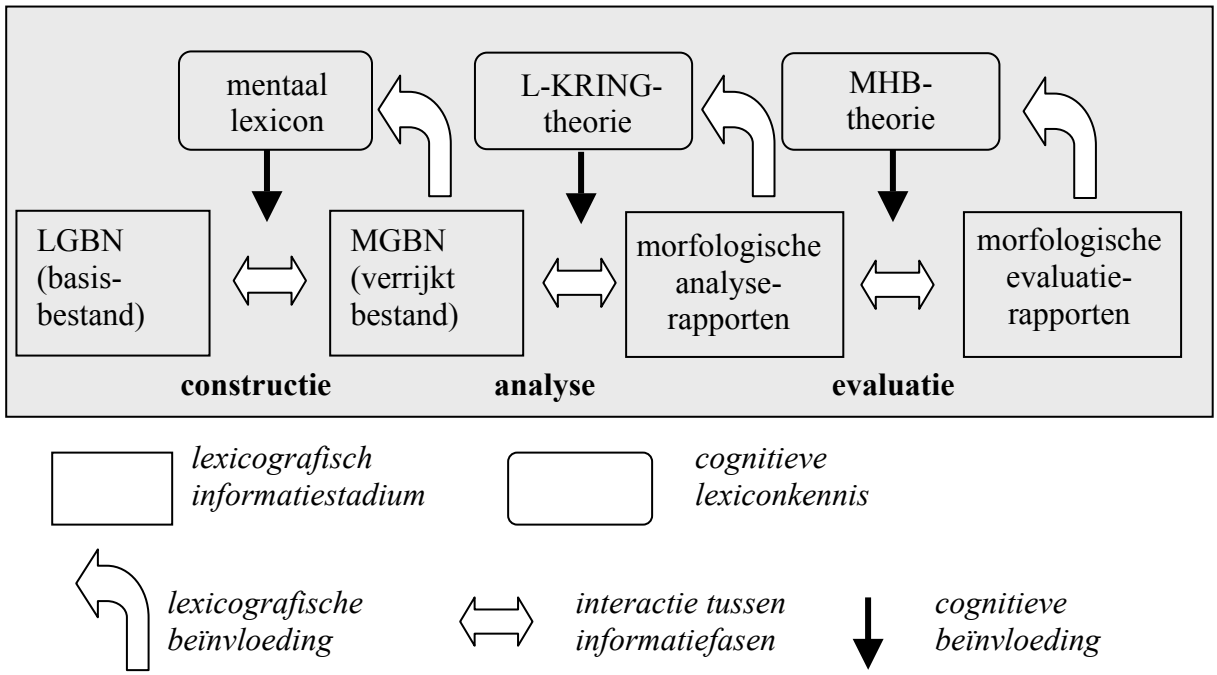
Het fasediagram in figuur 1-1 toont de opeenvolgende ontwikkelingsfasen van de MGBN. Hierbij zijn de volgende bewerkingsstappen te onderscheiden:

1) aanmaak en aanvulling van het lexicografische basisbestand, te weten de Lexicale Gegevensbank voor het Nederlands (LGBN); de LGBN-kenmerken komen uit VDL's Nederlandse WoordKenmerkenBank (WKB-Ned)
2) aanmaak en verrijking van de MGBN door cyclische toekenning van morfologische structuurrepresentaties (inclusief controles)
3) morfologische analyse door systematische inventarisatie van MGBN-kenmerken
4) morfologische evluatie door vergelijking van de analyserapporten met de MHB-kennis

Het fasediagram toont niet alleen deze bewerkingsstappen en de resulterende informatiefasen (in de onderste laag), maar ook hun interactie met een aantal cognitieve kennisbronnen over de morfologische structuur van het Nederlands (in de bovenste laag). Hierbij gaat het zowel om intuïtieve (niet-geanalyseerde) kennis als om theoretische (geanalyseerde) kennis: van links naar rechts gaat het om het mentale lexicon zelf, de L-KRING-theorie van de Nederlandse woordstructuur (die in interactie met de MGBN moet ontstaan) en de MHB-theorie van de Nederlandse morfologie. Elk van deze kennismodules heeft invloed op een bepaald aspect van de MGBN (zoals wordt aangegeven door de verticale pijlen). Omgekeerd kan de MGBN ook weer invloed uitoefenen op de inhoud van het mentale lexicon (aangezien de morfologische analysetaak tot nieuwe kennis leidt) of de twee morfologische theorieën; dit is aangegeven door een draaiende pijl. Ik zal deze interacties nu wat nader toelichten.

De eerste bewerkingsstap correspondeert met een segmentatieproces waarin alle basislexemen uit de MGBN semi-automatisch van morfologische structuur worden voorzien op basis van intuïtieve structuuroordelen (die op het mentale lexicon van de redacteur berusten). Na (partiële) voltooiing van de constructiefase kan de aangebrachte structuur systematisch worden geanalyseerd, wat een groot aantal morfologische analyserapporten oplevert. Deze informatie vormt (na evaluatie, al dan niet op basis van MHB-kennis) de basis voor de opbouw van een L-KRING-theorie van de Nederlandse woordbouw. Na deze analysefase volgt een evaluatiefase; hierin wordt nagegaan in hoeverre de MGBN-informatie aan nader te bepalen structuurcriteria voldoet, zoals de morfologische kennis in het Morfologisch Handboek van het Nederlands (MHB).[5] Dergelijke evaluaties leveren een morfologisch evaluatierapport op. Indien het evaluatierapport uitwijst dat er fouten en inconsistenties in de MGBN-analyses voorkomen, kan dit reden zijn om de MGBN aan te passen. Er is dan sprake van terugwaartse invloed van het evaluatierapport op de inhoud van de MGBN.

[5] Er zijn overigens ook andere (aanvullende) evaluatiemethodes denkbaar, bijvoorbeeld op basis van een lexicon met regelgebaseerde parseringen (zoals CELEX) of een corpusgebaseerde inventarisatie van neologismen.



Figuur 1-1: Fasediagram bij de MGBN; dit diagram toont de opeenvolgende informatiefasen bij de ontwikkeling van de MGBN en de interactie met de cognitieve kennisbronnen.

Doordat de MGBN langs inductieve weg van morfologische structuur wordt voorzien, hoeft de analyse niet te worden beperkt tot woorden met een compositioneel afleidbare betekenis, maar kunnen ook woorden met onregelmatige (distributief of etymologisch gemotiveerde) structuurkenmerken worden meegenomen, zoals woorden waarvan de stam allomorfie vertoont of woorden waarvan de stam slechts éénmaal voorkomt. Hierdoor bezitten de morfologische representaties in de MGBN een aanzienlijk grotere detailleringsgraad dan mogelijk is bij de toepassing van een regelgebaseerde parser (c.q. automatisch ontleedprogramma). Op dit punt is de MGBN dan ook completer dan CELEX (Baayen, Piepenbrock and Gulikers, 1995), want de morfologische representaties in het CELEX-lexicon berusten in beginsel op automatische (regelgebaseerde) structuurtoekenning, al is een deel van de representaties redactioneel gecontroleerd.[6] Op dit punt overtreft de MGBN ook de mogelijkheden van Word Manager (Domenig & Ten Hacken, 1992), want dit lexicografische ondersteuningsprogramma voor de toekenning van morfologische structuur is in essentie regelgebaseerd.

De unieke constructiemethode van de MGBN opent nieuwe mogelijkheden voor empirisch onderzoek naar de Nederlandse morfologie, en meer specifiek naar de combinatorische eigenschappen van Nederlandse affixen. Zo kan voor elk affix worden nagegaan in welke affixclusters het kan optreden en op welke stammen zo'n affix (of affixcluster) kan worden toegepast. Omgekeerd kan voor elke stam worden uitgezocht hoe het affixparadigma eruit ziet, dus welke affixen de stam allemaal kan selecteren, en welke stammen hetzelfde affixparadigma bezitten. Verder kan worden bekeken welke staminterne kenmerken de meeste invloed hebben op de selectie van een specifiek affix of affixparadigma. Zo vertonen inheemse stammen vaak ander selectiegedrag dan uitheemse. Maar er zijn nog vele andere factoren in het spel.

In deze studie worden de hier geschetste analysemogelijkheden concreet verkend. De inzichten die hieruit voortkomen zijn niet alleen van belang voor de morfologische theorievorming, maar kunnen ook bijdragen aan de verdere verfijning van de MGBN. Want de MGBN is geen statische inventarisatie van bestaande morfologische kennis, maar een dynamisch onderzoeksbestand waarmee langs inductieve weg morfologische kennis kan worden opgebouwd.

[6] Hetzelfde geldt voor de morfologische representaties in het Corpus Gesproken Nederlands (2004).

1.2 Lexicologische terminologie

1.2.1 Het empirische domein

De Grote Van Dale, d.w.z.: de 13e druk van Van Dale's *Groot Woordenboek der Nederlandse Taal* (of kortweg GWNT), definieert een *woordenboek* als volgt:

> woordenboek (het) = '*boek waarin woorden (met opgave van bep. grammaticale kenmerken) en de vaste verbindingen waarin ze gebruikt worden, met hun betekenis (in alfabetische volgorde) zijn opgenomen*'. (GWNT 13)

Deze omschrijving sluit goed aan bij de lexicografische praktijk, want de meeste Nederlandse woordenboeken voldoen eraan. Taalkundig gezien is deze omschrijving echter voor verbetering vatbaar, want volgens het GWNT-lemma *woord* correspondeert een woord met een 'kleinste geheel van spraakgeluiden dat op zichzelf een betekenis heeft en als zelfstandig taalelement gebruikt wordt'. Volgens deze definitie zouden alle teksteenheden die door spaties en/of leestekens van elkaar gescheiden zijn in aanmerking komen voor de toekenning van woordstatus. Toch is slechts een deel van deze woorden in de GWNT opgenomen. Dit komt niet zozeer doordat de GWNT woorden "gemist" heeft, maar omdat het gebruikelijk is dat een woordenboek de concreet aangetroffen woorden (zoals de orthografische woordvormen uit dit tekstfragment) systematisch tot een overkoepelend *trefwoord* (c.q. *lemma*) herleidt.[7] Zo'n trefwoord correspondeert met de citatievorm van een eenheid die sinds Matthews (1974) bekend staat als *lexeem*[8] en die de basis vormt voor regelmatig afleidbare inflectievormen (gegeven het standaardsjabloon voor werkwoordsvervoeging). Zo correspondeert het trefwoord *leven* met de citatievorm van de werkwoordstam die de basis vormt voor de inflectievormen *leef*, *leefde*, *leefden*, *leeft*, *leve*, *leven*, *levend* en *geleefd*. De klankvorm *leven* komt overigens ook voor als citatievorm van een naamwoordstam met enkelvoudsvorm *leven* en meervoudsvorm *levens*; ook de andere inflectievormen kunnen meestal meerdere functies vervullen. Ook daarom is het nuttig om onderscheid te maken tussen woordvorm en lexeem.

In deze studie definieer ik lexemen als een lexicale relatie tussen een arbitraire naam en een reeks lexicale kenmerken, waaronder een betekenis, een grammaticale categorie en een of meer klankvormen. Hierbij hanteer ik de conventie om lexemen met de kortste inflectievorm aan te duiden (met de extra conditie dat deze vorm goed moet aansluiten bij de gangbare citatievorm); het werkwoord *leven* krijgt dus de lexeemvorm LEEF.[9] Wegens de voorspelbaarheid van de aan een lexeem (c.q. trefwoord) verbonden inflectievormen beperken woordenboeken zich doorgaans tot de beschijving van de niet-voorspelbare kenmerken, zoals de hoofdbetekenissen, de syntactische eigenschappen, de syllabestructuur en de uitspraak. Maar ten aanzien van het inflectiegedrag wordt meestal volstaan met een samenvatting, zoals de (voorspelbare) informatie '(leefde, h. geleefd)' bij het werkwoord *leven* of de (onvoorspelbare) informatie '(sprak, h. gesproken)' bij het werkwoord *spreken*.

---

[7] De GWNT definieert de term *trefwoord* als een 'woord waardoor de stof van een geschrift wordt aangeduid en dat als titel dient om ernaar te verwijzen of om het in een catalogus te kunnen vinden' en de term lemma als 'titelwoord in een woordenboek of encyclopedie, hoofd van een artikel'.

[8] Zie hoofdstuk 3 voor nadere uitleg. Deze betekenis van de term *lexeem* wordt overigens niet in de GWNT vermeld; de GWNT geeft alleen de lexicologische definitie, die teruggaat op Lyons (1977): 'ben. voor minimale betekeniseenheid (van morfeem tot idioom): *de woorden komen, kwam, gekomen, komst worden opgevat als vier verschijningsvormen van het lexeem* KOM'.

[9] In mijn visie is een lexeem niet meer dan een index voor het bijeenhouden van verwante vormen. Daarom is de vorm van dit lexeem arbitrair. In de generatieve morfologie wordt echter aangenomen dat de inflectiestam met de "onderliggende" vorm van de inflectievormen correspondeert, en dat deze inflectievormen er op voorspelbare wijze van afgeleid kunnen worden, althans op het niveau van de klankvorm.

Doordat de GWNT alleen lexemen behandelt en geen inflectievormen, wordt het aantal benodigde lemma's sterk[10] gereduceerd. Maar het nadeel is dat slechts een deel van de bestaande woordvormen in het woordenboek is terug te vinden: zo geeft de GWNT geen (rechtstreekse) informatie over de aan deze sectie ontleende woordvormen *lexicografische*, *woordenboeken, correspondeert* en *aangeeft*, terwijl de woordvorm *sluit* wel vermeld wordt, maar alleen als (Surinaams) naamwoord met de betekenis 'zuinig' (want de betekenis van de werkwoordsvorm *sluit* wordt onder het trefwoord *sluiten* behandeld). In deze sectie komen ook woorden voor waarvoor nog geen trefwoord beschikbaar is, bijvoorbeeld de samenstellingen *woordgebaseerd*, *GWNT-lemma* en *taalelementen* (wat opmerkelijk is, aangezien dit laatste woord onderdeel uitmaakt van de hierboven aangehaalde GWNT-definitie van *woordenboek*).

Het ontbreken van deze woorden is niet ernstig, want het gaat om weinig voorkomende woorden met een sterk vaktaalkarakter. Bovendien bestaan al deze woorden uit kleinere delen waarvoor het woordenboek wel een trefwoord geeft (afgezien van de afkorting GWNT), namelijk *lemma*, *woord*, *gebaseerd* (via het trefwoord *baseren*), *taal* en *elementen* (via het trefwoord *element*), zodat de vorm- en betekeniskenmerken van de samenstellingen rechtstreeks valt af te leiden uit die van de samenstellende delen. Daarom acht de GWNT-redactie het voldoende om per woord een kleine reeks voorbeelden te geven van samenstellingen met een linker- of rechterdeel. Zo leest men bij het lexeem *woordenboek* dat dit woord ook voorkomt in samenstelling met linkerdelen als *beeld-*, *hand-*, *doorsnee-*, *valentie-*, *uitspraak-*, *zaak-* en nog twintig andere voorbeelden.[11] Door slechts een deel van de samenstellingen te behandelen, wordt (wederom) een aanzienlijke ruimtewinst geboekt. Het zou ook onbegonnen werk zijn om compleetheid na te streven, want er komen elke dag tal van nieuwe woorden bij.

De hier besproken observaties laten zien dat het beschrijvingsdomein van een woordenboek minder eenvoudig valt af te bakenen dan men op het eerste gezicht zou denken. Want de trefwoordenlijst van een woordenboek is het resultaat van een ingewikkeld proces van woordselectie en structuuranalyse, met als doel om de miljoenen woordvormen die in omloop zijn, terug te brengen tot een alfabetisch gesorteerde lijst van relevante trefwoorden. Hierbij wordt impliciet aangenomen dat het menselijke taalsysteem in staat is om woorden van morfologische structuur te voorzien, d.w.z. om woorden onder te verdelen in eenheden die langs compositionele weg bijdragen aan de woordkenmerken. Zo berust de mogelijkheid om de woordvormen *leven*, *leefde* en *geleefd* aan hetzelfde trefwoord te relateren op het feit dat deze woordvormen ook cognitief gezien een gemeenschappelijke lexeemstam bezitten, te weten $[\text{LEEF}]_{\text{V}}$ (V = verbum c.q. werkwoord). Op soortgelijke wijze kan men een morfologisch verband aanbrengen tussen de inflectievormen *leven* en *levens* van de lexeemstam $[\text{LEVEN}]_{\text{N}}$ (N is nomen c.q. naamwoord) of tussen de inflectievormen *levend* en *levende* van de lexeemstam $[\text{LEVEND}]_{\text{A}}$ (A = adjectief c.q. bepalingswoord).

In het taalkundige onderzoek naar de woordvorming worden van oudsher twee vormen van affixatie onderscheiden, namelijk inflectie en derivatie.[12] Dit onderscheid wordt gemotiveerd door de overweging dat inflectie betrekking heeft op voorspelbare, paradigmatisch beregelde lexeemtoepassingen (waardoor het woordenboek een groot aantal woordvormen kan weglaten zonder dat dit ten koste gaat van het informatieve gehalte), terwijl derivatie betrekking heeft op niet-voorspelbare, syntagmatisch beregelde lexeemtoepassingen (zodat woordenboeken op dit punt een zekere compleetheid nastreven). Van inflectie heb ik al enkele voorbeelden gegeven. Hiertegenover zijn lexemen als $[\text{LEVEN}]_{\text{N}}$, $[\text{LEVEND}]_{\text{A}}$ en $[\text{LEEFBAAR}]_{\text{A}}$ alledrie als een morfologische derivatie van de V-stam $[\text{LEEF}]_{\text{V}}$ te analyseren, namelijk als een stam-

[10] Uit de lexicografische gegevens van VDL kan worden opgemaakt dat dit zeker een factor 4 scheelt. En als men de inflectievormen als vorm-betekenis-eenheden definieert komt de reductiefactor nog veel hoger uit.
[11] In de toekomst zal de GWNT overigens ook informatie verstrekken over vaste woordgroepen.
[12] Hiernaast is er ook een woordvormingsproces dat zich kenmerkt door samenstelling van bestaande woorden.

suffix-combinatie: $[\text{LEEF}]_V$+EN_N, $[\text{LEEF}]_V$+END_A en $[\text{LEEF}]_V$+BAAR_A. Het suffixparadigma van $[\text{LEEF}]_V$ kent overigens een opvallende lacune, want er is geen derivatie met het suffix -ER beschikbaar. Hoewel het woord *lever* wel bestaat, wordt het uitsluitend als orgaannaam gebruikt. Voor een woordenboek is dit nuttige informatie. Het A-lexeem $[\text{LEEFBAAR}]_A$ kan zelf weer als basis dienen voor verdere derivaties, bijvoorbeeld *leefbaarder* en *leefbaarst*.[13]

De V-stam $[\text{LEEF}]_V$ leent zich ook voor derivaties met een prefix of partikel, bijv. BE+LEEF resp. OP+LEEF; hierbij heeft de resulterende stam zelf ook weer werkwoordstatus. Zoals ik later in deze studie zal toelichten,[14] beschouw ik deze V-lexemen als een V-derivatie van een gemeenschappelijke, categorieloze wortel $[\text{LEEF}]_0$ en analyseer ik het prefixloze V-lexeem $[\text{LEEF}]_V$ eveneens als een morfologisch complex lexeem, namelijk als $[0/\text{GE}]$+$[\text{LEEF}]_0$; hierbij is $[0/\text{GE}]$ een operator die in de meeste inflectievormen onzichtbaar is (*leeft*, *leefde*), maar die in de voltooide tijd de vorm *ge-* aanneemt (*geleefd*). Elk van deze V-stammen kan verschillende vervolgderivaties ondergaan (zoals de reeds genoemde suffixderivaties); hieronder bevindt zich ook een operatie waarmee een V-stam in een V-lexeem (c.q. inflectiestam) kan worden omgezet. Dit wordt geïllustreerd door tabel 1-1.

M0-stam	M1-stam	M2-stam	M3-stam / inflectie
$[\text{LEEF}]_0$	$[0/\text{GE}]$ + $[\text{LEEF}]_0$	M1-STAM + $V	M2-STAM + V-INFLECTIE
		M1-STAM + EN_N	M2-STAM + $N
		M1-STAM + END_A	M2-STAM + $A
		M1-STAM + BAAR_A	M2-STAM + $A
	[BE] + $[\text{LEEF}]_0$	M1-STAM + $V	M2-STAM + V-INFLECTIE
		M1-STAM + ING_N	M2-STAM + $N
		M1-STAM + ENIS_N	M2-STAM + $N
	[OP] + $[\text{LEEF}]_0$	M1-STAM + $V	M2-STAM + V-INFLECTIE
		M1-STAM + ING_N	M2-STAM + $N

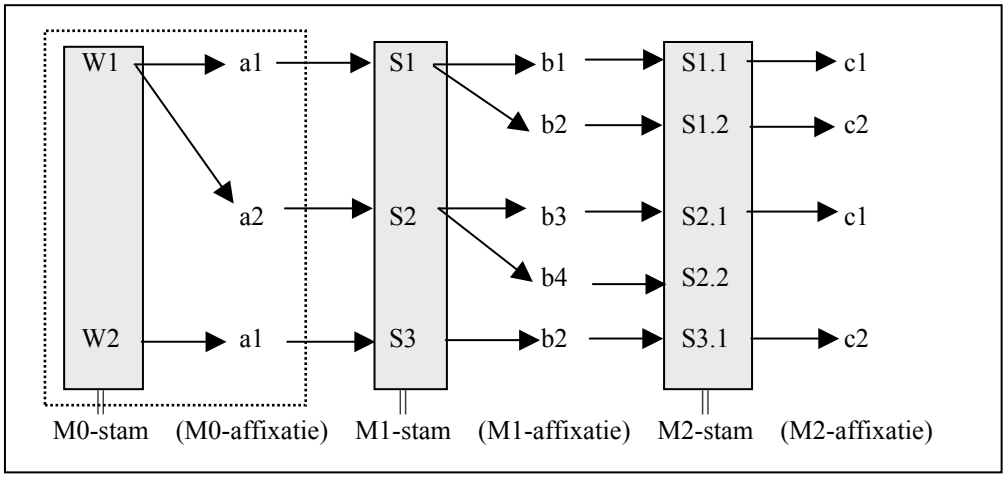
Tabel 1-1: Het derivatieparadigma van de M0-stam $[\text{LEEF}]_0$.

Deze tabel is als volgt opgebouwd. De eerste kolom specificeert de morfologische basisstam. De tweede kolom laat zien hoe men deze basisstam (M0) in een morfologisch complexe M1-stam kan omzetten door er een prefix aan toe te voegen. De derde kolom laat zien hoe men hier een nog complexere M2-stam van kan maken door een volgende operator toe te voegen, hetzij een $-markering (voor de aanmaak van een inflectiestam), hetzij een suffix waarmee een nieuw betekeniskenmerk wordt toegevoegd. In de vierde kolom wordt onder meer de inflectie van de $V-stammen gespecificeerd.

Indien men dergelijke derivatierelaties systematisch in kaart brengt, ontstaat een lexicaal netwerk van morfologisch gestructureerde M-stammen. In het onderstaande schema wordt op abstracte wijze weergegeven hoe zo'n netwerk eruit ziet. Dit netwerk bestaat uit een verzameling M-stammen (te weten M0, M1 en M2-stammen) die door affixaanhechting (c.q. M-affixatie) van elkaar worden afgeleid (dit wordt door pijlen gemarkeerd; hierbij corresponderen de a, b en c-variabelen met affixen). Het schema laat ook zien dat er een verschil is tussen de constructie van een M1-stam (die op een wortel W is gebaseerd) en de constructie van hogere M-stammen (die van een S-stam uitgaan). Het idee hierachter is dat W-stammen niet zelfstandig bruikbaar zijn en dat de overgang van W-stam naar M1-stam (bijvoorbeeld van STRU naar CONSTRU) deels onvoorspelbare lexeemkenmerken oplevert (in tegenstelling tot de overgang van de M1-stam CONSTRU naar de M2-stam CONSTRUEER). Daarom dient het lexicon de M1-stammen op een andere manier te verantwoorden dan de complexere M-

[13] Elk van deze A-stammen kent weer twee inflectievormen, namelijk de stamvorm met -e of zonder -e.
[14] Zie H3.4.2, H3.4.6 en H4.3.

stammen. Zo kan de M1-stam CONSTRU ook de vorm CONSTRUCT aannemen, die weer de



basis vormt voor M1-stammen als CONSTRUCTIE en CONSTRUCTOR.

Figuur 1-2: Abstracte weergave van een morfologisch gestructureerd lexicon.

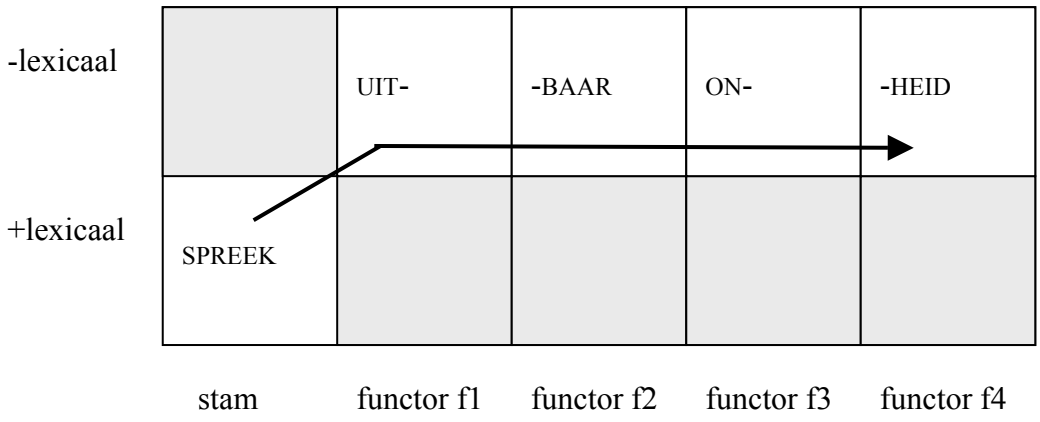
In het morfologische onderzoek bestaan twee verschillende analyseperspectieven, namelijk het syntagmatische (affixgebaseerde) perspectief, waarbij men de morfologische derivatiemogelijkheden als onafhankelijke processen probeert te beschrijven, en het paradigmatische (stamgebaseerde) perspectief, waarbij deze derivatiemogelijkheden juist in onderlinge samenhang (namelijk vanuit de stam) worden bekeken. In termen van figuur 1-2 betekent dit dat het syntagmatische perspectief met de horizontale structuurdimensie correspondeert, terwijl het paradigmatische perspectief zich op de verticale structuurdimensie richt. In de L-KRING-theorie zijn beide structuurdimensies even belangrijk.

1.2.2 Deductie versus inductie

Er bestaan zeer uiteenlopende theorieën over de rol van morfologische structuur bij de cognitieve representatie van woorden. Hierbij kunnen twee hoofdvisies worden onderscheiden, namelijk de deductieve (regelgebaseerde) visie, die als belangrijkste doel heeft om regels op te stellen waarmee de potentiële lexemen van een taal kunnen worden voorspeld,[15] en de inductieve (patroongebaseerde) visie, die als doel heeft om uit te leggen hoe men op basis van een compleet lexicon inzicht kan krijgen in de onderliggende woordvormingspatronen. Hieronder zal ik nader ingaan op deze twee visies, om vervolgens uit te leggen waarom ik de MGBN op een inductief lexiconmodel heb gebaseerd.

In de deductieve visie wordt aangenomen dat het lexicon uitsluitend basislexemen opslaat en dat de grammatica bepaalt welke derivaties deze morfemen kunnen ondergaan (in het bijzonder welke affixatiemogelijkheden er zijn). Hierbij worden regelmatig gevormde woorden niet integraal opgeslagen, maar steeds opnieuw van hun stammorfeem afgeleid. Ik zal dit toelichten aan de hand van het in figuur 1-3 afgebeelde processchema voor de opbouw van het lexeem *onuitspreekbaarheid*.

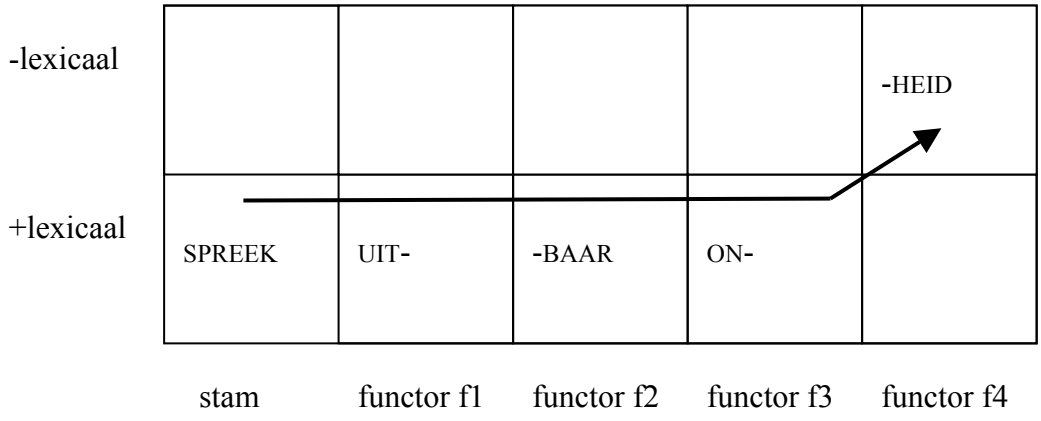
[15] In mijn optiek is het overigens niet mogelijk om voorspellingen te doen over de "grammaticaliteit" van potentiële woorden; men kan alleen berekenen hoe waarschijnlijk zulke nieuwvormingen zijn.



Figuur 1-3: Processchema voor de "deductieve" opbouw van het lexeem onuitspreekbaarheid.

Het schema laat zien hoe dit lexeem volgens de gangbare morfologische opvattingen van het stamlexeem SPREEK kan worden afgeleid. In de eerste stap wordt deze stam uit het lexicon gehaald, waarna het achtereenvolgens vier verschillende affixatiestappen ondergaat, te weten affixatie met het prefix UIT-, het suffix -BAAR, het prefix ON- en het suffix -HEID. In tegenstelling tot het stamlexeem corresponderen de door affixatie gevormde (tussen)producten per definitie met niet-lexicale lexemen.

De inductieve visie gaat er echter van uit dat het lexicon alle lexemen opslaat die men in het dagelijkse taalgebruik is tegengekomen. Deze kennis vormt een permanente basis voor de identificatie van morfologische combinatiepatronen c.q. redundantieregels. Deze redundantieregels worden meestal lexiconextern verantwoord en hebben als voornaamste functie om de herkenning en verwerving van binnenkomende lexemen te ondersteunen. Ik zal dit toelichten aan de hand van het processchema in figuur 1-4.



Figuur 1-4: Processchema voor de "inductieve" opbouw van het lexeem onuitspreekbaarheid.

In dit schema worden dezelfde affixatiestappen doorlopen, maar de inductieve analyse verschilt van de deductieve analyse doordat er geen sprake is van derivatieregels (die nieuwe lexemen genereren) maar van redundantieregels (die over lexicaal opgeslagen lexemen generaliseren). Hierdoor kunnen de meeste affixatieproducten integraal (d.w.z. inclusief alle lexicale kenmerken, zoals vormvarianten, betekenisvarianten, frequentiegegevens en selectiekenmerken) uit het lexicon worden opgehaald. In het hier uitgewerkte voorbeeld resulteert alleen de laatste affixatiestap in een niet-lexicaal lexeem; maar zodra dit lexeem gevormd is, kan het eveneens lexicale status krijgen.

In mijn optiek is de inductiebenadering zowel vanuit psychologisch als vanuit lexicografisch perspectief aantrekkelijker dan de deductiebenadering. Want terwijl de deductiebenadering

slechts een fractie van de parate woordkennis kan verantwoorden, probeert de inductiebenadering alle kennisdimensies te verantwoorden. Het lexicon van de inductiebenadering hoeft zich namelijk niet te beperken tot de verantwoording van basislexemen, maar kan zowel basislexemen als complexe lexemen opslaan (al dan niet in gecomprimeerde vorm). Hierdoor kan voor alle lexemen (zowel basislexemen als complexe lexemen) aanvullende informatie over betekenissen, vormvarianten en gebruiksfrequentie worden vastgelegd. Verder biedt de inductiebenadering een psychologisch plausibele oplossing voor de vraag hoe men binnenkomende data van structuur kan voorzien en hoe men de resulterende structuuranalyses kan benutten voor de productie en interpretatie van nieuwe woorden. Daarom vormt de inductiebenadering een aantrekkelijk uitgangspunt voor de morfologische analyse van de kennis die ten grondslag ligt aan de lemma's in een woordenboek.

1.3 Lexicografische achtergrond

1.3.1 Introductie

Deze sectie gaat nader in op de omstandigheden die het mogelijk hebben gemaakt om een morfologische gegevensbank te realiseren. Zoals al aan de orde kwam, is het in deze studie beschreven onderzoek voortgekomen uit de doelstelling om een bijdrage te leveren aan de systematisering van de woordkenmerken in de lexicale kennisbank van Van Dale Lexicografie (VDL). Voor dit doel heb ik intensief gebruik gemaakt van de bij VDL aanwezige infrastructuur voor de grootschalige ("industriële") bewerking van lexicografische gegevensbestanden. Daarom acht ik het nuttig om kort in te gaan op een aantal recente ontwikkelingen bij VDL. Eerst wordt wat verteld over de geschiedenis van het woordenboek, in het bijzonder die van de Grote Van Dale (in H1.3.2). Hierna wordt een beeld gegeven van de modernisering van het productieproces bij VDL (in H1.3.3). Tot slot wordt uitgelegd waarom VDL behoefte heeft aan een morfologische gegevensbank; hierbij komen zowel lexicografische als linguïstische toepassingen aan de orde (H1.3.4).

1.3.2 Woordenboeken in het pre-computer-tijdperk

De toonaangevende woordenboeken in het Europese taalgebied zijn veelal in de achttiende en negentiende eeuw ontstaan. Dit is niet toevallig, want als gevolg van de Verlichting ontstond toenemende behoefte aan systematisch vastgelegde kennis, wat zich in de ontwikkeling van verklarende woordenboeken en encyclopedieën vertaalde.[16] Deze naslagwerken waren niet alleen een voortvloeisel van het rationele wereldbeeld van die tijd, maar ze gaven ook uitdrukking aan de eigen culturele identiteit. En hoe kon die meer recht worden gedaan dan door de eigen taal zo gedetailleerd mogelijk in een woordenboek vast te leggen?

Dat er sindsdien zo weinig nieuwe uitgevers zijn bijgekomen, hangt samen met de grote investeringen die nodig zijn voor de ontwikkeling van een verklarend woordenboek. Daar komt bij dat een woordenboek meer gezag krijgt naarmate het een langere staat van dienst heeft. Het bijwerken van een bestaand woordenboek (door het schrappen van oude woorden en het toevoegen van nieuwe woorden) was daarom rendabeler dan het uitbrengen van een geheel nieuwe titel. Dit blijkt ook uit het feit dat de eerste edities van de door Van Dale en Koenen geredigeerde woordenboeken (beide uit 1872)[17] op een Frans voorbeeld waren geënt.

[16] De eerste naslagwerken zijn echter van veel ouder datum: zo publiceerde Kiliaan al in 1574 een woordenboek van de (Brabantse) volkstaal (het *Dictionarium teutonico-latinum*, beter bekend als *Etymologicum teutonicae linguae*), terwijl Van Maerlant in de 14e eeuw een natuurencyclopedie samenstelde (*Der Naturen Bloeme*).

[17] Dit was de tweede editie van het woordenboek dat nu bekend staat als de Grote Van Dale; de eerste editie (onder redactie van Calisch & Calisch) verscheen in 1864. Het is publiek beschikbaar via de website van de DBNL: http://www.dbnl.org/tekst/cali003nieu01/, en staat ook op de CD-ROM-versie van de GWNT (2005).

In deze begintijd hadden woordenboeken primair een didactische functie (Posthumus, 1997). Het hoofddoel was om informatie te geven over de spelling en de grammaticale kenmerken van veel voorkomende woorden. De betekenisspecificaties dienden voornamelijk ter desambiguering, en waren daarom zeer beknopt. Geleidelijk aan kregen de woordenboeken echter ook een verklarende functie, waardoor de betekenisomschrijvingen steeds uitvoeriger werden. Dit ging samen met een enorme toename van de omvang. Tegelijk met deze ontwikkeling maakte de didactische oriëntatie plaats voor een algemeen informatieve functie.

Tot de opkomst van de computer is er weinig veranderd aan het productieproces dat ten grondslag ligt aan het papieren woordenboek. Tot diep in de twintigste eeuw werd namelijk gebruik gemaakt van kaartenbakken met alfabetisch gesorteerde systeemkaarten (*fiches*). Voor elk trefwoord (d.w.z. lexeem met één of meer verwante betekenissen) werd een apart fiche aangemaakt, waarop behalve het trefwoord ook informatie over etymologie, grammaticale woordkenmerken, vaste verbindingen en betekenisdefinities werden genoteerd. Elke keer als er een nieuwe betekenis werd ontdekt, moest deze aan dit fiche worden toegevoegd.

Zo'n kaartenbaksysteem omvatte al gauw meer dan 100.000 trefwoorden, waardoor hersorteren praktisch ondoenlijk was. Als gevolg van deze beperking was het onmogelijk om een woordenboek systematisch op (semantische) consistentie te controleren, en hetzelfde gold voor de compleetheid op vormniveau. De kwaliteit van een woordenboek (in elk geval wat betreft de hier genoemde aspecten) was dan ook grotendeels afhankelijk van het geheugen van de redacteuren; niet voor niets danken veel woordenboeken hun bestaan aan de noeste arbeid van slechts één persoon (namelijk diens levenswerk). Deze werkwijze had als gevolg dat er steeds meer fouten, inconstenties en omissies in de woordenboeken slopen.

Ook de hedendaagse woordenboeken zijn voor verbetering vatbaar. Dit blijkt bijvoorbeeld uit een lexicografische studie van Verkuyl (1993a), die de twaalfde druk van de Grote Van Dale (uit 1992) heeft beoordeeld op basis van een aantal lexicografische kwaliteitscriteria (namelijk de C-criteria; zie ook H1.3.3), te weten consistentie, compleetheid, correctheid, courantheid en citatie. Dit onderzoek berust op een steekproef van 12 semantische domeinen, zoals het schaakdomein, het wiskundedomein, het rechtsdomein etc.

In deze steekproef bleek geen enkel domein aan Verkuyls kwaliteitscriteria te voldoen. In de onderzochte domeinen deed de Grote van Dale het bovendien slechter dan vergelijkbare woordenboeken uit het buitenland (zoals de Oxford Dictionary of English en Larousse), al lieten ook deze de nodige steken vallen. Ten tijde van het onderzoek berustte de inhoud van deze woordenboeken nog grotendeels op het oude kaartenbaksysteem, zodat men kan concluderen dat het zonder computationele hulpmiddelen blijkbaar onmogelijk is om aan de (strenge) kwaliteitscriteria van Verkuyl te voldoen. Ondanks de gebleken tekortkomingen blijken taalgebruikers een vanzelfsprekend vertrouwen te stellen in de autoriteit van de traditionele woordenboeken.

1.3.3 Van kaartenbak naar elektronisch informatiesysteem

De opkomst van de computer heeft grote gevolgen gehad voor het productieproces van woordenboekuitgevers zoals VDL. Sinds de elfde druk van de Grote Van Dale (1984) wordt hier steeds meer gebruik gemaakt van computerondersteuning, waardoor een vergaande automatisering van taken mogelijk is geworden. Tegelijkertijd worden steeds hogere eisen gesteld aan de kwaliteit van de lexicografische informatie, wat tot uitdrukking komt in een toenemende belangstelling voor taalkundige analysemethodes.

De modernisering van het productieproces begon met de overgang naar elektronisch zetwerk. Vervolgens werd eind jaren 80 een elektronisch woordenboek ontwikkeld, namelijk Lexitron (1988). Ondanks de hardwarebeperkingen van de toenmalige computers kende het bijzonder

krachtige zoekfuncties (zoals de combinatie van lexicografische en encyclopedische informatie; zoeken op klankvorm; geavanceerd zoeken ten behoeve van onderzoek etc.).[18]

De volgende stap was om een geheel nieuwe woordenboekreeks uit te geven die was gebaseerd op het uitgangspunt dat alle vertaalwoordenboeken hetzelfde basisbestand gebruiken, namelijk de woordenlijst die ten grondslag ligt aan VDL's Groot Woordenboek Hedendaags Nederlands (de WHN); verder moesten alle lemma's dezelfde structuur krijgen, zodat voor elk type kenmerk een apart veld beschikbaar was, en een strikte scheiding mogelijk werd van betekenisdefinitie en voorbeelden. De WHN is tot stand gekomen door een courante selectie te maken uit de woorden in de Grote Van Dale (die ook veel archaïsche, regionale en vakspecifieke woorden omvat). Daarnaast werd een begin gemaakt met de systematische inventarisatie van semantische relaties als synonymie en hyponymie. Hiermee was het fundament gelegd voor de opbouw van een meertalige, semantisch gestructureerde gegevensbank. Wat later werd ook een begin gemaakt met de systematische inventarisatie van vormkenmerken, zoals het coderen van samenstellingsgrenzen en afbreekposities, uitspraakrepresentaties en informatie over de verbuiging van werkwoorden en naamwoorden. In de jaren negentig werden deze werkzaamheden in toenemende mate geautomatiseerd.

De praktische inzet van taaltechnologie is van de grond gekomen toen de spellingswet van 1995 werd voorbereid. De invoering van deze wet dwong VDL om in korte tijd al haar woordenboeken om te spellen. Vooral de systematische invoering van de tussen-*n* vormde een probleem. Op zich beïnvloedde de nieuwe spellingsregel een betrekkelijk klein aantal woorden. Maar het opsporen van de aan te passen woorden was daardoor vergelijkbaar met het zoeken naar een speld in een hooiberg; bovendien zouden redacteuren makkelijk fouten kunnen maken doordat ze nog het "oude" spellingbeeld in hun hoofd hadden. Daarom werd gekozen voor een aanpak die intensief gebruik maakt van computationele analysetechnieken, onder meer voor de automatische identificatie van samenstellingsgrenzen. De hiertoe ontwikkelde computerprogramma's waren in staat om een groot deel van de woorden automatisch om te spellen, terwijl potentiële "omspellers" netjes apart werden gehouden. Deze omspellers werden bovendien in subklassen onderverdeeld die in verschillende "bakjes" werden gestopt. Pas daarna hoefde de redactie te worden ingeschakeld voor een handmatige beoordeling van de probleemgevallen. Hierna werd de resulterende lijst nog eens integraal gecontroleerd.

Het succes van deze computationele aanpak (namelijk een hoge kwaliteit in combinatie met een flinke werkbesparing) leidde tot het inzicht dat het nuttig was om de kenmerken van samengestelde woorden op te slaan op het niveau van de woorddelen; zo bestaat het woord *rekenmachine* uit de woorddelen *reken* en *machine*, die allebei in tal van andere samenstellingen voorkomen, maar vaak constante eigenschappen bezitten. Door de hele gegevensbank op deze manier te herstructureren werd het onderhoud vergemakkelijkt, want het werd eenvoudiger om veranderingen door te voeren (zoals een spellingsaanpassing of een nieuwe verbuigingsvorm), terwijl de woorden systematischer van kenmerken konden worden voorzien. De nieuwe structuurlaag kwam ook ten goede aan de compleetheid, consistentie en correctheid van de uitgebrachte woordenboeken. Dit blijkt bijvoorbeeld bij inspectie van het lemma *machine* in de dertiende druk van de Grote Van Dale: hier worden tal van woorddeeltoepassingen beschreven, met hele reeksen voorbeelden.

[18] Lexitron was zijn tijd te ver vooruit; commercieel werd het geen succes.

machine (de (v.); -tje, machientje)
[1693 Fr. ‹Lat. machina ((belegerings)werktuig, toneelmachine) ‹Gr. mèchanè (kunstvaardigheid, middel, werktuig)]
1) ieder uit delen bestaand toestel dat zekere werking of functie kan verrichten
2) complex werktuig waarmee handelingen verricht en voorwerpen vervaardigd worden, in de plaats komend voor het werk van de hand
2a) ook als eerste lid in samengestelde ww. ter aanduiding dat de in het tweede lid genoemde handeling machinaal verricht wordt; antoniem: hand-
machinegieten, machinenaaien, machineschrijven, machineweven, machinezetten
2b) ook als tweede lid in samenst. als de volgende, waarin het eerste lid een (machinaal te verrichten) handeling noemt
afkortzaagmachine, afreimachine, afweegmachine, bakkerijmachine, banderol-leermachine, betonstortmachine, borduurmachine, borstelmachine, bottel-machine, broodzaagmachine, capsuleermachine, clicheermachine, draadbuig-machine, draadtrekmachine, dresseermachine, ensileermachine, etiketteer-machine, filtreermachine, fineermachine, flensmachine, flotatiemachine, fotokopieermachine, gaufreermachine, geldtelmachine, glassmeltmachine, gommeermachine, graveermachine, harkmachine, hekelmachine, hoonmachine, houtbewerkingsmachine, houtschaafmachine, inkeepmachine, inpakmachine, kabeltrekmachine, katoenspinmachine, klinkmachine, lepmachine, lichtdruk-machine, lijmmachine, linieermachine, maalmachine, naaimachine, ontkorrel-machine, ontromingsmachine, ontstapelmachine, opzakmachine, persmachine, plaatbuigmachine, pletmachine, plukmachine, pompmachine, precisiezaai-machine, puddelmachine, radeermachine, raffineermachine, rangeermachine, reinigingsmachine, rilmachine, rimpelmachine, rondslijpmachine, rondzet-machine, schaakmachine, schilmachine, schoffelmachine, schudmachine, slijpmachine, smeermachine, soldeermachine, spitmachine, splijtmachine, staafbuigmachine, stansmachine, steenschraapmachine, strijkmachine, stuiklasmachine, tabletteermachine, tempereermachine, textielbewerkings-machine, uienschilmachine, vergaarmachine, verticuteermachine, vijlenkap-machine, vijlmachine, vlakschaafmachine, vlakslijpmachine, vlasspinmachine, vlastrekmachine, vlechtmachine, volmachine, walkmachine, wasserijmachine, wegenbouwmachine, wiedmachine, zaagslijpmachine, zakkennaaimachine, zakkenvulmachine, zandgraafmachine, ziftmachine, zuiveringsmachine
2c) ook als tweede lid in samenst. als de volgende, waarin het eerste lid een product noemt:
espressomachine, gimpmachine, kauwgomballenmachine, parfummachine, pasta-machine, rookmachine
3) (in 't bijzonder) (als verkorting van) *stoommachine*
11a) als tweede lid in samenst. als de volgende, waarin het eerste lid een (machinaal verrichte) taak, handeling of verrichting noemt: *vechtmachine, wetgevingsmachine*

Doordat het woorddeel *machine* verschillende functies kent, kan er verwarring ontstaan over de vraag wat nu de bedoelde betekenis is. Bovendien kunnen samenstellingen een gelexicaliseerde betekenis aannemen, zoals *stoommachine* (die Van Dale omschrijft als "machine die met behulp van een zuiger en toevoer van stoom drijfkracht ontwikkelt", een betekenis die taalkundig gezien niet erg voor de hand ligt. Op basis van de samenstellende delen had het ook een behangafstoom- of strijkijzer kunnen zijn). Daarom hebben veel samenstellingen met het rechterdeel *machine* toch een eigen woordingang gekregen (maar deze staan broederlijk tussen de regelmatige samenstellingen). Inspectie van deze samenstellingen leert dat de grammaticale kenmerken uniform, doch beknopt beschreven worden, namelijk als "(de (v.))", maar dat er aanzienlijke variatie zit in de betekenisomschrijving. Zo wordt slechts in een deel van de gevallen van een *machine* gesproken; in de andere gevallen vindt men meestal de aanduiding *toestel*, maar ook andere typeringen zoals *computer*, *apparaat*, *werktuig* en *hulpmiddel* komen voor, evenals synoniemen (bijvoorbeeld *vloerwrijver* voor *boenmachine*). Op dit terrein is dus nog geen "betekenismachine" aan het werk geweest.

Na de toekenning van woorddeelstructuur werd ook een omgekeerde werkwijze mogelijk: door alle woorddelen van een (zoveel mogelijk automatisch gegenereerde) uitspraakrepresentatie te voorzien, kon de uitspraak van de meeste samenstellingen automatisch worden afgeleid; dit leidde tot aanzienlijke tijdwinst. Zo'n regelgestuurde aanpak is ook gevolgd bij het verbeteren van de afbreekrepresentaties (Nunn, 2000). Kenmerkend voor deze aanpak is dat eerst een aantal hypotheses worden geformuleerd over de relatie tussen de spelvorm en het te beregelen woordkenmerk, zoals afbreekpatroon of uitspraakrepresentatie; gegeven een voorbewerkt testbestand kan deze relatie worden achterhaald door *reverse engineering*, d.w.z. het reconstrueren van verbanden door vergelijking van de aanvangsvorm en de bewerkte vorm. De op deze wijze verkregen taalregels kunnen worden getest door ze op te nemen in een computerprogramma dat concrete woordvormen van een nieuw woordkenmerk voorziet. Door dit programma los te laten op het testbestand, kan worden onderzocht hoe de regels in de praktijk uitwerken, en kunnen de instellingen worden verfijnd.

Na optimalisering van de regels (waarbij sommige regels overbodig kunnen blijken) kan het programma op de volledige gegevensbank worden toegepast. Hierbij kunnen nieuwe regelmatigheden of onverwachte uitzonderingen aan het licht komen die bijstelling van de regels nodig maken. Na enig proberen ontstaat echter ook op dit niveau een optimaal programma, dat in staat is om veruit het grootste deel van de woorden automatisch te verwerken; het residu dient ten slotte door een redactie te worden gecontroleerd. Op deze manier ontstaat dubbele winst: en men verkrijgt perfecte data, en er is een perfect voorspellend programma (bestaande uit een verzameling inductief tot stand gekomen taalregels).

In deze sectie is aangetoond dat de overheveling van woordkenmerken naar het niveau van de samenstellende delen nieuwe analysemogelijkheden biedt die ten goede komen aan de lexicografische consistentie van de WKB-Ned. Maar deze structuurlaag is niet gedetailleerd genoeg om alle consistentieproblemen op te lossen. Voor dit doel dient nog dieper in de woordstructuur te worden afgedaald, namelijk naar het niveau van de morfemen. Dit zal in de volgende sectie worden toegelicht.

1.3.4 Het nut van een morfologische gegevensbank

Zoals de eigenschappen van een samenstelling doorgaans afhankelijk zijn van de samenstellende woorddelen (c.q. basislexemen), zo zijn de kenmerken van deze basislexemen weer afhankelijk van de samenstellende morfemen. Hieruit volgt dat morfemen voorspellende waarde kunnen hebben voor de lexicale kenmerken op lexeemniveau, dus dat ze een nuttige bijdrage kunnen leveren aan de systematisering van de woordkenmerken. Ik zal dit toelichten aan de hand van een voorbeeld.

Indien een woord is afgeleid met het nominaliserende suffix -ER is het vrij zeker dat het een zelfstandig naamwoord is waarvan de betekenis kan worden getypeerd als de persoon of het voorwerp dat de handeling verricht of ondergaat die door het bijbehorende werkwoord wordt uitgedrukt; verder valt te verwachten dat het betreffende woord samengaat met het lidwoord *de* (bijv. *de strijder*), dat het (indien de semantiek dit toelaat) een meervoud op -S kiest (bijv. *strijders*) en dat de vrouwelijke vorm kan worden afgeleid door het suffix -ER voor het suffix -STER in te ruilen (wat in dit geval het woord *strijdster* oplevert). Voor het suffix -AAR gelden in beginsel dezelfde eigenschappen, behalve dat de vrouwelijke vorm van -AAR met de affixreeks -AAR+STER (bijv. *spijbelaarster*, *babbelaarster*) of de affixreeks -AAR+ES correspondeert (*lerares*, *bedelares*).[19]

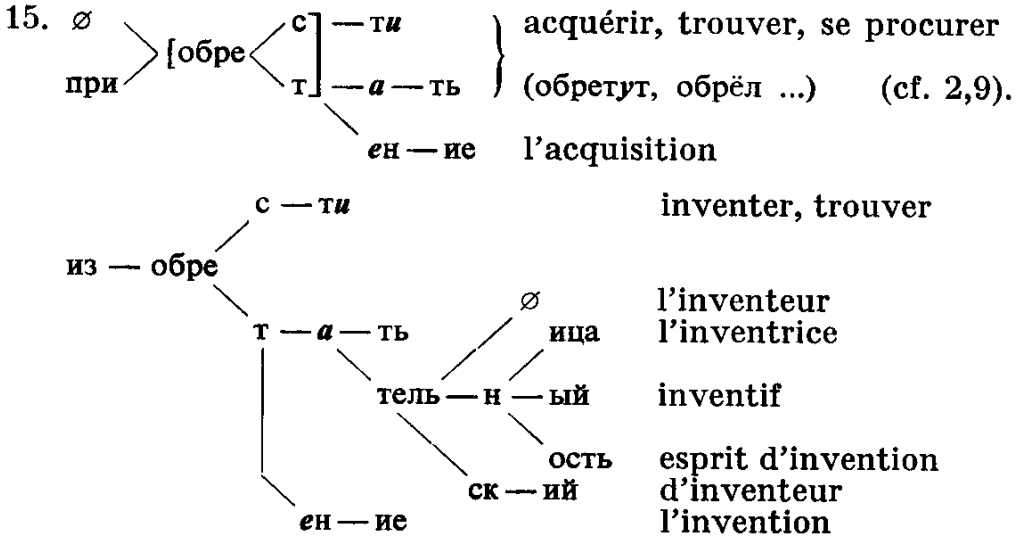
[19] Deze sequenties zijn alleen beschikbaar als het segment -AAR met een mannelijk persoonssuffix kan corresponderen; zo is het onjuist om het woord *palmares* ("erelijst") als PALM+AAR+ES te analyseren.

Lexicografisch gezien is het aantrekkelijk om dit soort eigenschappen rechtstreeks aan het bijbehorende suffix te koppelen. In dat geval hoeven de met -ER afgeleide woorden niet meer individueel te worden behandeld, maar kunnen ze automatisch worden afgeleid uit de eigenschappen van het stamwoord en het suffix -ER. Dit zou een toename van de consistentie kunnen opleveren. Bovendien kunnen dan woorden en inflectievormen worden verantwoord die wel mogelijk zijn, maar die nog niet in gebruik zijn of in elk geval niet in het woordenboek zijn opgenomen (gegeven het werkwoord *oplezen* kan bijvoorbeeld het naamwoord *oplezer* worden geconstrueerd, en vandaar de inflectievormen *oplezers*, *opleesster* en *oplezertje*).

Het aanbrengen van morfologische structuurinformatie biedt tal van redactionele voordelen. Hieronder volgt een puntsgewijs overzicht:

- nieuwe analysemogelijkheden
- systematisering van vorm- en betekeniskenmerken
- vereenvoudiging van het lexicale onderhoud
- voorspellen van nieuwe woordvormen
- automatische analyse van nieuwe woorden
- vergroting van zoekmogelijkheden

Bij de ontwikkeling van een morfologische structuurlaag dient onderscheid te worden gemaakt tussen de vormkenmerken, die doorgaans volstrekt regelmatig zijn, en de betekenis. Veel regelmatig gevormde woorden gaan in de loop van de tijd namelijk gelexicaliseerde (dus onvoorspelbare) betekenissen aannemen. Deze zullen daarom individueel geanalyseerd moeten worden. Maar door zoveel mogelijk uit te gaan van de regelmatige betekenisdefinities, kunnen de definities van woorden met een onregelmatige betekenis wel consistenter worden.



Figuur 1-5: Een Russisch paradigma uit 'Manuel de Russe' (Gentilhomme (1964), p. 580).

De informatie in de MGBN maakt ook vernieuwingen mogelijk met betrekking tot de ordeningswijze van woordenboeken. Zo zou men een woordenboek kunnen uitgeven waarin de woorden op stam zijn gesorteerd (zoals heel gebruikelijk is voor sterk paradigmatische talen als het Arabisch en het Hebreeuws). Bij het woord *zorg* zou de gebruiker dan afleidingen als *zorgelijk*, *bezorgd*, *zorgen*, *verzorgen*, *verzorger* en *ontzorging* moeten aantreffen (en wellicht het nog niet gangbare werkwoord *ontzorgen*). Het werkwoord *bezorgen* daarentegen zou een aparte ingang moeten krijgen, aangezien het geen semantische relatie met *zorg* ver-

toont.[20] Hierbij kan Gentilhomme's (1964) encyclopedie over de natuurwetenschappelijke terminologie van het Russisch als voorbeeld dienen. In deze encyclopedie wordt veel aandacht besteed aan de morfologische samenhang van Russische vaktermen; deze samenhang wordt zichtbaar gemaakt door woorden die tot hetzelfde paradigma behoren op structuralistische wijze te analyseren. Figuur 1-5 toont een voorbeeld van een op deze wijze geanalyseerd woordparadigma.

Tot slot biedt de MGBN een nieuwe basis voor empirisch onderzoek naar de Nederlandse woordbouw. Mogelijke onderzoeksthema's (met per thema een voorbeeldvraag):

- De afbakening van morfemen: hebben suffixen als -ERIJ en -ISEER interne structuur?
- De analyse van allomorfie: hoort -ION in FUNCT+ION+EEL bij de stam FUNCT of bij het suffix -EEL? of gaat *i* naar links en *on* naar rechts?
- Restricties op de clustering van affixen: wat kan er allemaal achter -EER komen, en wat achter -ISEER of -IVEER?
- Positierestricties op suffixen: welke suffixen staan altijd direct achter de stam; welke suffixen komen alleen op woordeinde voor?
- De interactie tussen prefigering en suffigering: werkwoorden met een inheems prefix kiezen relatief vaak de uitgang *ing*.
- Paradigmatische aspecten van woordvorming: zijn er suffixen die vaak met dezelfde stammen worden gecombineerd?

1.4 Het Ideale Woordenboek

1.4.1 Introductie

Deze sectie behandelt het Ideale Woordenboek-manifest van Verkuyl & al. (1998).[21] In dit manifest wordt uiteengezet hoe taalkundigen[22] zich verdienstelijk kunnen maken bij de structurering en ontsluiting van lexicografische gegevensbestanden. Het achterliggende idee is dat taalkundige theorieën vaak een economische opzet kennen, waardoor ze een goed vertrekpunt kunnen bieden voor de structurering van lexicografische informatiebestanden. Het manifest onderbouwt deze stelling door lexicografische vraagstukken te bespreken waarbij een taalkundige analyse behulpzaam kan zijn. Voor een deel raken deze vraagstukken aan fundamentele vragen uit het taalkundig onderzoek (zoals de verantwoording van niet-compositionele eigenschappen van woordgroepen). In deze gevallen kan de analyse van het probleem rechtstreeks bijdragen aan de taalkundige theorievorming. Maar er worden ook vraagstukken besproken die meer met de interactie tussen informatiebestand en gebruiker te maken hebben, zoals de opzet van slimme bedieningssystemen en gebruikersafhankelijke selectie van informatie. Alles bij elkaar gaat het om een rijk scala van vragen die één ding met elkaar gemeen hebben: ze hebben betrekking op taalgebruikskennis en vallen daardoor buiten het blikveld van grammaticale studies.

Deze sectie is als volgt opgezet. Eerst wordt het IW-model besproken (H1.4.2). Vervolgens wordt een toelichting geven op de lexicografische kwaliteitscriteria (H1.4.3) en worden deze

[20] Een soortgelijke systematiek treft men aan in het *Nieuw Volledig Zakwoordenboek* (uitgegeven in 1894): bij werkwoorden werden bijvoorbeeld ook de afleidingen op -ING, -ERIJ en -SEL vermeld; concurrent Koenen heeft deze systematiek niet overgenomen, want naar zijn (onderwijskundige) mening betrof het "afleidingen en samenstellingen bij werkwoorden, zelfst. nw. enz., die door niemand ooit gezocht worden" (Posthumus, 1997).

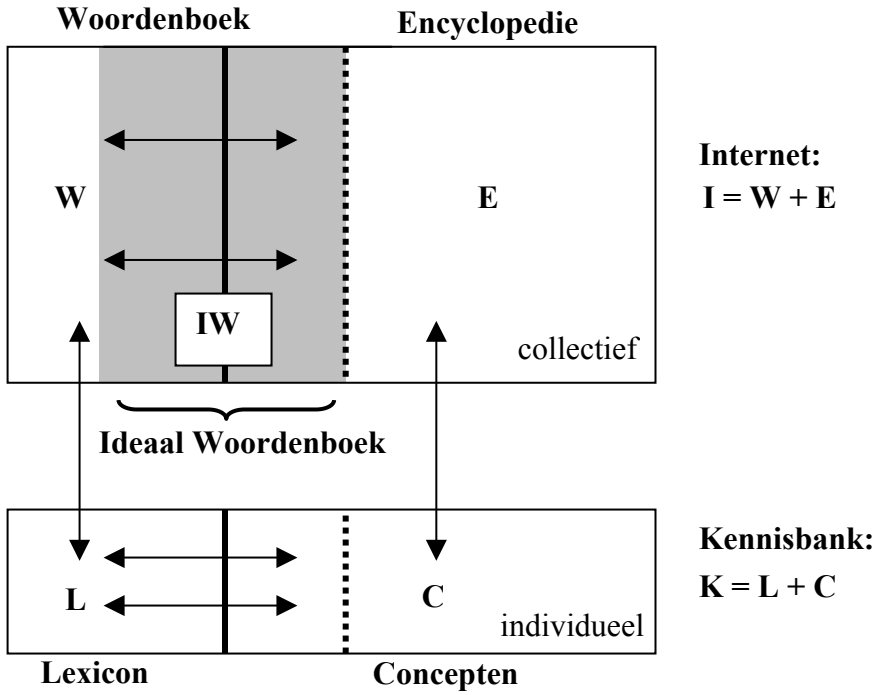
[21] De officiële naam luidt: "Work group lexicology and lexicography". Deze werkgroep hoort bij het Onderzoekinstituut voor Taal en Spraak (UiL OTS) van de Universiteit Utrecht (UU).

[22] In deze studie verwijst de term *taalkundige* naar onderzoekers die zich bezighouden met theorievorming over het cognitieve taalsysteem; in deze betekenis contrasteert de term met lexicografen, die zich bezighouden met de systematische inventarisatie van data uit een specifiek taaldomein ten behoeve van praktische toepassingen.

criteria bij wijze van voorbeeld op een deeldomein van de GWNT toegepast (H1.4.4). Hierna wordt uiteengezet wat voor eisen er aan het zoeksysteem kunnen worden gesteld (H1.4.5). In H1.4.6 ten slotte worden enkele beperkingen van het IW-model getoond, waarna een wat geavanceerder model wordt uitgewekt, te weten het Ideale Lexicon-model (= IL-model).

1.4.2 Het IW-model

Het IW-model is schematisch weergegeven in figuur 1-6. Het bestaat uit twee componenten, te weten de kennisbank K (waarmee het cognitieve systeem wordt aangeduid dat ten grondslag ligt aan de taalgebruikskennis van een individu) en het informatiesysteem I (waarmee de collectieve taalgebruikskennis wordt bedoeld).[23] De component K valt uiteen in een lexicale module L en een conceptuele module C, waarbij L en C complementaire informatie bevatten, dus K = L + C. Hierbij correspondeert L minimaal met kennis over woordvormen (of morfemen) en hun onderlinge combinatiemogelijkheden[24], terwijl C de aan deze vormen verbonden concepten specificeert en aanvult met "encyclopedische" kennis. In deze visie op het taalsysteem bevat L dus geen semantische informatie; in plaats daarvan worden woorden als interface-relaties tussen een woordvorm uit L en een concept uit C gedefinieerd.



Figuur 1-6: De structuur van het Ideale Woordenboek.

Parallel aan K kan de component I worden onderverdeeld in een woordenboek W en een encyclopedie E, dus I = W + E.[25] Elk van de vier componenten uit het IW-model heeft een interface met de twee aangrenzende componenten (aangeduid door tweezijdige pijlen). Deze interface legt relaties tussen de elementen uit beide componenten, zodat informatie-uit-

[23] Het manifest trekt hier een vergelijking met internet.

[24] Het manifest gaat ervan uit dat L ook informatie geeft over constructies van meerdere woorden (zoals vaste verbindingen en idioom), maar geeft helaas niet aan hoe dit dan technisch moet worden uitgewerkt. De centrale vraag in dit verband is of dergelijke constructies als autonome eenheden gelden of dat ze op lexicaal niveau ondergeschikt zijn aan het syntactische hoofd van de constructie.

[25] In dit geval zou men ook de formule I = W ∪ E kunnen overwegen. Deze is meer geëigend indien er overlap bestaat tussen W en E, tenzij men de + als een Boolese som interpreteert.

wisseling mogelijk wordt.[26] Volgens het manifest correspondeert het ideale woordenboek (IW) met één van deze interfaces, namelijk de interface tussen W en E. In schema 1-6 correspondeert deze interface met een dikke streep tussen W en E, terwijl het grijze gebied aangeeft wat het bijbehorende bereik is. Hieruit blijkt dat het ideale woordenboek slechts een deel van de gangbare woordvormen en concepten omvat, namelijk dat deel dat relevant is voor de functie van een woordenboek; want als een woordenboek toegang zou geven tot alle woordvormen (zoals niet-courante woorden, verkeerd gespelde woorden, eigennamen en conceptuele details) en concepten (inclusief allerlei encyclopedische subklassen) zou het voor de meeste gebruikers te informatief worden en daarmee zijn doel voorbijschieten. Het IW-model kan de functies van woordenboek en encyclopedie ook combineren; in dat geval zal het ideale woordenboek het complete I-domein omvatten. Het manifest gaat ervan uit dat het ideale woordenboek aan een contextgevoelig zoeksysteem is gekoppeld, zodat de gebruiker voor elke zoekterm uit W de best passende betekenis in E kan vinden en vice versa.

Het valt makkelijk in te zien dat de lexicografische betekenisdefinities uit een verklarend woordenboek niet buiten encyclopedische informatie kunnen. Zo kan bijvoorbeeld geen verschil worden gemaakt tussen een roodborstje en een eend zonder informatie te geven over uiterlijk, geluid, gedrag, voeding en leefgebied van deze vogels, bij voorkeur ondersteund door visuele en auditieve hulpmiddelen. Het IW-model gaat er dan ook vanuit dat er geen scherpe grens bestaat tussen lexicografische en encyclopedische betekeniskenmerken en dus ook niet tussen linguïstische en cognitieve betekeniskenmerken. In het schema zijn deze grenzen daarom door een onderbroken lijn weergegeven.

Ondanks het hier bedoelde afbakeningsprobleem geven de verklarende woordenboeken uit het Nederlandse taalgebied (zoals die van Van Dale, Koenen en Kramers) nauwelijks encyclopedische informatie. Ze beperken zich tot beknopte betekenisdefinities, naast taalkundige informatie over woordvorm, grammaticale kenmerken en vaste verbindingen. Omgekeerd richten encyclopedieën zich voornamelijk op niet-talige achtergrondinformatie bij de opgenomen termen en eigennamen.[27] In dit verband stelt Verkuyl (2000) dat Nederland geen traditie kent waarin serieus geprobeerd is encyclopedische informatie op te nemen in woordenboeken. In België daarentegen zijn al vele edities uitgebracht van het encyclopedische woordenboek Verschueren. Ook voor het Engelse, Franse en Duitse taalgebied zijn zulke woordenboeken beschikbaar (zoals Cobuild, Webster, Larousse en Meyer).

Door de opkomst van digitale informatiedragers zal het onderscheid tussen woordenboeken en encyclopedieën waarschijnlijk snel verdwijnen, want deze bieden inmiddels zoveel opslagcapaciteit dat ze deze functies makkelijk kunnen combineren. Toch blijft het traditionele onderscheid tussen lexicografische en encyclopedische informatie van belang, want wie een woord opzoekt wil niet bedolven worden onder encyclopedische details, maar eerst een beknopt overzicht krijgen van de meest voorkomende betekenissen (al dan niet in dezelfde taal). Dit kan worden opgelost door het IW-model met een interactieve zoekmodule uit te breiden; deze zoekmodule dient voor een gefaseerd informatieaanbod te zorgen, zodat de gebruiker stap voor stap kan inzoomen op de kenmerken waarover hij meer te weten wil komen.

1.4.3 Lexicografische criteria

In het IW-manifest worden vijf criteria besproken voor het vaststellen van de lexicografische kwaliteit van een (al dan niet elektronisch) woordenboek (of de hieraan ten grondslag liggende gegevensbank). Het betreft *compleetheid*, *consistentie*, *correctheid*, *courantheid* en

[26] De verbinding tussen E en C hoeft niet per se met een directe relatie tussen encyclopedische informatie en taalkundige concepten te corresponderen; de informatie kan bijvoorbeeld via plaatjes of geluid binnenkomen en dan pas in talige concepten worden omgezet.

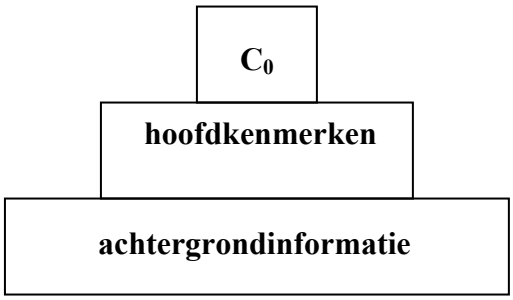
[27] Uitgezonderd de Grote Oosthoek Encyclopedie (1976-1978), die ook grammaticale informatie geeft.

citatie. Deze criteria waren oorspronkelijk op de kwaliteit van betekenisdefinities gericht, maar ze zijn ook bruikbaar voor de evaluatie van vormkenmerken. Hieronder zal ik deze evaluatiecriteria nader toelichten.

Compleetheid Er zijn twee niveaus van compleetheid, namelijk globale en locale compleetheid. Een woordenboek is *globaal compleet* als het alle woorden bevat die tot de beschreven taal kunnen worden gerekend en als het voor al deze woorden een complete inventarisatie van de woordeigenschappen geeft (zoals de beschikbare betekenissen). Dit is een moeilijk (of onmogelijk) te bereiken ideaal, aangezien de woordenschat een dynamisch karakter heeft: om te beginnen komen er steeds nieuwe woorden bij, terwijl er ook weer woorden verdwijnen (of in onbruik raken). Men heeft dus minimaal een methode nodig om deze veranderingen goed te kunnen volgen. Maar een bijkomende complicatie is dat talen over een morfologisch regelsysteem beschikken waarmee ze op voorspelbare wijze nieuwe woorden kunnen aanmaken door uitbreiding of samenstelling van bestaande woorden.

In het pre-computer-tijdperk was globale compleetheid een vrijwel onbereikbaar ideaal. Want zonder computers is het onmogelijk om automatisch woordvormen te genereren, terwijl ze ook nodig zijn om systematisch corpusonderzoek te doen. Daar komt bij dat papieren woordenboeken met ruimtetechnische beperkingen te maken hebben, zodat het weinig zin had om naar een complete inventarisatie te streven. Deze woordenboeken beperkten zich daarom tot de beschrijving van een normatieve selectie uit de werkelijk aangetroffen woorden. Deze beperking geldt niet voor het WNT, dat een nagenoeg complete inventarisatie biedt van Nederlandse woorden uit het schriftelijk taalgebruik tussen 1500 en 1970. Dit is een opmerkelijke prestatie, al heeft men er wel ruim 150 jaar voor nodig gehad. Door de komst van krachtige computers is het ideaal van globale compleetheid nu veel eenvoudiger te realiseren.

Volgens het IW-manifest dienen woordenboeken minimaal naar *locale compleetheid* te streven. Dat betekent dat voor elk betekenisdomein moet worden nagegaan of er een evenwichtige selectie is gemaakt van de beschikbare woorden, en of hun betekenisdefinitie dezelfde structuur heeft. Dus als een woordenboek het woord *koning* in de betekenis van "schaakstuk" vermeldt, dienen ook de termen voor de andere schaakstukken te worden opgenomen. Wat betreft de betekenisdefinitie moet het woordenboek op zijn minst het concept C_0 noemen (zie figuur 1-7), d.w.z. een koepelterm waar alle woordbetekenissen onder vallen, anders voldoet de definitie niet aan de informatieve ondergrens.



Figuur 1-7: De gelaagde betekenisopbouw van een lexicaal concept.

Bij de schaakspel-gerelateerde term *koning* kan dus niet worden volstaan met de omschrijving *ding*; het voldoet namelijk ook aan de definitie van een *schaakstuk*. Bij een wat uitgebreidere definitie zal ook een en ander over uiterlijk, samenstelling of functie worden meegedeeld (hoofdkenmerken). Maar er mag geen encyclopedische achtergrondinformatie worden vermeld, anders zou de informatieve bovengrens worden overschreden. Bovendien mogen geen details worden gegeven die reeds uit het basisdomein of de hoofdkenmerken volgen. Dus als *schaakspel* als een *bordspel* wordt gespecificeerd, is het niet nodig om te vermelden dat er een bord wordt gebruikt voor dit spel, noch dat er stukken op dit bord horen te staan. Zodra nadere

informatie wordt geven over het soort bord (bijv. bordspel dat op een schaakbord wordt gespeeld) en de aard van de stukken (bijv. bordspel waarbij wit en zwart allebei 16 stukken krijgen, te weten een koning, een dame, twee torens, een loper, een paard en acht pionnen) betreedt men het domein van de encyclopedische achtergrondinformatie. Dit geldt nog sterker voor informatie over de structuur van het bord of over de plaatsing, het uiterlijk en het gedrag van de afzonderlijke stukken; deze informatie staat namelijk los van hun semantische klasse en is daarom niet van belang voor de lexicale organisatie.

Consistentie Een woordenboek is *globaal consistent* als alle lemma's op een uniforme wijze zijn gestructureerd, ongeacht het betekenisdomein waartoe ze behoren (= "cross box"-consistentie). Men kan dit beoordelen door vergelijkend onderzoek te doen naar formele lemmakenmerken, zoals de wijze waarop hoofd- en subbetekenissen worden onderscheiden, de volgorde waarin vaste lemmakenmerken worden gepresenteerd (bijv. trefwoord, uitspraak, inflectiepatroon, functiewoorden en betekenissen) en de detailleringsgraad. Verder moet sprake zijn van inzichtgevende betekenisdefinities, d.w.z. van niet-redundante definities die woorden zoveel mogelijk in termen van superklassen of algemeen bekende synoniemen typeren.[28] Bij *locale consistentie* gaat het om de uniformiteit binnen een specifiek betekenisdomein. Dit domein dient dan niet alleen aan de eisen van globale consistentie te voldoen, maar ook aan de eis dat woorden die tot dezelfde hoofdklasse behoren expliciet aan die hoofdklasse worden gerelateerd, terwijl hun betekenisdefinities een vergelijkbare opbouw moeten bezitten. Zo is het domein van de schaakstukken (lokaal) consistent indien elk schaakstuk expliciet als schaakstuk wordt gedefinieerd en indien bij elk schaakstuk dezelfde betekenisdimensies worden gespecificeerd, zoals vorm en beweging.

Correctheid Het is zeer moeilijk om te bepalen of de informatie in een woordenboek correct is, want er is geen objectieve instantie die kan vertellen wat de betekenis is van een woord of welke grammaticale eigenschappen eraan moeten worden toegekend. Voor dit soort vragen grijpen taalgebruikers juist naar een woordenboek. Woordenboeken hebben op dit punt dus een normatieve functie. Om toch een indruk te krijgen van de betrouwbaarheid van deze informatie, zou men gebruik kunnen maken van een *descriptief* model van het taalgebruik, bijvoorbeeld een statistisch geanalyseerd tekstcorpus. Deze methode heeft als nadeel dat hij geen recht doet aan het *normatieve* karakter van een woordenboek: zo kan het woordenboek inflectievormen voorschrijven die in de praktijk vrijwel nooit voorkomen, terwijl het corpus weer woordvormen en constructies kan bevatten die niet in het woordenboek zijn terug te vinden, zoals het vooralsnog foutieve gebruik van de tussen-*n* in woorden als *gedachtenwisseling*, *woordenloos* en *zijdenlings*. In dit soort gevallen zullen toonaangevende taalexperts uitsluitsel moeten geven. Wat betreft de betekenisomschrijving moet een goede balans worden gevonden tussen beknoptheid (c.q. abstractheid) en volledigheid (zie figuur 1-3). Verder zou men bij vaktermen uit moeten gaan van het oordeel van vakspecialisten. Indien de betekenisdefinitie uit het woordenboek compatibel is met de vakdefinitie kan deze definitie correct worden genoemd; zo niet, dan zal de definitie waarschijnlijk verbeterd moeten worden.

Courantheid Woordenboeken hebben doorgaans niet genoeg ruimte voor een complete weergave van de bestaande woordenschat, zodat ze gedwongen zijn om hier een selectie uit te maken. Zo achten de meeste woordenboeken het niet nodig om woorden te vermelden waarvan vorm en betekenis op regelmatige wijze van een ander woord zijn af te leiden. Indien de resulterende woordenlijst nog steeds te lang is, zijn echter aanvullende selectiecriteria nodig. Volgens het IW-model zou hierbij voorrang moeten worden verleend aan courante woorden, d.w.z. woorden met een hoge gebruiksfrequentie (in de beschreven taalperiode). Men kan hier

[28] Een woorddefinitie in termen van synoniemen kan heel verhelderend zijn, bijv. in het geval van adjectieven. Maar indien de synoniemen alleen naar elkaar verwijzen, ontstaat een onwenselijke vorm van circulariteit.

informatie over verkrijgen door een groot tekstcorpus te analyseren, bijvoorbeeld een compleet krantencorpus. De hieraan ontleende frequentie-informatie kan ook worden benut om bij elk opgenomen woord een globale indicatie van de gebruiksfrequentie te geven (bijvoorbeeld *zeldzaam*, *normaal* of *hoogfrequent*). Het courantheidscriterium is echter niet zo bruikbaar voor woordenboeken die een culturele functie vervullen.

Citatie De wat grotere woordenboeken geven bij de meeste woordbetekenissen een concreet voorbeeld in de vorm van een citaat. Deze citaten zijn echter niet altijd even verhelderend, omdat de keuze van deze citaten meestal niet op taalkundige, maar op literaire overwegingen is gebaseerd. Dit is een gevolg van het uitgangspunt dat woordenboeken een functie hebben als hoeder van het literaire erfgoed. Volgens Verkuyl (1993) is dit uitgangspunt echter niet bevorderlijk voor de taalkundige helderheid, want schrijvers kunnen misschien wel mooie zinnen construeren, maar dat betekent niet dat deze zinnen ook bijdragen aan het begrip van het trefwoord. Daarom stelt Verkuyl dat de kwaliteit van een woordenboek mede kan worden afgelezen aan de mate waarin het verhelderende voorbeelden geeft, al dan niet gebaseerd op citaten van bestaande toepassingen.

Compositionaliteit In aanvulling op de kwaliteitseisen van het IW-model wil ik een zesde kwaliteitseis introduceren, namelijk de compositionaliteit van de betekenisdefinities (en evt. de vormrepresentaties). Bij deze eis gaat het om de vraag in hoeverre een woordenboek gebruik maakt van compositie of overerving. Bij de evaluatie van een woordenboek hoeft dit criterium minder strikt te worden toegepast dan bij een lexicografische gegevensbank, want een woordenboek heeft primair de taak om de gebruiker zonder heen-en-weer-geblader van adequate betekenisinformatie te voorzien; hierbij dienen triviale definities zoveel mogelijk te worden voorkomen. Zo is het niet erg informatief of zelfs onjuist om een *speelman* als een soort man te definiëren, maar heeft men meer aan een definitie waarin de speelman als een muziekgerelateerd beroep wordt gekarakteriseerd. Een lexicografische gegevensbank daarentegen wordt krachtiger naarmate deze een gedetailleerder beeld geeft van schijnbaar triviale betekenisrelaties tussen de hierin opgenomen woorden (en woordbetekenissen).

1.4.4 Demonstratie van de evaluatiemethode

De in H1.4.3 behandelde evaluatiecriteria zijn terug te voeren op het idee dat de kwaliteit van een woordenboek toeneemt naarmate de woorden meer in hun onderlinge samenhang worden beschreven, dus naarmate de lexicografische informatie een sterkere domeinstructuur vertoont. Om meer inzicht te krijgen in de uitvoerbaarheid van deze methode heb ik een proefevaluatie uitgevoerd op een concreet betekenisdomein uit de GWNT. Dit onderzoek wordt in appendix A besproken.

1.4.5 Zoekmogelijkheden

Een Ideaal Woordenboek beperkt zich niet tot een statische kennisinventarisatie (zoals het geval is bij papieren woordenboeken), maar is ook uitgerust met een contextgevoelig zoeksysteem. Deze "zoekassistent" dient onder meer aan de volgende eisen te voldoen:

1) De zoekassistent moet zoveel mogelijk lexicografische zoekmogelijkheden aan bieden. Hij moet bijvoorbeeld niet alleen op woordvorm kunnen zoeken, maar ook op klankvorm, betekeniskenmerken, morfologische kenmerken, woordcategorie en syntactische selectiekenmerken of via hyponiem-, synoniem- of antoniemrelaties, en ook op combinaties van deze kenmerken. De gebruiker moet kunnen kiezen tussen een zoekmodus (waarbij het zoeksysteem precies de woordinformatie geeft waar men naar op zoek is) en een bladermodus (waarbij het ook "buurwoorden" laat zien, gegeven het door de gebruiker gespecificeerde sorteercriterium).

2) De zoekassistent moet contextgevoelige antwoorden kunnen geven. Leest de gebruiker een tekst en stuit hij op een woord dat hij niet kent, dan kan hij de zoekassistent om een context-gebonden betekenisdefinitie vragen door het betreffende woord aan te wijzen. Het zoek-systeem dient vervolgens de meest waarschijnlijke betekenis te selecteren door rekening te houden met de syntactische omgeving van het opgegeven woord.

3) De zoekassistent moet flexibel zijn. Zo kan een gebruiker behoefte hebben aan informatie over de spelling van een woord waarvan hij alleen de uitspraak kent; in dat geval zal het zoeksysteem in staat moeten zijn om op basis van een weergave van de klankvorm toch bij het gewenste woord uit te komen, om vervolgens de gevraagde spelvorm te specificeren, of andere gevraagde kenmerken.

4) De zoekassistent moet de gebruiker begeleiden bij het formuleren van een bruikbare zoek-opdracht. Indien de ingevoerde zoekopdracht meerdere antwoorden oplevert, dient de zoek-assistent de gebruiker om een aanvullend criterium te vragen, en indien er helemaal geen antwoorden mogelijk zijn, zou de zoekassistent zelf aanpassingen moeten voorstellen.

5) De zoekassistent dient niet alleen lexicografische en encyclopedische zoekfuncties te ondersteunen, maar moet ook naar websites op internet kunnen doorverwijzen; hierbij dienen de geselecteerde informatiebronnen zo goed mogelijk aan te sluiten op de zoekopdracht.

1.4.6 Van Ideaal Woordenboek naar Ideaal Lexiconsysteem

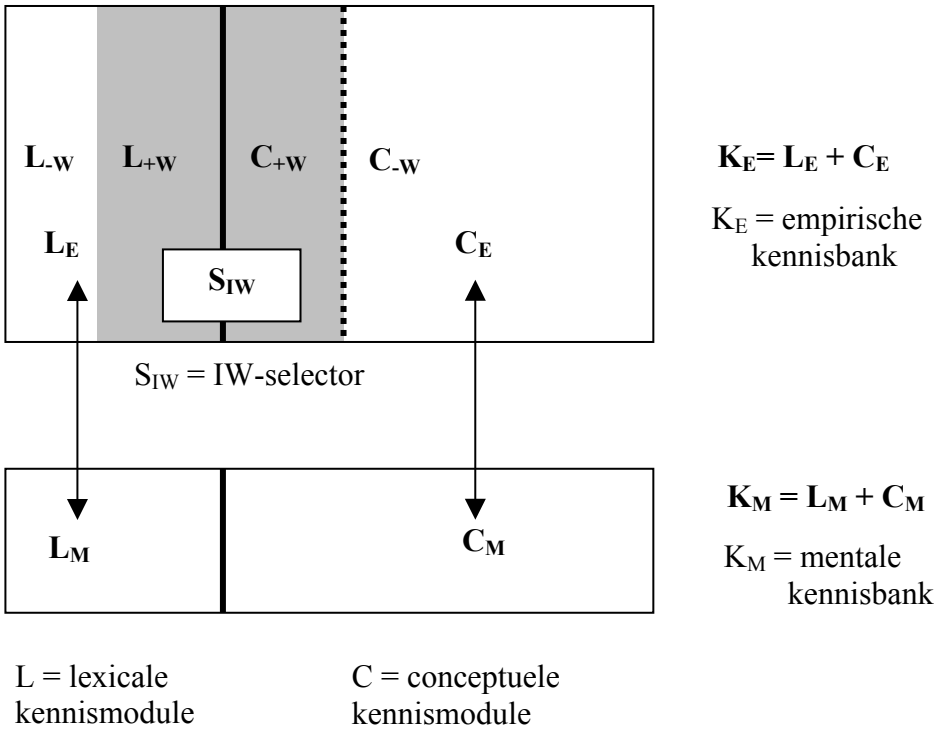
In mijn optiek biedt het IW-model een aantrekkelijk uitgangspunt voor de opzet van een lexicografisch informatiesysteem. Maar zowel inhoudelijk als terminologisch is dit model voor verbetering vatbaar. Zo houdt het IW-model geen rekening met de normatieve dimensie of met het feit dat de inhoud van een Ideaal Woordenboek afhankelijk is van de beoogde gebruikers. Verder heeft de term Ideaal Woordenboek geen eenduidige betekenis, want in het manifest worden er minstens drie verschillende functies aan toegekend, te weten:

1) interface I_W tussen tussen woordenlijst W en encyclopedie E
2) toegangsportaal tot alle via internet toegankelijke naslagwerken, zoals elektronische woordenboeken, encyclopedieën en catalogi.
3) zoekassistent bij het selecteren van informatie over een door de gebruiker opgegeven woord of begrip door gebruik te maken van de kennis in het Ideale Woordenboek.

Om deze verwarring weg te nemen definieer ik het Ideale Woordenboek liever als een gebruikerspecifieke selectie uit een Ideaal Lexicon (IL). Zo'n Ideaal Lexicon kan worden onderverdeeld in een Ideale Kennisbank (IKB) en een Ideale Zoekmachine (IZM). Het Ideale Woordenboek correspondeert dan met een redactioneel tot stand gekomen IZM-selectie uit de IKB, d.w.z. een optimaal op de gebruikersgroep toegesneden selectie van woorden, woord-kenmerken en betekenisdefinities uit de IKB, zoals een IW voor kinderen, een IW voor scholieren, een IW voor volwassenen, een IW voor tweedetaalverwervers, een IW voor taalkundigen en een IW voor cultuurminnende intellectuelen. Deze IW's kunnen zowel elektronisch worden gepubliceerd als in boekvorm. In de rest van deze sectie zal ik nader ingaan op de structuur van het hier voorgestelde informatiesysteem. Hiertoe zal ik eerst aandacht besteden aan de Ideale Kennisbank, om vervolgens de Ideale Zoekmachine te beschrijven.

Figuur 1-8 toont de IL-component met de Ideale Kennisbank. In dit schema wordt (net als in het IW-model) een structurele parallel getrokken tussen het cognitieve representatiesysteem, namelijk de Mentale Kennisbank K_M, en het computationele representatiesysteem (voor lexicografische toepassingen), namelijk de Empirische Kennisbank (K_E) (met informatie over concreet taalgebruik op het niveau van gebruikersgroepen). Beide componenten bestaan uit een lexicon L en een conceptueel systeem C. De kern van de Ideale Kennisbank correspondeert met de IW-Selector S_{IW}; deze bepaalt welk deel van de Ideale Kennisbank zichtbaar is

voor de gebruiker, waarbij de gebruikerswensen bepalend zijn voor de gemaakte selectie. Deze selectie is formeel gedefinieerd als een verzameling relaties (zoals woordrelaties) tussen eenheden uit de lexicale kennismodule L en de conceptuele kennismodule C. Een Ideaal Woordenboek is dus het resultaat van de toepassing van S_{IW} op L en C, wat een onderscheid oplevert tussen [+W]-eenheden (die deel uitmaken van IW) en [-W]-eenheden (die niet geselecteerd zijn). Voor de duidelijkheid is het [+W]-deel grijs gearceerd.



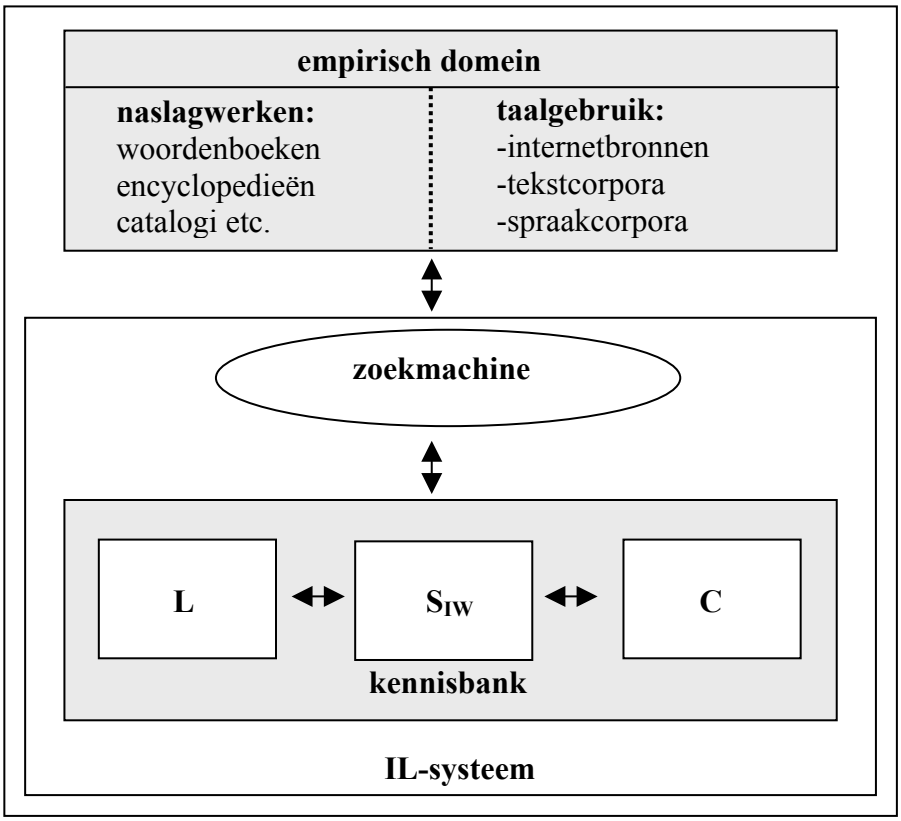
Figuur 1-8: Het IL-model. Het grijze gebied correspondeert met de door selector S geselecteerde data uit het Ideale Woordenboek (IW).

Deze nieuwe definitie van een Ideaal Woordenboek biedt structurele mogelijkheden voor de specificatie van gebruikerswensen, terwijl ook een normatieve dimensie kan worden ingebouwd. Zo zou men onderscheid kunnen maken tussen algemeen geaccepteerd taalgebruik en idiosyncratisch of incorrect taalgebruik door de S-functie afhankelijk te maken van een redactionele parameter [±R]. Ook zou onderscheid kunnen worden gemaakt tussen de bestaande woordenschat en de mogelijke woordenschat.

Figuur 1-9 toont de structuur van een Ideaal Lexicon-systeem, d.w.z. een informatiesysteem dat uitgaat van de principes van het IL-model. De bijbehorende zoekmachine biedt toegang tot twee verschillende zoekdomeinen, te weten het interne domein (c.q. de Ideale Kennisbank K) en het externe domein (c.q. het empirische domein E). In beide gevallen dient de zoekmachine naar een optimale zoekstrategie te streven door slim gebruik te maken van de informatiestructuur in de Ideale Kennisbank. Hierbij kan de gebruiker stap voor stap naar de gevraagde informatie worden geleid door hem een hele reeks keuzes voor te leggen waarvan de inhoud afhankelijk is van de reeds geactiveerde informatie.

Zo'n zoektocht zou als volgt kunnen verlopen. Stel dat iemand de zoekterm *vogel* opgeeft. De zoekassistent weet dan nog vrij weinig: eigenlijk niet meer dan dat de gebruiker informatie wil over een taalkundig of conceptueel aspect van het Nederlandse woord *vogel* (terwijl ook de mogelijkheid bestaat dat het hier om een naam gaat). De zoekassistent dient daarom te reageren met de vraag of de gebruiker iets wil weten over de taalkundige eigenschappen van dit woord, of over een inhoudelijk aspect, en in het laatste geval, of hij een betekenisdefinitie

verlangt of inhoudelijke informatie, bijvoorbeeld "encyclopedische" informatie of een overzicht van websites die iets met vogels te maken hebben. Zonder deze informatie kan het systeem geen verschil maken tussen de relevantie van een website over bedreigde vogels en die van een elektronisch woordenboek: het zijn immers allebei informatiebronnen over het woord *vogel*. Maar als de gebruiker aangeeft dat hij meer wil weten over de grammaticale eigenschappen van *vogel* is het woordenboek natuurlijk veel relevanter. In dat geval zou de zoekassistent de gebruiker kunnen doorverwijzen naar de verzamelcategorie 'woordenboek', met een overzicht van alle beschikbare titels.[29] Een andere optie is om gebruik te maken van de kennis in het interne domein. De zoekassistent kan dan verder gaan met de vraag of de gebruiker wil weten hoe het woord vertaald moet worden of dat hij informatie wil over de taalkundige kenmerken van dit woord. Zo kan het zoeksysteem steeds specifiekere keuzes voorleggen totdat duidelijk is wat de gebruiker nu eigenlijk wil weten, bijvoorbeeld welk Engels equivalent van *vogel* het beste in de context past of welk lidwoord men bij *vogel* kiest.



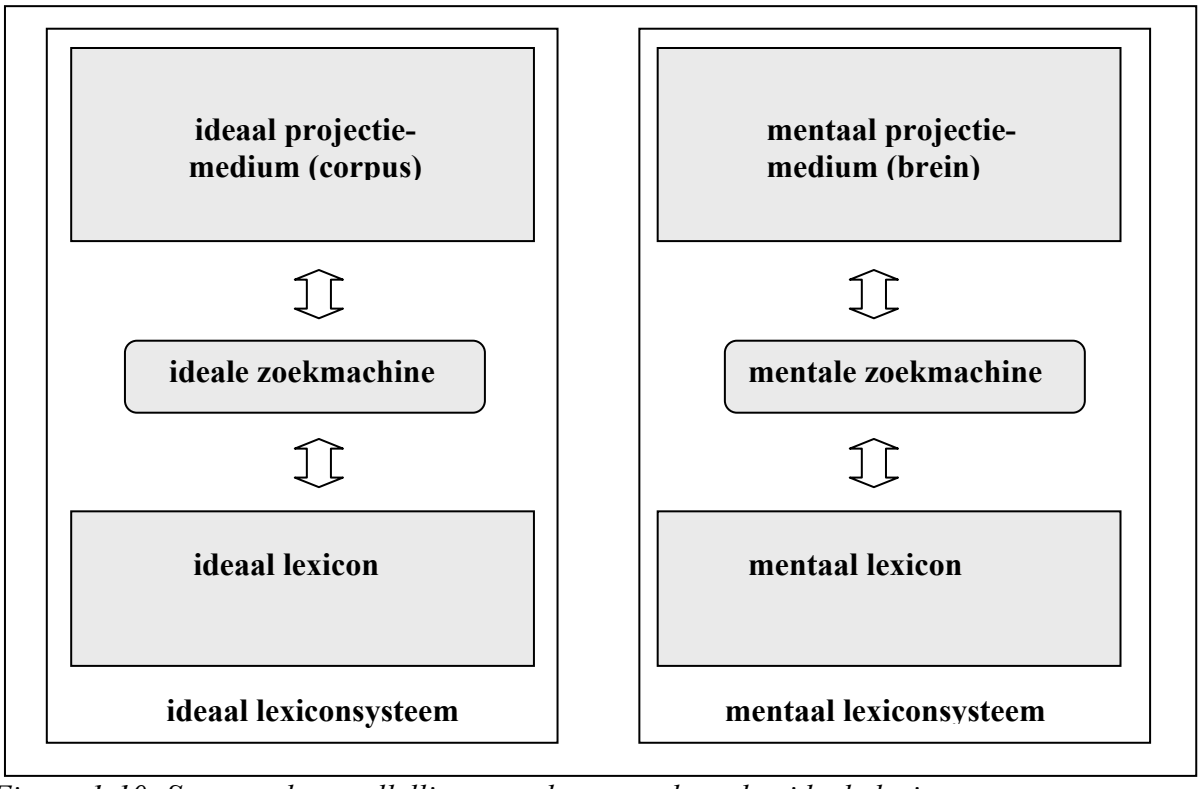
Figuur 1-9: Model van een Ideaal Lexicon-systeem (= IL-systeem).
L = lexicale module, C= conceptuele module, S_{IW} = selector (van IW-relaties)

Het hier voorgestelde model kan niet alleen een fundament bieden voor een computationeel informatiesysteem, maar leent zich ook voor een structurele koppeling tussen lexicografische en cognitieve kennisrepresentatie. Deze koppeling (die wordt uitgewerkt in figuur 1-10) moet ertoe bijdragen dat het lexicografische systeem dezelfde taalkennis kan opbouwen als menselijke experts; omgekeerd zou de op deze wijze opgebouwde kennis een interessante proeftuin kunnen bieden voor onderzoek naar cognitieve representatieprincipes. Vanzelfsprekend gaat het om een speculatief verband, maar voor zover ik hierover kan oordelen is het voorgestelde

[29] Het gebruik van verzamelcategorieën is een voor de hand liggende methode om de resultaten van een zoekopdracht overzichtelijk te presenteren; het is daarom verrassend dat deze functionaliteit (nog) niet standaard ingebouwd is in de zoekmachines op internet.

model zeker niet kansloos. Elke taalgebruiker beschikt immers over een individueel lexicon waarin hij lexicale en encyclopedische informatie kan opzoeken op basis van een woordvorm of betekenisomschrijving. Daarnaast beschikt hij over een mentaal projectiemedium (c.q. brein) waarin hij cognitieve representaties kan opbouwen van de buitenwereld, waaronder de door hem geraadpleegde websites op het internet en de hierop raadpleegbare documenten. Naar analogie van het brein zou men het (ideale) internet als een (ideaal) projectiemedium voor collectieve kennis kunnen aanduiden. Maar in tegenstelling tot de informatie op het internet zijn cognitieve representaties vaak van tijdelijke aard. Mensen zijn bijvoorbeeld niet goed in staat om integrale teksten te onthouden, maar tijdens het lezen van een tekst kan hun brein er wel een compleet beeld van opbouwen (inclusief de lay-out).

De parallellie tussen de ideale zoekmachine en het mentale taalsysteem is nog verder door te trekken. Zo kan men het formuleren van een zin als de uitkomst zien van een mentale zoekopdracht om een intern gestructureerd concept in woorden om te zetten, met de mogelijke nevencondities dat goed op de voorgaande zin moet worden aangesloten en dat een aantal stilistische eisen in acht moeten worden genomen. Dergelijke zoekopdrachten zijn zo complex dat het niet waarschijnlijk is dat dergelijke taken door computers kunnen worden gesimuleerd, maar dat betekent niet dat het onzin is om zo'n taak als een zoekopdracht op te vatten. Op dit punt is er weinig verschil met een zoekmachine die op basis van een aantal opgegeven woorden en randvoorwaarden als de citatiefrequentie en de status van het tijdschrift op zoek gaat naar een toonaangevend artikel over het onderwerp waar men in geïnteresseerd is. In beide gevallen is het zaak om door combinatie van gegevens systematisch mogelijkheden uit te sluiten, totdat er nog maar enkele kandidaten over zijn. Op een ander punt bestaat wel verschil: mensen lijken namelijk niet in staat hun mentale kennisbank integraal te doorzoeken. In plaats daarvan doorzoeken ze slechts de actieve gedeeltes, waarbij ze zo snel mogelijk met een bruikbaar antwoord proberen te komen; hierdoor kan het antwoord op een zoekvraag per keer verschillen. Voor een ideale kennisbank is dit een minder aantrekkelijke eigenschap.

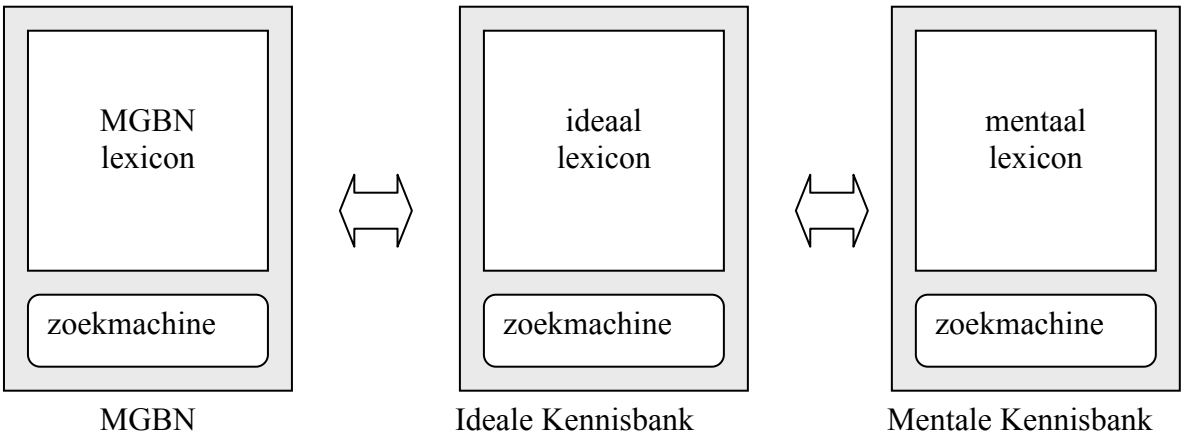


Figuur 1-10: Structurele parallellie tussen het mentale en het ideale lexiconsysteem.

De hier uiteengezette visie op het taalvermogen ligt ook ten grondslag aan de L-KRING-theorie, het door mij ontwikkelde systeem voor lexicale kennisrepresentatie (hoofdstuk 4). In deze theorie vormt de empirische kennisbank de kern van een semi-automatisch zoeksysteem met drie hoofdfuncties, te weten, het uitvoeren van een zoekopdracht, het rapporteren van de zoekresultaten en het aanpassen van het lexicon. Zonder dit zoeksysteem is de inhoud van het lexicon zo goed als onbruikbaar, want het lexicon van de L-KRING-theorie kenmerkt zich door een strikt-hiërarchische informatiestructuur, waarbij elke taaleenheid als een complexe verzameling van kleinere eenheden wordt opgevat die zelf ook weer het karakter van een verzameling hebben, wat doorgaat totdat de kleinst mogelijke informatie-eenheden zijn bereikt (zoals fonemen en basisconcepten). Bij directe aanschouwing is deze informatie even ondoorgrondelijk als de nullen en eentjes op een digitale geluidsdrager. Zoals een CD-speler nodig is om deze informatie als muziek te laten klinken, zo is een zoekassistent nodig om de informatie-eenheden uit het lexicon in leesbare woordrepresentaties om te zetten.

1.5 Opzet van de studie

In H1.4 is betoogd dat een Ideaal Woordenboek een kennisbank vereist waarvan de functionele structuur identiek is aan die van het mentale lexicon. Maar omgekeerd kan de inhoud van een lexicon dat aan de normen van een Ideaal Woordenboek, of meer specifiek, een Ideale Kennisbank voldoet (zie H1.4.6), ook inzicht geven in de structuur van het mentale lexicon. Want men hoeft niet te weten hoe het mentale lexicon is gestructureerd om de hierin aanwezige kennis in een empirische (c.q. lexicografische) kennisbank te kunnen onderbrengen: dit is mogelijk door systematisch taalintuïties te coderen, met als aangename bijkomstigheid dat het resulterende gegevensbestand als basis kan dienen voor empirisch onderzoek naar de structuurprincipes van het mentale lexicon. Deze visie ligt ook ten grondslag aan de opzet van de Morfologische Gegevensbank van het Nederlands (MGBN): de informatie in de MGBN is namelijk het resultaat van systematische raadpleging van de kennis in het mentale lexicon. Hierdoor kan statistisch onderzoek aan de MGBN bijdragen aan de kennis over het mentale lexicon, wat vervolgens weer tot verbetering van de MGBN kan leiden. De ontwikkeling van de MGBN loopt dan parallel aan de ontwikkeling van een model voor het mentale (en ideale) lexicon. Dit idee is schematisch uitgewerkt in figuur 1-11.



Figuur 1-11: De relatie tussen de mentale kennisbank, de ideale kennisbank en de in deze studie beschreven Morfologische Gegevensbank van het Nederlands (MGBN).

In dit schema bestaat elk van de drie kennisbanken uit een lexicon en een zoekmachine (conform de uitgangspunten van het IL-systeem). De equivalentiepijlen geven aan dat de kennisbanken een parallelle structuur hebben en dat er informatieuitwisseling mogelijk is.

Deze studie heeft als doel om een nieuw, taalonafhankelijk model over de structuur van het mentale lexiconsysteem te introduceren en motiveren, namelijk de L-KRING-theorie (hoofdstuk 2-4), en om te laten zien hoe de MGBN kan bijdragen aan de ontwikkeling van een L-KRING-model van het Nederlands (hoofdstuk 5-6). Hieronder wordt toegelicht hoe deze doelstelling zich tot de afzonderlijke hoofdstukken verhoudt.

Hoofdstuk 2 biedt een overzicht van de bestaande theorieën met betrekking tot de mentale representatie van woordkennis, dus modellen die inzicht geven in de structuur van het mentale lexicon. Deze inventarisatie dient ter voorbereiding op de introductie van een metamodel voor lexicale kennisrepresentatie, te weten het Integraal Dynamische Lexicon-systeem. Dit IDL-systeem, dat nader invulling geeft aan de principes van het Ideale Lexicon-model, brengt samenhang aan in de verschillende functies van het mentale lexicon. Hoofdstuk 3 biedt een overzicht van de bestaande kennis met betrekking tot de Nederlandse woordstructuur, waarbij het morfologiemodel uit het Morfologisch Handboek van het Nederlands als uitgangspunt dient. Hierbij wordt veel aandacht besteed aan de technische tekortkomingen van de bestaande beschrijvingsmodellen (gegeven de eisen van een Ideaal Lexicon, zoals uitgewerkt in het IDL-systeem), en komt per klasse van observaties een alternatieve benadering aan de orde. Deze uiteenzetting vormt een informele introductie tot de presentatie van de L-KRING-theorie, mijn in hoofdstuk 4 beschreven systeem voor lexicale kennisrepresentatie. In deze theorie wordt het mentale lexicon als een computationeel informatiesysteem voorgesteld dat langs inductieve weg structuur kan aanbrengen in de mentale kennis over de woordenschat.

In hoofdstuk 5 wordt uitgelegd hoe de Morfologische Gegevensbank is opgezet. In deze opzet fungeert het mentale lexicon als kennisbron voor de structurering van de lexeeminventarisatie in de MGBN. Er wordt dus een koppeling tot stand gebracht tussen de structuur van het mentale lexicon en de structuur van de MGBN, conform de uitgangspunten van de L-KRING-theorie. Bij de bespreking van de MGBN ga ik uitvoerig in op de analysemethode, die zich kenmerkt door een cyclische, semi-automatische werkwijze. In hoofdstuk 6 bespreek ik een reeks datarapporten die het resultaat zijn van morfologische structuuranalyses op een virtueel L-KRING-model van de MGBN. Hoewel de hierop gebaseerde structuuranalyses allereerst inzicht geven in de samenstelling van de MGBN zelf, ga ik ervan uit dat deze informatie ook een betrouwbaar beeld geeft van de morfologische eigenschappen van het mentale lexicon van de Nederlandse taalgebruiker en een empirische basis kan vormen voor onderzoek naar de onderliggende kennisprincipes. Om meer inzicht te krijgen in de houdbaarheid van deze aanname, heb ik de morfologische structuurkenmerken uit het MGBN-model in detail met de kennis in het Morfologisch Handboek vergeleken. De resultaten van dit onderzoek komen uitvoerig aan de orde. Hoofdstuk 7 ten slotte geeft een beknopt overzicht van de belangrijkste bevindingen van deze studie.

2 De modellering van het mentale lexicon

2.1 Introductie

Dit hoofdstuk heeft als doel om na te gaan wat er bekend is over structuur en functies van het mentale lexicon, hoe de bestaande kennismodellen zich onderling verhouden en hoe men deze modellen kan combineren tot een Ideaal Lexicon-model. Het hoofdstuk begint met een discipline-overschrijdende inventarisatie van morfologische kennismodellen (zie H2.2), d.w.z. een inventarisatie waarin niet het onderzoeksdomein of het lexiconperspectief centraal staat, maar de structuur van de voorgestelde modellen. Hierbij wordt veel aandacht besteed aan de classificatiecriteria, want deze zijn bepalend voor de vraag welke modelklassen men kan onderscheiden en hoe ver deze modellen van elkaar afstaan. Op deze manier hoop ik bij te dragen aan het inzicht in de fundamentele dilemma's uit de morfologische theorievorming, terwijl ik tegelijk wil aantonen dat er fundamentele overeenkomsten bestaan in de kennismodellen uit verschillende onderzoeksdisciplines. In H2.3 wordt vervolgens het grammaticale lexiconperspectief belicht. Hierbij worden eerst de uitgangspunten van het vigerende morfologiemodel besproken, waarna een aantal structurele tekortkomingen aan de orde komen. In H2.4 komt het psychologische lexiconperspectief aan de orde, d.w.z. het lexiconperspectief dat ten grondslag ligt aan psycholinguïstisch onderzoek naar de activatie van mentale woordkennis. Deze sectie begint met een bespreking van de centrale onderzoeksvragen, waarna uiteen wordt gezet welke modellen men bij dit onderzoek hanteert, wat voor inzichten inmiddels zijn opgedaan en welke problemen er zijn opgedoken. In H2.5 betoog ik dat zowel het grammaticale als het psychologische lexiconperspectief een aantal fundamentele beperkingen kent en dat men deze kan omzeilen door uit te gaan van de principes van een Integraal Dynamisch Lexicon-systeem, een lexicaal metamodel dat werk maakt van de uitangspunten van het Ideale Lexicon-model door de verschillende lexiconfuncties in onderlinge samenhang te beschrijven. Het hoofdstuk eindigt met een conclusie.

2.2 Inventarisatie van lexicale kennismodellen

2.2.1 Introductie

Deze sectie biedt een overzicht van lexicale kennismodellen, d.w.z. modellen die als doel hebben om inzicht te geven in inhoud en structuur van het lexicon dat ten grondslag ligt aan het menselijke vermogen om woorden te produceren en interpreteren en in de wijze waarop deze kennis wordt opgebouwd en geactiveerd. Met de term kennismodel wil ik uitdrukken dat de hier te presenteren modellen zich op een cognitief, symbolisch (of subsymbolisch) beschrijvingsniveau bevinden, ter onderscheiding van modellen die zich op een neuraal, fysisch beschrijvingsniveau bevinden. Hoewel er een klasse van kennismodellen bestaat die het taalsysteem in termen van een neuraal netwerk probeert te beschrijven (namelijk connectionistische netwerkmodellen), corresponderen deze modellen in essentie met data op cognitief (symbolisch) niveau, niet met data op neuraal (fysisch) niveau.

Bij de inventarisatie van lexicale kennismodellen zal ik me vooral bezighouden met de vraag welke hoofdklassen er kunnen worden onderscheiden, en op welke dimensies variatie bestaat in doelstellingen en beschrijvingsmethode. Alvorens hierop in te gaan, zal ik eerst een toelichting geven op de door mij gehanteerde classificatiecriteria, te weten: i) lexicografisch of cognitief, ii) monistisch of dualistisch, iii) formeel of informeel, iv) representatie, activatie of acquisitie. De volgorde waarin deze criteria aan de orde komen weerspiegelt hun invloed op de hoofdindeling: hoe later een criterium wordt behandeld, hoe minder invloed.

2.2.2 Algemene classificatiecriteria

i: lexicografisch of cognitief

Dit onderscheid heeft betrekking op de vraag of een kennismodel primair bedoeld is voor de systematische representatie van lexicografische kennis of voor de analyse van de cognitieve dimensie van het taalsysteem. Als gevolg van dit functieverschil besteden lexicografische kennismodellen meestal weinig aandacht aan de morfologische dimensie, terwijl deze dimensie juist centraal staat in cognitieve (waaronder generatieve) kennismodellen.

ii: monistisch of dualistisch

Dit onderscheid heeft betrekking op de relatie tussen de lexicaal vastgelegde woorden (die doorgaans uit verschillende deelrepresentaties bestaan, waaronder een klankvorm en een betekenis) en de morfologische structuurkenmerken van deze woorden. Er is sprake van een dualistisch morfologiemodel indien het lexicon van bestaande woorden losstaat van het morfologische regelsysteem (c.q. grammatica), dus als het woordlexicon (indien gespecificeerd) uitsluitend morfologisch ongelede representaties bevat, terwijl alle morfologisch gelede woorden via het regelsysteem moeten worden geconstrueerd. Er is sprake van een monistisch morfologiemodel indien lexicon en morfologisch regelsysteem vervlochten zijn in de zin dat de lexicale representatie van morfologisch complexe woorden informatie bevat over de met andere woorden gedeelde eenheden, zoals in het model van Bybee (1985; 1988), of zelfs uit deze eenheden moet worden opgebouwd, zoals in de op overervingsprincipes gebaseerde representatietaal DATR (cf. Gazdar & Evans (1996)).

iii: formeel of informeel

Los van hun andere eigenschappen kan men lexicale kennismodellen globaal in twee kampen onderbrengen, namelijk modellen met informele (primair descriptieve) opzet en modellen met een formele (computationeel implementeerbare) opzet. Hoe formeler een model, hoe beter berekend kan worden welke representaties wel en niet mogelijk zijn, en hoe meer het model correcte voorspellingen kan doen met betrekking tot het procesverloop bij de interpretatie en productie van concrete taaldata (gegeven de empirische resultaten van psycholinguïstisch onderzoek). Generatieve en cognitieve modellen kennen meestal een informele opzet (in de zin dat ze vaak een fragmentarisch karakter hebben en geen toetsbare voorspellingen doen), structuralistische en compositionele (categoriale en typelogische) modellen een formele opzet (waardoor ze altijd toetsbaar zijn binnen het beschreven domein).

iv: representatie, activatie of acquisitie

Lexicale kennismodellen worden meestal in nauwe samenhang met een concreet onderzoeksdomein ontwikkeld. Hierdoor zijn deze kennismodellen doorgaans niet in staat om rekening te houden met de resultaten uit een ander onderzoeksdomein. Hieronder zal ik de meest onderzochte domeinen kort bespreken.

Het onderzoek naar *lexicale kennisrepresentatie* richt zich op de ontwikkeling van een lexiconmodel dat antwoord geeft op de vraag hoe de bestaande woordenschat mentaal wordt gerepresenteerd en meer in het bijzonder hoe de lexicale kennis zich tot het grammaticale regelsysteem verhoudt, d.w.z. in hoeverre het lexicale opslagmechanisme gebruik maakt van de morfologische structuurregels die verantwoordelijk zijn voor de productie en interpretatie van nieuwvormingen (al dan niet op grammaticale grondslag).

Het onderzoek naar *lexicale kennisactivatie* richt zich op de ontwikkeling van een lexiconmodel dat antwoord geeft op de vraag welke stappen er worden doorlopen bij de productie en interpretatie van zowel bestaande als nieuwe woorden. Zo wordt veel onderzoek

gedaan naar de vraag wat de morfologische bouwstenen van het lexicon zijn, welke kwalitatieve en kwantitatieve factoren bepalen of een morfologisch geleed woord integraal wordt opgeslagen, of deze representatie beperkt blijft tot de fonologische of semantische dimensie en hoe deze representaties worden geactiveerd (rechtstreeks of via de morfemen). Bij onderzoek naar woordproductie probeert men onder meer inzicht te krijgen in de factoren die van invloed zijn op de keuze tussen twee verschillende constructiepatronen.

Het onderzoek naar *lexicale kennisacquisitie* richt zich op de ontwikkeling van een lexiconmodel dat antwoord geeft op de vraag hoe het cognitieve representatiesysteem kennis weet te verwerven over het morfologische regelsysteem. Hierbij kan onderscheid worden gemaakt tussen modellen die uitgaan van een aangeboren regelsysteem (c.q. Universele Grammatica) en modellen waarbij de regels langs inductieve weg uit de data moeten worden afgeleid; deze laatste klasse kan nog worden onderverdeeld in voorbeeldgestuurde acquisitiemodellen en zelforganiserende acquisitiemodellen.

2.2.3 Morfologische structuurcriteria

Tot op heden is verre van duidelijk welke structuurcriteria als basis moeten dienen voor de fundering van een morfologische grammatica. Enkele decennia geleden werd veel aandacht besteed aan dit probleem. In dit verband onderstreepte Chomsky (1970) het belang van een 'evaluatiematrix'. Bochner (1993) heeft dit concept aan een grondige studie onderworpen. Hij definieert de evaluatiematrix als een modelinterne verzameling criteria voor het vergelijken van op dit model (c.q. begrippenkader) gebaseerde theorieën van een gegeven verzameling morfologische observaties, met als doel om bij een keuze tussen twee theorieën die descriptief equivalent zijn (d.w.z. dezelfde kennis over bestaande en mogelijke woorden kunnen verantwoorden), aan te geven welke theorie het simpelst is; deze theorie zou altijd de voorkeur verdienen.[30] Er zijn inmiddels zeer uiteenlopende structuurcriteria in omloop. Hoewel deze criteria doorgaans op een informele (niet gekwantificeerde) wijze worden toegepast, is vaak wel een formele definitie beschikbaar. Het onderstaande overzicht is een poging om de belangrijkste structuurcriteria bijeen te brengen en elk van deze criteria semi-formeel te karakteriseren. Dit overzicht is onderverdeeld in twee hoofdklassen, te weten representatiecriteria (zie §1) en identificatiecriteria (zie §2). De representatiecriteria specificeren een algemene methode (c.q. evaluatiematrix) voor de representatie van morfologische structuurkenmerken, maar laten in het midden hoe men deze structuur kan achterhalen. Hiervoor zijn namelijk morfologische identificatiecriteria nodig, d.w.z. criteria die aangeven hoe men individuele morfemen kan identificeren. In dit verband kunnen minstens vier heuristische klassen worden onderscheiden, te weten definities op grond van productiviteit, transparantie, universaliteit en distributiekenmerken. Overigens ben ik mij ervan bewust dat de hier bijeengebrachte criteria zeer uiteenlopende definities kennen en dat er ook allerlei mengvormen bestaan. Toch denk ik dat mijn overzicht een redelijk compleet beeld geeft van de hedendaagse stromingen in de morfologische theorievorming.

2.2.3.1 Representatiecriteria

Minimalisatie van de grammaticale representatiekosten

Bij dit criterium gaat de voorkeur uit naar morfologische regels die tot minimale representatiekosten leiden bij de opslag van kennis over de bestaande woordenschat. Hiertoe dienen deze woorden zoveel mogelijk uit morfemen te worden opgebouwd, want de opslag van losse morfemen (zonder informatie over de bijbehorende woorden) is minder kostbaar dan de opslag van hele woorden. Dit analysecriterium (waarvan de formele definitie door

[30] Dit principe staat bekend als "het scheermes van Ockham": *entie non sunt multiplicanda praeter necessitatem*.

Chomsky (1970) is geïntroduceerd) zou een grammatica moeten opleveren die alle mogelijke woorden kan voorspellen.

Minimalisatie van grammaticale classificatiekosten

Bij dit criterium gaat de voorkeur uit naar morfologische regels die tot minimale classificatiekosten leiden bij de identificatie van bestaande woorden, gegeven een complete inventarisatie van bestaande woorden. Bij de bepaling van de classificatiekosten wordt nagegaan in hoeveel stappen een bestaand woord van een ander bestaand woord kan worden afgeleid, gegeven de reeds bestaande derivatieregels c.q. formeel gedefinieerde woordvormingspatronen waarvan de gebruiksfrequentie boven een nader te bepalen minimum ligt. Dit analysecriterium (waarvan de formele definitie door Jackendoff (1975) is geïntroduceerd) zou een grammatica moeten opleveren die alle mogelijke woorden kan voorspellen.

Maximale compressie van de bestaande woordenschat

Bij dit criterium berust de toekenning van morfologische structuur op het streven naar datacompressie zonder informatieverlies. Dit streven wordt formeel gedefinieerd door het Minimal Description Length-principe (MDL-principe) van De Marcken (1995). Gegeven twee of meer representatiemodellen voor een gegeven verzameling woorden, spreekt dit criterium een voorkeur uit voor het model waarin de som van de woordrepresentaties minimaal is; hierbij geldt de aanvullende eis dat het model reversibel moet zijn, dus dat er geen kennis verloren mag gaan. Dit principe ligt onder meer ten grondslag aan het morfologische compressiemodel van Goldsmith (2000; 2001).

2.2.3.2 Identificatiecriteria

Transparantie

Bij het transparantiecriterium geldt een woord als morfologisch complex indien de eigenschappen van het woord als geheel rechtstreeks (c.q. compositioneel) zijn af te leiden uit de (zonodig gedesambigueerde) bouwstenen c.q. morfemen, dus als elk vormsegment zowel syntactisch als semantisch gezien een voorspelbare functie bezit en als de combinatie van deze morfemen precies de informatie oplevert in het woord als geheel. Hoewel dit in de praktijk een vrij zware eis is, sluit het transparantiecriterium goed aan bij de intuïtieve definitie van een morfeem.

Productiviteit

Bij het productiviteitscriterium identificeert men morfologische derivatieregels door na te gaan of het bijbehorende affix een open toepassingsdomein heeft. Hierbij kan onderscheid worden gemaakt tussen modellen met een kwalitatief (intuïtief) beoordelingscriterium en modellen met een kwantitatief (corpusgebaseerd) beoordelingscriterium. In de kwalitatieve benadering moet worden nagegaan of de regel ook regelmatige nieuwvormingen toelaat, d.w.z. of het affix ook op "nieuwe" stammen kan worden toegepast, c.q. stammen die wel tot het potentiële stamdomein behoren, maar waarvoor de betreffende affixtoepassing nog geen deel uitmaakt van de bestaande woordenschat. In de kwantitatieve benadering moet worden nagegaan of het betreffende affix af en toe in nieuwe (niet eerder waargenomen) woorden opduikt, bijvoorbeeld door een actueel corpus te analyseren op het voorkomen van hapaxtreffers. Dit wordt nader toegelicht in H2.3.5.

Universaliteit

Het universaliteitscriterium is typerend voor generatieve kennismodellen. Bij dit criterium wordt beoordeeld of de voorgestelde structuurregels ook in andere talen voorkomen (mogelijk

in een gegeneraliseerde vorm). Zo ja, dan is er waarschijnlijk sprake van een universele regel, d.w.z. een regel waarvan wordt aangenomen dat deze tot de Universele Grammatica (UG) behoort. Zo nee, dan is er waarschijnlijk sprake van een idiosyncratische taalregel, al zou er ook sprake kunnen zijn van een nog niet opgemerkte generalisatiemogelijkheid.

Distributiekenmerken

In het kader van structuralistisch onderzoek naar de woordvorming is veel energie gestoken in de formulering van morfologische segmentatieregels (cf. Nida (1949); Harris (1955); Peters (1976)). Hiertoe werd systematisch onderzoek gedaan naar de formele distributiekenmerken van de morfologische bouwstenen in woorden met een handmatig aangebrachte (dus cognitief gemotiveerde) morfeemstructuur. Dergelijke heuristieken lenen zich goed voor een universele strategie voor de morfologische analyse van nieuw aangeboden woorden, en indirect voor de inductieve opbouw van een morfologisch regelsysteem.

2.2.4 Lexicografische kennismodellen

Bij de inventarisatie van lexicografische kennismodellen zal ik drie specifieke modelklassen onderscheiden, te weten de fonologisch gestructureerde lexiconmodellen, de semantisch gestructureerde lexiconmodellen en de morfologisch gestructureerde lexiconmodellen. Deze modelklassen corresponderen met de drie fundamentele structuurdimensies van een morfologisch geleed woord, te weten vorm, betekenis en woordinterne structuur. Hoewel het eerste lexicontype veruit het populairst is, bestaat steeds meer belangstelling voor de andere twee lexicontypes (die enkele decennia terug nog moeilijk realiseerbaar waren). Overigens zijn ook nog hele andere ordeningen mogelijk, bijvoorbeeld een ordening op frequentie, op thema (zoals in de *Wat & Hoe*-taalgidsen van de uitgever Kosmos Z&K) of op etymologische kenmerken (zoals jaar van eerste vermelding; cf. Van der Sijs, 2001).

Het fonologisch gestructureerde lexiconmodel

Het fonologisch gestructureerde lexiconmodel ordent woorden op basis van hun orthografische of fonetische vormrepresentatie. Het vormt de grondslag van alle traditionele woordenboeken uit het Nederlandse taalgebied (te weten de grote verklarende woordenboeken en de vertaalwoordenboeken). Ook uitspraakwoordenboeken (zoals het *Uitspraakwoordenboek* van Heemskerk & Zonneveld (2000)), retrograde woordenboeken en woordenboeken met een ordening op klankvorm (zoals de voor kinderen bedoelde *Lijsterbij 3* van Cranshoff & Zuidema (2002)) vallen in deze categorie.

Het semantisch gestructureerde lexiconmodel

Het semantische gestructureerde lexiconmodel ordent woorden op basis van trefwoorden die hun betekenis uitdrukken. Hierbij kan onderscheid worden gemaakt tussen lexica op onomasiologische grondslag, waarin elke betekenis naar één of meer woordvormen c.q. synoniemen leidt, en puur semantische lexica, d.w.z. lexica waarin een reeks trefwoorden in een ontologisch netwerkwordt ondergebracht (zoals een of een thesaurus). De realisatie van dit type ordening was lange tijd een complexe opgave. Voor dit laatste type lexicon is een goed gefundeerd betekenismodel nodig. Belangrijke pioniers op dit terrein zijn Lyons (1977) en Verkuyl (1978; 2000), die allebei een semantisch overervingsmodel presenteren.[31] Ook invloedrijk is het semantische lexiconmodel van Pustejovksy (1991). Een ontologisch lexicon kan voor een groot deel taalonafhankelijk worden ontwikkeld (mits sprake is van enigszins verwante talen). Zo vormt Wordnet[32] (een groot ontologisch lexicon met Engelse woorden) de

[31] Het model van Verkuyl (1978; 2000) komt uitvoerig aan de orde in H4.2.
[32] Zie de website van WordNet: http://wordnet.princeton.edu/w3wn.html/

conceptuele basis voor de lexicale ontologieën die deel uitmaken van EuroWordnet[33] (dat diverse Europese talen omvat, waaronder het Nederlands).[34]

In het Nederlandse taalgebied zijn nog maar weinig woordenboeken verschenen met een onomasiologisch ordeningsprincipe (al kennen elektronische woordenboeken soms wel de mogelijkheid om op betekenis te zoeken). Bestaande titels zijn *Het Juiste Woord* (Spectrum, 1993) en het *Groot Synoniemenwoordenboek* (Van Dale, 2001), dat behalve synoniemen en antoniemen ook hiërarchische conceptrelaties geeft (op basis van VDL's ontologische kennisbank). Maar de bekendste representant van de "omgekeerde woordenboeken" is ongetwijfeld de *Grote puzzelencyclopedie* van Dr. Verschuyl (2003). Hiernaast werkt het Instituut voor Nederlandse Lexicografie (INL) aan de opbouw van een systematisch betekeniswoordenboek, namelijk het Algemeen Nederlands Woordenboek (cf. Moerdijk, 2002).

Het morfologisch gestructureerde lexiconmodel

Doordat arabische en semitische talen van nature een sterk paradigmatische opbouw van de woordenschat kennen, zijn de voor deze talen ontwikkelde woordenboeken van oudsher op een morfologisch ordeningsprincipe gebaseerd: deze woordenboeken presenteren de lexemen namelijk als derivaties van de lexeeminterne stam. Voor woordenboeken van westerse talen werd dit ordeningsprincipe echter zelden toegepast; tot ver in de vorige eeuw beperkten deze pogingen zich tot de opzet van een klein, meestal etymologisch georiënteerd lexicon. Sinds enige tijd is echter een Duits woordenboek beschikbaar dat op een stamgebaseerd ordeningsprincipe is gebaseerd, namelijk het woordfamiliewoordenboek van Augst (1998). Hiernaast heeft het Zwitserse bedrijf Canoo Engineering AG een compleet Duits woordenboek (met 250.000 lemma's) van morfologische structuur voorzien. Gegeven een morfologische stam kan dit elektronisch raadpleegbare woordenboek alle hiervan afgeleide woorden tonen (samen met hun morfologische boomstructuur). Voor dit doel is gebruik gemaakt van Word Manager (Domenig & Ten Hacken, 1992), een tool voor semi-automatische woordanalyse.[35]

Op computationeel terrein wint het morfeemgebaseerde lexicon snel terrein. In Van Eynde & Gibbon (2000) worden bijvoorbeeld diverse computerlexica besproken met morfologische en fonologische informatie over het Nederlands; hierbij wordt veel aandacht besteed aan de computationele methodes waarmee deze lexica tot stand komen. Voor zulke digitale lexica zijn speciale representatiesystemen ontwikkeld, waaronder LFG (Bresnan (1982)), HPSG (Pollard & Sag (1987,1994)) en DATR (Evans & Gazdar (1996)). Deze representatiesystemen maken allemaal gebruik van lexicale overerving. Dit mechanisme maakt het mogelijk om een hele groep woorden van dezelfde grammaticale kenmerken te voorzien door deze eigenschappen aan een overkoepelend type te koppelen, en vervolgens aan te geven welke woorden allemaal onder dit type vallen. DATR is op dit punt het meest flexibel. Deze representatietaal kent namelijk geen grammaticale condities en gaat uit van overschrijfbare defaultregels. Hierdoor kan DATR voor elk gewenste taaleenheid (bijv. morfemen, lexemen of woordgroepen) een typehiërarchie definiëren; bovendien kan per eenheid worden aangegeven waar hij afwijkt ten opzichte van de defaultkenmerken.[36]

[33] Zie de website van EuroWordNet: http://www.illc.uva.nl/EuroWordNet/

[34] Zie Janssen (2002) voor een uitgebreide studie naar de mogelijkheid om zulke kennisnetten als uitgangspunt te nemen voor een automatisch vertaalsysteem.

[35] Het gaat om de *Dictionary of German Morphology* op Canoo-Net. Voor het Engels en Italiaans worden soortgelijke woordenboeken ontwikkeld. Deze lexica verschillen van het door mij ontwikkelde lexicon (de MGBN) doordat ze compleet regelgebaseerd zijn (wat mogelijk is gemaakt door Word Manager).

[36] Dit model vormt de basis voor een digitaal woordenboek met Russische morfologie (cf. Evans & al., 2003).

2.2.5 Cognitieve kennismodellen op dualistische grondslag

2.2.5.1 Introductie

Cognitieve kennismodellen kenmerken zich door het uitgangspunt dat het lexicon alleen structuurloze eenheden bevat en dat er een apart regelsysteem nodig is (hetzij lexicon-extern, hetzij lexicon-intern) om morfologisch complexe woorden te construeren. Simpel gezegd komt het erop neer dat dergelijke modellen een strikt onderscheid maken tussen lexicale eenheden en regelgebaseerde derivaties. Binnen deze klasse van modellen zijn zeer verschillende voorstellen gedaan. Ter bevordering van het overzicht heb ik deze voorstellen in drie hoofdklassen ondergebracht, te weten lexicongenererende en lexiconstructurerende grammaticamodellen en dualistische activatiemodellen. Deze indeling wordt in paragraaf 2.2.5.2 toegelicht en een aantal aanvullende classificatiecriteria bespreken.

2.2.5.2 Classificatiecriteria

Grammaticamodellen versus activatiemodellen

De dualistische kennismodellen kunnen worden onderverdeeld in grammaticamodellen en in activatiemodellen. De overeenkomst is dat beide modeltypes op een dualistisch representatiesysteem zijn gebaseerd (d.w.z. een systeem waarbij het lexicon als grammaticale basis dient voor de constructie van morfologisch complexe woorden). Het verschil is dat grammaticamodellen primair gericht zijn op de vraag hoe men regelmatige woordvormingsprocesssen kan verantwoorden, terwijl activatiemodellen voortkomen uit de doelstelling om antwoord te geven op de vraag hoe de lexicaal opgeslagen kennis geactiveerd wordt, en in het geval van morfologisch complexe woorden, welke factoren bepalend zijn voor de keuze tussen de compositionele activatieroute en de directe (lexicale) activatieroute. Met betrekking tot taalproductie bieden dergelijke modellen de mogelijkheid om te voorspellen wanneer regelmatige woordvormingsproccesen door een lexicaal alternatief worden verdrongen. Zoals gezegd zijn dualistische activatiemodellen rechtstreeks op een grammaticamodel gebaseerd, maar ze onderscheiden zich ervan doordat ze met een parallel zoekmechanisme zijn uitgerust en doordat het lexicon ook frequentie-informatie geeft.

Lexicongenererend versus lexiconstructurerend

Blijkens de studie van Bochner (1993) zijn er in de morfologische theorievorming twee fundamenteel verschillende evaluatiematrixen te onderscheiden, namelijk een benadering die uitgaat van minimale representatiekosten (door maximale toepassing van compositieprincipes) en een benadering die uitgaat van minimale classificatiekosten (door maximale toepassing van overervingsprincipes). De eerste benadering richt zich op de beschrijving van intuïtief gehanteerde woordvormingsregels die de basiswoordenschat (bestaande uit morfologisch ongelede woorden) uitbreiden met grammaticaal mogelijke woorden. De tweede benadering richt zich op de beschrijving van regelmatigheden c.q. redundantiepatronen in de bestaande woordenschat (die zowel ongelede als gelede woorden omvat). In het vervolg zal ik de modellen van de eerste benadering als lexicongenererende (c.q. deductieve) modellen (LG-modellen) aanduiden en de modellen van de tweede benadering als lexiconstructurerende (c.q. inductieve) modellen (LS-modellen). Hiernaast kunnen afgezwakte en gemengde modellen worden onderscheiden.

Syntagmatische versus paradigmatische regels

Het onderscheid tussen lexicongenererende modellen en lexiconstructurerende modellen gaat vaak samen met een tegenstelling tussen syntagmatische en paradigmatische woordvormingsregels. De woordvormingsregels uit LG-modellen kennen namelijk per definitie een lexicon-

onafhankelijke basis en zijn daarom niet in staat om informatie te geven over de vraag welk affixen vaak met dezelfde stam samengaan (dus welke affixen deel uitmaken van een morfologisch paradigma).[37] Dit is wel mogelijk bij LS-regels, want deze regels c.q. redundantiepatronen zijn juist rechtstreeks op relaties tussen bestaande woorden gebaseerd. Hierbij gaat het meestal om relaties tussen een ongeleed en een geleed woord (dus om dezelfde afleidingsrelatie als bij de LG-modellen), maar er is ook ruimte voor regels waarbij sprake is van affixsubstitutie; de op deze wijze bijeengebrachte affixen vormen samen een paradigma.

Productieve versus improductieve affixen

Een derde onderscheid dat van belang is bij de indeling van dualistische woordvormingsmodellen betreft de notie productiviteit. In de morfologische onderzoekspraktijk wordt de grammaticale relevantie van een morfologisch patroon vaak afhankelijk gemaakt van de vraag hoe productief dit patroon is. Dit wordt meestal op intuïtieve basis bepaald, conform de definitie van Schultink (1961):[38] volgens deze definitie is een morfologische regel productief als hij taalgebruikers in staat stelt om onbewust nieuwe woorden aan te maken op basis van bestaande stammen, waarbij de eigenschappen van het nieuwe woord volledig voorspelbaar moeten zijn uit de combinatie van stam en regel. Volgens deze definitie is een morfologische regel dus productief indien het mogelijk is om nieuwe woorden te verzinnen die volgens deze regel zijn gevormd en die niet kunstmatig aanvoelen. Bij de toepassing van dit criterium gaat men er vaak van uit dat woorden die door middel van productieve regels zijn gevormd geen lexicale representatie nodig hebben, maar steeds opnieuw worden afgeleid.

Hoewel het productiviteitscriterium een centrale rol speelt in de morfologische theorievorming, zijn er fundamentele problemen aan verbonden. Ten eerste zijn sommige nieuwvormingen "grammaticaler" dan andere, wat moeilijk is te verklaren indien de beoordeelde regel generiek toepasbaar is. Ten tweede vertonen morfologische grammaticaregels grote verschillen in productiviteit. Zo zijn er diverse corpusstudies verricht (waaronder Anshen & Aronoff (1988) en Baayen (1991)) waaruit blijkt dat sommige woordvormingsregels veel vaker worden toegepast dan andere. Naar aanleiding hiervan heeft Baayen (1989; 1990), die zelf weer voortbouwt op pionierswerk van Al & Booij (1981), voorgesteld om de productiviteit van een affix gelijk te stellen aan het aantal met dit affix gevormde hapax-woorden (c.q. neologismes) per miljoen woorden, wat men kan achterhalen door tellingen te verrichten aan een morfologisch geannoteerd tekstcorpus (zie ook H2.3.5). Een laatste probleem met betrekking tot het productiviteitscriterium is dat moeilijk valt hard te maken dat bestaande woorden steeds opnieuw worden afgeleid; dit is in elk geval strijdig met de observatie dat mensen onderscheid kunnen maken tussen gangbare en mogelijke woorden. In de praktijk wordt dit verantwoord door te stellen dat de grammaticale module met een kennismodel correspondeert; alle aanvullende kennis zou door een pragmatische module moeten worden verantwoord.

[37] Deze beperking geldt niet voor inflectie, want in tegenstelling tot derivatieregels worden inflectieregels vaak tot inflectieparadigma's samengevoegd. Deze paradigmatische organisatie berust op de observatie dat alle lexemen van eenzelfde categorie (zoals de verba en de nomina) in beginsel dezelfde inflectiemogelijkheden bezitten (ongeacht hun morfologische structuur). Bovendien heeft taalvergelijkend onderzoek uitgewezen dat de samenstelling van deze inflectieparadigma's tamelijk constant is (d.w.z. op functioneel niveau), wat als een aanwijzing geldt dat er sprake is van een UG-gebaseerd organisatiepatroon.

[38] Schultink (1994) meldt overigens dat de nu internationaal gangbare notie 'morfologische productiviteit' zijn oorsprong vindt in een Nederlandse onderzoekstraditie die terugvoert naar het werk van Uhlenbeck; cf. Uhlenbeck (1953; 1977), Schultink (1961).

Universele versus specifieke grammaticaregels

Zowel bij dualistische als bij monistische kennismodellen is een onderscheid mogelijk tussen generatieve modellen die uitgaan van een Universele Grammatica (UG) en overige modellen. UG-gebaseerde kennismodellen veronderstellen dat een kind aangeboren kennis heeft over de door hem te leren grammaticaregels. Dergelijke modellen kenmerken zich door een derivationele opzet, waarbij een structureel onderscheid wordt gemaakt tussen een waarneembare morfeemvorm (c.q. oppervlaktevorm) en een grammaticale morfeemvorm (c.q. onderliggende vorm). Hierdoor kan de bijbehorende grammatica niet rechtsteeks uit de data worden afgeleid en is het omgekeerd ook lastig om deze grammatica empirisch te toetsen.

2.2.5.3 Lexicongenererende grammaticamodellen

In de grammaticale traditie die teruggaat op Bloomfield (1933) dient de grammatica alle voorspelbare eigenschappen van een taal te beschrijven, terwijl het lexicon alle niet-voorspelbare eigenschappen moet verantwoorden. Toegepast op de woordenschat betekent dit dat het lexicon uitsluitend ondeelbare morfemen (namelijk stammen en affixen) opslaat, terwijl de grammatica aangeeft hoe deze eenheden gecombineerd moeten worden. Deze aanpak wordt gemotiveerd door het idee dat lexicale kennis kostbaar is, dus dat het lexicon zo compact mogelijk moet worden gehouden. Zoals Bochner (1993) uitlegt, correspondeert deze aanpak met een evaluatiecriterium waarbij de totale informatie-inhoud van het lexicon (bestaande uit stammen en affixen) minimaal is (gemeten in termen van bits), dus waarbij naar minimale representatiekosten wordt gestreefd. Hierdoor verdient een grammatica die geen morfologisch complexe woorden opslaat, maar alleen hun (atomaire) morfemen, de voorkeur boven een grammatica waarin het lexicon ook woordvormen opslaat die morfologisch complex zijn.

Structuralistisch

De morfologische visie van Bloomfield (1933) heeft veel navolging gekregen in structuralistische (door De Saussure (1916) geïnspireerde) studies van de woordvorming, al richten deze studies zich primair op de vraag welke morfologische patronen men kan waarnemen en pas in de tweede plaats hoe deze patronen in een overkoepelend regelsysteem kunnen worden ondergebracht. Belangrijke representanten van deze aanpak zijn Bolinger (1948), Newman (1948), Nida (1949), Jespersen (1949-1958), Harris (1955; 1967), Hockett (1958), Marchand (1969), Matthews (1972) en Peters (1976; 1983). Enkele Nederlandse studies die vermelding verdienen zijn Uhlenbeck (1953; 1977), Schultink (1961; 1962), Mattens (1970) en De Vries (1975).

Klassiek-Generatief

Het morfeemgebaseerde regelmodel is ook dominant in het generatieve (UG-gedreven) onderzoek naar de woordvorming. Hierbij wordt een lexicalistisch morfologiemodel gehanteerd, d.w.z. een model waarin de woordvorming als een afzonderlijke component van de grammatica wordt gezien. Deze onderzoeksrichting vindt zijn oorsprong in Chomsky & Halle (1968), en is nader uitgewerkt in Chomsky (1970). Het in deze studies gehanteerde kennismodel berust op het idee dat morfologische constructies net als syntactische constructies het resultaat zijn van een reeks transformaties, die als doel hebben om een dieptestructuur (namelijk de grondvorm van de stam) stap voor stap met affixen uit te breiden, hetgeen uiteindelijk in een uitspreekbare en interpreteerbare woordrepresentatie dient te resulteren. Voor dit doel worden gespecialiseerde grammaticaregels gedefinieerd. Deze grammaticaregels hebben een globaal karakter in de zin dat ze de hele derivatie kunnen overzien en op basis van deze informatie specifieke aanpassingen kunnen aanbrengen (zoals de toekenning van klemtoon of de selectie van een allomorf). Dit uitgangspunt heeft navolging gevonden in de klassiek-generatieve studies van (onder meer) Siegel (1974), Lieber (1980), Williams (1981), Kiparsky (1982) en Trommelen & Zonneveld (1986).

Compositioneel

In een fundamenteel verschillende uitwerking vormt het morfeemgebaseerde regelmodel ook de basis van morfologiemodellen die voortbouwen op de compositionele, modeltheoretisch gefundeerde grammaticavisie van Montague (1974), zoals de categoriale morfologiemodellen van Dowty (1979), Moortgat (1981), Hoeksema (1984), Van der Wouden (1988) en Heemskerk (1993), en de typelogische morfologiemodellen van Moortgat & Van der Hulst (1981) en Verkuyl (1993). Deze studies kenmerken zich door een modeltheoretische opzet.[39] Een belangrijk verschil met de grammaticamodellen uit de Bloomfieldiaanse traditie is dat de compositionele modellen geen afzonderlijke grammatica postuleren, maar alle combinatiepatronen via het lexicon verantwoorden. Hierbij wordt een modeltheoretische aanpak gehanteerd.

In zo'n modeltheoretische aanpak wordt voor elk affix (of meer in het algemeen, voor elke functor) een functie gespecificeerd. In de meest eenvoudige vorm legt zo'n functie een verband tussen een broncategorie (bijv. A) en een doelcategorie (bijv. B), waarbij beide categorieën met een open verzameling taaleenheden corresponderen, namelijk alle morfemen en morfeemcombinaties die de opgegeven categorie bezitten. De bijbehorende functie F kan worden gedefinieerd als F: A→B. Hierbij kunnen categorie A en B zowel met een simpele als met een samengestelde categorie corresponderen. In categoriale modellen corresponderen deze categorieën met de lexicale hoofdklassen, namelijk N, V, A of P (of een hiervan afgeleide combinatie, zoals <V,N>); in typelogische modellen bestaan deze categorieën uit combinaties van entiteiten en waarheidswaarden t, bijv. de functie <<e,t>,t>. Een functor kan ook meerdere broncategorieën specificeren. Zo vereist het suffix -ER in *hoogwerker* zowel een A (*hoog*) als een V (*werk*) om een N met de structuur $[A + V + ER]_N$ te genereren.

Het voordeel van de hier beschreven representatiemethode is dat hij een zeer exacte afbakening mogelijk maakt van het potentiële toepassingsdomein van de hiermee gelabelde functors. Het nadeel is dat de onderliggende types erg abstract zijn, al gauw complex (en daarmee onleesbaar) worden en geen informatie geven over de lexicale eigenschappen van de hiermee gelabelde morfemen, maar alleen over hun syntactische of semantische subcategorisatieframe.

Gesplitst

Een belangrijk bezwaar tegen het klassieke LG-model is dat de regelgebaseerde benadering van morfologisch complexe woorden problemen oplevert bij de verantwoording van onregelmatige inflectievormen. Want hoewel men ervan uitgaat dat inflectieprocessen per definitie productief zijn, blijken er in de praktijk allerlei afwijkingen voor te komen met betrekking tot de vormkeuze. Om die reden is voorgesteld om een principiële scheiding aan te brengen tussen de toekenning van een morfologische inflectiecategorie (als onderdeel van de grammaticale constructie van een predicaat) en de specificatie van de bijbehorende vorm (als onderdeel van het fonologische specificatieproces). Dit gesplitste morfologiemodel (c.q. *split morphology model*) vindt zijn wortels in het werk van Anderson (1982). Het leidde later tot de formulering van de Separation Hypothesis van Beard (1991). Het vormt ook een belangrijk uitgangspunt van het Distributed Morphology-model van Halle & Marantz (1993), dat veel invloed heeft gekregen. In hun interpretatie loopt de scheiding van grammaticale structuur en representatievorm echter tegen fundamentele problemen aan (cf. Booij, 1994). Dit geldt niet voor de interpretatie van Ackermann & Webelhuth (1998).[40] Het door hun voorgestelde HPSG-model is voortgekomen uit het streven om een adequate analyse van scheidbaar samengestelde werkwoorden te geven. In tegenstelling tot Halle & Marantz gaan ze ervan uit

[39] Gamut (1991) biedt een algemene introductie tot deze klasse van modellen.
[40] Hun theorie heeft navolging gevonden in Sadler & Spencer (2001).

dat de predicatiestructuur volledig losstaat van de morfosyntactische representatievorm (en meer in het bijzonder de keuze tussen een representatie als woord of als woordgroep). Hierdoor wint hun model aanzienlijk aan kracht. Maar net als andere gesplitste morfologiemodellen is dit model niet in staat om een directe relatie leggen tussen de morfologisch gelede predicatiestructuur en de bijbehorende representatievorm. In plaats daarvan gaan Ackermann & Webelhuth ervan uit dat er een aparte module bestaat die als taak heeft om voor elke morfologische stam een register bij te houden waarin kan worden opgezocht voor welke morfologische toepassingen van de geactiveerde woordstam een speciale representatievorm bestaat; bij de overige categorieën kan worden teruggevallen op de mogelijkheid om deze representatievorm op afroep te construeren. Een dergelijk systeem heeft als nadeel dat het veel redundantie oplevert bij de specificatie van morfologisch onregelmatige woordvormen.

Minimalistisch

Sinds Chomsky (1995) wordt het generatieve onderzoek gedomineerd door het idee dat de grammatica zo minimalistisch mogelijk moet worden ingericht, wat inhoudt dat de grammatica alleen de meest noodzakelijke structuurprincipes dient te specificeren en dat verder zoveel mogelijk combinatorische restricties via het lexicon moeten worden verantwoord (al dienen deze ook hier zoveel mogelijk achterwege te blijven); in dit opzicht vertoont de minimalistische aanpak conceptuele overeenkomst met de Montague-benadering.[41] Chomsky's minimalistische programma heeft de weg vrij gemaakt voor morfologische kennismodellen die uitgaan van een gedegenereerd (c.q. "impoverished") lexicon (bijv. Borer (2000) en Marantz (2001)). Deze benadering kenmerkt zich door een morfeemlexicon waarvan de ingangen uitsluitend informatie geven over niet-grammaticale basiskenmerken als de grondvorm en de betekenis; de lexicale categorie en andere grammaticale eigenschappen mogen dus ongespecificeerd blijven (wat een grote verandering is ten opzichte van de klassiek-generatieve morfologiemodellen).

De op deze wijze gedefinieerde morfemen kunnen vrijelijk in gefaseerd opgebouwde derivatieframes worden geïnserteerd, zolang hun kenmerken maar niet botsen met de selectierestricties van de frame-posities, met als resultaat een "anything goes"-model van de woordvorming. Hierdoor wordt een adequate verantwoording mogelijk voor het ontstaan van woorden waarbij de stam een nieuwe categorie of betekenisklasse lijkt aan te nemen, dus bij nieuwvormingen op basis van onvoorspelbare stamconversies (zoals in het werkwoord *geinen*, waarvan de stam met het nomen *gein* correspondeert, of het werkwoord *kukelekuen*, wat van de uitroep *kukeleku* komt) en nieuwvormingen op basis van een afkorting, bijv. *sms'en* en *ge-cc'd*. Hoewel de hier geschetste benadering interessante innovaties heeft opgeleverd, kleeft er een fundamenteel nadeel aan, namelijk dat dergelijke modellen niet of nauwelijks toetsbaar zijn, aangezien er (met een beroep op intuïtieve kennis) veel meer woordderivaties worden toegestaan dan er daadwerkelijk voorkomen, terwijl (gegeven de bestaande woordenschat) geen onderscheid wordt gemaakt tussen alledaagse en bijzondere derivaties.

Optimaliserend

Er zijn ook dualistische morfologiemodellen die uitgaan van optioneel toepasbare (c.q. schendbare) regels. Dit principe is onderdeel van de Optimaliteitstheorie (c.q. "Optimality Theory"; cf. Prince & Smolensky (1993); McCarthy & Prince (1993)). Deze theorie heeft als doel om een verklaring te geven voor het feit dat er grote variatie bestaat in de toepasbaarheid van grammaticale regels op concrete voorbeeldwoorden of voorbeeldzinnen, terwijl meestal interpersoonlijke overeenstemming bestaat over het grammaticale kwaliteitsniveau (in termen van goed, redelijk, matig, slecht). Dit wijst erop dat grammaticale patronen niet absoluut

[41] Zie Bierwisch (1996) voor een diepgaande beschouwing van de fundamenten van het minimalisme.

toepasbaar zijn, maar dat hun toepasbaarheid van tal van dataspecifieke factoren afhangt. Het Optimaliteitsmodel tracht dit type variatie te verantwoorden door clusters van condities (c.q. *constraints*) vast te stellen waarbinnen een bepaalde hiërarchie c.q. conditie-ordening bestaat. Hoe hoger een conditie geordend is, hoe minder de toepasbaarheid van deze conditie door andere factoren wordt beïnvloed (dus hoe robuuster de conditie is). Deze benadering werkt alleen als er voldoende condities worden geïntroduceerd, waarbij een deel van deze condities conflicterend gedrag dient te vertonen. Meestal wordt aangenomen dat deze condities aangeboren zijn en dat taallerende kinderen alleen maar hoeven na te gaan hoe ze onderling moeten worden geordend. De hier geschetste benadering leidt echter tot een proliferatie van ad hoc-condities, hetgeen fundamenteel strijdig is met de uitgangspunten van Bochner's evaluatiematrix. Niettemin is dit type verklaringsmodel in korte tijd zeer populair geworden.

2.2.5.4 Lexiconstructurerende grammaticamodellen

Naast lexiconmodellen die streven naar minimalisatie van de representatiekosten bestaat er ook een klasse van lexiconmodellen die naar minimalisatie van de classificatiekosten streven. Dit modeltype, dat voor het eerst werd voorgesteld door Jackendoff (1975)[42] en verder is uitgewerkt door Bochner (1993), berust op het uitgangspunt dat alle gangbare woorden integraal in het lexicon zijn opgeslagen en dat dit mentale lexicon als basis dient voor de opbouw van een systeem van redundantieregels. Deze redundantiegrammatica zou als enige taak hebben om de productie en herkenning van morfologische[43] nieuwvormingen te faciliteren (want de reeds bekende woorden kunnen rechtstreeks uit het lexicon worden gehaald).

Volgens Jackendoff (1975) dient zo'n grammaticamodel niet beoordeeld te worden op de totale informatie-inhoud van regels en lexicon, maar op het aantal classificatiekeuzes (c.q. keuzes tussen clusters van kenmerken) dat nodig is om alle aangeboden woorden uniek te kunnen identificeren. Zo kan men onderscheid maken tussen N-woorden en V-woorden, en binnen de N-woorden kan men subklasses onderscheiden van woorden met het suffix -ING, het suffix -ER en het suffix -SEL; bij elk suffix kunnen bovendien weer een aantal subklassen worden onderscheiden. Op deze wijze kan een al dan niet hiërarchisch opgebouwd stelsel van keuzemogelijkheden worden uitgewerkt waarin alle redundante woordkenmerken slechts eenmaal hoeven te worden gespecificeerd, zodat de hoeveelheid informatiekeuzes per woord minimaal kan blijven.

Het door Jackendoff (1975) voorgestelde representatiesysteem ligt in een iets andere vorm ook ten grondslag aan het morfologische kennismodel van Aronoff (1976). Een belangrijk verschil is dat het regelsysteem van Aronoff puur morfologisch is georiënteerd en dat deze regels aan een aantal grammaticale restricties zijn onderworpen. Bovendien gaat Aronoff ervan uit dat het lexicon geen plaats biedt aan woorden die op een strikt-productieve derivatie berusten. Door deze aannames is Aronoff's model wat "generatiever" van aard dan dat van Jackendoff, met als gevolg dat Aronoff's model meer navolging heeft gekregen; zo zijn diens ideeën direct of indirect terug te vinden in Booij (1977), Moortgat (1981) (zij het in een categoriale vertaling) en Van Santen (1992). Andere studies die uitgaan van een lexiconstructurerend kennismodel zijn Meijs (1985), Van Marle (1985; 1986), Ford en Singh' (1991; 1997), Neuvel (2000; 2002) en Neuvel & Fulop (2002).

Een belangrijk voordeel van het lexiconstructurerende kennismodel is dat het bijbehorende regelsysteem een inductieve basis (c.q. analogiebasis) heeft, dus dat geen aangeboren regelkennis hoeft te worden verondersteld, al zijn wel taalonafhankelijke segmentatiecriteria nodig

[42] Maar het fonologisch georiënteerde kennismodel van Venneman (1974) gaat reeds in dezelfde richting.

[43] De term "morfologisch" verwijst hier niet alleen naar afleidingen op basis van transparante affixen (te weten, eenheden die een vaste vorm-betekenis-relatie bezitten), maar ook naar afleidingen waar slechts een semantisch of fonologisch kenmerk wordt toegevoegd.

voor de extractie van redundantiepatronen; voor dit doel zou men gebruik kunnen maken van de heuristische principes van Nida (1949) of Harris (1955). Het nadeel van dit modeltype is dat er een sterk redundant (dus inefficiënt) lexicon voor nodig is, aangezien alle waargenomen woorden integraal moeten worden opgeslagen. Daar komt bij dat Jackendoff's model geen principieel onderscheid maakt tussen morfologische en niet-morfologische woordkenmerken, wat strijdig is met de taalkundige intuïtie dat morfemen een centrale rol spelen in de opbouw van het grammaticale systeem.

In mijn optiek worden grammaticamodellen die uitgaan van het minimale classificatiecriterium pas aantrekkelijk als aan de hier genoemde bezwaren tegemoet wordt gekomen. Hiertoe zou het classificatiesysteem geïntegreerd moeten worden met de informatie in het lexicon, zodat er ook representationeel gezien sprake is van informatiereductie. Lexiconmodellen die uitgaan van constructionistische principes, zoals HPSG en LFG (cf. subH2.2.4), komen deels aan deze kritiek tegemoet, maar beperken zich meestal tot overerving van de niet-fonologische (namelijk syntactische en semantische) kenmerken. Het enige mij bekende representatiesysteem dat volledige overerving van gedeelde woordkenmerken toelaat is DATR (Evans & Gazdar (1996)). Maar los van het feit dat DATR niet meer is dan een lexicale programmeertaal kent dit representatiesysteem een principieel statische opzet. Hierdoor zijn DATR-lexicons niet zonder meer verenigbaar met het streven naar een zelflerend representatiesysteem.

2.2.5.5 Dualistische activatiemodellen

In de voorgaande secties heb ik twee belangrijke klassen van dualistische morfologiemodellen besproken, namelijk het lexicongenererende en het lexiconstructurerende grammaticamodel. Het eerste modeltype heeft als fundamenteel nadeel dat het lexicon geen plaats biedt aan morfologisch complexe woorden, waardoor het niet mogelijk is om onderscheid te maken tussen gangbare en mogelijke woorden. Omgekeerd heeft het tweede modeltype het nadeel dat geen morfologische structuur beschikbaar is voor opgeslagen woorden (al kan deze wel via redundantieregels worden toegekend). Hierdoor is geen van beide modeltypes in staat om een verklaring te geven voor het bestaan van vormvarianten en lexicale blokkades, een fenomeen dat reeds door Uhlenbeck (1953) werd opgemerkt. Het gaat hierbij om een concurrentiestrijd tussen grammaticaal mogelijke woorden en gangbare woorden, waarbij de eerste soms verdrongen worden door de laatste: zo heet iemand die een vuurwapen gebruikt geen *schieter* maar een *schutter*. Maar als sprake is van een samenstelling blijkt het rechterlid ook met de vorm *schieter* te kunnen corresponderen, blijkens *harpoenschieter*, *lijnschieter*, *boutschieter* en *busschieter*. Wat betreft de betekenis "schietinstrument" geldt dat het lexeem *schieter* niet zelfstandig voorkomt, maar wel in *proppenschieter* en *pillenschieter*.

Ter verantwoording van dit soort beperkingen heeft Halle (1973) een kennismodel voorgesteld waarin naast het grammaticale morfeemlexicon ook plaats is voor een woordlexicon waarin taalgebruikers alle reeds waargenomen woorden kunnen opslaan. Hierbij probeert Halle het gebruik van de grammaticaal beschikbare woordenschat in te perken door een filter in te bouwen dat toegang heeft tot het lexicon met de opgeslagen woorden. Halle's voorstel heeft (evenals de hierna te noemen modellen) als nadeel dat het niet compatibel is met het streven naar lexicale economie. Dit impliceert dat de identificatie van morfemen niet langer gemotiveerd kan worden met een beroep op de economische opslag van lexicale kennis, maar alleen met een beroep op het bestaan van een productief woordvormingsmechanisme.

Ernstiger is dat dit voorstel op een aantal fundamentele bezwaren stuit (cf. Booij (1977) en Bochner (1993)). Halle's voorstel heeft dan ook geen school gemaakt. In plaats daarvan werden verscheidene voorstellen gepubliceerd voor de opzet van een grammaticaal concurrentiemodel (cf. Aronoff (1976), MacWhinney (1978) en Meijs (1985)). Zo'n model bestaat

uit twee componenten, namelijk een regelgebaseerd representatiesysteem (al dan niet op analogiebasis) en een woordgebaseerd representatiesysteem. Verder kent zo'n model een zoekmechanisme dat nagaat wat de aantrekkelijkste activatieroute is (meestal de route die het eerste resultaat oplevert).

Inmiddels heeft deze benadering steun in de rug gekregen vanuit corpusgebaseerd productie-onderzoek, zoals Anshen & Aronoff (1988), Baayen (1989; 1990), Baayen & Lieber (1991) en Plag (1999), vanuit psycholinguïstisch onderzoek naar woordherkenning en woordproductie, namelijk Frauenfelder & Schreuder (1992), Taft (1994), Baayen, Schreuder & al. (1995, 1997, 2000, 2002), Krott (2001), De Jong (2002) en Hay & Baayen (2002), en vanuit onderzoek naar tweede taalverwerving (cf. Lowie (1998)). Het concurrentiemodel is zo succesvol dat het ook invloed heeft op grammaticale studies; zo melden Booij & Van Santen (1998) dat een adequaat grammaticamodel van de woordvorming het best op een morfologisch concurrentie-model kan worden gebaseerd.

Volledigheidshalve dient te worden opgemerkt dat er ook andere voorstellen zijn gedaan. Zo zijn er tal van onderzoekers die ervan uitgaan dat alle gangbare woorden in het mentale lexicon staan en dat taalgebruikers daarom eerst hun lexicon raadplegen en pas regels gaan toepassen als het gezochte woord hier niet in blijkt te staan; dit idee is terug te vinden bij Butterworth (1983), Caramazza, Laudanna & Romani (1988) en Laudanna & Burani (1995), en in een wat andere vorm ook bij Giraudo & Grainger (2001; 2003). Er is ook een stroming (cf. Pinker & Prince (1988; 1994), Marcus (1995), Pinker (1998) en Clahsen (1999)) die ervan uitgaat dat er een strikte scheiding bestaat tussen productief gevormde woorden (die altijd via regels moeten worden gevormd) en overige woorden (die rechtstreeks uit het lexicon kunnen worden gehaald, of – in het geval van improductieve nieuwvormingen – analoog aan de al bekende woorden worden afgeleid). Uit Baayen, Schreuder & al. (2002) blijkt echter dat dit idee zowel theoretisch als empirisch onhoudbaar is.

2.2.6 Cognitieve kennismodellen op monistische grondslag

2.2.6.1 Introductie

Empirische kennismodellen op monistische grondslag kenmerken zich door het uitgangspunt dat het lexicon met een morfologisch gestructureerd netwerk correspondeert, dus dat de lexicale representaties van morfologisch complexe woorden interne structuur bezitten (al hoeft deze structuur niet rechtstreeks waarneembaar te zijn). Dit impliceert dat geen apart regelsysteem hoeft te worden geïntroduceerd, want een netwerk kan hetzelfde effect bereiken door analogisch gebruik te maken van lexicale patronen. De grote kracht van een netwerk schuilt in de mogelijkheid om het data-aanbod parallel te verwerken (aangezien elke knoop in een netwerk met een miniprocessor correspondeert).

Net als dualistische kennismodellen zijn monistische kennismodellen vaak voor een specifiek onderzoeksgebied ontworpen en kan zo'n model niet zomaar voor een andere functie worden gebruikt. Hierbij gaat het meestal om kennisverantwoording, kennisactivatie of kennisverwerving en in een enkel geval (ook) om taalverandering. Deze sectie biedt een overzicht van de belangrijkste modelvarianten. Het hierbij gehanteerde classificatiesysteem berust op de volgende drie opposities: 1) localistisch versus distributief, 2) voorbeeldgestuurd versus zelforganiserend en 3) symbolisch versus subsymbolisch. In §2 worden deze opposities kort toegelicht. Vervolgens worden vijf localistische netwerkmodellen besproken (in §3) en drie distributieve netwerkmodellen (in sectie §4).

2.2.6.2 Het classificatiesysteem

Localistisch vs. Distributief

Bij mijn bespreking van het netwerkmodel zal ik twee hoofdtypes onderscheiden, namelijk het localistische netwerkmodel en het distributieve netwerkmodel. Voor beide modeltypes geldt dat Rumelhart en McClelland de eerste taalkundige toepassing hebben gepubliceerd, respectievelijk in McClelland & Rumelhart (1981) (waarin de perceptiemogelijkheden van een localistische netwerkmodel worden beschreven) en Rumelhart & McClelland (1986) (waarin de leermogelijkheden van een distributief netwerkmodel worden beschreven). Deze modeltypes vertonen enkele cruciale verschillen.

Het eerste verschil is dat localistische netwerken de aangeboden gegevens aan een unieke geheugenplaats koppelen (zodat er een 1:1-relatie bestaat tussen gegevens en geheugenplaatsen), terwijl distributieve netwerken deze gegevens gespreid opslaan (zodat er een 1:n-relatie bestaat tussen gegevens en geheugenplaatsen); omgekeerd kan elke geheugenplaats vele woordrepresentaties ondersteunen. Deze gespreide opslag gaat samen met een generalisatieproces, met als gevolg dat de oorspronkelijke gegevens niet rechtstreeks zijn terug te vinden (dus niet reproduceerbaar zijn). Distributieve netwerken zijn dan ook veel sterker in spontane generalisaties dan in kennisreproductie. Een ander nadeel is dat ze uitsluitend taakgericht kunnen leren, terwijl localistische netwerken op dit punt flexibel zijn.

Voorbeeldgestuurd vs. zelforganiserend

Het onderscheid tussen localistische en distributieve netwerkmodellen gaat vaak samen met een contrast in de instructiewijze (c.q. leerproces): want distributieve netwerken vereisen meestal een voorbeeldgestuurde trainingsfase, terwijl een localistische netwerk permament kennis kan verwerven en deze zelf kan structureren (dus zelforganiserend is); hierbij kan de opgeslagen kennis interactief worden geactiveerd, wat inhoudt dat elk deelkenmerk van een gegeven representatie tot activatie van de representatie als geheel kan leiden.

Indien sprake is van een voorbeeldgestuurd leerproces dient het netwerk eerst een trainingsfase te doorlopen waarin het een representatieve verzameling voorbeelden of voorbeeldanalyses krijgt aangeboden; hierbij dient elk brongegeven aan een doelgegeven te zijn gekoppeld (bijvoorbeeld een taalkundig geanalyseerde representatie van het brongegeven). Na deze trainingsfase kan zo'n netwerk geen nieuwe kennis meer vastleggen (tenzij een nieuwe trainingsfase wordt ingelast).

Symbolisch vs. Subsymbolisch

Zowel bij monistische als bij distributieve netwerkmodellen kan men onderscheid maken tussen symbolische en subsymbolische modellen. Symbolische modellen kenmerken zich door een expliciete, computationeel implementeerbare representatie van woordinterne kenmerken, terwijl subsymbolische modellen doorgaans volstaan met een schema of met statistische gegevens.

2.2.6.3 Localistische netwerkmodellen

1. McClelland & Rumelhart (1981) waren de eerste onderzoekers die een taalkundige toepassing van een netwerkmodel presenteerden. Voor deze toepassing maakten ze gebruik van een interactief activatienetwerk op localistische basis. Een dergelijk netwerk kenmerkt zich door een verzameling geheugenlocaties die onderling zijn verbonden door paden met een dynamisch instelbare activatiewaarde. Dit model biedt de mogelijkheid om activatieprocessen te simuleren, want indien men één of meer van de opgeslagen gegevens activeert, kan dit via

een interactief proces (waarbij elk pad aanpasingen in zijn activatiewaarde ondergaat) tot activatie van een gecombineerd gegeven leiden.

2. Taft (1994) betoogt dat het interactieve activatiemodel zich goed leent voor de modellering van woordherkenningsprocessen. In het door Taft beschreven model kunnen morfologisch complexe woorden uitsluitend via hun morfemen worden geactiveerd (ook als deze woorden uit pseudomorfemen bestaan). Hierbij is de herkenningssnelheid van het woord als geheel een functie van de gebruiksfrequentie van de morfemen in het te activeren woord. Volgens Taft biedt dit model een verklaring voor de observatie dat de herkenning van morfologisch complexe woorden niet alleen afhangt van de frequentie van de samenstellende morfemen, maar ook van de frequentie van het woord als geheel. In het model van Taft zou activatie van het morfeem VOED bijvoorbeeld tot een verhoogde activatie van de woorden *voeding*, *voedsel* en *voeden* leiden, maar niet van *foedraal*. Als vervolgens een suffix wordt gepresenteerd kan de woordfrequentie doorslaggevend zijn voor de snelheid waarmee het woord herkend wordt. Dus hoewel -ING veel vaker voorkomt dan -SEL, zou het woord *voedsel* toch sneller herkend kunnen worden dan *voeding*. Taft's interactieve activatiemodel heeft tot nu toe weinig navolging gevonden, mogelijk omdat men over het hoofd ziet dat dit model ook is uitgerust met een duaal activatiesysteem (c.q. concurrentiemechanisme); dit impliceert dat Taft's model minstens even krachtig is als het activatiemodel van Schreuder & Baayen (1995).

3. Het localistische netwerkmodel speelt een centrale rol in het mentale lexiconmodel van Bybee (1985; 1988; 1995; 2001). Deze auteur gaat ervan uit dat het mentale lexicon met een morfologisch gestructureerd netwerk correspondeert, d.w.z. met een localistisch netwerk waarin elke knoop met een woordintern kenmerk correspondeert en waarbij relaties kunnen ontstaan tussen woorden met gemeenschappelijke kenmerken (in de vorm van een verhoogde activatiewaarde van de verbindende paden).[44] Bybee illustreert dit idee door middel van een grafisch schema. In essentie kent dit schema de volgende opzet:[45]

a	**b**	c	d
e	**b**	**f**	
	b	g	

Dit schema toont drie abstracte woordrepresentaties, te weten [abcd], [ebf] en [bg]. Hierbij zouden de letters bijvoorbeeld met morfemen kunnen corresponderen. Deze woorden bezitten dan één gemeenschappelijk morfeem, te weten b (bijv. de stam VOED). Als men nu het woord [ebf] aanbiedt (bijv. het woord OP+VOED+ING), zal dit eerst tot activatie van de lexicale representatie [ebf] leiden (wat is weergegeven door vetgedrukte letters), waarna het morfeem b weer invloed kan uitoefenen op het morfeem b in andere woordrepresentaties (wat eveneens door een vetgedrukte letter is weergegeven). Dit heeft weer als gevolg dat die andere woorden een hogere activatiewaarde krijgen en dus sneller herkend zullen worden. Dergelijke activatiemechanismes spelen een grote rol bij Bybee's analyse van semiproductieve en improductieve woordvormingsprocessen. De bijbehorende representaties maken alleen indirect gebruik van morfemen. Want net als Seidenberg & Gonnermann (2000) gaat Bybee ervan uit dat morfologische structuur niets anders is dan het virtuele product van convergerende codes in een localistisch netwerk, dus dat morfologische structuur als een epifenomeen kan worden gezien van relaties tussen niet-morfologische deelrepresentaties.

4. Skousen (1989) biedt een formeel uitgewerkt analogiemodel, waarbij hij voortbouwt op zijn eerdere ideeeën (Skousen, 1979). Net als Bybee (1985) gaat Skousen ervan uit dat het mentale lexicon met een localistisch netwerk correspondeert, waarbij de sterkte van de

[44] Dit dynamische taalleermodel is verwant aan het psychologische leermodel van Van Parreren (1971).
[45] Bybee's voorkeur voor constructieschema's is reeds zichtbaar in Bybee & Slobin (1982).

onderlinge verbindingen bepalend is voor de mate van morfologische verwantschap. Dit lexicon heeft de mogelijkheid om relevante constructiepatronen te identificeren door de opgeslagen woorden op alle mogelijke manieren in kleinere eenheden op te delen en vervolgens via een ingewikkeld telproces na te gaan welke subpatronen relevant zijn voor de woordvorming.[46] Deze kennis kan worden benut om nieuwe woorden te analyseren. Het door Skousen voorgestelde model wordt toegelicht en experimenteel getest in Derwing & Skousen (1989; 1994). Uit Krott (2001) blijkt dat Skousen's analogiemodel een adequate simulatie kan geven van haar experimentele resultaten bij enkele woordproductietaken.

5. Riehemann (1998) presenteert een HPSG-beschrijving van de derivationele mogelijkheden van het Duitse affix -BAR. Hiertoe heeft ze een overervingsmodel uitgewerkt waarin de selectierestricties van -BAR zeer gedetailleerd kunnen worden gespecificeerd. Hiermee toont ze aan dat het mogelijk is om een grammaticaal model op te zetten waarin op het oog onregelmatige derivatieprocessen toch langs structurele weg kunnen worden verantwoord. Deze analyse berust net als de voorgaande modellen op een localistisch netwerkmodel, maar verschilt ervan door de aanwezigheid van een overervingsmechanisme.

2.2.6.4 Distributieve netwerkmodellen

1. Rumelhart & McCelland (1986), die ook reeds aan de orde kwamen als pioniers op het gebied van netwerkgebaseerde taalmodellen, ondernamen een frontale aanval tegen het op een Universele Grammatica gebaseerde taalverwervingsmodel met de publicatie van hun studie naar automatische taalverwerving op basis van een distributief c.q. connectionistisch netwerkmodel. In deze tak van onderzoek wordt een distributief netwerk gericht getraind op het herkennen van een bepaalde klasse van patronen, waarna het netwerk de hierbij opgedane kennis zelfstandig naar nieuwe items dient uit te breiden, dus nieuwe kennis dient te genereren door over bestaande kennis te generaliseren.

Het door Rumelhart & McCelland (1986) gepubliceerde onderzoek richtte zich op de vraag of deze statistische leermethode geschikt is om inzicht te krijgen in de grammaticaprincipes die ten grondslag liggen aan de verledentijdsvorming van Engelse werkwoorden, d.w.z. of het mogelijk is om deze regels automatisch af te leiden uit een reeks voorbeelden. Ondanks de positieve uitkomst van dit onderzoek, werd de genoemde publicatie erg negatief ontvangen. Dit had deels te maken met het feit dat men een dergelijk verklaringsmodel erg bedreigend vond voor het grammaticale standaardmodel; maar technisch gezien was er ook het nodige op dit onderzoek aan te merken. Deze laatste bezwaren zijn grotendeels weggenomen door Plunkett & Marchman (1991) en MacWhinney & Leinbach (1991). Dit wordt uiteengezet in Rohde & Plaut (2003), die ook een beeld geven van recentere ontwikkelingen op dit onderzoeksterrein.

Een inherente beperking van het distributieve netwerkmodel is dat het op subsymbolische wijze is gestructureerd, waardoor geen rechtstreekse inspectie van de opgebouwde representaties mogelijk is. In dit verband is het interessant om kennis te nemen van de studie van Moscoso del Prado Martín (2003). Deze gaat namelijk uitvoerig in op de vraag in hoeverre de processen in een neuraal netwerkmodel zich laten analyseren. Indien inderdaad een oplossing wordt gevonden voor het blackbox-probleem, kunnen neurale netwerken ook interessant worden voor theoretisch onderzoek naar de cognitieve processen die (mogelijk) ten grondslag liggen aan woordherkenning en woordproductie.

2. Van den Bosch & Daelemans (1999a,b) hebben een symbolisch leersysteem ontwikkeld (te weten TIMBL) dat net als een connectionistisch netwerkmodel in staat is om op basis van

[46] Het probalistische analysemodel van Bod (1995) kent een soortgelijke analysemethode.

trainingsdata zelfstandig patronen te ontdekken en deze te generaliseren naar nieuwe gevallen. Hiertoe wordt elk gegeven uit de trainingsronde systematisch in een reeks samples opgedeeld door voor elk karakter (bijv. een letter binnen een woord) een linker- en een rechtercontext te selecteren. Ik zal de werking van dit leersysteem toelichten aan de hand van een eenvoudig voorbeeld. Stel dat men TIMBL wil leren om op basis van de spelvorm van een woord morfologische grenzen aan te brengen. In dat geval zal men dit systeem eerst moeten trainen door een dataverzameling aan te bieden waarin voor elke spelvorm (c.q. bronpatroon) wordt aangegeven wat het bijbehorende morfeempatroon is (doelpatroon). Zo zou men kunnen aangeven dat de spelvorm 'aandacht' in de analyse 'aan+dacht#' moet resulteren; hierbij correspondeert + met een morfeemgrens en # met de woordgrens. Deze informatie kan als volgt worden aangeboden (de haken staan voor begin en eind van de te verwerken eenheid).

[	a	a	n	d	a	c	h	t	]	-
[	0	0	+	0	0	0	0	#	]	-
0	1	2	3	4	5	6	7	8	9	10

Gegeven deze representatie zal het systeem een reeks samples gaan aanmaken door er een analysevenster overheen te schuiven en op elke positie een opname (c.q. sample) te maken van de locaal aanwezige informatie. Zo zou men een analysevenster kunnen definiëren dat uit drie posities bestaat, namelijk een focus-positie (F), een linkercontextpositie (L) en een rechtercontextpositie (R), hetgeen een LFR-venster oplevert. In onderstaande representatie wordt gedemonstreerd hoe men op basis van dit venster een sample kan maken van de locale informatie op positie 3 (die met de verticale kenmerkbundel {n,+} correspondeert).

[	a	**a**	**n**	**d**	a	c	h	t	]	-
[	0	**0**	+	**0**	0	0	0	#	]	-
0	1	2	3	4	5	6	7	8	9	10
		L	F	R						

Door alle gegevens op deze wijze te analyseren kan het systeem inzicht krijgen in de frequentie waarmee elk locaal patroon voorkomt. Op deze wijze zou het systeem kunnen achterhalen dat de letterstring 'and' in ca. 50% van de gevallen met een morfeemgrens tussen 'n' en 'd' correspondeert; indien het analysevenster wordt vergroot tot vier of vijf karakters zijn zelfs preciezere generalisaties mogelijk.

Uit experimenten van Van den Bosch & Daelemans (1999a,b) blijkt dat de hier beschreven leermethode resultaten oplevert die vaak net zo goed zijn als die van regelgestuurde systemen en deze soms overtreffen. Verder blijkt uit onderzoek van Daelemans, Van den Bosch & Zavrel (1999) dat uitbreiding van de dataverzameling in de trainingsfase tot betere generalisaties leidt, mits de trainingsdata een representatief beeld geven van het gangbare taalgebruik: het is dus niet raadzaam om laagfrequente woorden (of andere gegevens) en uitzonderingen weg te laten, want deze zijn juist cruciaal voor de precisie van de generalisaties. Hieruit blijkt dat onderzoek met taallerende machines inzichten kan opleveren die ook van belang zijn voor de theorievorming over het menselijke taalvermogen.

3. Goldsmith (2000; 2001) heeft een applicatie ontwikkeld (genaamd Linguistica) waarmee automatisch morfologische paradigma's kunnen worden geëxtraheerd uit een middelgroot tekstcorpus (10.000-100.000 woorden) van een willekeurige taal. Deze paradigma's vormen de basis voor een compressiemethode waarbij de aangetroffen woorden zoveel mogelijk als morfeemcombinaties worden gecodeerd. Dit resulteert in een morfologisch gestructureerd netwerk, waarin elk morfologisch complex woord deel uitmaakt van een morfologische signatuur, d.w.z. een paradigma van verwante morfeemcombinaties.

De door Goldsmith gehanteerde compressiemethode berust op het Minimal Description Length-principe (MDL-principe) van De Marcken (1995). Volgens dit evaluatiecriterium

wordt een lexicaal model aantrekkelijker naarmate de informatiewaarde van lexicon en grammatica tezamen kleiner wordt (d.w.z. minder bits in beslag neemt): er dient dus naar een optimale verhouding te worden gezocht tussen lexicale eenheden en structuurprincipes. Voor een optimaal resultaat kunnen deze structuurprincipes beter niet op traditionele grammaticaregels worden gebaseerd, maar dienen ze inductief van de te comprimeren data te worden afgeleid. Hiertoe heeft Goldsmith enkele heuristische criteria geformuleerd.[47]

De door Goldsmith gehanteerde criteria berusten op het uitgangspunt dat morfemen formeel herkenbaar zijn aan het feit dat ze met hoogfrequente lettercombinaties corresponderen. Door dit principe te combineren met de observatie dat veel morfemen onderling gesubstitueerd kunnen worden (dus deel uitmaken van een paradigma), is een (potentieel) universeel extractiemechanisme ontstaan. De door Goldsmith uitgevoerde experimenten wijzen uit dat dit extractiemechanisme vrij betrouwbaar is: bij inspectie van een Franse en een Engelse sample van 1000 woorden bleek 83% van de analyses correct te zijn.[48] Dit neemt niet weg dat de methode nog voor verbetering vatbaar is.

2.2.7 Conclusie

In deze sectie heb ik een systematisch overzicht gepresenteerd van de bestaande kennismodellen met betrekking tot de structuur, de activatie en de verwerving van lexicale kennis, en meer in het bijzonder de morfologische dimensie van deze kennis. Bij de uitwerking van deze inventarisatie heb ik geprobeerd om de grenzen van de hier genoemde onderzoeksdomeinen te doorbreken en de nadruk te leggen op de overeenkomsten in de gehanteerde modellen. Toch heb ik de lexicografische kennismodellen in een andere sectie ondergebracht dan de empirische kennismodellen, want deze onderzoeksdisciplines houden zich doorgaans met zeer verschillende vragen bezig. Later in dit hoofdstuk (in H2.5) zal ik een overkoepelend kennismodel presenteren waarin beide disciplines verenigd kunnen worden.

Met betrekking tot cognitieve kennismodellen kan worden geconcludeerd dat er ondanks de grote verscheidenheid aan voorstellen een aantal duidelijke hoofdklassen bestaan die in elke onderzoeksdomein terugkeren. Om te beginnen heb ik onderscheid gemaakt tussen dualistische en monistische kennismodellen, wat neerkomt op een onderscheid tussen morfologische grammaticamodellen (waarbij het lexicon niet meer is dan een lijst van morfemen of woorden) en morfologische netwerkmodellen (waarbij het lexicon met een morfologisch gestructureerd netwerk van woorden correspondeert). Binnen de klasse van dualistische kennismodellen heb ik nader onderscheid gemaakt tussen lexicongenererende modellen, lexiconstructurerende modellen en hybride activatiemodellen (die als een combinatie kunnen worden gezien van de twee andere modeltypes, in de zin dat ze geen keuze maken tussen woordopslag en woordconstructie, maar beide opties beschikbaar houden). Bij de monistische kennismodellen heb ik een nader onderscheid gemaakt tussen localistische netwerkmodellen en distributieve netwerkmodellen, waarna ik beide nog heb onderverdeeld in symbolische en subsymbolische netwerkmodellen.

In de nu volgende secties zal ik twee belangrijke perspectieven op het mentale lexicon bespreken, te weten het grammaticale lexiconperspectief en het psychologische lexiconperspectief. Hierbij zullen deels dezelfde modellen aan de orde komen als in H2.2, maar ik zal deze modellen nu wat diepgaander bespreken. Hoewel de kennismodellen van het psychologische perspectief vaak op grammaticale kennismodellen zijn gebaseerd, verschillen ze ervan doordat ze met een zoekmechanisme zijn uitgerust. Het psychologische onderzoek richt zich dan ook primair op de vraag hoe men gangbare woorden herkent of produceert, terwijl

[47] Hierbij bouwt Goldsmith voort op pionierswerk van Harris (1955; 1967).
[48] Zie Goldsmith (2001, p. 185).

het grammaticale onderzoek zich veel meer bezighoudt met de regels die ten grondslag liggen aan de mogelijke woordenschat.

2.3 Het grammaticale lexiconperspectief

2.3.1 Introductie

Veel grammaticaal onderzoek naar de woordvorming berust op de aanname dat het taalsysteem uit een lexicon en een grammatica bestaat en dat deze componenten complementaire informatie introduceren. Hierbij heeft het lexicon de taak om de kleinste compositionele bouwstenen van een taal te definiëren (namelijk de woorden of de morfemen), terwijl de grammatica aangeeft hoe deze bouwstenen tot grotere eenheden kunnen worden samengevoegd (zoals woorden en woordgroepen). Dit impliceert dat het lexicon geen polymorfemische woorden of syntactisch gevormde woordcombinaties kan opslaan, tenzij sprake is van idiosyncratische eigenschappen. De hier beschreven theorie vindt zijn oorsprong in de structuralistische taalvisie van Bloomfield (1933), en vormt sinds Chomsky & Halle (1968) tevens de basis van het morfologische standaardmodel uit de generatieve theorievorming.

In de lexicalistische visie van Chomsky (1970) kan de term *lexicon* niet alleen naar het morfologische basislexicon verwijzen, maar ook naar een specifieke submodule van de grammatica, namelijk de module die verantwoordelijk is voor de woordvorming. In dit verband maakt hij onderscheid tussen lexicon-interne structuurregels (c.q. morfologische grammaticaregels, die ten grondslag liggen aan de woordvorming) en lexicon-externe structuurregels (c.q. syntactische grammaticaregels, die onder meer aangeven hoe men woorden tot syntactische constituenten kan samenvoegen). In Chomsky's visie correspondeert de lexicon-interne grammatica met syntagmatische woordvormingsregels, d.w.z. grammaticaregels voor woordvorming door middel van affixaanhechting of samenstelling. Er bestaat echter ook een morfologiemodel[49] dat uitgaat van paradigmatische woordvormingsregels, d.w.z. grammaticaregels voor woordvorming door substitutie van affixen of stammen. In beide modellen gaat de woordvorming vooraf aan de opbouw van syntactische constituenten. Empirisch gezien is deze strikte ordening echter niet houdbaar (cf. Ackema, 1995).

In de nu volgende subsecties belicht ik een aantal inherente beperkingen van het morfeemgebaseerde morfologiemodel. Hierbij dient het generatieve morfologiemodel uit van Don & al. (1994), een algemeen introductieboek, als vertrekpunt. Dit versimpelde model kent de volgende aannames:

- De grammatica van een gegeven taal weerspiegelt de taalkennis (c.q. *competence*) van een homogene groep taalgebruikers. Hiertoe dient de grammatica te abstraheren van het empirisch waarneembare taalgebruik (c.q. *performance*), want deze variatie wordt tot het domein van de pragmatiek gerekend.
- Het lexicon correspondeert met een ongestructureerde lijst van morfemen, namelijk ongelede lexemen (c.q. woordstammen) en affixen; deze vormen de basis voor de afleiding van gelede woorden door de toepassing van morfologische grammaticaregels. De op deze wijze afgeleide woorden mogen niet in het lexicon worden opgeslagen (tenzij ze idiosyncratische vorm- of betekeniseigenschappen gaan vertonen).
- Het lexicon registreert uitsluitend niet-voorspelbare informatie over "grammaticale" morfeemkenmerken, d.w.z. over klankvorm, betekenis, syntactische categorie, inflectiegedrag en categoriale selectierestricties; het bevat dus geen informatie over grammaticaal afleidbare kenmerken, zoals voorspelbare klankvormen en de hierop gebaseerde spelvorm, of over "toevallige" gebruiksaspecten, zoals bestaande woordafleidingen, gebruiks-

[49] Deze benadering vindt zijn oorsprong in voorstellen van Jackendoff (1975) en Aronoff (1976).

frequentie, vaste morfeem- of woordcombinaties, pragmatische gebruikscondities en allerlei geassocieerde kennis (waaronder encyclopedische en etymologische kennis).

- De lexicale grammaticamodule bestaat uit productieve woordvormingsregels, d.w.z. regels waarmee in principe oneindig veel nieuwe woorden kunnen worden gevormd; zodra een regel improductief wordt, maakt deze niet langer deel uit van de grammatica (al kan hij wel nut hebben als lexicale redundantieregel). Productieve woordvormingsregels kunnen uitsluitend op lexicale eenheden worden toegepast.
- De lexicale grammaticamodule kent structuurprincipes die op een aangeboren taalvermogen (namelijk de "universele grammatica") zijn gebaseerd. Het hier bedoelde leervermogen dient algemeen toepasbare regelschema's te introduceren, zodat een taallerend kind gericht kan zoeken naar de taalspecifieke kenmerken van deze regels. Zonder dit speciale leervermogen zou moeilijk te verklaren zijn dat kinderen in staat zijn om onbewust en op basis van onvolledige informatie een complete taal te leren. Deze paradox staat ook wel bekend als "Plato's probleem".
- Het aangeboren taalleervermogen is maar tijdelijk beschikbaar, want na de vroege jeugd is het veel moeilijker om een nieuwe taal te leren. Deze tijdelijkheid heeft als gevolg dat taalkennis die pas op school wordt verworven (waaronder de spellingsdimensie) niet tot de kerngrammatica wordt gerekend. Dit geldt bijvoorbeeld voor het "Romaanse" deel van de morfologie, want dit domein bestaat voor een groot deel uit "geleerde" woorden die de meeste mensen pas op school verwerven.

2.3.2 Competence versus performance

Sinds Chomsky het generatieve grammaticamodel introduceerde, berust veel grammaticaal onderzoek op de aanname dat het direct waarneembare taalgedrag (c.q *performance*) van individuele taalgebruikers door een mentale module met universele grammaticaprincipes (c.q. *competence*) wordt aangestuurd.[50] Hoewel deze grammaticaprincipes vaak een taalspecifieke invulling kennen, zouden de principes zelf niet geleerd hoeven te worden, maar reeds in aanleg aanwezig zijn dankzij een aangeboren taalleervermogen; dit taalleervermogen (of de hieruit afleidbare verzameling van grammaticaprincipes) staat bekend als de universele grammatica (UG). Het generatieve taalonderzoek stelt zich ten doel om deze UG-module te reconstrueren door individuele talen zoveel mogelijk in termen van universele grammaticaregels te beschrijven en alle patronen die niet in dit formaat passen als subregelmatige, dus improductieve constructies af te doen. Volgens diezelfde redenering zou het niet nodig zijn om aandacht te besteden aan bewust geleerde regels (waaronder de taalkennis die men op school opdoet), stijlgerelateerde keuzes en verschillen in productiviteit en herkenbaarheid, want dergelijke kennis is taalspecifiek en daarom niet tot UG-principes te herleiden.[51]

Hoewel het competence-performance-onderscheid nog steeds een belangrijk uitgangspunt is van het generatieve taalonderzoek, is dit onderscheid nooit goed gefundeerd. Zo bestaat er geen bewijs voor de stelling dat het menselijk taalvermogen gebruik maakt van scherp afgebakende grammaticaregels, laat staan dat hierbij onderscheid wordt gemaakt tussen (UG-gebaseerde) competence-patronen en (taalspecieke) performance-patronen. Verder zijn geen criteria beschikbaar waarmee kan worden uitgemaakt welke taalfenomenen een grammaticale basis hebben, waardoor taalfenomenen die zich niet goed in harde grammaticaprincipes laten vangen (zoals gradaties in productiviteit en grammaticaliteit) al gauw worden genegeerd. Tot slot zijn competence-gebaseerde grammatica's niet verenigbaar met statistische methodes

[50] Het idee dat er een fundamenteel onderscheid bestaat tussen competence-kennis (met grammaticale status) en performance-kennis (zonder grammaticale status), ligt niet alleen ten grondslag aan het generatieve grammaticamodel, maar ook aan concurrerende modellen, zoals het categoriale grammaticamodel van Montague (1974).

[51] Er bestaat overigens een opvallende gelijkenis tussen competence-regels en normatieve regels: beide abstraheren namelijk van het concreet waarneembare taalaanbod (c.q. de performance).

voor regelextractie. Door deze beperkingen kennen competence-gebaseerde grammatica-modellen een beperkt empirisch bereik, wat niet alleen problematisch is met het oog op psychologisch en neurologisch georiënteerd taalonderzoek (cf. Levelt (1989); Jackendoff (2002)) en onderzoek naar kindertaalverwerving (Brown, Malmkjaer & Williams (1996)), maar ook met het oog op de ontwikkeling van systemen voor automatische taalverwerking (blijkens Scha (1990)). Chomsky's (1995) minimalistische programma biedt een opening om aan deze kritiek tegemoet te komen.[52]

Mijn eigen onderzoek berust op het uitgangspunt dat grammatica's geen autonoom bestaansrecht hebben, maar dat een grammatica slechts een handig hulpmiddel is om de meest voorkomende constructies van een taal op een overzichtelijke wijze te beschrijven. In mijn optiek dient een taaltheorie zich dan ook niet te beperken tot een inventarisatie van algemeen geldige structuurregels, maar dient zo'n theorie ook aan te geven hoe deze kennis uit de aangeboden taaldata kan worden afgeleid, dus hoe deze kennis kan worden aangepast.

2.3.3 Gelede woorden en woordgroepen

In het generatieve morfologiemodel blijft de functie van het lexicon beperkt tot de verantwoording van de kleinste combinatorische bouwstenen, namelijk morfemen. Dit impliceert dat polymorfemische woorden steeds opnieuw moeten worden afgeleid, tenzij er sprake is van idiosyncratische eigenschappen. Hieruit volgt dat het lexicon geen morfologisch gelede woorden (of grotere taaleenheden) kan opslaan.[53] Deze visie kent echter tal van taalkundige bezwaren. Allereerst is het morfeemgebaseerde lexicon strijdig met de observatie dat taalgebruikers van elk polymorfemisch woord kunnen aangeven of ze het reeds kennen en of het vaak wordt gebruikt. Ten tweede zijn er taalkundige aanwijzingen dat het lexicon paradigmatische structuur vertoont, wat impliceert dat het lexicon met een complete inventarisatie van bestaande woorden correspondeert (cf. Jackendoff (1975), Aronoff (1976), Verkuyl (1978), Booij (2002)). Een woordgebaseerd lexicon biedt ook een oplossing voor het feit dat veel polymorfemische woorden slechts partieel voorspelbaar zijn in de zin dat ze een idiosyncratische vorm of betekenis bezitten of ambigu zijn tussen compositioneel en idiosyncratisch gebruik (cf. Bybee (1988), Sandra (1994)).

Hoewel de hierboven genoemde studies ervan uitgaan dat het lexicon uitsluitend woorden opslaat, zijn er ook taalkundige studies waarin argumenten worden aangedragen voor het idee dat het lexicon woordoverstijgende informatie kan opslaan, zoals vaste woordcombinaties en syntactische constructies; hoewel dit idee reeds impliciet aanwezig is in de combinatorische benadering van Montague (1974), heeft het nog bijna twee decennia geduurd voor het expliciet werd uitgewerkt; cf. Pustejovsky (1991), Kamp & Reyle (1993), Kay (1997), Jackendoff (1997) en Dowty (2000)). Zo betoogt Pustejovksy (1991) dat de betekenis van adjectieven vaak wordt bepaald door het geselecteerde naamwoord, zoals blijkt uit de betekenisvariatie van *snel* in *een snelle schaatser*, *een snelle leerling* en *een snelle jongen*. Op soortgelijke wijze kan de interpretatie van werkwoorden als *beginnen* en *eindigen* van het syntactische object afhangen, zoals blijkt uit het contrast tussen *een lied beginnen* en *een boek beginnen*. Dit wijst erop dat betekenisbouwstenen soms aan grotere structuren dan woorden

[52] Marantz stelt in dit verband (cf. Marantz, 2003) dat het noodzakelijk is om de "performance"-aspecten van het taalsysteem in de "competence"-theorie te integrereren, en dat de competence-theorie experimenteel toetsbaar moet worden gemaakt. In zijn visie biedt het minimalistisch programma hier een goede basis voor.

[53] Deze problematiek staat centraal in het werk van Everaert (bijv. Everaert, 1993). Hij besteedt ook veel aandacht aan de tegenstelling binnentaal-buitentaal. Volgens Everaert (2002), die zelf naar Muysken (1999) verwijst, correspondeert de buitentaal met het domein van de lexicografie, terwijl de linguïstiek zich op het onderzoek naar de binnentaal richt. Dit onderscheid is nodig om te verklaren hoe het mogelijk is dat taalgebruikers veel kennis hebben over gangbare woorden en woordcombinaties, terwijl het onderliggende taalsysteem uitsluitend denkbare (niet per se bekende) woorden en woordcombinaties zou definiëren.

zijn verbonden. Dit volgt ook uit het bestaan van uitdrukkingen als *de sigaar zijn*, *een blauwtje lopen* en *zich een bult lachen*, waarvan de laatste twee ook syntactisch bijzonder zijn. Jackendoff (1997) spreekt in dit verband van een Wheel of Fortune-lexicon:[54] in deze theorie kan het lexicon niet alleen vaste vaste verbindingen opslaan (al dan niet met interne variabelen), maar ook complete gezegden en zelfs strofen. Onduidelijk is in hoeverre hierbij compositionele betekenisopbouw mogelijk is.

Inmiddels heeft psycholinguïstisch onderzoek uitgewezen dat het menselijk brein veel meer informatie opslaat dan in de grammaticale standaardvisie verondersteld wordt. Zo staat nu onomstotelijk vast dat het mentale lexicon niet alleen informatie over basiswoorden (c.q. morfemen) bevat, maar ook over morfologisch complexe woorden (mogelijk inclusief inflectie); cf. Stemberger & MacWhinney (1988), Sandra (1994) en Baayen, Dijkstra & Schreuder (1997)), terwijl er ook aanwijzingen zijn dat het menselijk brein frequent gebruikte woordcombinaties direct memoriseert (cf. Sandra & al. (1999), Harley (2002)).

Hoewel het dus problematisch is om aan te nemen dat het lexicon alleen morfemen kan opslaan, mag hieruit niet worden afgeleid dat morfemen geen psychologisch bestaansrecht hebben. Er zijn namelijk tal van psycholinguïstische publicaties (bijv. Baayen, Dijkstra & Schreuder (1997)) waarin op basis van woordherkenningsexperimenten aannemelijk wordt gemaakt dat morfologisch gelede woorden met een hoge woordfrequentie integraal uit het lexicon worden opgehaald, terwijl morfologisch gelede woorden met een lage woordfrequentie steeds opnieuw van een onderliggend basiswoord worden afgeleid. Alleen deze laatste groep leent zich voor een beschrijving in termen van productieve woordvormingsregels.

2.3.4 De orthografische dimensie

Het generatieve morfologiemodel gaat ervan uit dat de spellingsdimensie niet tot het domein van de grammatica behoort, maar deel uitmaakt van een externe regelmodule met bewust gemaakte afspraken. In een taal als het Nederlands zijn de spellingregels namelijk zo vormgegeven dat de spelling van een woord meestal rechtstreeks van zijn uitspraak kan worden afgeleid. Maar net als elk ander door mensen bedacht regelsysteem blijken ook de spellingsregels niet consequent te worden nageleefd en ruimte te laten voor niet-voorspelbare keuzes. Zo zijn er vele woorden waarvan de officieel vastgelegde spelling afwijkt van een spelling die aansluit bij de algemeen aanvaarde uitspraak, wat vaak te maken heeft met de overweging dat het vanuit technisch, didactisch of esthetisch oogpunt beter is om de etymologisch of morfologisch gemotiveerde spelvorm aan te houden. Een puur fonetische spelling zou ook onwerkbaar zijn, want zoals bekend vertoont de uitspraak van woorden zoveel variatie dat teksten al gauw onleesbaar zouden worden als iedereen van zijn eigen uitspraak uitgaat (zoals tot de invoering van de standaardspelling (in 1804) algemeen gebeurde). De uitspraak van een woord vormt daarom geen goed uitgangspunt voor de toekenning van een spelvorm.

Ook vanuit morfologisch perspectief is het van belang om spellingskenmerken te kunnen representeren. Zo is de spelling van Nederlandse werkwoorden sterk afhankelijk van morfologische overwegingen; het bekendste voorbeeld hiervan is de toevoeging van een |t| in de derde persoon enkelvoud van de tegenwoordige tijd, ook als men deze niet kan horen (zoals in |brandt|), met uitzondering van stammen op een |t| (zoals |praat|); soortgelijke regels gelden voor de spelling van de verleden tijd en het voltooid deelwoord. Daarnaast zijn er tal van stamvormalternanties die primair morfologisch zijn gemotiveerd; zo zijn alternanties van het type s/z, f/v en d/t afhankelijk van de vraag welk suffix er op deze letter volgt, al zijn er ook uitzonderingen (zo correspondeert het meervoud van het nomen |kruis| met de vorm |kruisen|

[54] Het Wheel of Fortune-lexicon ontleent zijn naam aan een Amerikaanse televisieshow waarin de deelnemers op basis van enkele letters een hele idiomatische constructie moeten zien te raden.

en niet met |kruizen| en schrijft men |naïevelijk| naast |liefelijk|). Om correct Nederlands te kunnen schrijven, dient de taalgebruiker dus van elk affix te weten hoe het moet worden gespeld en hoe het de spelling van het stamwoord beïnvloedt (cf. Neijt & Zuidema (1994)).

Uit deze observaties volgt dat het mentale lexicon minimaal in staat moeten zijn om de spellingsvorm van onregelmatig gespelde morfemen en morfeemcombinaties vast te leggen. Om die reden hebben verscheidene taalkundigen (waaronder Zonneveld (1980) en Nunn (1998)) voor een onafhankelijk representatieniveau voor orthografische structuur gepleit; in dit verband wordt meestal van autonome spellingregels gesproken (Van Oostendorp, 1998).

2.3.5 Gradaties in productiviteit

Het generatieve morfologiemodel gaat ervan uit dat er onderscheid moet worden gemaakt tussen productieve en improductieve woordvormingsregels; het verschil is dat productieve regels met woordvormingspatronen corresponderen die regelmatig gebruikt worden voor de constructie van nieuwe woorden, terwijl improductieve regels slechts generalisaties zijn over lexicale woordvormingspatronen, d.w.z. over de "morfologische structuur" van bestaande (lexicaal opgeslagen) woorden. Een affix kan dus productief worden genoemd als het een theoretisch onbegrensd toepassingsdomein kent. Dergelijke productiviteitsclaims berusten echter vaak op subjectieve oordelen over de mate waarin een affix nieuwe woorden kan vormen; zo wordt ten aanzien van Nederlandse suffixen vaak gesteld dat -ING, -BAAR en -ER volmaakt productief zijn, terwijl -SEL al minstens een halve eeuw geleden improductief zou zijn geworden. Het is echter onduidelijk of men bedoelt dat nieuwvormingen met -SEL ongrammaticaal "aanvoelen", of dat er zelden meer nieuwvormingen met -SEL worden aangetroffen. Indien de eerste claim waar is, zal waarschijnlijk ook de tweede claim gelden, maar de omgekeerde implicatie is zeker niet geldig. Een bijkomend probleem is dat men eigenlijk nooit kan bewijzen dat de reeds gangbare woorden (dus verreweg de meeste woorden die men tegenkomt) langs "grammaticale" weg zijn gevormd.

Baayen (1991b, 1992 enz.) heeft nieuw licht op deze problematiek geworpen door statistische maatstaven te ontwikkelen voor de productiviteitsnotie. Eén van deze maten (zie ook §2) drukt de productiveit van een affix uit in termen van het aantal met dit affix gevormde hapaxen (= eenmalig gebruikte woordvormen) per miljoen woorden in een tekstcorpus van een duidelijk afgebakende tijdsspanne. Dit criterium is gebaseerd op de aanname dat het aantal neologismes met een gegeven affix sterk correleert met het aantal eenmalig gebruikte woordvormen. Op basis van deze productiviteitsmaat kan voor alle affixen worden nagegaan hoe productief ze zijn, waarna een rangorde kan worden vastgesteld.

Bij een dergelijke aanpak kan het suffix -SEL zeker niet "dood" worden verklaard, want het tekstcorpus CELEX kent de nodige hapax-woorden met het suffix -SEL. Maar ook als -SEL geen productief suffix meer zou zijn, hoeft -SEL niet meteen zijn morfeemstatus te verliezen; er zijn immers allerlei voorspelbare eigenschappen aan verbonden, zoals de categorie N (van nomen), het meervoud -S, het lidwoord *het*, de betekenis "bedoeld of onbedoeld resultaat van de handeling die door de stam wordt uitgedrukt" enz. Wel zou bij een zorgvuldige analyse van alle bekende woorden met het suffix -SEL kunnen blijken dat het soms maar een deel van deze kenmerken ondersteunt; dit zou dan een aantasting zijn van de morfeemstatus van -SEL, aangezien niet voldaan wordt aan de eis dat er een constante relatie is tussen morfeemvorm en hieraan verbonden woordkenmerken.

2.3.6 Individuele variatie en niet-grammaticale gebruiksfactoren

Het generatieve morfologiemodel berust op de aanname dat talen met een homogene groep sprekers corresponderen, die allemaal dezelfde (kern-)grammatica hanteren. Maar het is veel aannemelijker dat elke taalgebruiker zijn eigen taalvariant spreekt, waarbij deze taalvariant

zich op een continuüm bevindt tussen de zogenaamde standaardtaal en een regionaal of sociaal dialect. Bovendien verschillen sprekers in hun lexicale en grammaticale kennis, terwijl ook veel invloed uitgaat van stilistische overwegingen. Het grammaticale morfologiemodel is niet in staat om dit variabele taalgedrag te verantwoorden, want dit model gaat ervan uit dat alle sprekers van een gegeven taal dezelfde woordvormingsregels hanteren, terwijl niet-grammaticale factoren tot het terrein van de "pragmatiek" worden gerekend. Daarom zijn er altijd mogelijkheden en beperkingen die buiten het bereik van de model vallen.

Zo zijn er tal van potentiële inflectievormen die in de praktijk toch niet (of zelden) voorkomen, meestal om een semantische of fonologische reden. Dit kan worden toegelicht aan de hand van de trappen van vergelijking. Bij korte woorden vormt men de vergrotende trap door een onverbogen adjectief (bijv. *mooi*) te combineren met het suffix -ER (bijv. *mooier*), en de overtreffende trap door combinatie met het suffix -ST (bijv. *mooist*). Bij een woord als *logisch* is de vorm *logischst* echter minder aantrekkelijk, omdat deze vorm er gek uitziet en slechts ten dele uitspreekbaar is; men gebruikt daarom liever de syntactische parafrase *meest logisch*. Iets soortgelijks geldt voor *bezorgdst* versus *meest bezorgd*. Er zijn ook adjectieven die (meestal) een niet-gradeerbare betekenis uitdrukken, zoals *dood*. Toch is *doder* als vorm best aanvaardbaar. Bij langere woordvormen kan men zowel de morfologische als de syntactische variant van de vergrotende of overtreffende trap aantreffen; zo kan men spreken over *de plausibelste verklaring*, maar ook over de *meest plausibele verklaring*. Hoe precies de verhouding ligt, hangt sterk van het basiswoord af. Dit is een aanwijzing dat deze keuze door stilistische factoren wordt bepaald.[55]

Meer in het algemeen kent het Nederlands een tendens om ingewikkelde samenstellingen en grote suffixstapelingen te vermijden door de voorkeur te geven aan een syntactisch alternatief. Hierdoor hoeft de luisteraar niet vermoeid te worden met een complex, moeilijk te interpreteren woord. Men zal bijvoorbeeld niet snel de vorm *verontreinigbaarheid* of zelfs *onverontreinigbaarheid* (bijv. van leidingwater) aantreffen, hoewel hier formeel niets op aan te merken valt; men kan echter ook spreken over *de (on)mogelijkheid iets te verontreinigen*. Bij uitheemse derivaties kan men in theorie nog veel grotere suffixstapelingen bereiken, zoals *constitutionaliseerbaarheid*, bijvoorbeeld als aanduiding van "de mogelijkheid om iets constitutioneel te regelen". Maar in de praktijk wordt de voorkeur gegeven aan een syntactische omschrijving. Dergelijke keuzes worden sterk beïnvloed door stilistische overwegingen en zijn daarom moeilijk in een grammaticaal regelmodel te vangen. Indien men afstapt van het idee dat er universele woordvormingsregels bestaan, kan men dit probleem oplossen door morfologische regels te heranalyseren als stilistische gebruiksconventies. Een dergelijke benadering is in overeenstemming met de reeds besproken observatie dat er gradaties bestaan in de productiviteit van de woordvomingsregels.

2.3.7 Leerbaarheidsvragen

Hoewel het generatieve morfologiemodel uitgaat van een aangeboren taalleervermogen, is nog weinig onderzoek gedaan naar de vraag hoe een gegeven verzameling woordvormingsregels zonder kennis vooraf uit de beschikbare taaldata kan worden afgeleid. Er zijn wel de nodige studies verschenen waarin voor afzonderlijke grammaticaprincipes is nagegaan in welke leerfase het verworven wordt en hoe het verwervingsproces verloopt (cf. Berko (1958), Brown (1973), Schaerlakens & Gillis (1987), Frijn & De Haan (1990)). Hierbij gaat men ervan uit dat dit speciale taalleervermogen slechts in de vroege kinderjaren actief is. Er is dus een apart leervermogen nodig om uit te leggen hoe de taalkennis van een kind kan worden

[55] Bij VDL is de beschikbaarheid van deze vormen integraal in kaart gebracht. Hiertoe is voor alle bij VDL bekende adjectieven onderzocht of de grammaticaal te verwachten vormen (die automatisch werden aangemaakt) daadwerkelijk mogelijk zijn, en zo niet, of dit een fonologische of semantische oorzaak had.

uitgebreid tot de taalkennis van *volwassen* taalgebruikers. Gegeven zo'n aanvullend leervermogen zou een kind uiteraard ook kennis kunnen verwerven over de morfologische structuur van het Romaanse deel van de woordenschat. Dit impliceert dat er geen reden is om het domein van het morfologisch onderzoek in te perken tot de door (jonge) kinderen verworven morfologie of om het Romaanse deel van de Nederlandse morfologie te negeren.

Vanuit het generatieve perspectief is het overigens goed mogelijk dat "Plato's probleem" (zie Kerstens & al., 1997), d.w.z. de vraag hoe het mogelijk is dat kinderen op basis van een minimale verzameling voorbeelden in staat zijn om een taal te leren, niet relevant is voor de verwerving van de woordenschat. Het menselijk brein heeft namelijk een enorme opslagcapaciteit, zodat er weinig reden is om te bezuinigen op de hoeveelheid informatie in het lexicon; in dat geval zijn geen morfologische regels nodig voor de verantwoording van de bestaande woordenschat, maar alleen voor de sporadisch benutte mogelijkheid om een nieuw woord te vormen. Er is echter toenemende evidentie dat het heel goed mogelijk is om deze nieuwvormingen door middel van analogieprincipes te verklaren, dus dat er helemaal geen grammaticale woordvormingsregels nodig zijn. Dit uitgangspunt ligt bijvoorbeeld ten grondslag aan de lexicale theorieën van Jackendoff (1975) en Bybee (1988).

Zo'n analogiebenadering heeft als voordeel dat hij ook bruikbaar is voor de analyse van subregelmatige patronen. Verder kan het analogieprincipe ook een verklaring bieden voor de verbijsterende woordcreativiteit van kinderen (cf. de inventarisatie in Fikkert (2003). Een mogelijk nadeel van deze benadering is dat hij geen ruimte lijkt te bieden voor zogenaamde default-affixen; dit zijn affixen die onbeperkt productief zouden zijn binnen hun selectiedomein. Maar uit de discussie in Clahsen (1999) blijkt dat het hier om een controversieel concept gaat. Er zijn namelijk tal van empirische studies die aantonen dat elk affix analogiegedrag vertoont ten opzichte van de bekende woordenschat, of er nu wel of niet een defaultstatus aan wordt toegekend. Het analogiemodel is waarschijnlijk ook bruikbaar voor de verwerving van syntactische constructies. Want hoewel er oneindig veel zinnen mogelijk zijn, berusten de meeste zinnen op een constante basisverzameling van syntactische constructies en vaste woordcombinaties (c.q. idioom), terwijl nieuwe constructies vaak van bestaande constructies zijn afgeleid (cf. Peters (1983), Kay (1997), Jackendoff (2002)).

2.3.8 Conclusie

In deze sectie heb ik een aantal fundamentele beperkingen van het grammaticale morfologiemodel besproken. Hoewel ik mijn kritiek primair op het generatieve model van Don & al. (1994) heb gericht, zijn de door mij naar voren gebrachte beperkingen meestal een direct of indirect gevolg van de morfeemgebaseerde opzet van het onderliggende lexicon. Deze aanpak blijkt ertoe te leiden dat het lexicon slechts een deel van de gangbare woorden representeert, terwijl er een zeer streng criterium bestaat voor de identificatie van morfologisch gelede woorden. Dergelijke woorden dienen namelijk aan de eis te voldoen dat hun representatie volledig uit de samenstellende morfemen kan worden opgebouwd. Bij de evaluatie van deze eis beperkt men zich vaak tot de vraag of er een directe relatie bestaat tussen klankvormsegmenten en morfosyntactische functies. Hierbij blijft de spelvorm dus buiten beschouwing (aangezien deze dimensie tot de pragmatische kennis wordt gerekend), waardoor een hoop onregelmatigheden onzichtbaar blijven. Vaak wordt ook voorbij gegaan aan een analyse van de betekenisdimensie, met als gevolg dat onvoldoende oog bestaat voor het feit dat morfologisch complexe woorden doorgaans tal van gelexicaliseerde betekenissen kennen. Een nadere analyse van de orthografische en semantische woorddimensie leidt al gauw tot de conclusie dat het morfeemgebaseerde representatiesysteem uit het grammaticale standaardmodel onhoudbaar is. Daarom geef ik de voorkeur aan een lexicaal netwerkmodel met een op redundantie gebaseerd overervingssysteem. Maar dat neemt niet weg dat het morfeemgebaseerde

onderzoek naar de Nederlandse woordvormingsregels waardevolle kennis heeft opgeleverd en dat nog steeds doet vanwege de inzichten die het verschaft in de morfologische bouwprincipes van de bestaande woordenschat.

2.4 Het psychologische lexiconperspectief

2.4.1 Introductie

Hoewel sinds de jaren zeventig veel onderzoek is gedaan naar de vraag welke universele principes ten grondslag liggen aan de mogelijkheid om nieuwe woorden te vormen, was er tot begin jaren negentig weinig bekend over de vraag in hoeverre morfologische structuur een rol speelt in de psychologische processen die ten grondslag liggen aan de opslag, verwerving en raadpleging van morfologisch complexe woorden.[56] In een inventarisatie van bestaand onderzoek naar deze vraag stelt Sandra (1994) dat er niet aan hoeft te worden getwijfeld dat het brein in staat is om morfologische bouwstenen te herkennen, want alle talen kennen de mogelijkheid om nieuwe woorden te produceren door recombinatie van bestaande bouwstenen, terwijl taalgebruikers geen noemenswaardige moeilijkheden hebben om deze nieuwvormingen te interpreteren (zoals ze ook geen moeite hebben om steeds weer nieuwe zinnen te interpreteren). Daarom ligt het voor de hand om dit fenomeen niet alleen vanuit grammaticaal perspectief te onderzoeken (d.w.z. vanuit de doelstelling om per taal na te gaan welke regels ten grondslag liggen aan de aanmaak van nieuwe woorden), maar ook vanuit psychologisch perspectief (dus vanuit de doelstelling om inzicht te krijgen in de mentale aspecten van morfologische complexiteit). Het onderstaande citaat geeft een indruk van de mogelijke onderzoeksvragen:

> "What is the nature of the processes involved in producing and understanding novel polymorphemic words? What insights can be gained from the psycholinguistic study of existing polymorphemic words? Do language users draw on their perception of morphological relations between words for storing these words in their mental lexicon? If so, how does this translate into representational language, that is, what do the representations of polymorphemic words look like? Furthermore, what are the precise consequences for lexical processing? More particularly, at what processing level is morphological structure involved and what purpose does it serve there? Does the mental lexicon treat polymorphemic words as a monolithic class or does it accommodate different word types in different ways? If the latter possibility is correct, which linguistic variables are relevant for defining those word types? Is it possible to generalise across languages or are language-specific properties of morphological structure important determinants of the way polymorphemic words are represented and processed? How do the properties of the various language modalities – reading, writing, listening and speaking – affect the answers to the above questions?" (Sandra 1994, p. 228)

Sandra (1994) acht een theorie over de morfologische aspecten van het taalsysteem pas compleet als al deze vragen beantwoord kunnen worden. Met het oog hierop is Sandra nagegaan wat voor opvattingen er over de morfologische dimensie van het mentale lexicon bestaan en in hoeverre deze opvattingen door empirische evidentie worden ondersteund.

[56] Harley (2002) biedt een grondige introductie tot dit onderzoeksgebied, waarbij alle lexicale processen (zoals lezen, schrijven, luisteren en spreken) en structuurniveaus (van fonemen tot woorden en zinnen) systematisch aan bod komen. Levelt (1989) geeft inzicht in de psychologische processen die ten grondslag liggen aan het vermogen om een gegeven propositie (c.q. boodschap) in een coherente zin om te zetten en deze uit te spreken.

2.4.2 Psycholinguïstische aspecten van morfologische structuur

Sandra (1994) richt zich op onderzoek naar de morfologische aspecten van woordherkenning bij leestaken ("visual word recognition"). Uit diverse leesexperimenten blijkt dat het morfologische structuurtype van de aangeboden woorden invloed heeft op de snelheid waarmee de woorden herkend worden. Zo maakt het uit of men gelede of ongelede woorden aanbiedt, of het gangbare woorden zijn of nieuwe woorden, of de woorden het resultaat zijn van inflectie, derivatie of samenstelling en (in het geval van derivaties) of er sprake is van een prefix of een suffix. Vanuit grammaticaal perspectief leidt Sandra's onderzoeksdomein tot een verrassende uitbreiding van de morfologische problematiek. Want bij leesonderzoek houdt men zich primair bezig met de vraag welke mechanismes ervoor zorgen dat visueel aangeboden woorden tot de activering van mentale woordrepresentaties leiden, wat impliceert dat het hierbij gehanteerde taalmodel expliciet moet aangeven hoe orthografische woordrepresentaties zich tot de kennis in het mentale lexicon verhouden. In het grammaticale perspectief geldt de orthografie echter als een triviaal, namelijk fonologisch gedetermineerd aspect van het taalsysteem, want de spellingregels zouden geen natuurlijke oorsprong hebben, maar op bewust gemaakte codeerafspraken berusten. In deze visie dienen alle onregelmatigheden in de spelvorm langs pragmatische weg te worden verantwoord.

Uit de bespreking van Sandra blijkt dat de grammaticale benadering van mentale woordrepresentaties onverwachte problemen ontmoet bij de verantwoording van morfologisch complexe woorden, want een deel van deze woorden bestaat uit orthografisch ambigue morfemen. Dit impliceert dat de kennis over deze woorden gedeeltelijk door het mentale lexicon moet worden verantwoord. Orthografisch gezien is het ideaal van een volledig regelgestuurde morfologie dan ook onhoudbaar. Dit probleem is een rechtstreeks gevolg van het feit dat grammaticale modellen ervan uitgaan dat morfemen met een vaste combinatie van een fonologisch en een semantisch kenmerk corresponderen, ongeacht hun orthografische kenmerken. Verder kent de grammaticale morfeemdefinitie ook complicaties van fonologische en semantische aard, maar deze hebben een minder systematisch karakter dan de spellingsproblemen. Meer in het algemeen is de grammaticale benadering extreem gevoelig voor kleine afwijkingen van de voorspelde woordkenmerken. Hierdoor kan slechts een klein deel van de gangbare woorden via regels worden verantwoord en zal de rest integraal in het mentale lexicon moeten worden opgeslagen. Een mogelijke uitweg is om morfologisch complexe woorden als partieel gelede eenheden op te slaan, namelijk als eenheden waarvan één of meer kenmerken uit (partiële) morfemen worden opgebouwd. In dit verband kan men onderscheid maken tussen partieelmorfologische activatiemodellen, d.w.z. modellen die het lexicale belang van morfologische structuur tot één of meer vaste woordkenmerken beperken, namelijk tot de spelling, de klankvorm en de betekenis (cf. Stanners & al. (1979a,b), Taft (1988), Giraudo & Grainger (2001; 2003)), en modellen met een volledige compositieroute. Deze tweede visie is kenmerkend voor integraal-morfologische activatiemodellen (cf. Caramazza, Laudanna & Romani (1988), Anshen & Aronoff (1988), Taft (1994), Schreuder & Baayen (1995)).[57]

Sandra (1994) zelf hanteert een classificatie die uitgaat van de vraag wat de functie is van morfemen. Hierbij onderscheidt Sandra drie verschillende morfeemfuncties, te weten economische kennisopslag, efficiënte kennisactivatie en verwerving van nieuwe woorden. Voor elk van deze functies gaat Sandra na in hoeverre de voorgestelde modellen worden ondersteund of weerlegd door empirische evidentie. Hiernaast besteedt Sandra ook aandacht aan de mogelijkheid dat morfologische structuur een epifenomeen is van interacties tussen fonologische en semantische woordkenmerken (cf. Bybee (1985; 1988), Seidenberg & Gonnerman (2000)). Sandra's vraagstelling suggereert dat morfemen slechts één (hoofd)functie bezitten, en dat

[57] Deze indeling (met de bijbehorende termen) is overigens geen voorstel van Sandra (1994), maar van mijzelf.

nader onderzoek moet uitwijzen welk van de genoemde functies hier het dichtst bijkomt. In mijn optiek correspondeert elk van de door Sandra genoemde morfeemfuncties echter met een relevant aspect van het mentale lexicon, en zijn deze functies moeilijk te scheiden. Zo is de leerfunctie onverbrekelijk verbonden met de kennisopslagfunctie, want anders zou de opgeslagen kennis nooit morfologische structuur kunnen bezitten. En zonder morfologisch gestructureerde kennisopslag is geen morfeemgebaseerde woordactivatie mogelijk. Het is dan ook niet verrassend dat de lexiconmodellen die ten grondslag liggen aan de door Sandra besproken studies vrijwel altijd uitgaan van partieel gestructureerde woordrepresentaties (als compromis tussen dynamisch geconstrueerde morfeemcombinaties en ongelede woordrepresentaties), ongeacht de veronderstelde morfeemfunctie.

Later in deze studie (namelijk in hoofdstuk 4) zal ik een theorie presenteren die de drie genoemde functies op een algemeen, data-onafhankelijk leermechanisme terugvoert. Dit leermechanisme stelt de taalgebruiker in staat om de bestaande woordenschat incrementeel en langs inductieve weg van morfologische structuur te voorzien op basis van een structuurcriterium dat globaal kan worden getypeerd als het streven om de lexicale informatie zo compact mogelijk op te slaan, maar zonder dat er informatie verloren gaat.[58] Dit structuurcriterium leidt tot een compleet, efficiënt gestructureerd en snel toegankelijk lexicon, namelijk een lexicon waarin de opgeslagen woorden zoveel mogelijk uit gemeenschappelijke bouwstenen worden opgebouwd, te weten morfemen, pseudo-morfemen en submorfologische eenheden, zoals syllaben en fonemen. Deze eenheden corresponderen met hiërarchisch gestructureerde kennisindexen (die aanzienlijk minder opslagruimte vergen dan complete morfemen). Vanuit dit perspectief zijn de bestaande morfologiemodellen niet meer dan partiële specificaties van het cognitieve systeem dat ten grondslag ligt aan het mentale lexicon.

2.4.3 De structuur van het mentale lexicon

2.4.3.1 Morfologische complexiteit

In het psycholinguïstische onderzoek naar het mentale lexicon kunnen (los van de onderzochte modaliteit, en los van de vraag of er sprake is van een productiemodel, een perceptiemodel of een functie-onafhankelijk lexiconmodel) vier soorten activatiemodellen worden onderscheiden, namelijk:

1) lexicale activatiemodellen, d.w.z. activatiemodellen waarbij het mentale lexicon geen morfologische structuur bezit, maar alle morfologisch gelede woorden integraal opslaat, hetzij in ongestructureerde vorm (cf. Butterworth (1983), Kostić (1995)), het zij in gestructureerde vorm (cf. Bybee (1985; 1988)).
2) grammaticale activatiemodellen, d.w.z. activatiemodellen waarbij het mentale lexicon precies dezelfde structuur bezit als de grammatica, in de zin dat het lexicon alleen ongelede woorden opslaat, terwijl alle morfologisch gelede woorden op afroep uit hun samenstellende morfemen worden geconstrueerd; cf. Pinker & Prince (1994), Clahsen (1999).
3) partieel-morfologische activatiemodellen, d.w.z. activatiemodellen waarbij morfemen slechts als partiële bouwsteen dienen van de vormrepresentatie of de betekenisrepresentatie van de hiermee opgebouwde woorden. Indien de morfemen de basis vormen van de vormrepresentatie, bezitten deze uitsluitend een klankvorm (cf. Taft (1988)) en in modellen waarbij deze morfemen de basis vormen van de betekenisrepresentatie bezitten deze alleen betekeniskenmerken (cf. Giraudo & Grainger (2001; 2003)).

[58] Dit idee ligt (in een wat andere vorm) ook ten grondslag aan het autonome, taalonafhankelijke analysemodel van Goldsmith (2001); hiervan bestaat ook een computerimplementatie, te weten Linguistica.

4) integraal-morfologische activatiemodellen, d.w.z. activatiemodellen waarin morfologisch gelede woorden op twee manieren beschikbaar zijn, namelijk als integrale woordvorm en als een op afroep te vormen morfeemcombinatie; binnen deze hoofdklasse kunnen de volgende subklassen worden onderscheiden:

   a) Het parallelle zoekmodel c.q. dual route-model (Schreuder & Baayen (1995)): in dit model is sprake van een zoekproces waarbij tegelijk morfeemcombinaties en integrale woorden worden geactiveerd; hierbij wint de route die als eerste een complete woordrepresentatie oplevert.
   b) Het interactieve zoekmodel (Taft, 1994): in dit model zijn morfologisch complexe woorden alleen toegankelijk via hun morfemen; indien sprake is van een productief gevormd woord gaat, kan dit woord op afroep worden geconstrueerd.
   c) Het sequentiële zoekmodel (Caramazza & al. (1988)): in dit model wordt eerst het lexicon doorzocht; indien het opgegeven woord niet snel genoeg wordt gevonden, wordt de regelmodule ingeschakeld.

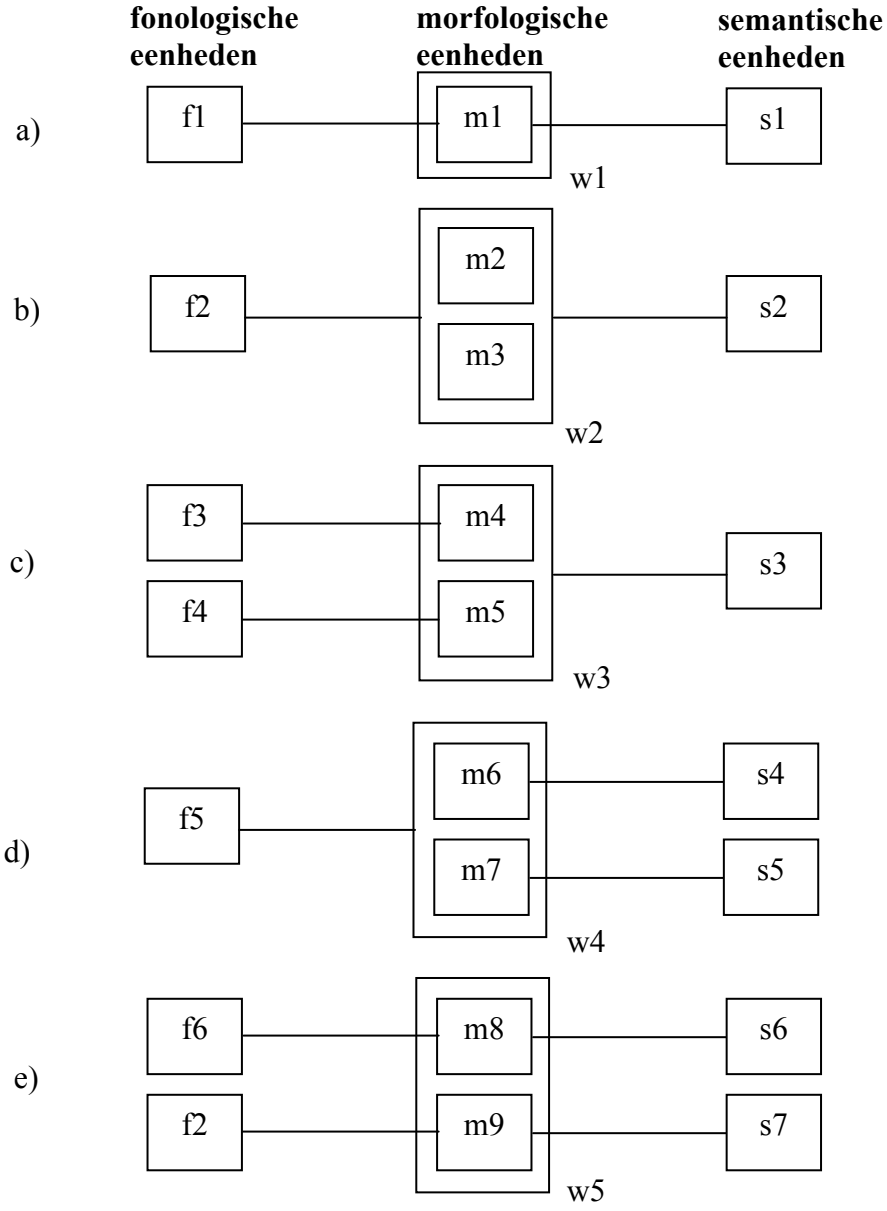
Ter onderbouwing van deze modellen is veel experimenteel onderzoek gedaan naar de activatie van morfologisch gelede woorden. Deze experimenten draaien om de vraag in hoeverre specifieke klassen van morfologisch complexe woorden (waaronder inflectievormen) in het mentale lexicon worden opgeslagen. Ik zal dit toelichten aan de hand van de structuurschema's in figuur 2-1. Deze schema's tonen vijf mogelijke basisrelaties tussen de fonologische[59], morfologische en semantische bouwstenen van lexicale woordrepresentaties en kunnen zowel van links naar rechts (bij perceptieprocessen c.q. luister- en leesprocessen) als van rechts naar links (bij productieprocessen c.q. spreek- en schrijfprocessen) worden doorlopen.[60] Indien men van links naar rechts gaat (dus indien er sprake is van een perceptieproces), kan men de vraag stellen of de gezochte betekenis rechtstreeks vanuit de fonologische woordvorm moet worden geactiveerd, of dat men de woordvorm eerst in morfologische vormsegmenten moet opdelen, om vervolgens de bijbehorende morfemen (c.q. morfeemlemma's) te activeren. Het antwoord op deze vraag hangt sterk af van het woordtype, maar zou niet afhankelijk mogen zijn van het analyseperspectief, want theoretisch gezien is het niet aantrekkelijk om meerdere structuurrepresentaties per woord aan te nemen (al gebeurt dit wel in hybride activatiemodellen, zoals het concurrentiemodel van Schreuder & Baayen (1995)). Maar doordat elk analyseperspectief slechts op een deelvraag is gericht, wordt in de bijbehorende experimenten ook slechts een deel van de structuur zichtbaar. Ik zal de mogelijke structuurschema's nu kort toelichten.

Structuurschema a) correspondeert met woorden zonder interne morfeemstructuur, wat betekent dat er grammaticaal gezien geen enkel vorm- of betekeniskenmerk is dat potentieel als een morfeem kan worden geïdentificeerd. Deze woorden zullen dus in elk perspectief als ongeleed gelden. Structuurschema b) is sterk verwant aan schema a); het enige verschil is dat het woord nu uit twee morfemen bestaat. Er is echter geen enkele relatie met de interne structuur van de semantische of fonologische representatie, zodat de weergegeven structuur alleen langs grammaticale weg (dus op basis van abstracte representatieprincipes) kan worden gemotiveerd (bijvoorbeeld door onzichtbare agreement-features te postuleren). De structuurschema's in c) en d) weerspiegelen woorden met een partieel compositionele structuur: in

[59] In linguïstische theorieën wordt deze fonologische representatie meestal als de onderliggende vorm van een reeks auditieve (c.q. fonetische) representatievarianten gedefinieerd en is er geen directe relatie met de orthografische woordkenmerken. In mijn optiek corresponderen de auditieve en de orthografische representatie echter met gelijkwaardige representaties, die allebei toegang geven tot een integrale (modaliteitsonafhankelijke) vormrepresentatie, namelijk de fonologische representatie.

[60] In grammaticale morfologiemodellen speelt deze procesrichting echter geen rol: dergelijke modellen zijn er immers alleen op uit om de beschikbare kennis te verantwoorden, ongeacht de toepassing van deze kennis.

structuurschema c) correspondeert de morfologische structuur met een fonologische geleding; in dergelijke gevallen spreekt men meestal van een formeel geleed woord, bijv. *ontmoeten* (met voltooide tijd *ontmoet*, niet *ge-ontmoet*) of *deftig* (adjectief, wegens *-ig*). In structuurschema d) correspondeert de morfologische structuur met een semantische geleding, wat op morfologische conversie duidt, bijv. *duw* (N of V) of *gek* (N of A). Structuurschema e) ten slotte toont een volledig compositionele structuur: voor elk morfeem is namelijk een fonologische en een semantische representatie beschikbaar.



Figuur 2-1: Mogelijke structuurschema's van lexicaal opgeslagen woorden.

Grammaticale studies gaan er meestal vanuit dat formeel gelede woorden met een idiosyncratisch betekenisaspect morfologisch ongeleed zijn en daarom integraal in het lexicon moeten worden opgenomen. Maar dit impliceert waarschijnlijk dat *alle* door een gebruiker verwerkte woorden in het lexicon moeten worden opgenomen, want in de praktijk ontwikkelt bijna elk morfologisch gevormd woord één of meer idiosyncratische (c.q. "gelexicaliseerde") betekenissen, zodat de onderliggende woordvorm niet langer afgeleid kan worden, maar net als de andere woordkenmerken integraal (dus ongeanalyseerd) in het lexicon moet worden op-

genomen.[61] Aan de andere kant wordt vaak een onhoorbaar 0-morfeem gepostuleerd bij woorden die in een regelmatige conversierelatie staan. In deze grammaticale benadering levert de introductie van morfemen dus geen enkele bijdrage aan de reductie van de lexicale opslagruimte, ondanks het feit dat Chomsky & Halle (1968) dit als centraal argument voor de introductie van morfeemstructuur hanteerden. Maar in hedendaagse grammaticastudies ontlenen morfemen hun bestaansrecht uitsluitend aan het feit dat ze als productie-eenheid kunnen optreden, d.w.z. als bouwsteen bij de constructie van niet eerder gebruikte woorden.

In psycholinguïstische morfologiestudies richt men zich meestal op de vraag in hoeverre grammaticaal gelede woorden een lexicale representatie bezitten en of deze representatie (al dan niet partieel) uit morfemen is opgebouwd. De onderliggende experimenten richten zich vooral op woorden met de structuurschema's c) en d). Want indien men uitgaat van een vroeg (c.q. "pre-access") decompositiemodel (zoals Taft (1988)), moet worden aangetoond dat semantisch gelede woorden met een onregelmatige vorm (zoals *schot* en *dacht*), dus woorden met schema d), niet via hun semantische bouwstenen kunnen worden geactiveerd, terwijl voor een laat ("post-access") decompositiemodel (cf. Giraudo & Grainger (2001; 2003)) moet worden aangetoond dat formeel gelede woorden met een idiosyncratische betekenis (zoals *schutter* en *bezitten*), dus woorden met schema c), niet via hun formele morfemen kunnen worden geactiveerd.

Voor beide modeltypes geldt dat doorgaans meer morfologische structuur aanwezig is dan er kan worden verantwoord. Want in een vroeg decompositiemodel valt de semantische morfeemdimensie weg, terwijl een laat decompositiemodel blind is voor de fonologische morfeemdimensie, of het nu om partieel gelede woorden (met structuurschema c of d) of integraal gelede woorden (met structuurschema e) gaat. Dit roept de vraag op of deze asymmetrische decompositiemodellen niet inherent incompleet zijn. Een bijkomend probleem is dat er steeds meer aanwijzingen komen dat partieel gelede woorden even snel herkend worden als transparant gelede woorden, gegeven een equivalent priemwoord[62], dus dat het mentale lexicon geen discrete morfemen kent, maar dat er sprake is van een morfologisch continuüm van totaal ongelede woorden naar transparant gelede woorden (cf. McKinnon & al. (2003), Voga & Grainger (2004), Rastle, Davis & New (2004) en Gonnerman, Seidenberg & Andersen (2004, ms.). De genoemde auteurs zien dit als aanwijzing voor de hypothese dat het mentale lexicon de structuur van een connectionistisch netwerk bezit (d.w.z. een netwerk met distributief opgeslagen, dus niet exact localiseerbare structuurinformatie). Maar deze gegradeerde structuur is waarschijnlijk ook verenigbaar met een localistische (c.q. symbolische) benadering; dit blijkt bijvoorbeeld uit een studie van Krott (2001).

2.4.3.2 Kwantitatieve aspecten

In het experimentele onderzoek naar de lexicale representatie van morfologisch complexe woorden richt men zich vaak op frequentie-effecten. Hierbij probeert men aan te tonen dat de reactiesnelheid van de proefpersoon systematisch correleert met frequentiekenmerken van het aangeboden woord, zoals de frequentie van de integrale woordvorm (= tokenfrequentie), de (cumulatieve) frequentie van de stam of een combinatie van beide factoren. Voor dit doel dienen de verschillende woordklassen zo uniform mogelijk te worden samengesteld met be-

[61] Deze gedachte is ook terug te vinden bij Aronoff (1976), Booij (1977) en Bochner (1993).

[62] De term *priemwoord* is een vernederlandsing van de Engelse term *prime* (c.q. prime-woord), d.w.z. een aan het te beoordelen woord voorafgaand woord waarvan vorm en/of betekenis deels of geheel met het doelwoord (c.q. target-woord) overeenkomt.

trekking tot token- en typefrequentie van woordvorm en samenstellende morfemen.[63] Dit type onderzoek vormt de basis voor het lexicale kennismodel van Schreuder & Baayen (1997).

	<X,V>-derivatie	<X,V>-derivatie	<X,V>-derivatie	<X,V>-derivatie
V-prefix	[{0/ge}$_{\|V}$ + bouw]$_V$	[be$_{\|V}$ + bouw]$_V$	[ver$_{\|V}$ + bouw]$_V$	[P$_{\|V}$ + bouw]$_V$
	V-inflectie	V-inflectie	V-inflectie	V-inflectie
Infl. OTT	[bouw]$_V$ > OTT	[bebouw]$_V$ > OTT	[verbouw]$_V$ > OTT	[P + bouw]$_V$ > OTT
Infl. OVT	[bouw]$_V$ > OVT	[bebouw]$_V$ > OVT	[verbouw]$_V$ > OVT	[P + bouw]$_V$ > OVT
0$_{\|GW}$	[bouw]$_V$+0$_{\|GW}$	[bebouw]$_V$+0$_{\|GW}$	[verbouw]$_V$+0$_{\|GW}$	[P + bouw]$_V$+0$_{\|GW}$
en$_{\|INF}$	[bouw]$_V$+en$_{\|INF}$	[bebouw]$_V$+en$_{\|INF}$	[verbouw]$_V$+en$_{\|INF}$	[P + bouw]$_V$+en$_{\|INF}$
end$_{\|TDW}$	[bouw]$_V$+end$_{\|TDW}$	[bebouw]$_V$+end$_{\|TDW}$	[verbouw]$_V$+end$_{\|TDW}$	[P + bouw]$_V$+end$_{\|TDW}$
d$_{\|VDW}$	[ge+bouw]$_V$+d$_{\|VDW}$	[bebouw]$_V$+d$_{\|VDW}$	[verbouw]$_V$+d$_{\|VDW}$	[P + ge+bouw]$_V$+d$_{\|VDW}$
	<V,Y>-derivaties	<V,Y>-derivaties	<V,Y>-derivaties	<V,Y>-derivaties
0$_{\|VN}$	[bouw]$_V$+0$_{\|VN}$	-	[ver+bouw]$_V$+0$_{\|VN}$	[P + bouw]$_V$+0$_{\|VN}$
0$_{\|N}$	[ge+bouw]$_V$+0$_{\|N}$	-	-	[P + bouw]$_V$+0$_{\|N}$
-ing	-	[be+bouw]$_V$+ing$_{\|VN}$	[ver+bouw]$_V$+ing$_{\|VN}$	-
-er	[bouw]$_V$+er$_{\|N}$	?	[ver+bouw]$_V$+er$_{\|N}$	[P + bouw]$_V$+er$_{\|N}$
-baar	?	[be+bouw]$_V$+baar$_{\|A}$	?	-
-sel	[bouw]$_V$+sel$_{\|N}$	-	-	?
	[totaal: 4 klassen]	[totaal: 2 klassen]	[totaal: 3 klassen]	[totaal: 3 klassen]

P = Prepositie = {aan, af, bij, door, in, mee, na, om, onder, op, over, rond, uit, voor, etc.}
OTT = Onvoltooid Tegenwoordige Tijd, OVT =Onvoltooid Verleden Tijd
GW = Gebiedende Wijs, INF = Infinitief, TDW/VDW = Tegenwoordig/Voltooid Deelwoord
$0_{|VN}$ resp. $0_{|N}$ = N met dynamische (V-gebaseerde) resp. statische (N-gebaseerde) betekenis
<X,V>-derivatie = derivatie van een V-lexeem op basis van een X-stam
<V,Y>-derivatie = derivatie van een Y-lexeem op basis van een V-lexeem

Tabel 2.1: De morfologische derivatiemogelijkheden van de wortel BOUW$_0$.

Het experimentele onderzoek van De Jong & al. (2001) bouwt voort op de inzichten van Schreuder & Baayen (1997). Hun nieuwe onderzoek heeft aannemelijk gemaakt dat de volgende frequentiematen psychologisch relevant zijn:

- surface frequency
- base frequency
- family size
- family frequency
- cumulative family frequency

Ik zal deze frequentiematen toelichten aan de hand van het woordparadigma van de morfologische stam BOUW$_0$ (zie tabel 2-1). Deze tabel is als volgt opgebouwd: uitgaande van het morfologische wortelmorfeem BOUW worden vier verschillende V-derivaties gespecificeerd, namelijk de V-lexemen die ten grondslag liggen aan de werkwoorden *bouwen*, *bebouwen*, *verbouwen* en P+*bouwen* (zoals in *aanbouwen*, *bijbouwen* en *opbouwen*). Voor elk van deze lexemen worden de belangrijkste inflectieklassen gespecificeerd (in totaal zes subklassen, die elk weer verdere inflectiemogelijkheden vertonen); voorts wordt voor elke lexeemstam een overzicht gegeven van de bijbehorende derivatiemogelijkheden (c.q. lexeemtoepassingen).

Hieronder wordt elk van de reeds genoemde frequentiematen van een definitie voorzien; deze zal worden toegelicht aan de hand van een op tabel 2.1 gebaseerd voorbeeld:[64]

[63] Hierdoor zijn de resultaten van ouder onderzoek (zoals van Taft & Forster (1975)) niet langer betrouwbaar.

- vormfrequentie ("surface frequency"): gebruiksfrequentie van een concrete woordvorm, bijv. de inflectievorm *verbouwde* van het V-lexeem VERBOUW.
- lexeemfrequentie ("base frequency"): cumulatieve vormfrequentie van een lexeem (meestal inclusief stamallomorfie); zo correspondeert de lexeemfrequentie van het V-lexeem BOUW met de som van de vormfrequenties van de traditionele inflectievormen (ongeacht hun specifieke betekenis), te weten de infinitiefvorm *bouwen*, de tegenwoordige-tijdsvormen *bouw*, *bouwt*, *bouwen*, de verleden-tijdsvormen *bouwde*, *bouwden* en de participia *bouwend* en *gebouwd*.
- familieomvang ("family size"): type-frequentie van de morfologische stam, d.w.z. het aantal lexemen waarvan de (directe) stam overeenkomt met die van het referentielexeem; zo kent het V-lexeem *bouwen* (met stam BOUW) de volgende (directe) familieleden: *bouwen* (V), *bouwer* (N), *gebouw* (N), *bouwsel* (N), *bouw* (V/N). Het laatst genoemde lexeem komt niet alleen zelfstandig voor (bijv. *de bouw*), maar ook als linkerdeel van *bouwval*, *bouwwerk*, *bouwproject* en *fortenbouw*. Bij de bepaling van de familie-omvang beperkt men zich meestal tot de directe stamfamilie, d.w.z. tot derivaties die in één keer van de stam kunnen worden afgeleid, zonder toevoeging van een extra prefix. Dit impliceert dat werkwoorden die door middel van een prefix van de stam zijn afgeleid (en hun directe lexeemfamilie), niet meetellen bij de bepaling van de familieomvang. De subfamilies *verbouwen* (met de stam [VER+BOUW], *bebouwen* (met de stam [BE+BOUW]) en P-werkwoorden zoals *aanbouwen* en *uitbouwen* corresponderen dan met andere families dan het (ongelede) werkwoord *bouwen*.
- familiefrequentie ("family frequency"): som van de gebruiksfrequenties van alle woorden die tot de (directe) familie behoren van de stam van het referentielexeem. Gegeven het V-lexeem *bouwen* correspondeert de familiefrequentie bijvoorbeeld met de gebruiksfrequentie van alle woorden die tot de (directe) familie van de stam BOUW in het V-lexeem $[\{0/GE\} + BOUW]_V$ behoren; het V-lexeem zelf telt echter niet mee.
- cumulatieve familiefrequentie ("cumulative family frequency"): zelfde als de familiefrequentie, maar dan inclusief de gebruiksfrequentie van het referentielexeem; in dit geval dient het V-lexeem $[\{0/GE\} + BOUW]_V$ dus wel te worden meegeteld.

Mijns inziens ontbreekt hier nog een frequentiemaat, namelijk de stamfrequentie:

- stamfrequentie: som van de gebruiksfrequenties van alle lexemen die direct of indirect van een gegeven stam zijn afgeleid, bijvoorbeeld de morfologische basisstam BOUW.

Uiteraard is men bij de bepaling van de hier besproken frequentiematen afhankelijk van de morfologische theorie die ten grondslag ligt aan de uitgevoerde analyses. Het door mij gepresenteerde woordparadigma berust bijvoorbeeld op het idee dat een woord als *bouwer* niet rechtstreeks van een werkwoord (c.q. V-lexeem) is afgeleid, maar van de stam van dit V-lexeem (dus BOUW, een morfeem zonder categoriespecificatie). De onderzoeksgroep van Baayen baseert de gebruiksfrequentie en familieomvang echter op de informatie in CELEX. De hierin opgenomen morfeeminformatie weerspiegelt de morfologische criteria van het grammaticale standaardmodel, zodat het N-lexeem *bouwer* hier als een directe derivatie van het V-lexeem *bouw* geldt, net als de V-lexemen *verbouwen* en *bebouwen*.

[64] De hier gegeven termdefinities weerspiegelen mijn eigen interpretatie van de Engelse termdefinities van Baayen & al.; wegens de complexiteit van het morfologische domein zouden formele definities beter zijn.

2.4.3.3 De bovengrens van het lexicon

Er is veel onderzoek gedaan naar de vraag wat de bovengrens is van het mentale lexicon, d.w.z. wat de grootste taaleenheden zijn die systematisch in het lexicon worden opgeslagen. Vanuit een grammaticaal lexiconperspectief zou men namelijk verwachten dat het lexicon uitsluitend morfemen opslaat en geen plaats biedt aan morfologisch complexe woorden (want hiervan wordt aangenomen dat ze door middel van regels worden geconstrueerd). Toch is al heel lang bekend dat het mentale lexicon ook informatie over morfologisch complexe woorden en inflectievormen kan opslaan. Daarom is het interessant om na te gaan of alle taaleenheden die met enige regelmaat voorkomen in het lexicon zijn terug te vinden, of dat er toch een bovengrens bestaat. In dit verband gaat de aandacht vooral uit naar de overgang van lexeem naar inflectievorm. Want de lexeeminventarisatie van een taal is betrekkelijk constant, en zou dus in aanmerking kunnen komen voor opslag. Maar dit ligt anders voor inflectie, want in een taal als Turks bestaan zoveel inflectievormen dat het onmogelijk zou zijn om deze allemaal op te slaan.

Inmiddels is duidelijk dat deze laatste redenering ongeldig is. Want Aksu & Slobin (1984) hebben aangetoond dat de inflectievormen van het Turks op grote schaal worden gememoriseerd. Lukatela & al. (1980) komen tot dezelfde conclusie voor het Servisch, MacWhinney (1978) voor het Hongaars en Berman (1981) voor het Hebreeuws. Met betrekking tot talen als Engels, Duits en Nederlands zijn genuanceerdere resultaten gepubliceerd. Voor deze talen is aangetoond dat hoogfrequente inflectievormen worden opgeslagen, maar dat laagfrequente inflectievormen meestal regelgebaseerd zijn (cf. Stemberger & MacWhinney (1988), Baayen & al. (1997; 2002). Voorts hebben Bertram & al (2000) aangetoond dat de keuze tussen opslag en derivatie bepaald wordt door een samenspel van verschillende factoren (waaronder gebruiksfrequentie en woordvormingstype).

Deze resultaten laten zien dat het mentale lexicon mogelijk geen bovengrens kent, maar dat de keuze tussen wel of niet opslaan eerder door frequentie-overwegingen wordt bepaald. Dit idee wordt versterkt door een studie van Hay & Baayen (2002), die laten zien dat het Nederlands een parseerlinie ("parsing line") kent, in de zin dat de keuze tussen opslag of derivatie van morfologisch complexe woorden rechtstreeks afhangt van de verhouding tussen de gebruiksfrequentie van het complexe woord en het grondwoord: alleen als het complexe woord frequenter is dan het grondwoord zal het rechtstreeks uit het lexicon worden gehaald, anders zal het via het grondwoord worden geconstrueerd. Dit impliceert dat er geen absoluut frequentiecriterium bestaat, maar alleen een relatief frequentiecriterium. Dit resultaat zou overigens vrij ernstige gevolgen kunnen hebben voor eerdere experimenten, want deze berusten vaak op de aanname dat er wel een absoluut frequentie-effect bestaat.

2.4.3.4 Het transparantiecriterium

Schreuder & Baayen (1995) gaan ervan uit dat alleen woorden met een transparante structuur in aanmerking komen voor een compositieroute. Dit is echter minder vanzelfsprekend dan het lijkt. Want in de praktijk krijgen bestaande (lexicaal opgeslagen) woorden allerlei onvoorspelbare (maar wel gerelateerde) betekenistoepassingen. Hierbij kan men een continuüm waarnemen van sterk compositionele naar sterk opake (gelexicaliseerde) betekenissen. Ik zal dit toelichten aan de hand van het N-lexeem *schrijver*. In de standaardvisie geldt *schrijver* als een productieve afleiding van het V-lexeem *schrijven* (of liever, van de stam SCHRIJF), waarbij de betekenisbijdrage van het suffix -ER wordt omschreven als "de persoon die de handeling uitgedrukt door V verricht"), en secundair (indien de eerste omschrijving niet beschikbaar is of slecht in de context past) "het voorwerp dat de handeling uitgedrukt door V verricht". Bij nadere beschouwing is dit echter een erg algemene betekenisomschrijving die hooguit een deel van de contextueel begrepen betekenis dekt of zelfs incompatibel is met deze betekenis.

Zo correspondeert het woord *schrijver* zelden met iemand die (op het moment van spreken) aan het schrijven is, maar gaat het meestal om iemand die een compleet artikel of een boek heeft geschreven (als incidenteel *schrijver*, waarbij nog onderscheid kan worden gemaakt tussen echte auteurs en fictieve auteurs)[65] of die dat regelmatig doet (als habitueel *schrijver*). In het geval van een romancier (dus een habituele romanschrijver) kan de betreffende persoon ook buiten de context van zijn schrijfactiviteiten als *schrijver* worden aangeduid, want onder invloed van de Romantiek wordt er alom vanuit gegaan dat schrijvers en andere kunstenaars over bijzondere inspiratie beschikken die onverbrekelijk met hun identiteit als mens is verbonden. Hierdoor is het zelfs mogelijk om de paradoxale status van "onsterfelijke schrijver" te verwerven, dus om voort te leven in de boeken die men tijdens zijn leven heeft geschreven (of ze nu wel of niet zijn voltooid)[66]. Maar het woord *schrijver* kent ook de nodige interpretatierestricties. Zo zal een ambtenaar die beleidsstukken voorbereidt zelden als *schrijver* bekend worden (hooguit als *schaduwschrijver*). Verder is het ongebruikelijk om het woord *schrijver* als zelfstandige aanduiding van een pen, typemachine of registreertoestel te gebruiken.[67]

Er blijken dus allerlei bijzondere eisen te worden gesteld aan de toepassing van het woord *schrijver*, en dit geldt ook voor andere woorden met het suffix -ER, zoals men zelf kan nagaan voor woorden als *bakker*, *speler*, *dromer*, *spreker* en *schutter* (als afleiding van het werkwoord *schieten*). Hoewel voor elke betekenistoepassing geldt dat een taalgebruiker deze op één of andere manier moet kunnen herkennen (en er dus ook een mentale representatie voor moet kunnen aanmaken), duidt het feit dat taalgebruikers heel goed weten wat standaard beschikbare betekenismogelijkheden zijn (getuige de definities in een woordenboek) erop dat ze deze woorden niet steeds opnieuw van een contextueel passende betekenis voorzien, maar dat ze elk van deze toepassingen integraal in het geheugen opslaan. Hierbij is niet duidelijk in hoeverre dergelijke woordtoepassingen (d.w.z. niet-voorspelbare vorm-betekenis-relaties) interne structuur bezitten.

Het onderliggende probleem is dat dualistische activatiemodellen per definitie uitgaan van het idee dat polymorfemische woorden langs productieve weg uit hun samenstellende morfemen kunnen worden afgeleid en dus een compositionele betekenis bezitten. Indien geen sprake (meer) is van een compositionele betekenis (zoals bij de genoemde interpretaties van het formeel gelede *schrijver*) verandert zo'n woord in een monomorfemisch woord en is geen concurrentie mogelijk tussen de compositionele representatie en de lexicale representatie. Dit is een ernstig probleem, want in de praktijk is er onder de bestaande woorden bijna geen enkel polymorfemisch woord waarvan de betekenis volledig transparant is. De transparantie beperkt zich namelijk bijna altijd tot de morfosyntactische kenmerken. Er bestaan ook woordvormen die zowel een gelede als een ongelede analyse toestaan; zo kent het woord *koper* een gelede toepassing met de betekenis ("iemand die iets koopt") en een ongelede toepassing met de betekenis "materiaalsoort"; in de dualistische activatiebenadering kunnen deze niet gescheiden worden. Het morfologische concurrentiemodel kent dus fundamentele beperkingen die het ongeschikt maken als model van het mentale lexicon. De enige uitweg is om een nieuwe definitie van morfologische transparantie te introduceren, namelijk een definitie waarbij de

[65] Zo is "Logic, Language and Meaning" (Gamut, 1991), een introductieboek tot de formele semantiek, het product van een schrijverscollectief met de naam L.T.F. Gamut (= Groningen, AMsterdam, UTrecht). En in politieke kringen verschijnen vaak publicaties in naam van een minister of een politieke voorman, terwijl de publicatie in werkelijkheid het product is van een naaste medewerker of een groep ambtenaren.

[66] Er zijn namelijk tal van schrijvers die wereldberoemd zijn geworden met een nooit voltooid boek (vaak hun laatste boek), blijkens de beschouwing ["Het onvoltooide boek" in NRC-H, juni 2004].

[67] Deze betekenissen zijn wel mogelijk indien *schrijver* als rechterdeel van een samenstelling optreedt, blijkens *fijnschrijver*, *viltschrijver, meerpuntschrijver* ("type pen"), *magneetbandschrijver* ("type typemachine") en *penschrijver*, *puntschrijver*, *stjftschrijver*, *papierschrijver*, *storingschrijver* ("type registreertoestel").

transparantie zich mag beperken tot een voorspelbare relatie tussen klankvorm en morfosyntactische kenmerken.

Het hier gesignaleerde probleem heeft ook gevolgen voor de schatting van de relatieve gebruiksfrequentie van de morfemen. Want indien men blijft vasthouden aan de eis van semantische transparantie, zal men het corpus dat ten grondslag ligt aan de frequentiecijfers moeten heranalyseren met betrekking tot de semantische transparantie van de morfologisch gelede woordvormen. Hierbij moet voor elke morfologisch gelede woordvorm worden nagegaan of deze morfosyntactisch en/of semantisch transparant is. Aan de andere kant zullen tal van niet-gecodeerde pseudomorfemen (d.w.z. niet-productieve segmenten met een voorspelbaar effect op de morfosyntactische eigenschappen) alsnog morfeemstatus moeten krijgen en dus ook mee moeten worden genomen in het experimentele onderzoek naar de mentale structuur van bestaande woorden. De psychologische relevantie van pseudomorfemen blijkt overigens ook uit een studie van Baayen zelf, namelijk Schreuder, Burani & Baayen (2001) en uit de studies die aan het eind van §1 zijn genoemd.

Een en ander kan grote gevolgen hebben voor de uitkomsten van het experimentele onderzoek. Zo maken veel experimenten met Nederlandse en Engelse woordherkenning gebruik van CELEX-frequenties. CELEX berust echter op een annotatiemethode waarbij alle woorden die qua vormkenmerken als morfologisch geleed kunnen gelden als zodanig zijn gecodeerd, dus ongeacht hun semantische transparantie. Dit zou kunnen verklaren waarom de cumulatieve familiefrequentie geen effect blijkt te hebben op de reactietijden, terwijl er wel een effect is van de familiegrootte.

De hier gesignaleerde problemen gelden overigens niet voor het activatiemodel van Taft (1994), want in dit model blijft de bijdrage van de morfemen (en pseudomorfemen) in eerste instantie beperkt tot de vormdimensie: in deze benadering wordt het woord *schrijver* bijvoorbeeld altijd via de toegangseenheden SCHRIJF en ER geactiveerd. Tezamen geven deze toegang tot de lexicale eenheid SCHRIJVER, waarvoor een of meer idiosyncratische betekenissen kunnen worden vastgelegd. Maar indien men dit woord in de compositionele betekenis wil gebruiken, dient men deze betekenis uit de optioneel activeerbare betekenissen van de samenstellende morfemen te construeren. Taft geeft overigens niet aan hoe het semantische compositieproces verloopt.

2.4.4 Conclusie

In deze sectie heb ik een beeld proberen te geven van de centrale onderzoeksvragen van studies met een psycholinguïstisch lexiconperspectief. Hierbij is duidelijk geworden dat het psycholinguïstische perspectief een welkome aanvulling biedt op het grammaticale lexiconperspectief, doordat het zich primair met de bestaande woordenschat bezighoudt. In dit verband is een belangrijke vraag in hoeverre morfologisch complexe woorden in het mentale lexicon worden opgeslagen en hoe deze kennis is gestructureerd. Inmiddels is duidelijk dat het lexicon veel meer kennis opslaat dan in grammaticale studies verondersteld wordt: want naast eenvoudige basiseenheden (c.q. morfemen) blijkt het lexicon ook morfologische derivaties en inflectievormen te kunnen opslaan. Volgens de huidige generatie kennismodellen, waaronder het hybride activatiemodel van Schreuder & Baayen (1995) is de keuze tussen de lexicale route en de compositionele route rechtstreeks afhankelijk van de tokenfrequentie van de benodigde eenheden (te weten de samenstellende morfemen en het lemma als geheel). Hoe deze interactie precies verloopt is nog steeds een belangrijke onderzoeksvraag waarover tot op heden geen concensus is bereikt. Op dit punt ondervindt het model van Schreuder & Baayen dan ook concurentie van andere modellen, waaronder de modellen van Caramazza & al. (1988), Taft (1994), Laudanna & Burani (1995), Giraudo & Grainger (2001; 2002) en Baayen & Hay (2002).

Desondanks is nu al duidelijk dat de meeste activatiemodellen ernstig tekortschieten, want een compleet activatiemodel moet zich niet beperken tot de beschrijving van morfologisch transparante woorden, maar ook gradaties in morfologische verwantschap kunnen verantwoorden (zoals gradaties in semantische transparantie). Hoewel connectionistische netwerkmodellen dit standaard lijken toe te staan, kennen dergelijke modellen twee belangrijke beperkingen: ten eerste kunnen connectionistische modellen geen woorden (en dus ook geen gebruiksfrequenties) opslaan, maar alleen regelmatige patronen verantwoorden; ten tweede zijn dergelijke modellen bijzonder intransparant, waardoor ze geen inzicht geven in de representatiestructuur. Dit gaat niet alleen ten koste van hun verklarende kracht, maar ook van de mogelijkheden tot systematische kennisverantwoording in een computationeel lexicon. Daarom geef ik de voorkeur aan een symbolisch representatiemodel dat expliciet inzicht geeft in de morfologische (fonologische en semantische) structuurkenmerken van bestaande woorden. In dit opzicht biedt het activatiemodel van Taft (1994) een interessante aanzet. In de volgende sectie zal ik aangeven aan wat voor eisen zo'n model precies moet voldoen.

2.5 Naar een Integraal Dynamisch Lexiconsysteem

2.5.1 Introductie

In de voorgaande secties heb ik twee verschillende perspectieven op morfologische structuur behandeld, namelijk het grammaticale lexiconperspectief en het psychologische lexiconperspectief. In het grammaticale perspectief bleek de aandacht vooral uit te gaan naar de vraag welke grammaticaregels beschikbaar zijn voor de productieve aanmaak van nieuwe (al dan niet bestaande) woorden. In het psychologische perspectief daarentegen lag het accent meer op de vraag in hoeverre bestaande woorden werkelijk morfologische structuur bezitten, d.w.z. in hoeverre de grammaticaal gemotiveerde morfeemstructuur een rol speelt bij de mentale representatie van deze woorden. Het aan dit perspectief verbonden onderzoek wordt gedreven door de veronderstelling dat het mentale lexicon in staat is om behalve basiswoorden ook kennis over morfologisch complexe woorden op te slaan.

Met betrekking tot monistische opslagmodellen kan onderscheid worden gemaakt tussen modellen die uitgaan van eenvoudige woordopslag en modellen waarbij de opgeslagen woorden interne structuur (kunnen) bezitten. Hoewel dit laatste modeltype nog nauwelijks is benut,[68] heeft het zowel theoretisch als empirisch de beste papieren. Theoretisch gezien is een model waarin slechts één representatie per taaleenheid nodig is immers te verkiezen boven een model waarin twee verschillende representaties moeten worden verondersteld. En empirisch gezien is het noodzakelijk om een model te ontwikkelen dat recht kan doen aan de toenemende evidentie dat de herkenning van bestaande woorden gevoelig is voor de omvang van de stamfamilie en de gebruiksfrequentie van de samenstellende morfemen. De introductie van een intern gestructureerd lexicon is echter niet verenigbaar met het grammaticale standaardmodel. Dit heeft ook gevolgen voor de empirische onderbouwing van dit voorstel, want veel psycholinguïstische studies berusten op een model dat sterk beïnvloed is door inzichten uit grammaticaal onderzoek.

Dergelijke problemen kunnen worden voorkomen als men het grammaticale lexiconperspectief van begin af aan met het psychologische lexiconperspectief combineert en een Integraal Dynamisch Lexicon-systeem probeert te construeren. Dit voorstel wordt toegelicht door aan te geven aan welke eisen een theorie moet voldoen om als IDL-systeem te kunnen

[68] Het eerste en voorzover mij bekend meest complete voorstel dat uitgaat van morfologisch gestructureerde woordrepresentaties is het informele lexiconmodel van Bybee (1985; 1988). Voor een meer formele uitwerking van dit principe is men (vooralsnog) aangewezen op statische kennismodellen, zoals DATR en HPSG.

worden aangemerkt (H2.5.2) en door aan te geven hoe zo'n systeem concreet kan worden vormgegeven (H2.5.3); hierbij wordt ook de lexicografische praktijk besproken.

2.5.2 Basiseisen voor een Integraal Dynamisch Lexiconsysteem

In deze subsectie bespreek ik een aantal eigenschappen bespreken die cruciaal zijn voor een IDEL-systeem, d.w.z. een kennismodel dat zich niet beperkt tot de beschrijving van een lexicaal deeldomein, zoals het domein van het lexicografische, grammaticale of psychologische lexiconperspectief, maar dat de betreffende kennis ook in onderlinge samenhang kan beschrijven en dat bovendien recht doet aan de dynamische dimensie van deze kennis. Om de status van IDL-systeem te kunnen verwerven, dient een lexicaal kennismodel aan de volgende eisen te voldoen:

1) de lexicon-component moet zo zijn opgezet dat het in potentie alle parate kennis over de bestaande woordenschat van een individuele taalgebruiker kan representeren. Dit betekent:

a) dat elk bestaand woord (en elke woordinterne component) een lexicale ingang moet kunnen krijgen (ook als het een gelede structuur bezit);
b) dat bij elke ingang informatie wordt gegeven over de spelvorm, de klankvorm, de betekenis, de lexeemcategorie, het inflectieparadigma, de morfologische en syntactische selectiekenmerken en de gebruiksfrequentie;
c) dat voor elk morfologisch complex woord (of kleinere eenheid) informatie wordt gegeven over de morfologische structuurkenmerken, ook indien sprake is van onregelmatige vorm- of betekeniskenmerken. Meer in het bijzonder dient gecodeerd te worden welke woorden van dezelfde morfologische stam zijn afgeleid (op basis van intuïtieve kennis over de morfologische samenhang van de woordenschat);

2) het kennismodel moet inzicht geven in het mechanisme dat ten grondslag ligt aan de generalisatie van de redundantiepatronen in de lexicale kennis, met als extra randvoorwaarde dat het (in potentie) alle bestaande (al dan niet "productieve") generalisatiemogelijkheden moet kunnen verantwoorden (inclusief hun waarschijnlijkheid);

3) het kennismodel moet inzicht geven in het mechanisme dat ten grondslag ligt aan de verwerving van de lexicale kennis; dit verwervingsmechanisme moet een taalonafhankelijke opzet krijgen, wat inhoudt dat het moet uitgaan van taalonafhankelijke leerprincipes;

4) het kennismodel moet inzicht geven in het mechanisme dat ten grondslag ligt aan de activatie van lexicale kennis, bijvoorbeeld door een duaal zoekalgoritme te definiëren (d.w.z. een zoekalgoritme dat tegelijk het lexicon kan doorzoeken en constructieregels kan activeren); net als het leermechanisme dient dit zoekmechanisme zo taalonafhankelijk mogelijk gedefinieerd te worden;

5) het kennismodel moet een dynamische opzet hebben, wat inhoudt dat alle kennis in het lexicon permanent en incrementeel kan worden aangepast aan nieuwe gebruiksinformatie; meer in het bijzonder dient er een mechanisme te worden gedefinieerd dat in staat is om de morfologische structuurpatronen en de hiervan afgeleide generalisaties inductief uit de kwalitatieve en kwantitatieve eigenschappen van de bijbehorende taaleenheden af te leiden (tenzij sprake is van een bewust afgesproken taalregel).

Met betrekking tot de relatie tussen lexicon en regelsysteem kunnen vier hoofdklassen worden onderscheiden, namelijk statische kennismodellen, kwantitatief-dynamische kennismodellen, kwalitatief-dynamische kennismodellen en integraal-dynamische kennismodellen. Hiernaast kan een hybride klasse worden onderscheiden. Ik zal elk van deze klassen kort toelichten:

- statische kennismodellen: bij deze modellen vertoont het morfologische representatiesysteem noch kwalitatieve, noch kwantitatieve gevoeligheid, wat impliceert

dat deze modellen op een puur deductief regelsysteem berusten. In generatieve kringen wordt dit gerechtvaardigd door erop te wijzen dat de grammatica een kennissysteem is dat grotendeels op aangeboren structuurprincipes en categorieën berust en dat reeds in de kinderjaren verworven wordt. Bovendien zou dit verwervingsproces op een taalspecifiek leermechanisme berusten. Lexicale representatietalen als HPSG, LFG en DATR kennen eveneens een statische opzet.

- kwantitatief-dynamische kennismodellen: bij deze modellen is het morfologische representatiesysteem alleen gevoelig voor de gebruiksfrequentie van de reeds opgeslagen taaleenheden; hierbij blijft in het midden hoe men deze taaleenheden zou moeten ontdekken. Dit uitgangspunt is kenmerkend voor de morfologische concurrentiemodellen van Taft (1994), Schreuder & Baayen (1995) en Lowie (1998).
- kwalitatief-dynamische kennismodellen: bij deze modellen is het morfologische representatiesysteem alleen gevoelig voor de kwalitatieve eigenschappen van de opgeslagen woorden, in het bijzonder hun semantische transparantie en hun intuïtieve productiviteit; hierbij blijft in het midden hoe men kan voorkomen dat laagfrequente eenheden (zoals hapax-eenheden) als volwaardig morfeem worden geanalyseerd. Dit uitgangspunt is kenmerkend voor lexiconstructurerende modellen zoals Jackendoff (1975), Bochner (1993) en Ford & Singh (1991; 1997). De eerste twee auteurs zijn zich overigens heel goed bewust van de noodzaak van een kwantitatief gevoelige evaluatiematrix, maar komen niet verder dan een schetsmatige uitwerking van deze matrix.
- integraal-dynamische kennismodellen: bij deze modellen berust de identificatie van een morfeem (c.q. morfologisch combinatiepatroon) op een combinatie van kwalitatieve en kwantitatieve criteria. Dit uitgangspunt vormt de basis voor diverse monistische kennismodellen, waaronder Bybee (1985; 1988), Riehemann (1998; 2001) en Seidenberg & Gonnerman (2000).
- hybride kennismodellen: deze klasse correspondeert met voorbeeldgestuurde leermodellen, waaronder het connectionistische netwerkmodel van Rumelhart & McClelland (1986) en het symbolische netwerkmodel van Van den Bosch & Daelemans (19989). Hoewel dergelijke modellen over een inductief leervermogen beschikken, zijn ze afhankelijk van voorgestructureerde trainingsdata. Hierdoor zijn de hieruit voortkomende generalisaties indirect op grammaticale structuurprincipes gebaseerd. Een tweede beperking is dat dergelijke systemen niet continu kunnen bijleren, maar een afzonderlijke leerfase nodig hebben.

6) Het kennismodel dient een psychologische basis te hebben, wat inhoudt dat het qua structuurprincipes compatibel moet zijn met wat er bekend is over de cognitieve bouwprincipes en dat het qua lexicale gegevens in overeenstemming moet zijn met experimentele resultaten uit psycholinguïstisch onderzoek naar de lexicale kennis van individuele taalgebruikers;

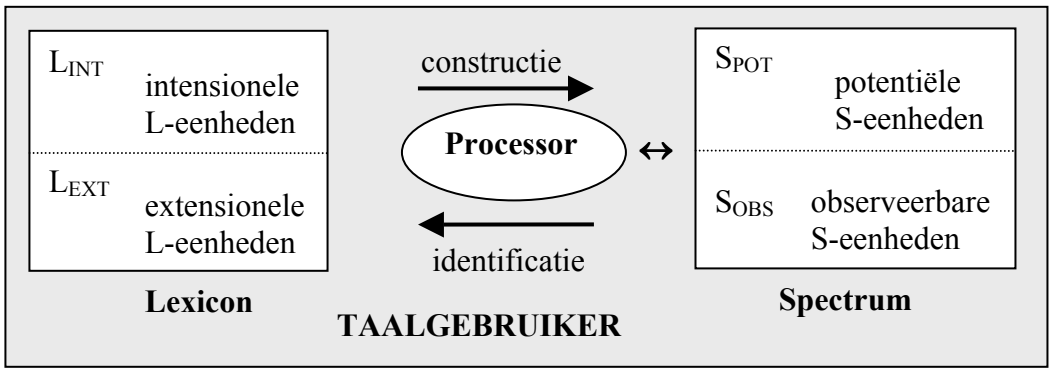
7) Het kennismodel dient empirisch toetsbaar te zijn, wat impliceert dat het model exact gedefinieerde representaties kent en dat het een compleet beeld geeft van het onderzochte domein.

8) Het kennismodel dient lexicografisch implementeerbaar te zijn, wat impliceert dat het formeel moet zijn opgezet en dat er heldere classificatieprincipes worden gebruikt (iets wat in grammaticale morfologiemodellen doorgaans niet het geval is).

2.5.3 De structuur van een Integraal Dynamisch Lexiconsysteem

2.5.3.1 Algemene beschrijving

Deze sectie introduceert een algemeen bouwplan voor een Integraal Dynamisch Lexiconsysteem (IL-systeem). Bij de opzet van dit bouwplan, dat is weergegeven in figuur 2-2, ben ik uitgegaan van de in H2.5.2 besproken criteria. Het door mij voorgestelde systeem bestaat uit een Lexicon (L), een Processor[69] en een Spectrum (S). Het spectrum is een virtueel kennisdomein dat alle taaleenheden omvat die de Processor ooit heeft verwerkt of potentieel uit L-eenheden kan construeren. De processor draagt zorg voor de aansturing van lexicale activatieprocessen, inclusief de constructie en identificatie van nieuwvormingen.



Figuur 2-2: De structuur van een Integraal Dynamisch Lexiconsysteem (IDL-systeem).

Het Spectrum S vormt de verbindende schakel tussen de collectieve en de individuele woordenschat. Dit spectrum bestaat uit twee subcomponenten, namelijk het direct observeerbare deel (S_{OBS}), dat met bestaande woorden correspondeert, en een potentieel deel (namelijk S_{POT}), dat met niet-bestaande, maar wel voorspelbare woorden correspondeert (namelijk woorden die men via regels of analogieprincipes uit bestaande eenheden kan opbouwen). Het observeerbare woordspectrum (S_{OBS}) vormt tevens de empirische basis voor de opbouw van de mentale kennis over bestaande én mogelijke woorden.

Zoals al aan de orde kwam, moet een IDL-model van het mentale lexicon alle bestaande woorden lexicaal kunnen vastleggen (inclusief informatie over woordfrequentie, betekenis, gebruikscontext en eventuele bijzonderheden), waarbij deze kennis zo moet worden gestructureerd dat hij de dynamische basis kan vormen voor de constructie van nieuwe woorden (wat equivalent is aan de inductieve opbouw van een morfologisch regelsysteem). Om dit mogelijk te maken kent het lexicon van een IDL-systeem eveneens twee componenten:[70]

- een component voor extensionele L-eenheden (L_{EXT}), die ofwel met bestaande woorden corresponderen ofwel met bouwstenen van bestaande woorden (namelijk stammen en functors, waarbij het toepassingsdomein van de functor uit een opsomming bestaat van alle stammen waarmee het een bestaand woord kan vormen).
- een component voor intensionele L-eenheden (L_{INT}), te weten een inventarisatie van woordinterne functors (zoals affixen) waarvan het toepassingsdomein met een open verzameling stammen correspondeert; gegeven zo'n functor kan elke stam die aan de selectie-eisen voldoet de basis vormen voor een nieuw geconstrueerd woord, zelfs als dezelfde combinatie ook al met een bestaand woord correspondeert.

[69] De combinatieprincipes van de Processor kunnen desgewenst op UG-principes worden gebaseerd

[70] De termen *extensioneel* en *intensioneel* ontleen ik aan de modeltheoretische semantiek. In dit vakgebied is de extensie van een woord (bijv. *student*) gelijk aan de verzameling entiteiten die aan het bijbehorende predicaat voldoen ("individuen die studeren"); de intensie correspondeert met een verzameling van mogelijke extensies (bijv. studentpopulaties in een reeks van studiejaren). In mijn toepassing van deze termen correspondeert het verwijzingsdomein echter niet met *semantische* entiteiten, maar met *lexicale* entiteiten (zoals woordvormen).

Het verschil tussen L_{EXT}-eenheden en L_{INT}-eenheden kan het makkelijkst worden toegelicht aan de hand van een concreet voorbeeld. Neem bijvoorbeeld het suffix -ER in de betekenis "uitvoerder (persoon/instrument) van de handeling uitgedrukt door de stam X in [X + ER]"; dit suffix bezit zowel een extensioneel als een intensioneel gedefinieerd toepassingsdomein. Het eerste domein correspondeert met een opsomming van alle stammen X die in combinatie met het suffix -ER een bestaand woord kunnen vormen (c.q. S_{OBS}-woorden), waaronder de stammen SCHRIJF (van *schrijver*), SPREEK (van *spreker*) en KIJK (van *kijker*). Door over deze stammen te abstraheren kan een intensioneel combinatiepatroon worden afgeleid waarbij X met een intensionele stam (c.q. stamtype) correspondeert (bijv. transitieve en unergatieve V-stammen); dit intensionele patroon vormt de basis voor nieuwvormingen (c.q. S_{POT}-woorden), zoals *zapper* (van *zappen*), *surfer* (van *surfen*) en *chiller* (van *chillen*).

De Processor is de drijvende kracht achter de interactie tussen L en S, want gegeven een taalgebruiker T zorgt de Processor voor de identificatie van de aan T aangeboden S-eenheden (door de S-eenheden als een combinatie van één of meer L-eenheden, zoals morfemen, te analyseren) en voor de constructie van door T zelf te uiten S-eenheden (door één of meer L-eenheden, zoals morfemen, te selecteren en hier een uitspreekbare eenheid van te maken). Indien sprake is van een voor T bekende S-eenheid (dus een S_{OBS}-eenheid), kan de Processor volstaan met de selectie van L_{EXT}-eenheden. Maar indien sprake is van een nieuwvorming (dus een S_{POT}-eenheid), is minimaal één L_{INT}-eenheid nodig.

Het is van belang om lexicon L en spectrum S niet door elkaar te halen: in het IL-systeem is L namelijk een theoretisch instrument om S-eenheden te coderen, wat impliceert dat de inhoud van L niet rechtstreeks waarneembaar is. In plaats daarvan is men aangewezen op intuïtieve of experimentele observaties aan een deeldomein van woordspectrum S, want zoals aan de orde kwam in H2.4 kan men hieruit aflezen welke kennis via L_{EXT} (c.q. als ongeleed woord) en welke kennis via L_{INT} (c.q. als morfeemcombinatie) verantwoord dient te worden.[71] In de onderzoekspraktijk correspondeert S vaak met een zo representatief mogelijke selectie uit het empirische domein waarvoor men een taalmodel wil opstellen. Deze S dient dan als toetsingskader voor dit deelmodel. Maar in het grammaticale onderzoek wordt de relatie tussen taalmodel en spectrum (c.q. toetsingskader) zelden geëxpliciteerd: men vertrouwt er namelijk op dat enkele goedgekozen voorbeelden wel zullen volstaan.

De explicatie van S is wel een standaardonderdeel van experimenteel taalonderzoek, want de samenstelling van S is hier bepalend voor de validiteit van de onderzoeksopzet. Maar voor zover mij bekend bestaan er geen kennismodellen waarin het spectrum S als fundamenteel onderdeel van het taalsysteem zelf wordt gerepresenteerd, hoewel dit naar mijn mening noodzakelijk is voor een complete verantwoording van het taalsysteem. Dat hier weinig aandacht voor is, hangt waarschijnlijk samen met de dominante status van het grammaticale taalmodel. Want in dit model correspondeert de kern van het taalvermogen met een statisch (mogelijk aangeboren) regelsysteem dat de deductieve basis vormt van de door taalgebruikers voortgebrachte woorden en zinnen en dat niet door deze data beïnvloed wordt. In mijn visie daarentegen correspondeert het taalvermogen met een dynamisch kennissysteem waarvan de morfologische patronen langs inductieve weg uit de voor taalgebruikers waarneembare taaldata moet worden afgeleid (namelijk de taaldata in Spectrum S).

[71] De tussen haakjes toegevoegde specificaties hebben betrekking op het duale activatiemodel van Schreuder & Baayen (1995) of soortgelijke modellen; in dergelijke modellen wordt (nog) geen onderscheid gemaakt tussen extensionele en intensionele morfeemcombinaties, zodat lexicaal opgeslagen woorden altijd ongeleed zijn.

2.5.3.2 Een voorbeeld

Ter verduidelijking van het IL-systeem bespreek ik nu een concreet voorbeeld, namelijk de lexicale representatie van de woorden uit het volgende deelspectrum (S1):

> S1 = {*schouw, schouwbaar, schouwen, schouwer, schouwing, schouwtoneel, beschouwen, beschouwelijk, beschouwer, beschouwing, welbeschouwd*}

Ik ga ervan uit dat S1 uit voor taalgebruiker T bekende woorden bestaat, dus dat S1 met S_{OBS}-eenheden (c.q. observationele S-eenheden) correspondeert. Dit impliceert dat deze woorden tot het extensionele deel van het lexicon behoren (ook als ze morfologisch geleed zijn) en dat het lexicon informatie kan geven over hun gebruiksfrequentie (die minimaal de waarde 1 moet bezitten) en de niet-compositionele betekenissen. Zo geeft de GWNT aan dat het naamwoord *schouw* wel vier verschillende betekenissen kent, waaronder 'stookplaats', 'inspectie', 'bepaald type vaartuig' en toepassing als adjectief in de betekenis 'schuw' of 'scabreus'; van deze vier betekenissen kunnen er natuurlijk meerdere in het lexicon van de taalgebruiker voorkomen. Bij *beschouwing* moet onderscheid worden gemaakt tussen een interpretatie als handeling en een interpretatie als object (namelijk handelingsplan) en bij *beschouwelijk* moet melding worden gemaakt van de connotatie 'contemplatief', 'nadenkend'. Tot slot moet bij *schouwtoneel* worden aangeven dat dit woord alleen in vaste uitdrukkingen voorkomt.[72]

woord	klasse	morfeemstructuur	betekenissen	frequentie	opm.
$schouw_1$	N	[schouw]	b1, b2, b3	122, 18, 4	
$schouw_2$	A	[$schouw_a$]	b4	1	
schouwbaar	A	[schouw]baar	b1	1	
schouwen	V	[schouw]en	b1	23	
schouwer	N	[schouw]er	b1	17	
schouwing	N	[schouw]ing	b1	10	
schouwtoneel	N	[schouw][toneel]	b1	15	Vondel
beschouwen	V	be[schouw]en	b1, b2	388, 24	
beschouwelijk	A	be[schouw]elijk	b1	156	
beschouwer	N	be[schouw]er	b1	276	
beschouwing	N	be[schouw]ing	b1, b2	254	
welbeschouwd	A	wel;be[schouw]d	b1	57	

Tabel 2-2: Voorbeeld van een morfologisch gestructureerd lexicon.

Intuïtief gezien zijn de woorden uit S1 allemaal op dezelfde stam gebaseerd, namelijk de stam SCHOUW. Toch zijn hun betekenissen lang niet altijd compositioneel van deze stam af te leiden. Het lexicon zal dus op een of andere manier moeten coderen welke woorden tot dezelfde stam behoren. Dit kan worden verantwoord door alle woorden als morfologisch gestructureerde eenheden op te slaan (voor zover ze structuur vertonen, uiteraard). Tabel 2-2 geeft een concreet voorbeeld van deze representatiemethode.

2.5.3.3 Lexicale submodules

Hoewel veel morfologiemodellen onderscheid maken tussen productieve (primair intensionele) en improductieve (primair extensionele) woordvormingspatronen, bestaan er grote verschillen in de mate waarin ze extensionele informatie opslaan. Om deze variatie goed te kunnen verantwoorden, dient het lexicondeel uit het IL-systeem in kleinere modules te worden opgedeeld. Hierbij correspondeert elke module met een specifiek type informatie,

[72] De GWNT citeert bijvoorbeeld een gevleugelde uitspraak van de zeventiende-eeuwse dichter Vondel: "*De wereld is een schouwtoneel, elk speelt zijn rol en krijgt zijn deel.*"

zoals zichtbaar in tabel 2-3.[73] Deze tabel specificeert aparte lexiconmodules voor morfemen, lexemen en woorden. Bovendien is elke module onderverdeeld in 1-tupels (die met zelfstandige eenheden corresponderen) en 2^+-tupels (die hier met een vaste combinatie van twee of meer eenheden corresponderen). Dit onderscheid is niet alleen mogelijk op het niveau van de morfemen, maar ook op het niveau van de lexemen (want naast enkelvoudige lexemen kan men ook samenstellingen aantreffen) en de woorden (want naast enkelvoudige woorden bestaan ook gelexicaliseerde woordgroepen, bijv. *met rode koontjes*, *een snelle jongen* of *normen en waarden*). Tot slot toont de tabel dat men per klasse van eenheden onderscheid kan maken tussen extensionele kennis (d.w.z. eenheden die rechtstreeks uit het lexicon komen) en intensionele kennis (d.w.z. eenheden die alleen beschikbaar zijn door over bestaande eenheden te generaliseren, bijv. door affixatie van een bestaand lexeem).

Voor elk van de onderscheiden submodules is een aparte code gespecificeerd; zo correspondeert de code L-1e met extensionele eenheden die uit precies één lexeem bestaan (ter onderscheiding van samenstellingen) en de code W-2i met intensionele eenheden die uit twee of meer woorden zijn opgebouwd (dus met gelexicaliseerde woordgroepen). Omdat er nog geen term bestaat waarmee naar elk van de hier onderscheiden eenheidsklassen kan worden verwezen, heb ik er zelf een term voor bedacht, namelijk *taxeem* (c.q. plaatsingseenheid).[74]

submodule:	morfemen		lexemen		woorden	
tupel-type:	1-tupels	2^+-tupels	1-tupels	2^+-tupels	1-tupels	2^+-tupels
L (intens.)	M-1i	M-2i	L-1i	L-2i	W-1i	W-2i
L (extens.)	M-1e	M-2e	L-1e	L-2e	W-1e	W-2e

Tabel 2-3: Inventarisatie van lexicale modules. Elke submodule (te weten morfemen, lexemen en woorden) bestaat uit een specifieke klasse van eenheden; deze zijn horizontaal onderverdeeld in 1-tupels en 2^+-tupels en verticaal in extensionele en intensionele eenheden.

In de vier deeltabellen van tabel 2-4 worden de hier geïntroduceerde submodules (namelijk morfemen, lexemen en woorden) geconcretiseerd door per tupeltype (namelijk voor 1-tupels en voor 2^+-tupels) een aantal voorbeelden te geven. Bovendien wordt per categorie een aanvullend onderscheid gemaakt tussen inheemse en uitheemse morfemen. Tabel 2-4a geeft voorbeelden van lexicale basismorfemen (M1-tupels), terwijl tabel 2-4b voorbeelden geeft van lexicale morfeemcombinaties (M2-tupels). Om de stammen goed van de affixen te kunnen onderscheiden, zijn ze vetgedrukt.

M1-tupels (morfemen)	M-1e/i (inheems):	M-1e/i (uitheems):
stammen [+autonoom]	**werk**, **speel**, **hand**, **tuin**, **groot**	**legaal**, **foto**, **bureau**, **portret**
stammen [-autonoom]	√**aam**, √**schied**, √**deft**, √**ham**	√**form**, √**duct**, √**bio**, √**graaf**
affixen [+autonoom]	op-, af-, in-, uit-, -kunde	sub-, ex-, post-, -soof, -graaf
affixen [-autonoom]	-er, -aar, -ij, -ig, ge-, -te, ver-	-iteit, -eer, -ie, -ist, in-, re-

Tabel 2-4a: Voorbeelden van lexicale basismorfemen (M1-tupels).

M2-tupels (morfemen)	M-2e/i (inheems):	M-2e/i (uitheems):
stam-stam-combinaties	**hand**+**werk**, **speel**+**goed**, √**aam**+**beeld**, √**okker**+**noot**	**aqua**+√**duct**, **bureau**+√**craat** **foto**+√**graaf**, √**bio**+√**graaf**,

[73] De door mij voorgestelde partitionering vertoont veel overeenkomsten met die van Ten Hacken (1994).

[74] Ik heb deze term afgeleid van de wortel TAX in *syntaxis*; dit morfeem, dat etymologisch gezien ook de basis vormt van woorden als *tactisch* en *tactiek*, kenmerkt zich door het betekenisaspect 'plaatsing', 'opstelling'. De hier voorgestelde term is overigens niet nieuw, blijkens een vermelding in (o.a.) het online-glossarium op www.tiscali.co.uk/reference/dictionaries/difficultwords/, waar *taxeme* wordt gedefinieerd als "linguistic feature (e.g. difference in stress, pronunciation or word-order) differentiating otherwise identical utterances".

stam-affix-combinaties	**hand**-ig, **speel**-er, uit-ver-**groot** ge-√**schied**-enis, √**deft**-ig	**legaal**-iteit, **portret**-eer, in-√**form**-eer, re-√**duct**-ie
affix-affix-combinaties	-er-ij, -ar-ij, -er-ig, ge-..-te	-at-or, -is-eer, -iv-at, -ion-al

Tabel 2-4b: Voorbeelden van lexicale morfeemcombinaties (M2-tupels).

De eenheden in tabel 2-4a zijn inherent extensioneel, want om intensionele morfeemcombinaties te kunnen vormen zijn lexicaal geïntroduceerde basiseenheden nodig. Deze redenering geldt niet per se voor hogere taxeemniveaus (zoals de eenheden in tabel 2-4b). Zo gaan grammaticale morfologiemodellen er meestal vanuit dat regelmatig afleidbare lexemen niet extensioneel (langs lexicale weg) hoeven te worden verantwoord, al heeft dit als nadeel dat geen frequentiegegevens kunnen worden vastgelegd.

De linkerkolom van tabel 2-4a toont vier morfologische subklassen, namelijk alle subtypes die beschikbaar zijn op basis van het onderscheid stammen vs. affixen en het onderscheid autonome (c.q. vrije) vs. niet-autonome (gebonden) eenheden. Het eerste onderscheid (stammen vs. affixen) kan formeel worden uitgewerkt in termen van variabele-vrije eenheden (c.q. stammen) en variabele-introducerende eenheden (c.q. affixen). Het tweede onderscheid (vrij vs. gebonden) correspondeert met een distributiecontrast tussen stammen en affixen: want net als in andere Germaanse talen kunnen vrijwel alle inheemse stammen van het Nederlands zelfstandig worden gebruikt, in tegenstelling tot affixen die doorgaans alleen in gebonden positie voorkomen. Toch kennen zowel de stammen als de affixen eenheden die zich aan deze vuistregel onttrekken, namelijk wortels en geïncorporeerde preposities (die een sterk affixkarakter hebben, blijkens hun woordinterne distributiegedrag en hun moeilijk te specificeren betekenisbijdrage). Door deze subklassen in het lexicale model op te nemen, wordt het mogelijk om na te gaan in hoeverre de bestaande morfologiemodellen in staat zijn om deze subklassen te onderscheiden en het bijbehorende gedrag te verantwoorden.

De hier onderscheiden subklassen zijn niet zomaar toepasbaar op hogere structuurniveaus. Dit hangt samen met het feit dat de morfemen met de kleinste bouwstenen van de grammatica corresponderen. Hierdoor zijn deze bouwstenen relatief goed bestudeerd, met als gevolg dat er ook een fijnmazig netwerk van subklassen beschikbaar is. Op hogere structuurniveaus kunnen echter andere klassen nodig zijn. De vergelijking van deze niveaus kan wel bijdragen aan een dieper inzicht in de overeenkomsten en verschillen tussen het morfeemniveau en de hogere structuurniveaus.

L-tupels (lexemen)	L-e/i (inheems):	L-e/i (uitheems):
L1-tupels, 1 morfeem	**werk**, **speel**, **hand**, **tuin**, **groot**	**legaal**, **foto**, **bureau**, **portret**
L1-tupels, >1 morfeem	**hand**-ig, **speel**-er, √**deft**-ig	**legaal**-iteit, in-√**form**+eer,
L2-tupels	**tuin**+**werk**, **speel**-er-s+ver-**blijf** ge-√**schied**-enis+**boek**	**foto**+**portret**, **portret+foto** **legaal**-iseer-ing-s+**bureau**

Tabel 2-4c: Voorbeelden van basislexemen (L1-tupels) en lexeemcombinaties (L2-tupels).

W-tupels (woorden)	W-e/i (inheems):	W-e/i (uitheems):
W1-tupels, 1 lexeem	**werk**-(t/en/te/ten), **hand**-(en)	**curieus**-(ze), con-**trast**-eer-(t)
W1-tupels, >1 lexeem	**hand**-ig-(e/er/st), **speel**-er-(s)	**legal**-iteit, **politie**+**bureau**-(s)
W2-tupels	in een **speelse bui**, op **handen**	**cum laude**, **femme fatale**-(s)

Tabel 2-4d: Voorbeelden van basiswoorden (W1-tupels) en woordcombinaties (W2-tupels).

De tabellen 2-4c en 2-4d hebben betrekking op resp. het lexeemniveau en het woordniveau. Hierbij correspondeert de verticale dimensie met dezelfde soort van subklassen, te weten:

i) L1-tupels met 1 morfeem resp. W1-tupels met 1 lexeem

ii) L1-tupels met 2 of meer morfemen resp. W1-tupels met 1 of meer morfemen
iii) L2-tupels (bestaande uit twee lexemen), W2-tupels (bestaande uit 2 of meer woorden)

Merk op dat de L1-tupels (ongeacht het aantal morfemen) met dezelfde klankvormen kunnen corresponderen als de M1-tupels. Er is alleen een onzichtbaar verschil in de grammaticale kenmerken, want bij de L1-eenheden is sprake van basislexemen (met inflectiecategorie), terwijl de M1-eenheden met categorieloze en dus niet-verbuigbare basismorfemen corresponderen. Dit hergebruik van lexicale eenheden is ook zichtbaar bij de W1-tupels.

2.6 Conclusie

In dit hoofdstuk heb ik uiteengezet wat reeds bekend is over de structuur en de functies van het mentale lexicon en hoe deze eigenschappen zich onderling verhouden. Met het oog op deze vragen is een systematisch overzicht gegeven van bestaande kennismodellen van het mentale lexicon, waarbij veel aandacht is besteed aan de structurele overeenkomsten en verschillen tussen deze modellen. Hierna ben ik in detail ingegaan op vraagstelling, uitgangspunten, resultaten en beperkingen van twee belangrijke onderzoeksperspectieven, te weten het grammaticale lexiconperspectief en het psychologische lexiconperspectief. Uit deze inventarisatie is naar voren gekomen dat de beperkingen van de bestaande kennismodellen grotendeels het gevolg zijn van het feit dat elk model slechts een deelaspect van het mentale lexicon bestrijkt. In H2.5 heb ik betoogd dat dit probleem kan worden opgelost door de onderzoeksdomeinen van begin af aan in onderlinge samenhang te beschrijven. Voor dit doel heb ik een bouwplan gepresenteerd voor een ideaal model van het mentale lexicon, te weten een Integraal Dynamisch Lexiconsysteem.

Er is sprake van een IDL-model als het model potentieel in staat is om alle functies van het mentale lexicon te verantwoorden, in het bijzonder lexicale kennisopslag, lexicale kennisactivatie en lexicale kennisverwerving. Bovendien dient het lexicon zo te worden opgezet dat het alle parate kennis over de bestaande woordenschat kan verantwoorden, waaronder kennis over gebruiksfrequenties, semantische en fonologische eigenaardigheden, morfologische en syntactische combinatiemogelijkheden en morfologische verbanden met andere woorden. Voorts dient het systeem in staat te zijn om de parate kennis over de potentiële woordenschat te verantwoorden, waartoe het systeem moet kunnen generaliseren over de woordvormingspatronen in de bestaande woordenschat. Bij de opbouw en structurering van deze kennis moet het systeem in staat zijn om inductief te werk gaan (dus zonder terug te vallen op aangeboren regels), wat impliceert dat het over een zelforganiserend leervermogen moet beschikken; in dat geval is sprake van een dynamisch kennismodel.

Eén van de weinige modellen die in aanleg aan alle eisen van een Integraal Dynamisch Lexiconsysteem kan voldoen, is het lexicale netwerkmodel van Bybee (1985; 1988). Dit model is echter zo informeel opgezet dat het moeilijk is te toetsen. Om dezelfde reden leent dit model zich minder goed voor de implementatie in een computationeel systeem. Op dit punt bieden computationele overervingsmodellen als DATR en HPSG meer perspectief; deze zijn echter primair lexicografisch georiënteerd en hebben daardoor een statisch karakter. Er is dus behoefte aan een geavanceerder representatiemodel. Om in deze behoefte te voorzien, zal ik in hoofdstuk 3 en 4 een lexicale representatiemethode uitwerken die volgens mij als fundament kan dienen voor de ontwikkeling van lexiconmodellen die in alle opzichten aan de eisen van een IDL-systeem voldoen.

3 De Nederlandse woordbouw

3.1 Introductie

Dit hoofdstuk biedt een systematisch overzicht van de morfologische aspecten van de Nederlandse woordbouw. Hierbij bespreek ik eerst het traditionele, op grammaticaregels gebaseerde classificatiesysteem uit het Morfologische Handboek van het Nederlands (MHB) van De Haas en Trommelen (1993), om vervolgens een nieuw, op distributieprincipes gebaseerd classificatiesysteem voor te stellen. Dit morfologische classificatiesysteem (dat zich kenmerkt door een paradigmatische analysemethode) wordt gemotiveerd door te laten zien dat de traditionele aanpak (die zich kenmerkt door een syntagmatische analysemethode) een aantal fundamentele problemen kent die eenvoudig zijn op te lossen in het nieuwe systeem.

Het hoofdstuk is als volgt ingedeeld. In H3.2 zal ik uitvoerig ingaan op de MHB-theorie over de Nederlandse woordbouw. Met deze sectie beoog ik enerzijds een beeld te geven van een aantal kenmerkende eigenschappen van de Nederlandse woordbouw en de bijbehorende terminologie, terwijl ik anderzijds kritisch wil ingaan op de grammaticale aannames van het MHB (en de onderliggende literatuur). In H3.3 zal ik nader ingaan op de MHB-classificatie van Nederlandse affixatie- en samenstellingspatronen. Hierbij zal ik ook een aantal kritische opmerkingen maken over de gekozen opzet.

In H3.4 vindt men een integrale inventarisatie van Nederlandse lexeemklassen en de bijbehorende affixkenmerken. Met deze inventarisatie beoog ik een nieuw licht te werpen op de functie van affixen. En passant fungeert deze inventarisatie als een aanvulling op het MHB, dat zich doorgaans beperkt tot de bespreking van affixen die een lexeem met categorie V, N, A of B vormen, maar voorbij gaat aan affixen die een rol spelen bij de vorming van functiewoorden (met categorieën als C, D en P). Het lexeemperspectief biedt volgens mij een beter uitgangspunt voor een complete morfologische beschrijving van het Nederlands, want als men morfologie definieert als de verzameling structuurpatronen die de basis kunnen vormen voor een lexeem uit een van de Nederlands lexeemklassen, zal men eerst moeten nagaan welke lexeemklassen er zijn.

H3.5 bespreekt diverse klassen van lexicale structuurrelaties, namelijk allomorfie, affixpotentiatie, paradigmatische woordvorming en affixconcurrentie. Hierbij zal duidelijk worden dat een morfologisch model dat uitgaat van vrije morfeemcombinaties, zoals het geval is bij het syntagmatische regelmodel, weinig recht doet aan de realiteit. Want zoals ik in deze sectie aannemelijk zal maken, berusten alle morfologische "regels" op generalisaties over lexicaal vastgelegde morfeemcombinaties.

Tot slot (in H3.6) behandel ik de hiërarchische structuurdimensie. Deze sectie draait om de vraag hoe men de eigenschappen van een woord als geheel uit de eigenschappen van de woordinterne bouwstenen kan afleiden. In grammaticale studies van de Nederlandse woordbouw wordt meestal aangenomen dat deze interactie aan een of andere variant van de Rechterhoofdregel (RHR) voldoet. Er zijn echter veel problemen met dit principe, die ik uitvoerig zal belichten. Deze bespreking is mede bedoeld om de voordelen van een compositionele aanpak te demonstreren. Want bij een compositionele theorie is het laatst aangehechte morfeem automatisch het hoofd van de hele constructie, wat impliceert dat deze eenheid ook bepalend is voor de eigenschappen van het lexeem als geheel. Wat in deze benadering echter niet wordt verantwoord is waarom de aanhechtingsvolgorde is zoals die wordt voorgesteld. Vanuit een paradigmatisch perspectief is dit een zeer wezenlijke vraag, die niet eenvoudig kan worden beantwoord. Deze vraag speelt ook een belangrijke rol bij de formele uitwerking van mijn lexicale representatiemodel in hoofdstuk 4.

3.2 De MHB-theorie van de Nederlandse woordbouw

3.2.1 Algemene achtergrond

Het Morfologisch Handboek van het Nederlands (MHB) biedt een uitvoerige inventarisatie van productieve en improductieve woordvormingspatronen uit het hedendaagse standaard-Nederlands. Deze inventarisatie is voor een groot deel gebaseerd op bestaande literatuur over de Nederlandse morfologie, maar voor de meeste morfemen is ook aanvullend onderzoek gedaan op basis van lexicografische naslagwerken, namelijk Van Dale (1984a), Van Dale (1984b), Nieuwborgs *Retrograde woordenboek van de Nederlandse taal* en het geautomatiseerde woordenbestand *Lexitron* (Van Dale, 1988), dus op basis van de bestaande woordenschat.[75] Ter inleiding van deze inventarisatie zal ik eerst het onderliggende analysemodel bespreken en waar nodig van kritisch commentaar voorzien. Dit analysemodel, dat sterk beïnvloed lijkt door de ideeën van Siegel (1974), kenmerkt zich door een syntagmatische visie op de morfologie. In deze visie bestaat elke taal uit een lexicon en een grammatica en dienen deze componenten complementaire informatie te bevatten. Hierbij heeft het lexicon de taak om de kleinste bouwstenen (c.q. morfemen) van een taal op te slaan, terwijl de grammatica aangeeft hoe deze tot grotere eenheden kunnen worden samengevoegd. Deze morfologisch gelede woorden kunnen niet in het lexicon worden opgeslagen, maar dienen steeds opnieuw via regels te worden afgeleid van een al opgeslagen woord.

3.2.2 De afbakening van het morfologische domein

Het MHB (zie pag. 9) onderscheidt twee verschillende opvattingen ten aanzien van het onderzoeksdomein van de *morfologie*. In de eerste opvatting gaat het om de tak van taalkunde die zich bezighoudt met de bestudering van de interne structuur van woorden, in de tweede opvatting om de tak van taalkunde die de (morfologische) procédés of operaties bestudeert met behulp waarvan gelede woorden gevormd worden. Hierbij correspondeert de eerste morfologie-opvatting met een morfeemgebaseerde benadering en de tweede met een procesgebaseerde benadering. Volgens het MHB is het voornaamste verschil dat in de eerste opvatting geen plaats is voor conversierelaties, aangezien dit type woordrelatie niet aan een hoorbare klankeenheid kan worden gerelateerd. Om die reden zou men geen morfeem mogen introduceren, maar gebruik moeten maken van een derivationele operatie. Neem bijvoorbeeld het woord *inzet*. Dit woord kan zowel met de stam van het (scheidbaar complexe) werkwoord *inzetten* als met het nomen *inzet* corresponderen. In navolging van Don (1993) stelt het MHB dat dergelijke nomina als een 0-derivatie van de werkwoordstam kunnen worden geanalyseerd: $[\text{verzet}]_V + 0_N = [\text{verzet-0}]_N$. Deze nomina hebben met elkaar gemeen dat hun betekenis is te omschrijven als uitvoering of resultaat van de handeling die bij de V-stam hoort (in dit geval 'inspanning'), dat ze het lidwoord *de* kiezen en (indien van toepassing) een meervoud op -EN. Dergelijke kenmerken zouden niet goed te verenigen zijn met de omgekeerde derivatierichting. Indien deze redenering klopt, kent het Nederlands dus minstens één klasse van derivaties die niet door een overt affix wordt gemarkeerd.[76] Dit hoeft echter niet te betekenen dat er geen affix aanwezig is, want er zijn tal van onderzoekers, waaronder Don[77] zelf, die een theorie verdedigen waarin klankloze affixen c.q. 0-affixen zijn toegestaan. In dit licht bezien is het theoretische onderscheid uit het MHB enigszins triviaal.

[75] Dit is een opvallende bron, want in de morfologievisie van het MHB (en de onderliggende literatuur) wordt slechts een klein deel van de bestaande woorden lexicaal opgeslagen, namelijk het deel dat niet morfologisch geleed is. Alle overige woorden behoren tot de potentiële woordenschat, al kan men deze op lexicografisch niveau onderverdelen in bestaande (mentaal opgeslagen) woorden en niet-bestaande woorden.

[76] Maar het is niet moeilijk om tegenvoorbeelden te verzinnen, bijv. *het ver+zet*, *het be+leg* en *het ont+haal*.

[77] Don verklaart deze eigenschappen overigens door een procesvariant op het 0-morfeem te introduceren.

Het hier besproken onderscheid kan echter aan een veel fundamentelere tegenstelling worden verbonden. De eerste morfologie-opvatting kan namelijk aan een inductieve doelstelling worden gerelateerd, d.w.z. de beschrijving van structurele overeenkomsten tussen bestaande woorden in de vorm van *lexicale redundantieregels*. De tweede opvatting daarentegen zou men aan een deductieve doelstelling kunnen verbinden, namelijk het voorspellen van de verzameling van mogelijke woorden door het opstellen van *productieve* grammaticaregels. Zoals ik in hoodstuk 2 uiteen heb gezet, leidt dit verschil in doelstellingen tot een verschillende afbakening van het domein van de morfologie; lexicale redundantieregels zijn immers op de bestaande woordenschat gebaseerd, terwijl productieve grammaticaregels primair gemotiveerd worden door potentiële woorden, d.w.z. woorden die afleidbaar zijn van bestaande woorden maar nog niet tot de bestaande woordenschat kunnen worden gerekend (althans, niet tot het grammaticale lexicon behoren). Het MHB lijkt beide kampen tevreden te willen stellen, want het streeft enerzijds naar een complete inventarisatie van lexicale redundantieregels (die zowel productieve als niet-productieve regels omvat), maar anderzijds wordt ook aangegeven welke woordvormingspatronen productief zijn. Deze combinatie is alleen mogelijk doordat het MHB voor een relatief neutraal, descriptief perspectief kiest.

Los van de hier aangekaarte problematiek bieden de bovenstaande morfologiedefinities nog niet zoveel houvast voor de afbakening van het door het MHB geanalyseerde feitendomein; hiertoe moet eerst bekend zijn wat precies onder een woord wordt verstaan, en hoe men gelede woorden van ongelede woorden kan onderscheiden. Ter beantwoording van de eerste vraag introduceert het MHB een onderscheid tussen *fonologische* woorden en *morfologische* woorden. Voor morfologische woorden hanteert het MHB (p. 3) de volgende definitie:

> **Definitie 3.1:** "Een *morfologisch woord* duidt een taaleenheid van vorm en betekenis aan die zelfstandig functioneert in een zin en in deze syntactische constructie geïsoleerd en verplaatst, maar niet intern gescheiden kan worden; deze eenheid dient aan de eisen van een fonologisch welgevormd woord te voldoen en één hoofdaccent te bezitten (waarmee het kan worden onderscheiden van partikels, woorddelen en woordgroepen)."

Volgens deze definitie kan elk fonologisch welgevormd woord met een vaste (lexicale) betekenis dus een morfologisch woord worden genoemd. Zo niet, dan is sprake van een wortel, waarbij men onderscheid kan maken tussen lexicale wortels[78] (die de stam vormen van een morfologisch geleed woord, bijvoorbeeld √*wust* in *bewust*) en niet-lexicale wortels (die niet herkenbaar zijn als een morfologisch relevante eenheid, zoals **zaun*). Hoewel er veel discussie bestaat over de theoretische status van lexicale wortels als uitgangspunt voor morfologische derivaties, rechtvaardigt het MHB het gebruik van deze structuureenheid met het argument dat het de morfologische beschrijving van het Nederlands aanzienlijk vergemakkelijkt; anders zouden wortelgebaseerde derivaties bijvoorbeeld via truncatieregels moeten worden verantwoord (zie pag. 19-20 van het MHB voor nadere informatie).

Ik zal dit toelichten aan de hand van een voorbeeld, namelijk het nomen *concretisatie*. Op het eerste gezicht is dit woord een regelmatige afleiding van het werkwoord *concretiseren*, maar deze afleiding kan niet als een combinatie van de werkwoordstam CONCRETISEER en het suffix -ATIE worden geanalyseerd, want dit zou in de vorm **concretise(e)ratie* resulteren. Daarom is in de morfologische literatuur voorgesteld om in dit soort gevallen een truncatie toe te passen, d.w.z. inkorting van de stam door deletie van één of meer eindletters. In dit specifieke geval zou de truncatieregel moeten resulteren in de wortelstam √CONCRETIS, waarna het suffix -ATIE zonder problemen kan worden aangehecht; deze truncatie zou kunnen worden afgedwongen door bij het affix -ATIE aan te geven dat het geen stammen met de uitgang *eer* kan selecteren.

[78] In het MHB worden vormen die naar morfologische wortels verwijzen voorafgegaan door een √-teken.

Het MHB geeft echter niet aan wat de juiste oplossing is. Een voor de hand liggend alternatief is bijvoorbeeld om aan te nemen dat uitheemse suffixen ook productief aan wortelstammen zoals √CONCRETIS kunnen worden aangehecht (zonder dat deze een categorie hoeven te hebben).

Het MHB noemt een Nederlands woord *fonologisch welgevormd* indien het uit één of meer lettergrepen bestaat, waarbij elke lettergreep op zijn minst een klinker moet bevatten. Bovendien dient elke klinker aan weerskanten door één of meer medeklinkers begrensd te worden, waarbij naar buiten toe sprake moet zijn van een afnemende sonorantie en waarbij de articulatiefeatures van aangrenzende fonemen compatibel moeten zijn; hierdoor kunnen klankvormen als /raakm/, /tuinp/ en /maasr/ niet als fonologisch woord fungeren.[79] Indien sprake is van een fonologisch welgevormd woord zonder morfologische geleding volgt de plaats van het hoofdaccent meestal uit onderstaande regel voor accenttoekenning:[80]

> **Hoofdaccentregel**: Plaats het hoofdaccent op de voorlaatste lettergreep van het woord, tenzij de laatste lettergreep bestaat uit een lange klinker gevolgd door één of meer consonanten, of uit een korte klinker gevolgd door minstens twee consonanten; in die gevallen valt het accent op de laatste lettergeep (Booij & Van Santen 1998, p. 204).

In het geval van samenstellingen dient eerst te worden vastgesteld welk woorddeel het hoofdaccent draagt (normaal gesproken het linkerdeel); vervolgens kan met behulp van de hoofdaccentregel de precieze locatie van het hoofdaccent worden berekend. Zo valt het hoofdaccent in *watervoorziening* (een samenstelling met de structuur *water+voorziening*) op de lettergreep *wa*, want het linkerdeel van deze samenstelling is *water*, terwijl het hoofdaccent van *water* op de eerste lettergreep ligt.

Een woord is *intern onscheidbaar* als klanken en woorddelen een vaste plaats en volgorde bezitten; er kan onderscheid worden gemaakt tussen drie niveaus van onscheidbaarheid of integriteit, te weten:

i) fonologische integriteit: de volgorde van de klanken binnen een woorddeel is bepalend voor de betekenis (zo bestaan *soep* en *poes* allebei uit de klanken /s/, /oe/ en /p/)[81]
ii) morfologische integriteit: de volgorde van de woorddelen is bepalend voor de betekenis (zo bestaan *groentesoep* en *soepgroente* allebei uit de woorddelen *groente* en *soep*)
iii) lexicale of syntactische integriteit: er mogen geen woorddelen uit een woord worden verplaatst, en er kan ook niet naar woordinterne delen worden terugverwezen (zoals blijkt uit de foutieve verwijzing in *Hij voegde het$_i$ aan de groente$_i$soep toe*)

De eis van morfologische integriteit leidt tot een probleem bij samenkoppelingen zoals *aanlopen* en *overschrijven*, want de partikels *aan* en *over* kunnen worden losgekoppeld van de woorddelen *lopen* en *schrijven*, hoewel ze er semantisch gezien nauw mee verbonden zijn en qua vorm ook meer op een woord dan op een woordgroep lijken; het MHB behandelt samenkoppelingen daarom als morfologische constructies. Omgekeerd zijn idiomatische woordgroepen zoals *de lelijke eend* en *de snelle jongen* syntactisch en semantisch gezien onscheidbaar, maar vertonen ze qua vorm meer overeenkomst met een woordgroep dan met een woord; daarom classificeert het MHB ze als *gelexicaliseerde woordgroep*, dus als een

[79] Overigens gelden in het Nederlands veel meer (en strengere) restricties; zo zijn er consonenten die niet tegelijk in de tweede onset-positie en de voorlaatste coda-positie kunnen voorkomen, ook al wordt aan de sonorantie-eisen voldaan; dit blijkt uit voorbeelden als **troert* en **klijlk*; evenmin mogelijk zijn onsets als **tl*, **ng*, etc.
[80] Een mogelijke uitzondering zijn woorden die op *-loos* eindigen, zoals *háveloos*, *róekeloos* en *rédeloos*. Maar als men het segment *-loos* niet als suffix, maar als woorddeel analyseert (vgl. -ACHTIG), verdwijnt het probleem.
[81] Dit verschijnsel wordt ten volle uitgebuit in het Opperlandse *Oeps, er zit een poes in de soep* (Battus, 2002).

syntactische constructie. Dit impliceert dat het lexicon niet alleen woorden, maar ook grotere constructies kan opnemen, namelijk alle syntactische configuraties waarvan de betekenis en/of de formele eigenschappen niet uit de vorm voorspeld kunnen worden (gegeven de grammaticaregels van een taal). Het MHB beperkt zich overigens tot de analyse van woordinterne structuurkenmerken.

3.2.3 Woordinterne structuureenheden

Het MHB gaat er (conform de grammaticale traditie) vanuit dat morfologische woorden kunnen worden onderverdeeld in *gelede* en *ongelede* woorden. Voor dit doel wordt een criterium gehanteerd dat als volgt kan worden gedefinieerd (zie p. 6):

> **Definitie 3.2:** Een woord is *morfologisch geleed* indien er een systematische koppeling bestaat tussen een vormaspect van het woord en een functioneel moment (d.w.z. een morfologische, syntactische of semantische eigenschap). De kleinste woordeenheden waarvoor een vaste koppeling bestaat tussen vorm en functioneel moment heten *morfemen.*

O*ngelede* woorden kenmerken zich doorgaans door een arbitraire relatie tussen vorm en betekenis van het woord; bij *gelede* woorden kan de woordvorm echter in kleinere eenheden worden gesplitst die elk met een eigen betekenisaspect en/of grammaticale functie corresponderen. Zelfstandige eenheden die niet verder analyseerbaar zijn (zoals *fiets*) worden *vrije morfemen* of *stammen* genoemd; niet-zelfstandige eenheden die niet verder analyseerbaar zijn (zoals BE- en -ER) worden *gebonden morfemen* of *affixen* genoemd. Deze affixen (die niet met morfologische woorden hoeven te corresponderen) kunnen op hun beurt worden onderverdeeld in *prefixen* (die links van de stam worden aangehecht, zoals BE- in *bestaan*), *suffixen* (die rechts van de stam worden aangehecht, zoals -END in *levend*), *discontinue affixen* (zoals GE-[..]-TE in *gebeente*) en *bindmorfemen*[82] (die tussen de woorddelen uit een samenstelling kunnen worden gevoegd, bijv. het morfeem -S- in *bestaansgrond*).

Indien een morfologisch woord uit meerdere deelwoorden is opgebouwd, spreekt men van een *samenstelling*. De betekenis van *fietswiel* kan bijvoorbeeld worden afgeleid uit de betekenis van de morfologische woorden *fiets* en *wiel*. Indien een morfologisch woord echter één of meer niet-zelfstandige eenheden telt, spreekt men van een *afleiding* c.q. *derivatie*; zo bestaat het woord *fietser* uit het vrije morfeem FIETS en het gebonden morfeem -ER. Er bestaan ook mengvormen: zo is het woord *hoogslaper* een samenstellende afleiding, namelijk een afleiding van de woorden *hoog en slaap* (die een niet-zelfstandige samenstelling vormen) en het gebonden morfeem -ER. Verder zijn er tal van woorden die wel formeel zijn samengesteld, maar die zich semantisch gezien als een ongeleed woord gedragen, bijvoorbeeld *oorlog*, *juffrouw*, *ledikant*, *faliekant*, *tsjoeketsjoeke*, *simsalabim* en *hieperdepiep*. Deze categorie blijft onzichtbaar indien men vasthoudt aan een strikt compositioneel analysecriterium.

In een ouder taalstadium kende het Nederlands nog een vijfde woordvormingsprocédé, namelijk de toekenning van woordinterne affixen c.q. *infixen*; deze manifesteren zich doorgaans als een klinkeraanpassing op de stam, zoals klinkerwisseling in sterke werkwoorden (*kijk*/*keek)* en sterke nomina (*stad/steden*). In het hedendaagse Nederlands corresponderen dergelijke alternanties echter niet langer met een regel. Volgens Booij & Van Santen (1998) wordt daarom voorgesteld om ze als een vorm van *stamallomorfie* te analyseren; hetzelfde geldt voor consonant-alternanties, zoals de s/t-alternantie in het woordpaar *chaos-chaotisch* of de e/i-alternantie in *meubel-meubilair*. Soms is zelfs sprake van suppletie, d.w.z. van een morfologisch verband tussen totaal verschillende stamvormen, bijvoorbeeld *goed-beter-best*.[83] Ook

[82] Het MHB spreekt hier overigens van *bindfonemen*, ondanks de overeenkomst met morfemen. Ikzelf gebruik de term *bindfoneem* alleen voor fonemen op de overgang van stam naar affix (bijv. de /d/ in *verstaander*).
[83] Het formeel gelede adjectief *na+bij* lijkt zelfs locale suppletie toe te (kunnen) staan: *nabij*, *naderbij*, *naastbij*.

affixen kunnen allomorfie vertonen. Hierbij kan onderscheid worden gemaakt tussen affix-geconditioneerde allomorfie, zoals de *eel-aal*-alternantie in het woordpaar *origineel-originaliteit* en stamgeconditioneerde allomorfie, zoals de zes vormvarianten van het diminutief-suffix (te weten -PJE, -KJE, -JE, -TJE, -ETJE en -EKE). Suffixparen als -ER/-AAR en -EUR/-OR vallen ook in deze categorie.

Naast gelede woorden onderscheidt het MHB ook woordparen die in een conversierelatie staan, zoals de conversie van het adjectief *gek* in het nomen *gek*. Bij dit type conversie vertoont het nomen altijd een voorkeur voor lidwoord *de* en meervoudssuffix -EN, wat erop wijst dat het nomen van het adjectief is afgeleid en niet andersom;[84] deze relatie wordt overigens vaak door een suffix -E gemarkeerd (bijvoorbeeld $rood_{A}$ – $rode_{N}$), maar deze markering is blijkbaar niet altijd nodig. Soortgelijke verbanden bestaan ook tussen nomina en werkwoorden, of tussen adjectieven en bijwoorden. In de literatuur worden dergelijke conversierelaties vaak verantwoord door een klankloos morfeem c.q. 0-morfeem te postuleren. Een andere mogelijkheid is om aan te nemen dat een formeel ongelede vorm structureel met verschillende betekenissen kan samengaan (die al dan niet in een afleidingsrelatie staan). Dit leidt wel tot een verdere aantasting van het basisidee dat morfologische structuur zich manifesteert als een vaste koppeling tussen vorm en betekenis: blijkbaar is de vorm net zo min noodzakelijk als de betekenis. Het MHB beperkt zich echter tot de beschrijving van de functionele eigenschappen van (overte) affixen; de fonologisch ongelede woorden blijven grotendeels buiten beschouwing.

In alle subklassen van morfologisch gelede woorden kan onderscheid worden gemaakt tussen systematisch gelede en formeel gelede woorden. Het verschil is dat bij formeel gelede woorden geen direct verband hoeft te bestaan tussen morfeemstructuur en woordbetekenis. Dit is bijvoorbeeld het geval bij afleidingen waarvan de stam niet (meer) als zelfstandig woord bestaat, zoals de wortels √GIN in *beginnen* en √DEFT in *deftig*. Ook vrije stamvormen kunnen een niet-transparante toepassing krijgen, zoals de stam STAAN in *verstaan*. Het MHB spreekt in beide gevallen van *formeel gelede* of *formeel afgeleide* woorden (ter onderscheiding van *systematisch gelede* vormen). Bij dergelijke woorden bestaat er geen koppeling tussen *vormmoment* en *betekenismoment*, maar wel tussen vormmoment en morfologische eigenschappen (zoals derivationele beperkingen) of syntactische eigenschappen (zoals woordcategorie en transitiviteit). Indien er geen enkele koppeling bestaat tussen woordvorm en functionele eigenschappen, kan er nog wel sprake zijn van *etymologische* structuur, d.w.z. een morfeemstructuur uit een ouder taalstadium of uit de brontaal van een leenwoord. Dergelijke structuur wordt niet relevant geacht voor het grammaticale systeem.

De onderstaande tabel (die ontleend is aan het MHB; zie p. 19) geeft een overzicht van alle door het MHB onderscheiden subklassen van morfologisch gelede lexemen:

prefigering	*be+loop, on+diep, aarts+vader, ante+dateer, in+actief, deci+liter*
suffigering	*duik+el, groen+ig, lui+heid, triomf+eer, kolos+aal, modern+isme*
samenstelling	*zweef+vlieg, dood+ziek, fiets+wiel*
samenkoppeling	*af+loop, piano+speel, goed+keur, voor+aan, tegen+over*
conversie	*gekA-gekN, loopV-loopN, computerN-computerV*
allomorfie	*apostel-apostol+air, meubel-meubil+air, sentiment+eel, president+ieel*
suppletie	*zie-zag, goed-beter*

Een laatste structuureenheid die aandacht verdient is het *morfologische hoofd*.[85] Deze term verwijst naar de eenheid die bepalend is voor de combinatorische kenmerken van een

[84] Er bestaan echter uitzonderingen, blijkens de woordparen $vuil_{A}$ - *het* $vuil_{N}$ en *het* $oranje_{N}$ – *de* $oranjes_{N}$.

[85] De notie *hoofd* is oorspronkelijk uit de syntaxis afkomstig, waar deze term gebruikt wordt om de eenheid te benoemen die bepalend is voor de syntactische categorie van een gegeven woordgroep (zoals de V van een VP).

morfologisch woord, zoals de syntactische categorie, het inflectieparadigma, functiewoorden (zoals lidwoord en hulpwerkwoord) en de betekenisklasse. In navolging van Trommelen & Zonneveld (1986) gaat het MHB ervan uit dat het hoofd altijd met het meest rechtse morfeem van een morfologisch geleed woord correspondeert. Bij derivaties wordt de hoofdfunctie meestal door een suffix uitgeoefend (zoals het suffix -ER in *fietser*). Indien er geen suffix aanwezig is wordt deze functie normaal gesproken door het stammorfeem overgenomen (zoals de stam DRACHT in *opdracht*).[86] Alleen indien de stam met een wortel correspondeert (bijv. de stam √GIN van *beginnen*), kan de hoofdfunctie door een prefix worden uitgeoefend (in dit geval het prefix BE-). In het geval van samenstellingen correspondeert het hoofd in principe met het rechterwoorddeel en daarbinnen met het meest rechtse morfeem.[87] De hier beschreven hypothese, die oorspronkelijk als generalisatie over de Engelse woordvorming is geformuleerd (cf. Williams, 1981), staat in de Nederlandse morfologie bekend als de *Rechterhand Hoofd Regel* (of kortweg RHR).

3.2.4 Inheemse en uitheemse morfologie

Bij de classificatie van Nederlandse morfemen maakt het MHB systematisch onderscheid tussen *inheemse* ("Germaanse") en *uitheemse* ("niet-Germaanse" c.q. Romaanse) morfemen; hiernaast kunnen ook *geleende* morfemen worden onderscheiden (zoals in *Blitzkrieg* en *oldtimer*), maar deze morfemen behoren (nog) niet tot het grammaticale systeem van het Nederlands en blijven daarom buiten beschouwing.[88] Het onderscheid tussen inheemse en uitheemse morfemen is primair grammaticaal van aard, hoewel het duidelijk taalhistorische wortels heeft. Wat betreft de morfologische basisstam kan dit onderscheid als volgt worden getypeerd: een inheemse basisstam bestaat typisch uit één of twee lettergrepen, waarbij de eerste lettergreep altijd met een volle klinker correspondeert (CVC, bijv. *boom*), en de tweede met een schwa, gevolgd door een consonant (@C, bijv. *rommel*). Een uitheemse basisstam heeft meestal twee of drie lettergrepen, waarbij elke lettergreep met een volle vocaal correspondeert (bijv. *patiënt*, *politiek*).[89]

Inheemse stammen komen vrijwel altijd als zelfstandig woord voor, maar uitheemse stammen voldoen vaak niet aan de eisen van een fonologisch welgevormd woord (bijv. √SPECTR in *spectraal*); het zijn dus meestal *wortels*. Soms kennen zulke uitheemse stammen meerdere verschijningsvormen (c.q. *allomorfen*), namelijk een vrije (inheemse) vorm en een gebonden (uitheemse) vorm; dit blijkt uit woordparen als *filter*/ *filteraar*/ *filtreren*, *meubel*/ *meubelen*/ *meubilair*, *profijt*/ *profijtelijk*/ *profiteer*. Hierbij is het vaak onduidelijk of de allomorfie via de stam of via het aangehechte suffix moet worden verantwoord. Er treden ook allomorfie-effecten op bij de aanhechting van prefixen, zoals blijkt uit de varianten van het uitheemse prefix CON- in het rijtje *content*, *compensatie*, *collaboratie*, *corresponderen*.

Booij & Van Santen (1998) laten zien dat de stamalternantie in woordparen als *poneren*/ *positie* historisch verklaard kan worden uit het inflectieparadigma van Latijnse werkwoorden. In het moderne Nederlands berusten dergelijke verbanden echter in de eerste plaats op semantische overwegingen. Maar sommige alternanties zijn zo regelmatig, dat van een morfologisch patroon kan worden gesproken. Daarom maakt het MHB bij elk affix melding van min of meer systematisch optredende alternanties. Hiertoe worden soms aparte affix-allomorfen gepostuleerd, zoals de vormen -ISEER, -IEER, -ULEER en -ONEER bij het verbaliserende

[86] Het Nederlands kent overigens verscheidene lexemen waarbij het suffix geen eigen betekenis bezit of ondergeschikt is aan de betekenis van het geheel. Men denke aan het suffix -egge in *dievegge* of -esse in *secretaresse* (dat niet als vrouwelijke vorm van *secretaris* geldt, zoals een vrouwelijke *tuinman* ook geen *tuinvrouw* heet).
[87] Ook deze generalisatie is te sterk: zo heeft *het luchtbelwaterpas* helemaal geen hoofd, want qua betekenis is dit meetinstrument geen soort *pas*, maar een *waterpas*, maar desondanks is het bijbehorende lidwoord *het*.
[88] Zie Van der Sijs (2001) voor nadere informatie over de opname en aanpassing van geleende morfemen.
[89] Zie ook Van Heuven, Neijt & Hijzelendoorn (1994).

hoofdsuffix -EER; er zijn echter ook vormeffecten die als suffix-geïnduceerde stamalternanties worden beschreven.

Het onderscheid tussen inheemse en uitheemse morfemen gaat meestal samen met een complementaire selectie van affixen: *inheemse* affixen hechten zich bij voorkeur aan *inheemse* stammen, terwijl *uitheemse* affixen een voorkeur hebben voor *uitheemse* stammen; veel inheemse affixen kunnen echter ook productief aan uitheemse stammen worden gehecht. Net als bij de stammen vertonen inheemse en uitheemse suffixen een fonologisch contrast: inheemse suffixen bestaan ofwel uit een zwakke klinker of schwa gevolgd door een medeklinker (bijv. -EL, -ER, -IG) ofwel uit een volle klinker die wordt omgeven door medeklinkers (bijv. -BAAR, -ZAAM). Bij woordfinale toepassing dragen deze suffixen nooit klemtoon, met uitzondering van de N-modificerende suffixen -IER (*herbergier*), -IN (*boerin*) en -ES (*prinses*). Adjectiverende suffixen, zoals -IG, -LIJK en -LOOS, kunnen de klemtoon wel verschuiven (blijkens de paren *daadkracht* / *daadkrachtig* en *werkwoord* / *werkwoordelijk*). Voor uitheemse suffixen geldt dat de eerste lettergreep vrijwel altijd met een volle klinker begint, die gevolgd wordt door een medeklinker (bijv. -AAL, -EEL, -ON, -IEN en suffixoïden als -SOFIE); ze kunnen eventueel door een tweede lettergreep gevolgd worden. Uitheemse suffixen zijn doorgaans niet klemtoon-neutraal, maar dragen zelf klemtoon of trekken de klemtoon naar achteren. Het MHB noemt één suffix dat geen klemtoon aantrekt, maar zich verder sterk uitheems gedraagt, te weten -ISCH; dit suffix heeft namelijk een sterke voorkeur voor uitheemse stammen en dwingt stamfinale klankalternanties af.[90]

3.2.5 Inflectie versus derivatie

Bij de classificatie van affixen hanteert het MHB een strikte scheiding tussen *inflectie* en *derivatie*. In formeel opzicht onderscheidt inflectie zich van derivatie doordat inflectie-affixen zich wel aan derivatie-affixen kunnen hechten, maar niet andersom; men zou ook kunnen zeggen dat inflectie met lexeemexterne affixatie correspondeert en derivatie met lexeeminterne affixatie. Ten tweede is inflectie per definitie niet categorie-veranderend, terwijl derivaties dat meestal wel zijn. Een derde verschil is dat inflectie-affixen vaak deel uitmaken van een affix-paradigma, terwijl derivatie-affixen zich veel vrijer gedragen. Een laatste contrast is dat inflectiepatronen veel productiever zijn dan derivatiepatronen; hierdoor correspondeert inflectie doorgaans met een *open stamklasse* (met een niet-opsombaar aantal leden), terwijl derivaties vaak met een *gesloten stamklasse* corresponderen (met een opsombaar aantal leden). Om derivatie-affixen duidelijk te kunnen onderscheiden van inflectie-affixen, noemt het MHB de eerste *affixen* en de laatste *uitgangen*.

Er kan een aanvullend onderscheid worden gemaakt tussen *inherente inflectie* (namelijk de aanhechting van affixen die een betekeniskenmerk toevoegen) en *contextuele inflectie* (namelijk de aanhechting van affixen die alleen een syntactisch gemotiveerde congruentierelatie uitdrukken).[91] Het eerste type inflectie vertoont grote overeenkomsten met derivatie; het voornaamste verschil is dat inherente inflectie categorieneutraal is en met rechtsperifere suffixen correspondeert. In het geval van nomina gaat het om meervoudsvorming (namelijk inflectie met -EN of -S), in het geval van adjectieven om de vergrotende en overtreffende trap (namelijk -ER resp. -ST) en in het geval van werkwoorden om markeringen van tijd, aspect en diverse gebruiksmodi. Van contextuele inflectie kent het Nederlands slechts enkele toepassingen, namelijk de persoons- en getalkenmerken van werkwoorden en de verbuiging van attributief gebruikte adjectieven. Van de vier lexicale categorieën moeten alleen bijwoorden het zonder

[90] Zie Heynderickx & Van Marle (1994) voor een inzichtgevende analyse.
[91] Zie bijvoorbeeld Booij (2002); het betreft een traditioneel onderscheid uit het grammaticale onderzoek naar de indo-europese woordvormingsprincipes.

inflectievormen stellen; dit geldt ook voor functionele categorieën, zoals lidwoorden, voorzetsels en voegwoorden.

3.2.6 Woorden versus lexemen

Het MHB gaat ervan uit dat morfologische afleidingen in beginsel altijd van een bestaand, mogelijk geleed woord dienen uit te gaan. Dit basiswoord wordt de *morfologische basis* genoemd; indien de morfologische basis met een ongeleed woord correspondeert, wordt ook vaak de term *stam* gebruikt (namelijk sinds Bloomfield (1933); cf. Beard (1995), Don & al (1994)).[92] Het MHB maakt geen terminologisch onderscheid tussen onverbogen woorden (zoals de V-stam DRIJV) en verbogen woorden (zoals de werkwoordsvormen *drijf*, *drijft* en *drijven*). Toch is het normaal gesproken niet toegestaan om een verbogen woordvorm als uitgangspunt te nemen voor een morfologische afleiding, blijkens de onmogelijkheid van derivaties als *DRIJFT+ING en *DRIJFT+ER.[93] Om die reden heeft Matthews (1974) een aparte term geïntroduceerd voor onverbogen, categoriedragende eenheden, namelijk *lexeem*.[94] Hierdoor kan bijvoorbeeld onderscheid worden gemaakt tussen het lexeem WERK (met de categorie V) en het woord *werk* (dat met de eerste persoon enkelvoud presens van het lexeem WERK correspondeert), ondanks het feit dat woord en lexeem hier dezelfde klankvorm bezitten. Sinds Aronoff (1994) bestaat er ook een conventie om de (onderliggende) klankvorm(en) van een lexeem als *stam* aan te duiden. Zo kent het V-lexeem DRIJV twee verschillende stammen, namelijk een stam met de klankvorm *drijf* (resp. *drijv*) en een stam met de klankvorm *dreef (*resp. *dreev)*. De stam correspondeert dus met de klankvorm van een woord minus zijn inflectie-uitgangen.

De introductie van lexemen leidt tot een complicatie van het onderscheid tussen gebonden en niet-gebonden morfemen, want lexemen kunnen per definitie niet zelfstandig voorkomen: ze dienen eerst verbogen te worden. Maar in tegenstelling tot affixen hoeven lexemen niet eerst aan een stam te worden gehecht, maar zijn ze direct beschikbaar voor de toekenning van inflectie, dus voor gebruik als zelfstandig woord. In talen als het Nederlands en het Engels zijn de meeste (inheemse) lexeemvormen bovendien identiek aan een zelfstandige woordvorm; in deze talen zou men lexemen dus ook kunnen definiëren als morfologische eenheden waarvan de klankvorm als zelfstandig woord kan worden gebruikt. In talen met rijke inflectie zoals het Latijn en veel Slavische talen corresponderen alle lexeemtoepassingen echter met een overte uitgang, zodat men de klankvorm van het lexeem nooit als zelfstandig woord zal aantreffen (tenzij het toevallig om de verbogen vorm van een ander lexeem gaat). In semitische talen, zoals het Arabisch en het Hebreeuws is de status van lexemen nog abstracter, aangezien het lexeem hier geen stam met een vaste klankvorm is, maar een fonologisch basisschema (bijv. /k.t.b/) dat door intercalatie (d.w.z. verweving met een complementair klankschema, bijv. /.a.i./) in een uitspreekbare woordvorm moet worden omgezet (bijv. /katib/).[95] Deze talen kennen nauwelijks onderscheid tussen derivationele en inflectionele morfologie, want beide zijn in beginsel op paradigmatische wijze georganiseerd.

---

[92] De term *stam* kent ook andere invullingen: in schoolgrammatica's duidt deze term bijvoorbeeld vaak de romp van een werkwoordsvorm aan (bijv. /werk/ in *werkt*), terwijl lexeemgerichte benaderingen van de morfologie de term *stam* voor de klankvorm van een *lexeem* reserveren (zie verderop); in de taalkundige literatuur is de term *stam* meestal inwisselbaar met de term *lexeem*.

[93] Dit geldt niet voor derivaties. Zo kent het werkwoord *drijven* de nominalisatie *drift*, die zelf weer de (formele) basis vormt voor de afleidingen *driftig* en *drifter*. Dus gestapelde derivaties zijn niet onmogelijk.

[94] De term *lexeem* kent twee verschillende definities: in de definitie van Matthews (1974), die sinds Aronoff (1994) breed ingang heeft gevonden in de morfologische literatuur (cf. Booij (2002)), dient een lexeem altijd met een woordinterne eenheid te corresponderen. In de lexicologische literatuur daarentegen (cf. Lieske, 1994) wordt meestal de definitie van Lyons (1977) aangehouden; in deze definitie kan een lexeem ook naar idiomatische constructies verwijzen.

[95] Oehrle (2000) laat zien hoe dit type morfologie in een categoriaal framework kan worden ingepast.

In de rest van deze studie zal ik de term *stam* voor derivationele contexten reserveren en de term *lexeem* voor inflectionele contexten; verder ga ik ervan uit dat stammen geen syntactische categorie bezitten, maar een morfologische categorie. Voor ongelede woordstammen zal ik meestal de term *wortel* gebruiken, ongeacht de vraag of de stam een zelfstandige betekenis introduceert; in mijn optiek is dit namelijk een irrelevant criterium.

3.2.7 Samenvatting

Het Morfologisch Handboek (MHB) biedt een uitgebreid overzicht van bestaande en nieuwe observaties ten aanzien van de woordstructuur van het Nederlands. Hierbij gaat de aandacht vooral uit naar de vraag welke affixen het Nederlands kent, wat hun combinatorische eigenschappen zijn en welke woordkenmerken aan deze affixen zijn verbonden. Verder wordt aangegeven welke affixen productief gebruik toestaan. Het beschrijvingsmodel van het MHB (c.q. het MHB-model) kenmerkt zich door de volgende eigenschappen:

i. Het MHB-model berust op een syntactische morfologiebenadering. In deze benadering correspondeert het lexicon met een opslagplaats van kleinste morfologische bouwstenen c.q. morfemen en dient de grammatica alle potentiële (afleidbare) woorden te definiëren door aan te geven hoe deze morfemen tot grotere eenheden kunnen worden gecombineerd. Hierbij worden drie soorten morfemen onderscheiden, namelijk stammen (c.q. ongelede woorden), wortels (niet-zelfstandig bruikbare stammen) en affixen (c.q. niet-zelfstandige segmenten die zich aan een ander morfeem kunnen vasthechten). Het lexicon biedt geen plaats aan morfologisch gelede lexemen (bijv. DRAAI+ING) of vaste affixcombinaties (bijv. IEF+EER).
ii. In het MHB-model zijn de grammaticale eigenschappen van morfologisch complexe woorden rechtstreeks af te leiden uit de eigenschappen van het meest rechtse morfeem of woorddeel in een woord. Dit principe staat bekend als de RechterHoofdRegel (RHR).
iii. Het MHB-model kent een categoriale grondslag in de zin dat alle affixen als een relatie tussen twee syntactische categorieën zijn gedefinieerd, namelijk als de omzetting van een lexeem uit categorie x naar een lexeem uit categorie y, waarbij categorie x gelijk mag zijn aan categorie y en waarbij elke syntactische basiscategorie (N, V, A, B, P en T) kan worden gesubcategoriseerd door specificatie van semantische en fonologische selectiekenmerken.[96]
iv. Het MHB-model beperkt zich niet tot de beschrijving van transparant gelede woorden, maar biedt ook inzicht in de structuur van formeel gelede woorden. Het verschil is dat transparant gelede woorden voorspelbare woordkenmerken hebben en dus niet in het lexicon hoeven te worden opgenomen. Maar zodra een morfologisch geleed woord een kenmerk bezit dat niet aan de eis van compositionele voorspelbaarheid voldoet, krijgt het de status van formeel geleed woord en is lexicale opslag nodig.
v. Het MHB-model gaat er vanuit dat transparant gelede woorden niet in het lexicon worden opgeslagen en dus geen invloed hebben op de productiviteit van de onderliggende woordvormingsregels; dit impliceert dat de notie productiviteit geen gradaties kent.

[96] Het MHB-model is echter geen categoriale grammatica in de Montagoviaanse zin (cf. Montague, 1974). Voor (aanzetten tot) een categoriale grammatica van de Nederlandse morfologie kan men onder meer terecht bij Van der Hulst & Moortgat (1980), Hoeksema (1984) en Heemskerk & Van Heuven (1993).

3.3 De morfologische classificatieprincipes van het MHB

3.3.1 Introductie

Het MHB biedt een uitvoerig overzicht van de affixatie- en samenstellingspatronen die ten grondslag liggen aan de opbouw van Nederlandse lexemen.[97] De patrooninventarisatie blijft echter beperkt tot lexemen met de categorie V, N of A. Wat de andere woordsoorten betreft beperkt het MHB zich tot een kleine sectie over adverbiale suffixen en een lijst van prepositieparen. Of de door het MHB gepresenteerde affixinventarisatie echt compleet is, hangt af van het beoordelingscriterium. Zeker is dat de affixen met een wat hogere typefrequentie (zeg 50) allemaal zijn vertegenwoordigd. Of dit ook geldt voor hun semantische of syntactische subtypes is echter moeilijk te bepalen. Wat de affixen met een lage typefrequentie betreft is onduidelijk welke criteria de auteurs hebben gehanteerd. Zij geven hier zelf geen informatie over (behalve dat ze veel gebruik hebben gemaakt van woordenboeken en van hun eigen intuïties over bestaande en mogelijke woorden). Maar een nadere bestudering van de affixinventarisatie in het MHB leert dat deze ook extreem infrequente affixen omvat, waaronder -EGGE (in *dievegge*) en -ITSA (in *tsaritsa*), alsmede affixsplinters, waaronder *s-* (in *sloom*) en *d-* (in *dwars*). In dit deeldomein is het MHB echter verre van compleet.

In het MHB worden drie derivatietypes onderscheiden, te weten *prefigering* (c.q. prefixaanhechting), *suffigering* (c.q. suffixaanhechting) en *samenstelling* (c.q. stamkoppeling en lexeemkoppeling).[98] Deze onderverdeling is rechtstreeks doorgevoerd in de hoofdstukindeling: het MHB wijdt namelijk afzonderlijke hoofdstukken aan prefixen, suffixen en samenstellingen. Binnen elk hoofdstuk van het MHB is een onderverdeling aangebracht op basis van categoriale criteria, d.w.z.: een hoofdordening op basis van de syntactische categorie van het afgeleide woord (en dus van het morfeem dat als morfologisch *hoofd* functioneert) en een subordening op basis van de categorie van de stam (die zowel met een zelfstandig woord als met een wortel kan corresponderen). Verder worden de inheemse en uitheemse affixen gescheiden behandeld. Bij de suffixen is nog een semantische subclassificatie aangebracht door onderscheid te maken tussen persoonsnaamvormende, zaaknaamvormende en plaatsnaamvormende suffixen. Tot slot wordt per affix vermeld of het productief is.

Hoewel de eigenschappen van de woordstam (in de betekenis van derivationele basis) een belangrijke rol spelen bij de vraag welke affixen er mee kunnen worden gecombineerd, heeft het MHB er geen apart hoofdstuk aan gewijd. Het is ook onmogelijk om een complete inventarisatie van woordstammen te geven, aangezien het een open klasse betreft. Per woordcategorie kan men echter wel algemeen voorkomende stamtypes onderscheiden op basis van overeenkomsten in hun syntactische, semantische en morfologische eigenschappen. Bovendien kan per woordcategorie worden aangegeven welke inflectiepatronen er bestaan. Het MHB heeft ervoor gekozen om de stamgerelateerde informatie in de hoofdstukken over prefixen en suffixen onder te brengen, overeenkomstig de algemene taakverdeling tussen prefixen en suffixen: het prefixhoofdstuk begint daarom met een sectie over V-stammen (aangezien prefixen vaak een werkwoord selecteren), terwijl het suffixhoofdstuk is uitgebreid met secties over N-stammen en A-stammen (namelijk aan het begin van de secties over resp. N-selecterende en A-selecterende suffixen). Hierbij gaat de meeste aandacht uit naar de inflectiepatronen.

---

[97] Er zijn enkele publicaties waarin de inhoud van het Morfologisch Handboek kritisch wordt beoordeeld, te weten een korte recensie van De Schutter (1994) en een uitgebreide recensie van Van Santen (1994). Beide auteurs tonen zich redelijk tevreden over de informatie zelf; de kritiek gaat vooral uit naar de ordeningswijze.

[98] Naast het mechanisme voor *woordverbinding* lijkt overigens ook een mechanisme voor *wortelverbinding* te bestaan (bijv. *aqua+duct*); zie Ten Hacken (1994) of Iacobini (2000) voor nadere motivatie.

Zoals eerder werd opgemerkt door Van Santen (1995) is de ordeningswijze van het MHB niet erg overzichtelijk. Want door een basisonderscheid te hanteren tussen prefixen en suffixen,[99] is het niet mogelijk om affixen van dezelfde functionele categorie bijeen te houden. Dit effect wordt verergerd door de beslissing om V-informatie aan de prefixen te koppelen en N- en A-informatie aan de suffixen. Het gevolg is dat de categoriegerelateerde affixinformatie sterk versnipperd wordt behandeld; zo staat de informatie over V-inflectie in de inleiding van het prefixhoofdstuk, maar zijn de V-vormende suffixen in het suffixhoofdstuk opgenomen; verder worden N- en A-vormende prefixen op een andere plaats behandeld dan N- en A-vormende suffixen. Lastig is dat de informatie over samenkoppelingen niet in het hoofdstuk over samenstellingen staat, maar (op grond van de relatie met werkwoordsvorming) in het prefixhoofdstuk; en omdat complexe preposities als samenkoppeling gelden, is de bijbehorende informatie ook maar in het prefixhoofdstuk ondergebracht. Een laatste nadeel van de MHB-ordening is dat discontinue affixen soms onder de prefixen worden behandeld en soms onder de suffixen. Het is niet gemakkelijk om dit alles te doorzien. Van Santen zelf zou (geïnspireerd door Marchand's (1969) morfologische beschrijving van het Engels) het liefst een indeling zien die uitgaat van semantisch en syntactisch gemotiveerde affixclusters (zoals een cluster voor affixen met een vrouwelijk betekenisaspect). Deze ordening maakt het mogelijk om per functie te generaliseren over prefixen en suffixen en om per affix informatie te geven over de gebruiksmogelijkheden (zoals de keuze van de syntactische categorie). Binnen elke cluster zouden de affixen op klankvorm moeten worden geordend.

In de nu volgende subsecties volgt een bespreking van de door het MHB gehanteerde classificatiecriteria. Er wordt achtereenvolgens aandacht besteed aan criteria voor prefixen (3.3.2), suffixen (3.3.3), discontinue affixen (3.3.4) en samenstellingen (3.3.5).

3.3.2 Prefixen

Voor de prefixen hanteert het MHB de volgende basisclassificatie:

i. categorie-neutrale Germaanse prefixen:
  a) nomen-modificerend: AARTS-, ON-, OER-, ...; bijv. *aartslui*, *onaardig*, *oerkracht*
  b) werkwoord-modificerend: alleen HER-, bijv. *herhalen*
  c) pseudoprefixen: B-, D-, GE-, S-, bijv. *b-uiten*, *d-rammen, ge-trouw*, *s-loom*
ii. categorie-neutrale niet-Germaanse prefixen:
  a) naamwoord-modificerende prefixen: ANTI-, ULTRA-, SUB-, MONO-, A-, ...
  b) werkwoord-modificerende prefixen:AD-, CON-, DE-, TRANS-, -IN, ...
  - subklasse 1: *[prefix + stam], *[stam + suffix], bijv. *absorberen* (*absorb, *sorberen)
  - subklasse 2: [prefix + stam], *[stam + suffix], bijv. *concretiseren* (bevat *concreet*)
  - subklasse 3: *[prefix + stam], [stam + suffix], bijv. *deblokkeren* (bevat *blokkeer*)
iii. categoriebepalende Germaanse prefixen (met subklassen voor elke stamcategorie)
  a) verbaliserende prefixen: BE-, VER-, ONT-, bijv. *bespreken*, *verzetten*, *ontnemen*
  b) nominaliserende prefixen: GE-, [0]-, bijv. *gelach*, *verkoop*
iv. prefigering met een Germaans partikel (adpositie of adverbium)
  -samenkoppeling: partikel is scheidbaar en draagt hoofdaccent;
  bijv. *tegenwerken*, *voorzeggen*, *overschrijven*, *uitlachen*, *weggooien*, *volhouden* etc.
  -vaste constructie: partikel is niet scheidbaar en draagt geen hoofdaccent;
  bijv. *omsingelen*, *doorzoeken*, *ondergraven*, *overmeesteren* etc.

Het MHB gaat uitvoerig in op de syntactische en semantische effecten van categoriebepalende prefixen, die altijd Germaans zijn. In het algemeen blijkt aanhechting van een verbalise-

[99] Deze indeling lijkt gemotiveerd te zijn door het Rechterhoofdprincipe; want volgens dit principe zijn suffixen wel, maar prefixen niet in staat om de woordcategorie te bepalen; cf. Zonneveld en Trommelen (1986).

rend prefix in een transitief werkwoord te resulteren, zelfs als de afleiding zelf niet meer doorzichtig is. Maar indien de onderliggende stam een intransitief werkwoord is, kan het resultaat ook een intransitief werkwoord zijn. Er zijn ook enkele werkwoorden die van valentieklasse veranderen als gevolg van klinkermodificatie, zoals door de volgende woordparen wordt geïllustreerd : *liggen*/*leggen*, *zitten*/*zetten*, *vallen*/*vellen*.

3.3.3 Suffixen

Bij de classificatie van suffixen let het MHB niet alleen op *categoriale* eigenschappen, maar ook op *subcategoriale* en semantische eigenschappen. Zo wordt bij nominaliserende suffixen een onderverdeling gehanteerd met persoonsnaamvormende suffixen (met een aparte sectie voor vrouwelijke persoonsnamen), zaaknaamvormende suffixen, geografische suffixen, stofnaamvormende suffixen en diminutieven. Het nadeel van deze indeling is dat veel suffixen in meerdere subcategorieën kunnen voorkomen. Het MHB heeft dit opgelost door de hoofdclassificatie op de dominante subcategorie te baseren, en vervolgens alle gebruiksmogelijkheden langs te lopen. Zo wordt het inheemse suffix -AAR geclassificeerd als een persoonsnaamvormend suffix (bijv. *molenaar*) dat incidenteel zaaknaamvormend is (*schakelaar*). Deze zaaknaamtoepassing wordt dan als apart subtype behandeld in de sectie over -AAR. Bij een uitheems suffix als -AAL ontstaat de extra complicatie dat het zowel nominaliserend (bijv. *nationaal*) als adjectiverend (*kwartaal*) kan worden gebruikt; daarom wordt dit suffix op twee verschillende plaatsen behandeld, hoewel het MHB aangeeft dat er misschien sprake is van een conversie-relatie (bijv. bij *koloniaal*).

De systematische inventarisatie van suffixen wordt nog verder bemoeilijkt door het feit dat met name uitheemse suffixen vele varianten kennen, waarbij het vaak lastig is om de grens tussen stam en suffix aan te wijzen. Voor het onbeklemtoonde suffix -IE (in de functie van nominaliserend suffix bij een werkwoordstam met eindsuffix -EER) onderscheidt het MHB bijvoorbeeld de varianten -ATIE, -ETIE, -ITIE en -UTIE, -TIE, -ANTIE en -ENTIE; deze suffixvarianten worden gemotiveerd door lexemen als *adoptie, demonstratie, deletie, notitie, elocutie*, *produktie*, *dominantie* en *correspondentie* (in dezelfde volgorde als de reeds genoemde suffixvormen). Met betrekking tot de morfeemvarianten -ATIE, -ETIE, -ITIE en -UTIE merkt het MHB op dat de syllabe voor het segment *ie* (dus de syllabes *at*, *et*, *it* en *ut*) best met een (betekenisloos) suffix zouden kunnen corresponderen; in het geval van -ANTIE en -ENTIE acht het MHB deze suffix-optie zelfs heel waarschijnlijk, aangezien de gedeeltes *ant* en *ent* ook in gebruik zijn als adjectiverend suffix (resp. -ANT en -ENT); hierdoor is niet bij voorbaat duidelijk hoe een lexeem als *dominantie* is gestructureerd: als [DOMINEER]$_V$+ANTIE of als [DOMINANT]$_A$+IE. In het eerste geval zou de uitgang *antie* in zijn geheel met een suffix corresponderen, in het tweede geval daarentegen zou het gedeelte *ant* ofwel met een afzonderlijk suffix corresponderen, ofwel met een stamuitgang (indien *dominantie* een morfologisch ongeleed woord blijkt te zijn). Hoewel dit soort dilemma's op grote schaal voorkomt, biedt het MHB geen systematische inventarisatie van mogelijke en onmogelijke suffixverbindingen. Op dit punt beperkt de informatie zich tot de meest in het oog springende relaties.

Eveneens problematisch, want lexicaal van karakter, zijn uitheemse wortels die zich als suffix gedragen c.q. suffixoïden, waaronder -SCOOP, -GRAAF, -NAUT. Het MHB doet geen poging de suffixoïden (of prefixoïden) systematisch in kaart te brengen, maar beperkt zich tot een korte beschrijving van hun morfologische kenmerken. Volgens deze beschrijving gedragen suffixoïden zich qua klemtoongedrag als uitheemse suffixen, kunnen ze meestal door het beklemtoonde suffix -IE worden gevolgd, en is hun betekenis soms ambigu tussen een persoons- en een zaaknaam (zoals blijkt uit het paar *fotograaf*/*fonograaf*).

3.3.4 Discontinue affixen

Het MHB spreekt van een discontinu affix als een affix A uit een prefix-deel (P_A) en een suffix-deel (S_A) is opgebouwd, dus als een stam X is ingebed in de structuur P_A-[X]-S_A. Zo gaat het MHB ervan uit dat de voltooide tijd van niet-geprefigeerde werkwoorden altijd door middel van een discontinu affix wordt gevormd, namelijk GE-[..]-D/T (in het geval van zwakke werkwoorden) of GE-[..]-EN (in het geval van sterke werkwoorden).[100]

Dezelfde affixen kunnen ook een puur adjectiverende functie hebben, namelijk bij toepassing op een [-V]-basis; zo kan het adjectief *gelaarsd* (GE-[LAARS]$_N$-D) uitsluitend van het nomen *laars* zijn afgeleid, want er bestaat geen werkwoord *laarzen*. Deze adjectiverende functie is overigens ook beschikbaar voor affixcombinaties met andere V-prefixen; voorbeelden zijn *betoeterd* (BE-[TOETER]$_N$-D), *ontheemd* (ONT-[HEEM]$_X$-D) en *verliefd* (VER-[LIEF]$_A$-D).

Volgens het MHB dient ook de prefix-suffix-combinatie GE-[..]-TE, die nomina oplevert, als een discontinu affix te worden geclassificeerd. Dit affix is zichtbaar in woorden als *gebladerte* (GE-[BLADER]$_N$-TE) en *gebeente* (GE-[BEEN]$_N$-TE).

Het laatste discontinue affix[101] dat het MHB bespreekt is het V-vormende affix BE-[..]-IG, dat wordt aangetroffen in werkwoorden als *beëdigen* en *beangstigen*. Met betrekking tot het tweede voorbeeld kan de vraag worden gesteld of er sprake is van een deadjectivale V-afleiding (BE-[ANGSTIG]$_A$) of van een denominale V-afleiding (BE-[ANGST]$_N$-IG); op basis van semantische overwegingen neigt het MHB naar de tweede optie, maar bij de V *beëindigen* prefereert het MHB de structuur BE-[EINDIG]$_V$.

Het idee dat er discontinue affixen bestaan is moeilijk te verenigen met de derivationele benadering van morfologie. Het lijkt een direct gevolg van de aanname dat derivaties altijd een lexeembasis moeten hebben; hierdoor is het niet toegestaan om derivaties op potentiële lexemen of categorieloze stammen te baseren. Indien men deze restrictie loslaat is wel degelijk een sequentiële analyse mogelijk. Hierbij zorgt het prefixdeel ervoor dat de stam morfologisch en semantisch wordt voorbereid op specifieke derivatiemogelijkheden.

3.3.5 Samenstellingen

Het MHB wijdt een apart deel aan de analyse van lexeemgebaseerde samenstellingen. Hierbij onderscheidt het MHB de volgende hoofdtypes: *endocentrische samenstellingen, exocentrische samenstellingen, copulatieve samenstellingen, gelexicaliseerde woordgroepen*, *reduplicatieve samenstellingen, pseudo-samenstellingen* en *samenstellende afleidingen*. Hieronder volgt voor elk type samenstelling een samenvatting van de in het MHB verstrekte informatie.

i) er is sprake van een *endocentrische* samenstelling als alle woorddelen een zelfstandige betekenis dragen en als de samenstelling voldoet aan het schema AB = B; dit schema zegt dat de rechterconstituent bepalend dient te zijn voor de categoriale, morfologische en semantische eigenschappen van de samenstelling als geheel. Bij het samengestelde nomen *huisdeur* (met structuur HUIS+DEUR) zijn deze eigenschappen bijvoorbeeld volledig af te leiden uit die van de constituent *deur*: een huisdeur is immers een soort deur, draagt categorie N, kiest een meervoud op -EN en correspondeert met het lidwoord *de*. En bij het samengestelde werkwoord *opspringen* (met structuur OP+SPRINGEN) berusten de belangrijkste eigenschappen op de constituent *springen*: want beide bezitten categorie V, beide vertonen een sterk vervoegingsschema (*spring*/*sprong*/*gesprongen*) en semantisch gezien is *opspringen* een subklasse van de *spring*-gebeurtenissen.

[100] Volgens Drijkoningen (1995) is deze analyse zelfs verenigbaar met de RHR (gegeven de theorie van Kaynes).
[101] Wel zij opgemerkt dat het MHB overweegt om het bijwoord *beneden* als be+[neer]+en te analyseren.

Bij nominale samenstellingen bestaat de mogelijkheid dat er tussen het linkerdeel en het rechterdeel een bindsegment verschijnt, zoals het segment *-en* in *boekenkast*, het segment *-er* in *kinderopvang* en het segment *-s* in *groepsleider*. Hoewel het MHB in dit verband van een bindfoneem spreekt, geef ik zelf de voorkeur aan de term bindmorfeem of affix, wegens de sterke parallellie met de affixen voor het meervoud. In het algemeen geldt dat het bindmorfeem -EN alleen mogelijk is indien het linkerdeel met een nomen correspondeert dat een meervoud op -EN selecteert. Het bindmorfeem -ER is nog selectiever: het duikt alleen op bij N-stammen waarvan de meervoudsvorm op *-eren* eindigt; zo vormt het nomen *kind* (meervoud *kinderen*) de basis voor samenstellingen als *kinderkamer* en *kinderboek*. Het bindmorfeem -S daarentegen kan met bijna alle stammen worden gecombineerd.

Bij gelede linkerdelen berust de keuze van het bindmorfeem vaak op het laatste suffix.[102] Zo vertonen nomina met het eindsuffix -IST, -IN en -LING, waarvan de meervoudsvorm met het suffix -EN wordt gevormd, een voorkeur voor het bindmorfeem -EN (bijv. *componistenconcours*, *vriendinnenclub* en *vreemdelingenhaat*). Nomina agentis op -ER, -AAR -IER en -EUR, die een meervoud op -S selecteren, vertonen echter een voorkeur voor het bindmorfeem -S (bijv. *schippersvakbond*, *goochelaarskunst*, *koetsiershuis*, *ingenieursbureau* en *potjeslatijn*).[103] Dit geldt nog sterker voor nomina waarvan de uitgang met een diminutiefsuffix (-JE) correspondeert (bijv. *potjeslatijn*). Er zou ook een categorie eindsuffixen bestaan waarbij de selectie van het bindmorfeem altijd samengaat met het betekenisaspect meervoud; dit geldt bijvoorbeeld voor zaakaanduidende nomina op -THEEK, -IER, -SEL, -TE, -IE, -ER, -AAR en -EUR (bijv. *bibliotheekboek*, *portierkruk*, *uittrekselboek*, *breedtegraad*, *adoptiekind*, *wijzerplaat*, *schakelaarknop* en *mitrailleurpatroon*. De invoeging van bindmorfemen blijft overigens niet beperkt tot samenstellingen waarvan het linkerdeel een nomen is. Ze komen ook voor bij:

V+N-samenstellingen (bijv. *scheidsmuur*, *hebbedingetje*, *huilebalk, hinkepoot* en *wiegelied*)

A+N-samenstellingen (bijv. *hogeschool*, *jongeman, wildebras*),

V+V-samenstellingen (bijv. *spelevaren*, *trekkebekken, schuddebollen*).

Maar deze subklassen zijn niet productief.

Volgens het MHB corresponderen nominale en adjectivische samenstellingen allebei met zeer productieve woordvormingsprocédés (in tegenstelling tot verbale samenstellingen). Verder kennen nominale samenstellingen de bijzondere mogelijkheid van recursie, in de zin dat beide woorddelen zelf ook weer een samenstelling kunnen zijn, zoals *(huis+werk)+(op+dracht)* en *water+(over+last)*. Voor zowel nominale als adjectivische samenstellingen geldt bijna altijd dat het linkerdeel als *determinant* (c.q. *modificeerder*) fungeert en het rechterdeel als *determinatum* (c.q. *gemodificeerde*). Bij verbale samenstellingen hoeft dit niet het geval te zijn; zo geldt voor werkwoorden als *stampvoeten*, *klappertanden* en *trekkebekken* dat het rechterdeel juist de modificeerder is van het linkerdeel. Wat categorie betreft zijn dergelijke samenstellingen dus linkshoofdig. Een werkwoord als *trekkebekken* wordt echter zwak vervoegd, ondanks het feit dat *trekken* zelf een sterk werkwoord is. Dit wijst erop dat deze samenstelling een exocentrische structuur bezit.

ii) bij *exocentrische* samenstellingen bestaat geen directe relatie tussen de categorie of betekenis van de woorddelen en de categorie of betekenis van de samenstelling; het gaat meestal om metaforisch gebruikte nomina, zoals *wijsneus*, *spleetoog*, *halfbloed* etc. De op deze wijze gevormde nomina nemen doorgaans het lidwoord *de* en bezitten vaak een negatieve connotatie. Tussen de samenstellende delen kan een bindmorfeem verschijnen.

[102] Overigens zijn er tal van andere factoren in het spel, blijkens een studie van Krott (2001).
[103] N-stammen op -IER kunnen echter ook het suffix -EN selecteren, blijkens *scholierengeweld*, *arabierenhater* en *bankierenoverleg*. In deze voorbeelden verwijst het linkerdeel steeds naar een plurale entiteit.

iii) *copulatieve* samenstellingen zijn samenstellingen waarbij er semantisch gezien sprake is van nevenschikking, d.w.z. de determinant/determinatum-relatie is ofwel wederkerig ofwel afwezig. Voorbeelden van deze constructie zijn *kind-ster*, *radio-cassetterecorder*, *minister-president*, *rechter-commissaris*. Normaliter berust het meervoud op de vervoegde vorm van het rechterlid, zoals in *radio-cassetterecorders* en *rechter-commissarissen*. In enkele archaïsche gevallen is het echter ook mogelijk om beide delen te vervoegen (bijv. *tolken-vertalers*) of alleen het eerste deel (bijv. *secretarissen-generaal*).

iv) *gelexicaliseerde woordgroepen* vormen een restgroep waarin allerlei gelexicaliseerde woordcombinaties zijn opgenomen; deze categorie omvat bijvoorbeeld versteende syntactische constructies zoals *derde wereld*, *zwarte markt* en *lelijke eend*; verder treft men hier op eigennamen gebaseerde constructies aan zoals *jansaliegeest, pietsnot, hansworst, jantje-van-leiden*, *mallebabbe* en *gekkehenkie*.[104] Voorts zijn er meerwoordige constructies van het type *vergeet-me-nietblauw*, *spring-in't-veld* en *broodje-aapverhaal*, partikelconstructies als *flapuit*, *bemoeial, weetniet* en partikelclusters als *voorop*, *achterover* en *tussendoor*.

v) *reduplicatieve* samenstellingen komen relatief weinig voor in het Nederlands; ze manifesteren zich meestal als samenstellingen waarvan het eerste lid identiek of bijna identiek is aan het tweede deel. Voorbeelden van de eerste soort zijn klanknabootsende woorden als *klopklop* en *wafwaf*, en leenwoorden als *couscous* en *gadogado*. Samenstellingen met bijna identieke woorddelen kunnen worden onderverdeeld in gevallen met klinkeralternantie (bijv. *wipwap*, *bimbam*, *wissewasje* en *prietpraat*) en gevallen met medeklinkeralternantie (bijv. *ietsiepietsie*, *hocuspocus*, *haaibaai* en *jemigdepemig*). Soms worden hierbij tussenklanken ingevoegd.[105]

vi) *pseudosamenstellingen* onderscheiden zich van echte samenstellingen doordat ze één of meer woorddelen bevatten die geen zelfstandige betekenis dragen, bijvoorbeeld doordat ze niet meer als zelfstandig woord in gebruik zijn. Een belangrijke overweging om toch van samenstellingen te spreken is het feit dat het hoofdaccent van deze woorden, dat steevast op het eerste woorddeel valt, meestal niet aan de hoofdaccentregel voldoet, maar wel aan de accentregel voor samenstellingen. In het MHB worden drie hoofdklassen onderscheiden, namelijk samenstellingen waarvan alleen het rechterdeel betekenis draagt (*aalbes*, *antwoord*, *sperzieboon* en *argwaan*), samenstellingen waarvan alleen het linkerdeel betekenis draagt (*eerbied*, *paspoort*, *diefstal*, *autoped*) en samenstellingen waarin geen van beide woorddelen een herkenbare betekenis bezit (*aalmoes*, *eekhoorn*, *oorlog*, *plankton*, *potlood*, *olifant*, *hospitaal*). Een probleem voor deze benadering is dat het rechterdeel lang niet altijd als morfologisch hoofd fungeert, d.w.z. niet bepalend is voor categorie, meervoudsvorm of lidwoordkeuze. Dit probleem kan wellicht worden ontzenuwd door in zulke gevallen van exocentrische pseudosamenstellingen te spreken. Men kan zich ook afvragen in hoeverre reguliere samenstellingen daadwerkelijk compositioneel worden geïnterpreteerd: een analyseerbaar woord als *vulpen* lijkt op dit punt niet wezenlijk te verschillen van het onanalyseerbare *potlood*.

vii) ook *samenstellende afleidingen* vormen een probleemcategorie: in het MHB worden deze gedefinieerd als een combinatie van een samenstelling en een afleiding, waarbij noch de samenstelling noch de afleiding zelfstandig voorkomt (althans niet in dezelfde betekenis). Zo kan *driewieler* niet worden geanalyseerd als een afleiding op basis van de samenstelling *driewiel*, en ook niet als een samenstelling van *drie* en *wieler*; het is dus een samenstellende afleiding met de woorddelen *drie* en *wiel* en het suffix -ER.[106] Dit geldt ook voor *bevelhebber*,

[104] Deze woorden vallen buiten het bereik van de spellingregels; hun spelvorm vertoont veel variatie.
[105] Daniels (2001) heeft deze klasse van woorden (c.q. herhalingswoorden) voor diverse talen in kaart gebracht.
[106] Het woordcluster *driewiel* kan echter ook als modificerend deel van een samenstelling voorkomen. Want in een radiobericht op 29 maart 2002 hoorde ik dat de consumentenbond niet erg te spreken was over de kwaliteit van *driewielwagens* (of zelfs *driewielkinderwagens*).

want noch *bevelheb* noch *hebber* zijn bestaande (c.q. acceptabele) woorden van het Nederlands. Het woord *boekverbrander* kan echter wel als een samenstelling worden opgevat, want zowel *boek* als *verbrander* zijn bestaande woorden van het Nederlands. Qua semantische structuur bestaat er echter grote gelijkenis met de samenstellende afleiding *bevelhebber*. Dit wijst erop dat de samenstellende afleiding een speciaal soort betekenisstructuur codeert, en dus bestaansrecht heeft als aparte constructie (ook indien er meerdere analyses mogelijk zijn). Daarom behandelt het MHB samenstellende afleidingen en afleidingen op basis van samenstelling als één categorie. Ze worden bovendien als endocentrische samenstellingen opgevat, wat impliceert dat het eindsuffix als hoofd fungeert, en dus categoriebepalend is.

3.4 De classificatie van lexemen en lexeeminterne eenheden

3.4.1 Introductie

Bij de beschrijving van de Nederlandse morfologie hanteert het MHB een affixgebaseerd derivatieperspectief op syntactische grondslag (verder aan te duiden als syntactisch classificatiemodel). Hierbij veronderstelt het MHB dat de morfologische regels van het Nederlands in beginsel lexeemgeoriënteerd zijn in de zin dat het functies zijn van bestaande lexemen naar nieuwe lexemen.[107] Maar men kan de door het MHB geboden affixinventarisatie ook herinterpreteren als een lexeemclassificatie, dat wil zeggen, als een (partiële) inventarisatie van de morfologische structuurtypes binnen elke lexeemklasse. Want elk affix correspondeert in feite met een morfologisch subtype van de bijbehorende lexeemklasse.[108] Zo correspondeert het suffix -ING met een operatie die V-lexemen in N-lexemen omzet. In het lexeemgeoriënteerde classificatieperspectief gaat het echter om een operatie die de klasse van N-lexemen systematisch uitbreidt met eenheden van het subtype [V + ING].

Het belang van deze alternatieve benadering schuilt in het feit dat de interne structuur van een lexeem doorgaans sterk van invloed is op zijn functionele kenmerken. Zo impliceert de aanwezigheid van het hoofd -ING niet alleen dat het hiermee afgeleide lexeem de categorie N draagt, maar ook dat dit lexeem het meervoudssuffix -EN en het bepaalde lidwoord *de* kiest. Er zijn echter ook affixen die polyfunctioneel zijn, wat inhoudt dat deze affixen meerdere syntactische categorieën kunnen selecteren en dus met meerdere lexeemklassen corresponderen. Dit is het geval bij affixen als -ER (ambigu tussen N, A en V) en -IEK (ambigu tussen N en A). Dergelijke affixen vormen een grote complicatie voor het syntactische classificatiesysteem van het MHB, want dit systeem veronderstelt dat morfemen (en lexemen) slechts één syntactische categorie kunnen bezitten, dus dat er geen polyfunctionele affixen bestaan. Indien een gegeven affixvorm meerdere affixfuncties kan vervullen, zal elke affixfunctie dus door een aparte affixingang moeten worden verantwoord. Deze aanpak resulteert in een model waarin geen structureel verband kan worden gelegd tussen lexemen die op dezelfde (mogelijk gelede) lexeemstam zijn gebaseerd, maar waarvan de syntactische categorie verschilt. Naar aanleiding van het hier beschreven probleem zal ik een alternatieve analyse uitwerken waarin morfemen niet langs syntactische, maar langs semantische weg worden getypeerd. In deze benadering is de semantische categorie bepalend voor de vraag welke syntactische functies een lexeem kan ondersteunen. Deze functies corresponderen vaak met een eigen inflectiekenmerk en hoeven niet per se onder dezelfde syntactische categorie te vallen. Dit analyse-

[107] Het MHB hanteert hier de term *woord*; het maakt geen onderscheid tussen woorden en lexemen.
[108] Deze vorm van morfologische subclassificatie is een vast onderdeel van de lexeemklassebeschrijvingen in de ANS; dit perspectief ligt ook ten grondslag aan het paradigmatische morfologiemodel van Schultink (1962) en aan het declaratieve grammaticamodel van Neef (1999). Zie Don (2003) voor een kritische beschouwing over Neef's analysemethode.

model biedt een rechtstreekse verklaring voor het feit dat lexemen vaak tot meerdere syntactische categorieën behoren.

Deze sectie is als volgt opgezet. H3.4.2 biedt een systematische bespreking van het syntactische classificatiesysteem van het MHB en de ANS (dat tevens de basis vormt van de meest gangbare grammaticamodellen); in dit kader zal ik een complete inventarisatie van traditionele lexeemklassen presenteren. In H3.4.3 zal ik dit classificatiemodel kritisch bespreken door uitvoerig in te gaan op systematische overeenkomsten in de semantische, syntactische en morfologische eigenschappen van de traditionele lexeemklassen. Hierbij probeer ik aan te tonen dat de geconstateerde overeenkomsten moeilijk verklaarbaar zijn als men vasthoudt aan het gebruikelijke syntactische classificatiesysteem, maar dat ze eenvoudig zijn te verantwoorden vanuit het hier voorgestelde morfologische classificatiesysteem. Vervolgens zal ik in H3.4.5 aandacht besteden aan een aantal fundamentele problemen bij de syntactische classificatie van lexeeminterne constituenten en wederom de voordelen van mijn eigen classificatiesysteem belichten. H3.4.6 ten slotte behandelt de classificatie van gebonden stammen, waarbij uitvoerig aandacht is voor de vraag in hoeverre de Nederlandse morfologie een paradigmatisch karakter heeft. Hiertoe ga ik na in hoeverre er overeenkomsten bestaan in het derivationele paradigma van een uitheemse en een inheemse stamverzameling met een vergelijkbare interne structuur. Op basis van deze data probeer ik aan te tonen dat het Nederlandse lexicon een paradigmatische structuur kent en dat dit heel goed verenigbaar is met een morfologisch classificatiesysteem. In H3.4.7 wordt het hoofdstuk kort samengevat.

3.4.2 Traditionele lexeemklassen

3.4.2.1 Introductie

In deze subsectie zal ik de traditionele lexeemklassen bespreken. Hierbij baseer ik me primair op de informatie in het MHB en de ANS, maar waar nodig zal ik deze informatie aanvullen met kennis uit andere bronnen (met name Booij (2002) en Beard (1995)) en met eigen observaties. Het MHB beperkt zich tot de morfologische analyse van lexemen met de categorie N, V, Adj en Adv; bij deze categorieën wordt ook aandacht besteed aan de syntactische plaatsingsmogelijkheden en aan het inflectiegedrag. Er komen alleen lexemen uit andere categorieën aan de orde als ze het linkerdeel van een samenkoppeling of samenstelling kunnen vormen, naast een incidentele vermelding als stam van een specifiek affix.[109] Via deze achterdeur wordt enige aandacht besteed aan partikels (P), telwoorden (T) en XP's (geïncorporeerde woordgroepen); andere categorieën, zoals connectieven (C) en determiners (D), waaronder clitica, blijven echter geheel buiten beschouwing. In de ANS worden deze categorieën wel behandeld, waarbij enkele morfologische constructiepatronen aan de orde komen die niet in het MHB worden besproken. Mijn eigen inventarisatie omvat bijna alle lexeemklassen uit de ANS (die er bij enkele klassen een iets andere ordening op nahoudt); deze inventarisatie bestaat uit lexeemklassen met achtereenvolgens categorie V, N, A, B, T, P, D, C. Tot slot zal ik kort ingaan op de classificatie van lexemen uit de restklasse R.

3.4.2.2 V-lexemen (verba)

De meeste V-lexemen (c.q. verba) kenmerken zich door de eigenschap dat ze een gebeurtenis of toestand aanduiden en deze temporeel kunnen localiseren via een inflectionele markering op de indicatiefmodus van de stam, namelijk tegenwoordige of verleden tijd, gevolgd door markeringen voor getal en persoon. Er bestaan echter nog vele andere V-stam-modi, zoals infinitief, participium presens, participium perfectum, conjunctief en imperatief. Bovendien bestaan er vaste hulpwerkwoorden die in combinatie met deze basis-modi complexere V-modi

[109] Dit blijft beperkt tot de suffixen -ER, -LING, -HEID, -TJE(S) en -ET, waarvan wordt vermeld dat ze ook in combinatie met een telwoord kunnen worden gebruikt, blijkens *tiener*, *tweeling*, *eenheid*, *vijfje*(*s)* en *kwartet*.

kunnen construeren, zoals het perfectum (met *hebben* of *zijn*)[110], het passivum indicatief (met *worden*), het passivum perfectum (met *zijn*) en de toekomende tijd (met *zullen*); elk van deze V-modi kent weer een tegenwoordige en een verleden tijd (via het hulpwerkwoord).

Sommige V-stam-modi zijn ambigu tussen een temporele (werkwoordelijke) toepassing en een niet-temporele (adjectivale of nominale) toepassing. Dit geldt bijvoorbeeld voor de infinitief, het participium presens en het participium perfectum. Tabel 3.1 demonstreert deze functionele ambiguïteit voor het werkwoord *spreken*. Dit type ambiguïteit roept de vraag op of de hier gedemonstreerde V-stam-modi wel inflectievormen zijn; men kan met evenveel recht stellen dat het om derivaties van een abstracte V-stam gaat die optioneel een temporele betekenis dragen. In H3.4.3 zal ik deze mogelijkheid verder uitdiepen.

V-stam-modus	V-toepassing	N-toepassing	A-toepassing
infinitief	te spreken	het spreken	-
participium presens	sprekend	de sprekende	sprekend gelijken
participium perfectum	heeft gesproken	het gesprokene	het gesproken woord

Tabel 3.1: De V-stam-modi infinitief, participium presens en participium perfectum: syntactische toepassingsmogelijkheden met voorbeelden.

Het voert te ver om hier een compleet overzicht van de werkwoordsmodi en hun inflectie- en conversiemogelijkheden te geven. Hiervoor raadplege men de ANS of Booij (2002). Ik zal me beperken tot het aanstippen van enkele eigenschappen die relevant zijn met het oog op de morfologische analyse van V-lexemen. De meeste werkwoorden kennen een zwakke vervoeging, waarbij de stam van de infintief als directe basis dient voor de vorming van presens, imperfectum en perfectum; hierbij wordt of een d-schema of een t-schema gevolgd.[111] In bijgaande voorbeelden worden voor elk werkwoord de volgende vormen gespecificeerd: infinitief, stamvorm tegenwoordige tijd (c.q. presens), stamvorm verleden tijd (c.q. imperfectum) en stamvorm voltooide tijd (c.q. perfectum). Deze stamvormen zijn al wel gemarkeerd voor de werkwoordstijd (via affixen, waarbij de 0-functor (F_0) aangeeft dat er geen hoorbare markering is), maar nog niet voor persoonskenmerken (getal, persoon en geslacht).

V-stammen met een stemhebbende uitgang selecteren het d-schema:

leven	[leev]	[leev]+de	ge+[leev]+d

V-stammen met een stemloze uitgang selecteren het t-schema:

werken	[werk]	[werk]+te	ge+[werk]+t

Indien de stam zelf op een d of een t eindigt, is een erop volgende d of t onhoorbaar:

branden	[brand]	[brand]+(d)e	ge+[brand]+(d)
spotten	[spot]	[spot]+(t)e	ge+[spot]+(t)

Er zijn circa 100 ongelede V-stammen die allomorfie (meestal klinkeralternantie) vertonen bij verandering van de werkwoordstijd. De bijbehorende werkwoorden kenmerken zich door een sterk vervoegingsschema. In dit schema krijgt de stamvorm van de verleden tijd een (onhoorbaar) 0-suffix (in plaats van -DE of -TE), terwijl de stamvorm van de voltooide tijd niet met -D/-T maar met -EN wordt gevormd. Hieronder volgen enkele voorbeelden:

strijken	[strijk]+0	[streek]+0	ge+[streek]+en
breken	[breek]+0	[bra(a)k]+0	ge+[brook]+en

[110] Traditioneel wordt aangenomen dat de keuze tussen deze hulpwerkwoorden lexicaal bepaald is, maar in Lieber & Baayen (1997) wordt betoogd dat deze keuze meestal op aspectuele overwegingen berust.
[111] Indien de d/t-keuze via een aparte aanpassingsregel wordt verantwoord, zou slechts één schema nodig zijn.

Zowel zwakke als sterke V-stammen kunnen vooraf worden gegaan door een gebonden prefix (bijv. BE-, VER- of ONT-) of een partikel (d.w.z. een prefix dat etymologisch verwant is aan een prepositie of adverbium). Hierbij kunnen de partikelwerkwoorden (PV's) worden onderverdeeld in een scheidbare klasse ([+SPV], met beklemtoond partikel, bijv. *overlopen* en *aansteken*) en een onscheidbare klasse ([-SPV], met onbeklemtoond partikel, bijv. *achterhalen* en *ondernemen*).[112] Hieronder geef ik per klasse enkele voorbeeldparadigma's.

[-PV]-klasse: complexe werkwoorden met een [-partikel] prefix

verwerken	ver+[werk]	ver+[werk]+te	ver+[werk]+t
bespreken	be+[spreek]	be+[spra(a)k]+0	be+[sprok]+en

[-SPV]-klasse: complexe werkwoorden met een [-scheidbaar,+partikel] prefix

overleggen	over+[leg]	over+[leg]+de	over+[leg]+d
onderwerpen	onder+[werp]	onder+[wierp]+0	onder+[worp]+en

[+SPV]-klasse: complexe werkwoorden met een [+scheidbaar,+partikel] prefix

aantonen	aan+[toon]	aan+[toon]+de	aan+ge+[toon]+d
uitspreken	uit+[spreek]	uit+[spra(a)k]+0	uit+ge+[sprok]+en

Al deze werkwoordklassen volgen het basisschema [prefix + W]$_V$, waarbij W voor een wortel van categorie V staat (d.w.z. voor een V-stam). Dit schema is ook toepasbaar op wortels met een andere categorie, zoals nominale wortels (bijv. de N-stam PLANT in *beplanten en aanplanten*), adjectivische wortels (bijv. de A-stam SCHOON in *verschonen* en *opschonen*) en categorieloze wortels (bijv. de X-stam BRUIK in *verbruiken*). Hiernaast kan W ook met een werkwoord corresponderen dat op hetzelfde constructieschema is gebaseerd (wat neerkomt op het stapelen van prefixen; dit blijkt uit vormen als *be+geleiden*, *her+overwegen*, *uit+betalen*, *over+verhitten* en *uit+onderhandelen*; het prefix kan ook zelf geleed zijn, zoals VOOR+OP in *voorop+stellen* en ONDER+UIT in *onderuit+halen*. Tot slot zijn er vele complexe werkwoorden waarvan het linkerdeel met een stam of (complex) lexeem correspondeert. Hiervoor geldt vaak dat het vormenparadigma lacunes vertoont: zo kennen samengestelde werkwoorden vaak alleen maar een infinitiefvorm (bijv. *zweefvliegen* en *stagediven*).

Voor complexe werkwoorden met een V-wortel geldt vrijwel altijd dat de resulterende V-stam dezelfde inflectie-eigenschappen vertoont als de geïncorporeerde stam, ongeacht of de nieuwe eenheid een compositionele betekenis heeft. Als de ingebedde wortel met een andere categorie correspondeert, kent het resulterende werkwoord echter een zwakke vervoeging. Los van dit contrast geldt dat de voltooide tijd geen prefix GE- toelaat als de wortel direct vooraf wordt gegaan door een onscheidbaar prefix (zoals BE-, VER-, ONT- of een niet-beklemtoond partikel); bij dergelijke lexemen kan de modus *voltooide tijd* dus alleen worden afgelezen aan de stamvorm en/of zijn suffix. Bij werkwoorden met een beklemtoond (scheidbaar) partikel daarentegen gedraagt de ingebedde stam (mits prefixloos)[113] zich als een zelfstandige V-stam: de voltooide tijd wordt hier gewoon door het prefix GE- gemarkeerd, wat zichtbaar is in vormparen als *aantonen* / *aangetoond* (aan+ge+[toon]+d]) en *uitspreken* / *uitgesproken* (uit+ge+[sprok]+en]). Dit wordt meestal uitgelegd als een bewijs dat scheidbaar complexe werkwoorden zijn opgebouwd uit een zelfstandig werkwoord en een partikel.[114] In mijn visie is het echter noodzakelijk om onderscheid te maken tussen wortels en stammen (zoals ik nog zal toelichten; zie H3.4.3 en verder). Daarom zal ik in H3.4.6 een alternatieve analyse voorstellen (zie ook H3.6). Hierbij krijgen stammen pas V-status in combinatie met een V-vormende functor. Deze functor correspondeert ofwel met een prefix of met de (coverte) operator [0/GE].

[112] Zie Blom (2005) voor een synchrone en diachrone studie naar de eigenschappen van deze werkwoorden.
[113] Als de ingebedde stam het resultaat is van een prefix-wortel-combinatie, is normaliter geen voltooide tijd met *ge-* mogelijk. Dit blijkt uit vormen als *heroverwogen* en *uitonderhandeld* naast. **hergeoverwogen* en ?*uitgeonderhandeld* (waar *uit* een iteratief proces begrenst, analoog aan de adjectieven *uitgepraat* en *uitgeouwehoerd*).
[114] Deze analyse vindt steun bij o.a. Neeleman & Schippers (1992), Drijkoningen (1995) en Booij (2002).

3.4.2.3 N-lexemen (nomina)

N-lexemen (c.q. nomina) kenmerken zich door de eigenschap dat ze (na specificatie van hun morfologische en syntactische variabelen) als argument kunnen dienen van een predicaat (te weten een eigenschap of relatie die door een N, V of A wordt geïntroduceerd). Onder de N-lexemen kunnen twee hoofdklassen worden onderscheiden, namelijk conceptuele nomina (C-nomina), d.w.z. nomina die een concept introduceren en standaard een lidwoord of telwoord vereisen, en referentiële nomina (R-nomina) c.q. eigennamen. Ik zal eerst ingaan op de eigenschappen van de C-nomina. De tweede klasse zal verderop aan de orde komen.

De C-nomina kunnen worden onderverdeeld in nomina die naar (semi-)permanente ([-T]) concepten (zoals personen, objecten en materialen) verwijzen en nomina die naar tijdsafhankelijke ([+T]) concepten (zoals gebeurtenissen en toestanden) verwijzen. Voor [+T]-nomina geldt dat ze vaak van een werkwoordstam zijn afgeleid (zo kunnen de N-lexemen *draai* en *draaiing* als een nominalisatie van het werkwoord *draaien* worden gezien). Los van het [+T]/[-T]-onderscheid kunnen C-nomina worden onderverdeeld in telbare concepten en niet-telbare concepten. De meeste telbare N-lexemen kennen naast hun basisvorm ook een meervoudsvorm en een verkleinvorm; verder is bij persoonsnamen en dierennamen soms een vrouwelijke vorm beschikbaar; tot slot bezit elk nomen dat in het enkelvoud kan worden gebruikt (inclusief een deel van de niet-telbare nomina) een woordgeslacht, dat onder meer bepalend is voor de keuze van het bepaalde lidwoord.[115]

Bij inheemse N-lexemen correspondeert de meervoudsvorm normaal gesproken met het suffix -S of -EN.[116] Hierbij is vaak een vaste relatie zichtbaar tussen de klankstructuur van de stamuitgang en het meervoudssuffix (zie MHB, p. 158-160).[117] Voor een aantal suffixen (waaronder -AAR, -IER en -AARD) geldt echter dat ze systematisch een andere meervoudsvorm en/of een ander lidwoord kiezen dan op basis van hun klankvorm verwacht mag worden:

	[-suffix]	[+suffix]
N-lexeem op /aar/ resp. -AAR	*sigaar-sigaren*	*tekenaar-tekenaars*
N-lexeem op /ier/ resp. -IER	*rivier-rivieren*	*koetsier-koetsiers*
N-lexeem op /aard/ resp. -AARD	*paard-paarden*	*rijkaard-rijkaards*

Tabel 3.2: De relatie tussen N-uitgang en suffixstatus, met voorbeelden.

Een deel van de uitheemse N-lexemen kent afwijkend inflectie-gedrag met betrekking tot getal, verkleinvorm en vrouwelijke vorm. Wat betreft de uitdrukking van getal specificeert het MHB (p. 163) de volgende relaties tussen stamuitgang en meervoudsvorm:

a) persoonsnaam op -a, dan meervoud op -ae of op -'s
   collega-collegae-collega's
b) persoonsnaam op -us/icus, dan meervoud op -i (of evt. op -en)
   historicus-historici, doctorandus-doctorandi-doctorandussen
c) zaaknaam op -ex/-ix, dan meervoud op -ices of op -en
   index-indices-indexen, appendix-appendices-appendixen
d) zaaknaam op -um, dan meervoud op -a of op -s
   centrum-centra-centrums, museum-musea-museums
e) Italiaanse stam op -o, dan meervoud op -i of op -'s

[115] Hierbij geldt de vuistregel dat persoonsnamen een voorkeur voor *de* hebben en objectnamen voor *het.* Het is niet altijd duidelijk of pluralia tantum (zoals *hersenen*) ook een woordgeslacht bezitten (vgl. [*?de/*het*] *hersen*).
[116] In een enkel geval wordt het suffix -EN echter vooraf gegaan door het segment -er, bijv. *kind/kind+er+en*. Ook zijn er flink wat stammen waarvan de meervoudsvorm klinkerverlenging kent, bijv. *gebed/gebeed+en*.
[117] Indien zo'n uitgang dezelfde vorm heeft als een regulier suffix, gedraagt de stam zich vaak alsof hij geleed is. Zo selecteren N-stammen met de uitgang *-ing*, zoals *koning* en *haring*, vrijwel altijd het lidwoord *de* en het meervoudssuffix -EN (net als N-afleidingen met het suffix -ING). Zie Trommelen & Zonneveld (1986).

saldo-saldi-saldo's, solo-soli-solo's

f) zaaknaam op -ma, dan meervoud op -mata of op -'s

lemma-lemmata-lemma's, paradigma-paradigmata-paradigma's

Hoewel de meervoudsvorm hier als afleiding van de enkelvoudsvorm wordt beschreven, zou men ook kunnen stellen dat deze N-lexemen een paradigmatisch inflectieschema bezitten, waarbij zowel het enkelvoud als het meervoud door een suffix wordt gecodeerd. Zo zou men een lexeem √CENTR kunnen postuleren waarvan het enkelvoud met het suffix -UM en het meervoud met het suffix -A correspondeert. Desgewenst kan men deze uitheemse structuur ook negeren en de enkelvoudsvorm herinterpreteren als een ongelede N-stam, d.w.z. een stam die zich wat betreft meervoudsvorming als inheemse N-lexemen gedraagt; in dat geval verkrijgt men de meervoudsvorm door de stam met het suffix -S uit te breiden, met als resultaat *lemma's*. In H3.5 zal ik nader ingaan op de vraag in hoeverre de Nederlandse morfologie paradigmatisch is georganiseerd.

De tweede klasse van N-lexemen bespreken, de referentiële nomina c.q. eigennamen, kenmerkt zich door het feit dat ze een unieke entiteit (namelijk persoon, dier, ding of abstract concept) of een unieke locatie aanduiden; ze zijn dan ook niet telbaar, al kunnen ze wel betrekking hebben op een meervoudig concept. Eigennamen bevinden zich dus op de grens van telbare en niet-telbare concepten. Als gevolg van deze eigenschap vertonen ze heel ander gedrag dan conceptuele nomina, hoewel ze syntactisch gezien dezelfde functie vervullen (namelijk de functie van argument bij een predicaat). Nederlandse eigennamen kennen meestal maar één vorm, namelijk óf een enkelvoudsvorm zonder lidwoord (bijv. *Hendrik* of *de Veluwe*) óf een meervoudsvorm met lidwoord (bijv. *de Verenigde Naties* of *de Wadden*). Indien nodig kan men persoonsnamen echter toch in het meervoud zetten; in dat geval duidt de eigennaam niet langer één entiteit aan (bijv. *Louis Andriessen*), maar alle entiteiten die in een bepaald domein dezelfde naam dragen (bijv. *de Andriessens*).

In het MHB worden veel suffixen besproken die eigennamen manipuleren, zoals het afleiden van een inwonersnaam uit een lands- of stadsnaam of de afleiding van het bijbehorende adjectief. Vooral bij uitheemse namen blijkt dit vaak samen te gaan met allerlei aanpassingen in de stamvorm (bijv. *Arist*ó*teles* - *Aristoteliáans*). Dit is een groot probleem voor de syntagmatische benadering van het MHB. Booij (2002) laat zien dat veel van deze aanpassingen zich eenvoudig laten verklaren indien men een paradigmatisch perspectief inneemt.

3.4.2.4 A-lexemen (adjectieven)

De ANS definieert A-lexemen (c.q. adjectieven) als woorden die een nadere bijzonderheid van een zelfstandigheid, in het bijzonder een nomen, kunnen specificeren. Hierbij wordt onderscheid gemaakt tussen modificatie van een eigenschap, bijv. *tenger*, *vierkant* en *lastig*, en modificatie van een toestand, bijv. *dronken*, *ziek* en *nat*. Naast deze adnominale (N-modificerende) eigenschap kunnen de meeste adjectieven ook een adverbiale (V-modificerende) functie vervullen.[118] Dit laatste houdt in dat het adjectief een nadere bijzonderheid specificeert bij een werkwoord (V) of een hier omheen geformeerde woordgroep (bijv. een VP), zoals het adjectief *snel* in het zinnetje *De atleet rende snel* (naast *de snelle atleet*). In morfologische derivaties (bijv. *groenig*, *snelheid*) en bij linkerdelen van samenstellingen (*groenvoer*, *snelkookpan*) valt ook niet uit te maken welke functie de voorkeur verdient.

Als gevolg van deze overlap in functies besteedt de ANS geen aparte aandacht aan adverbiale adjectieven; ook het MHB maakt hier geen onderscheid in. Dat toch vaak een categorieonderscheid wordt gemaakt tussen de hier besproken functies, komt waarschijnlijk doordat

[118] Men spreekt meestal van *adverbia* en *adverbiale* eigenschappen, maar eigenlijk is de aanduiding *adverbaal* inzichtelijker, omdat deze een expliciete relatie met verba legt (analoog aan adnomina); cf. Beard, 1995.

het Engels een speciale markering kent voor lexemen met een adverbiale functie, namelijk het suffix -LY. Maar in het Nederlands onderscheidt de adverbiale functie zich alleen doordat deze geen inflectie toelaat. Desondanks worden adverbiaal gebruikte adjectieven vaak als een subklasse van de bijwoorden beschreven, want morfologisch en syntactisch gezien vertonen ze hetzelfde gedrag als niet-adjectivale ("echte") bijwoorden (zie verder paragraaf 3.4.2.4). Deze analyse heeft als nadeel dat er een aparte lexeemingang nodig is om de adverbiale toepassing van een A-lexeem te verantwoorden.

Adjectieven kunnen nader worden onderverdeeld in absolute en gradeerbare adjectieven. In tegenstelling tot absolute adjectieven (zoals *dood* en *eeuwig*) kennen gradeerbare adjectieven naast hun basisvorm een vergrotende trap en een overtreffende trap. Deze trappen kunnen op twee verschillende manieren worden gevormd, namelijk langs morfologische weg (met behulp van de suffixen -ER (OF -DER) EN -ST (OF -T), bijv. MOOIER en MOOIST) of langs syntactische weg (met behulp van de woorden *meer* en *meest*, bijv. MEER BELEZEN en MEEST BELEZEN); in het algemeen vertonen korte, ongelede adjectieven een voorkeur voor de morfologische weg en lange adjectieven voor de syntactische weg. Onder de gradeerbare adjectieven zijn er enkele die suppletie vertonen, namelijk *veel*, *weinig* en *goed.*

Bij adjectieven met een adnominale functie kan ook onderscheid worden gemaakt tussen adjectieven in attributieve positie en adjectieven in predicatieve (in feite adverbiale) positie. Alleen in attributieve positie is een buigings-*e* mogelijk, bijvoorbeeld *het mooie boek*; adjectieven in predicatieve positie zijn altijd onverbogen (bijv. *het boek is mooi*), evenals adjectieven in een adverbiale functie (bijv. *hij schrijft mooi*). Er bestaat ook een gebruiksmodus waarin het adjectief door de uitgang -S wordt gevolgd; deze modus correspondeert met constructies van het type *iets* A-*s (bijv. Zij zocht iets groens*) of *weinig* A-*s* (bijv. Er is *weinig opmerkelijks gebeurd*). Vrijwel alle adjectieven kunnen in een nomen worden omgezet door ze een buigings-*e* te geven (zo kan iemand die *mooi* is worden aangeduid als *de mooie* en iets wat *mooi* is als *het mooie*). Bij sommige adjectieven kan de buigings-*e* zelfs achterwege blijven, bijv. bij *gek* (blijkens *de gek*).

Bij adjectieven die zijn afgeleid van een stofnaam is meestal geen buigings-*e* mogelijk; dit geldt bijvoorbeeld voor adjectieven met het suffix -EN, zoals *houten*, *koperen* en *aluminium*; dezelfde beperking geldt voor geconverteerde leenwoorden als *nylon* (in *een nylon kous*) en *plastic* (in *een plastic pop*). Deze adjectieven kunnen alleen adnominaal (d.w.z. in combinatie met een nomen) worden gebruikt. Deze beperking geldt ook voor adjectieven die van een telwoord zijn afgeleid (door middel van de uitgang -DE of -STE), bijvoorbeeld *tiende* of *honderdste*, en voor allerlei morfologisch gelede adjectieven, waaronder *zeldzaam*, *vermeend*, *voormalig*, *onmetelijk*, *ontelbaar*, *onherroepelijk* en *onverkwikkelijk.* Deze functiebeperking lijkt vaak een semantische reden te hebben.

Tot slot kan worden opgemerkt dat de meeste adjectieven zowel een positieve als een negatieve vorm kennen, die beide een basis kunnen vormen voor verdere afleidingen; de negatieve vorm wordt meestal uitgedrukt door een adnominale stam met het prefix ON- te combineren, bijv. *eerlijk* - *oneerlijk*, maar er zijn ook adjectiefparen die zich door suppletie kenmerken, bijv. *mooi* - *lelijk* (**onmooi*). In de adverbiale toepassing is soms ook een verkleinvorm mogelijk, zoals in *losjes*, *zachtjes*, *stilletjes*. Deze leent zich niet voor adnominaal gebruik.

Om de parallellie tussen de adnominale en de adverbiale adjectieffunctie zo goed mogelijk te laten uitkomen, zal ik de belangrijkste derivatie- en verbuigingsmogelijkheden nog eens per functie op een rijtje zetten:

adnominale adjectieffunctie:
a. *de snelle atleet; de hoge sprong; de verre gooi*
b. *de atleet is snel; de lat is hoog; de afstand is ver*

adverbiale adjectieffunctie:
a. *Alfred {liep snel, sprong hoog, gooide ver}*
b. *Alfred {liep sneller / sprong hoger / gooide verder} dan Bernard*
c. *Alfred {liep het snelst / sprong het hoogst / gooide het verst} van allen*

3.4.2.5 B-lexemen (bijwoorden)

De term bijwoord (c.q. *adverbium*)[119] verwijst meestal naar woorden die informatie geven over een (specifiek aspect van een) werkwoordelijk of adjectivisch predicaat. Zoals al aan de orde kwam in paragraaf 3.4.2.3 onderscheidt de ANS (conform de traditie) twee subklassen, te weten de deadjectivische bijwoorden, die zowel nomina als verba (c.q. verbale predicaten) kunnen modificeren en daarom bij de adjectieven worden behandeld, en de "echte" bijwoorden, die zich van de eerste klasse onderscheiden doordat ze normaliter geen nomina kunnen modificeren, maar meestal gespecialiseerd zijn in een specifiek aspect van een verbaal of adjectivisch predicaat. In het Engels wordt alleen de eerste subklasse door het suffix -LY gemarkeerd. Dat deze twee bijwoordklassen toch tot dezelfde categorie worden gerekend, komt doordat ze geen van beide inflectie toestaan en syntactisch gezien hetzelfde gedrag vertonen. In deze studie zal ik de deadjectivische bijwoorden als V-modificerende adjectieven (A_V) aanduiden (ter onderscheiding van N-modificerende adjectieven: A_N) en de overige bijwoorden als B-lexemen.

De B-lexemen kunnen in tal van subklassen worden verdeeld (cf. de ANS), waaronder:

B(G):	bijwoord van graad	*tamelijk*, *behoorlijk, ongelooflijk*, *zeer*
B(F):	bijwoord van frequentie	*steeds*, *vaak*, *nooit, altijd*
B(T):	bijwoord van tijd	*gisteren*, *nu*, *straks*
B(L):	bijwoord van locatie	*hier*, *ergens*, *nergens*, *overal*
B(P):	bijwoord van padtype	*voort*, *heen*, *terug, rond*, *weg*
B(M):	bijwoord van modaliteit	*misschien*, *mogelijk*, *zeker*

Een belangrijk verschil tussen A_V- en B-lexemen is dat de laatste zelden gradeerbaar zijn; enkele uitzonderingen zijn het frequentie-aanduidende bijwoord *vaak*, dat ook de vormen *vaker* en *vaakst* kent, en het modale bijwoord *waarschijnlijk*, dat ook de vormen *waarschijnlijker* en *waarschijnlijkst* kent. Een ander verschil is dat B-lexemen meestal niet als linkerdeel van een scheidbaar samengesteld werkwoord c.q. samenkoppeling kunnen optreden; deze optie staat alleen open voor de pad-additieven, bijv. *voortlopen* en *weglopen*.

3.4.2.6 T-lexemen (telwoorden)

Zowel de ANS als het MHB (p. 395-396) behandelt telwoorden als een aparte syntactische klasse. Deze klasse wordt, conform de traditie, onderverdeeld in hoofdtelwoorden en rangtelwoorden, die beide weer in bepaalde en onbepaalde telwoorden uiteenvallen. Volgens het MHB kunnen alleen de bepaalde telwoorden als basis van een afleiding of samenstelling dienen. Daarom zal ik me in deze paragraaf beperken tot bespreking van de bepaalde telwoorden; de onbepaalde telwoorden zullen bij de D-lexemen worden behandeld.

De bepaalde telwoorden vormen een open morfologische klasse, omdat op basis van enkele tientallen basismorfemen (zoals *een*, *twee, drie, ...*, *tien*, *elf*, *twaalf*, *twintig*, *dertig*, *honderd* etc.) systematisch nieuwe getallen kunnen worden geconstrueerd.[120] Semantisch gezien lijken de hoofdtelwoorden het meest op eigennamen, aangezien ze uniek verwijzen naar een specifiek getal. Ook syntactisch gezien lijken ze op nomina, want ze kunnen vooraf worden gegaan

[119] Adverbia worden meestal aangeduid als ADV; maar ik kies liever voor een enkele letter: de B van bijwoord.
[120] De verzameling der telwoorden is echter niet oneindig groot, want voor elke factor 1000 is een nieuwe basiseenheid nodig en het algemeen bekende namenstelsel reikt niet verder dan de basiseenheid *triljard*.

door een lidwoord, terwijl de kleine getallen ook meervoudsvorming toestaan (bijv. *de twee - de tweeën* en *de negen - de negens*).[121] Maar hoofdtelwoorden kunnen zich ook adjectivisch gedragen in de zin dat ze meestal een nomen modificeren, bijv. *vier symfonieën*; alleen kunnen ze geen verbuiging ondergaan.[122]

Rangtelwoorden lijken nog sterker op adjectieven (beperkt tot de adnominale functie), want zij staan bij voorkeur voor een nomen en bezitten standaard een verbogen vorm (bijv. *de vierde symfonie*). Anderzijds kennen rangtelwoorden geen vergelijkingstrappen. Voorts kunnen rangtelwoorden altijd zelfstandig worden gebruikt (zonodig met een extra *-n*), blijkens *de vierde is het mooist* en *deze maat bestaat uit vierden*. Deze observaties laten zien dat het moeilijk is om hoofd- en rangtelwoorden syntactisch te classificeren. Dit verklaart mogelijk waarom ze traditioneel als aparte categorie worden behandeld.

3.4.2.7 P-lexemen (partikels)

Partikels, of meer specifiek *adposities*[123], zijn herkenbaar aan het feit dat ze een syntactische relatie kunnen leggen tussen een verbaal of nominaal predicaat en een nominale constituent; in de VP *naar de wedstrijd kijken* legt de adpositie *naar* bijvoorbeeld een relatie tussen de V *kijken* en de NP *de wedstrijd*. De combinatie *naar iets kijken* heeft in dit geval een gelexicaliseerde status, maar dit hoeft niet altijd zo te zijn. Hoewel de adpositie-klasse in principe onbeperkt uitbreidbaar is (blijkens duidelijk gelede vormen als *benoorden, hangende, overeenkomstig, uitgezonderd, richting* en *tegenover*), bezitten de meeste talen slechts een klein aantal ongelede adposities, die meestal een ruimtelijke of temporele relatie uitdrukken.[124] Voor het Nederlands ziet deze verzameling er als volgt uit (beperkt tot ruimtelijke partikels):

> *aan; achter; af; bij; binnen, boven, buiten, door; in; langs; mee; met; na; naar; neer; om; onder; op; over; rond; tegen; te; toe; tot; tussen; uit; van, voor; zonder*

Deze lijst kan eventueel worden uitgebreid met strikt temporele adposities, zoals *sinds*, *sedert* en *tijdens*. In het Nederlands kunnen adposities zowel voor als na de NP staan; in het eerste geval heten ze preposities, in het tweede postposities. Postposities drukken doorgaans een richting uit, terwijl preposities op dit punt vrij zijn. Een deel van de adposities kan zowel prenominaal als postnominaal worden gebruikt, al gaat dit soms samen met een vormverandering (zo verandert *met* in *mee* en *tot* in *toe*).

In het MHB worden alleen adposities besproken die deel kunnen uitmaken van een lexeem. Hierbij gaat het om twee soorten constructies, namelijk de scheidbaar samengestelde partikelwerkwoorden ([+SPV]-lexemen), bijv. *óverlopen*, en de onscheidbaar samengestelde partikelwerkwoorden ([-SPV]-lexemen), bijv. *overdénken*; deze constructies verschillen in het feit dat [+SPV]-lexemen een beklemtoond partikel bezitten dat los van de stam kan staan, terwijl [-SPV]-lexemen zich als normale prefix-werkwoorden gedragen. Volgens het MHB zijn de meeste [-SPV]-lexemen echter uit [+SPV]-lexemen voortgekomen.[125] Dit impliceert dat alle adposities die deel kunnen uitmaken van een [-SPV]-lexeem ook in een [+SPV]-lexeem kunnen voorkomen. Het MHB behandelt deze [±SPV]-adposities gescheiden van de adposities die alleen in [+SPV]-lexemen kunnen voorkomen. Deze laatste adpositie-klasse

[121] Een soortgelijk fenomeen treft men aan in adverbiale constructies als "in X-en" (bijv. "in negenen") en "met zijn X-en" (bijv. "met zijn negenen'). Zie Booij (2005b, ms.) voor een nadere beschouwing.
[122] Hierop bestaat een kleine uitzondering, namelijk de vorm *ene* van het telwoord *een* (bijv. in *Ene Harry...*).
[123] De term *adpositie* ontleen ik aan Beard (1995); deze term generaliseert over preposities en postposities (zoals affixen over prefixen en suffixen generaliseren). In het MHB en de ANS dient de term *prepositie* vaak als verzamelterm voor prenominale en postnominale adposities; de term *adpositie* daarentegen komt er niet voor.
[124] Het proefschrift van Nard Loonen (2003) biedt een zeer gedetailleerde classificatie.
[125] Dit blijkt echter veel genuanceerder te liggen. Zie Blom (2005) voor een uitvoerige studie.

wordt nog onderverdeeld in strikte preposities, strikte postposities en flexibele adposities. Het onderstaande overzicht specificeert voor elk van de genoemde subklassen welke adposities het MHB eraan toekent.

[+/-klemtoon]:	*aan, achter, door, om onder, over, voor*
preposities:	*bij, binnen, boven, buiten, na, tegen*
postposities:	*af, mee, toe*
pre/post-posities:	*in, langs, op, rond, uit, voorbij*

Vergelijking met de voorgaande adpositielijst leert dat *met, naar, van, te* en *zonder* blijkbaar geen onderdeel kunnen zijn van een V-lexeem. Dit geldt ook voor strikt temporele adposities als *sinds*, *sedert* en *tijdens*.

3.4.2.8 D-lexemen (determiners)

Syntactisch gezien is het moeilijk om een grens te trekken tussen lidwoorden, voornaamwoorden, onbepaalde voornaamwoorden, onbepaalde bijwoorden en telwoorden. Sinds Abney (1987) is het dan ook gebruikelijk om al deze functieklassen in één verzamelcategorie onder te brengen, namelijk de determiner (D).[126] Dit heeft als voordeel dat geen categoriaal onderscheid hoeft te worden gemaakt tussen:

a) lidwoorden, namelijk *de*$_{SG}$, *het*, *een*, [0]$_{SG}$, *de*$_{PL}$, [0]$_{PL}$;[127]

b) verwijswoorden voor personen, dieren en dingen (c.q. persoonlijke voornaamwoorden), zoals *ik*, *jij*, *hij, het, wij, jullie, men, haar, hun, deze, dit, die, dat, gene, iemand, niemand* en daaraan gerelateerde vraagwoorden zoals *wie*, *wat* en *welke*. Van veel voornaamwoorden bestaat ook een clitic-variant, bijvoorbeeld *'m*, *'r*, *'t*, *me*, *je*, *ze*, *m'n*, *d'r* en *z'n*; deze worden vaak als suffix aan een werkwoord vastgehecht, bijvoorbeeld in *ze heeft 't 'm verteld*.

c) verwijswoorden voor plaats, tijd en toestand (c.q. onbepaalde bijwoorden), zoals de basislexemen *hier, daar, ergens, toen, dan, ooit, nooit, zo,* de samengestelde lexemen *hiermee, daarop, eronder* en de vraagwoorden als *waar, wanneer* en *hoe*;

d) kwantitatieve verwijswoorden (c.q. onbepaalde telwoorden), zoals *geen, elk, alle, beide, sommige, de meeste/minste, veel, weinig, enkele, verschillende, talloze*, *ontelbare*, *meer, minder, meerdere*, *laatste*, *hoeveelste.*

Nederlandse determiners (excl. de bepaalde telwoorden) kunnen over het algemeen niet als basis van een morfologische derivatie of samenstelling fungeren, maar de persoonlijke voornaamwoorden en de onbepaalde telwoorden vertonen wel adjectivische inflectie (namelijk een buigings-*e*). In het MHB wordt overigens geen aandacht besteed aan determiners.

3.4.2.9 C-lexemen (connectieven)

Connectieven (c.q. voegwoorden) hebben als functie om een neven- of onderschikkingsverband te leggen tussen twee opeenvolgende constituenten, waarbij de gemodificeerde constituent vrijwel altijd met een (deel)zin (S) correspondeert; dit geldt niet voor *en* en *of*, die constituenten van een willekeurig type kunnen verbinden. Voorbeelden van nevenschikkende voegwoorden zijn *en*, *of*, *dus*, *daarom*, *maar* en *toen*; ze hebben doorgaans geen effect op de woordvolgorde en kunnen niet als basis van een morfologische constructie worden gebruikt. Voorbeelden van onderschikkende voegwoorden zijn *dat*, *of*, *om*, *terwijl*, *omdat*, *doordat* en *nadat*; ze leiden altijd tot bijzinsvolgorde. Onderschikkende voegwoorden vertonen veel overeenkomsten met adposities; het verschil is dat connectieven een zin als complement nemen,

[126] Het is echter niet zeker of de bepaalde telwoorden D-status moeten krijgen; cf. Verkuyl (1993).

[127] Voor het lidwoord *de* bestaan overigens ook nog inflectievormen, namelijk *der*, *den* en *des*. Dergelijke oude vormen treft men ook nog aan bij voornaamwoorden als *deze* (*dezer*/*dezes*), *gene* (*gener*) en *ons* (*onze*, *onzer*).

terwijl adposities een nonimale constituent modificeren. Veel connectieven bestaan echter uit een combinatie van een adpositie en het onderschikkende voegwoord *dat*, bijv. *omdat, doordat, voordat* en *nadat*. Connectieven als *onderwijl*, *ondertussen, alvorens, hierdoor* en *daarom* lijken eveneens van een adpositie te zijn afgeleid. Het MHB besteedt echter geen aandacht aan dit type samenstellingen.

3.4.2.10 Restklasse R

Elke poging tot een complete classificatie van lexemen lijkt gedoemd te mislukken. Er komen namelijk altijd woorden bij die zich volstrekt eigenzinnig gedragen, want de werkelijkheid laat zich nu eenmaal niet helemaal in een vooraf gedefinieerd begrippensysteem vangen, ook de taalwerkelijkheid niet. Relatief vertrouwde, maar niettemin moeilijk te vangen fenomenen zijn lexemen die met een uitroep (*oeps!*), vloek (*potverdriedubbeltjes!*), afkortingen (*m.u.v.*, *s.v.p.*, *p.s.*, *resp.*, *etc.*) of een onomatopee (*oehoe*, *ia, boink, ring*) corresponderen. Indien men een woordenboek doorneemt kan men nog tal van andere types tegenkomen. Syntactisch en morfologisch gezien vertonen dergelijke lexemen nogal idiosyncratisch gedrag, zodat geen coherente categorie kan worden gedefinieerd. Dergelijke lexemen kunnen, in afwachting van nader onderzoek, in de restklasse R worden ondergebracht.

3.4.3 Naar een morfologisch gestructureerd classificatiesysteem

Het traditionele systeem voor lexeemclassificatie berust op het uitgangspunt dat lexemen één op één met een syntactische categorie (c.q. bundel van syntactische functies) corresponderen en dat deze categorie (c.q. elk van de hierdoor omvatte functies) vaak samengaat met een categoriespecifiek inflectieparadigma. In deze benadering corresponderen N-lexemen met syntactische eenheden waarvan het plaatsingsgedrag en het inflectiegedrag grotendeels uit de categorie N volgen. Deze N-categorie omvat onder meer de functies "complete NP" (bijv. *het grote huis*) en "kale N" (bijv. de N *viool* in *Zij speelde viool*). Meer in het algemeen wordt aangenomen dat een lexeemklasse X gedrag vertoont dat uit categorie X volgt, al worden binnen zo'n categorie vaak weer subklassen geïntroduceerd. Zo wordt bij de nomina onderscheid gemaakt tussen telbare en niet-telbare nomina, wat nodig is om te verantwoorden dat de tweede klasse normaliter geen meervoudsuitgang toelaat en niet door een telwoord kan worden voorafgegaan (bijv. *water* vs. *drie waters*). En binnen de klasse van telbare nomina zijn subklassen nodig voor nomina met een *s*-meervoud, nomina met een *en*-meervoud, nomina met zowel een *s*- als een *en*-meervoud (zoals *lade*) en nomina met een uitzonderlijke meervoudsvorm (zoals *museum* en *neerlandicus*). Als gevolg van deze differentiatie is het niet mogelijk om de in H3.4.2 onderscheiden lexeemklassen scherp te definiëren: het zijn eerder bundels van stammen met overlap in betekeniskenmerken, syntactische functies, morfologisch gedrag en inflectiegedrag. Bovendien bestaan sommige lexeemklassen, in het bijzonder V-lexemen, uit een zeer heterogene verzameling subfuncties. Dit alles heeft als gevolg dat de verschillende lexeemklassen op diverse terreinen overlap vertonen.

In de nu volgende subsecties worden deze syntactische afbakeningsproblemen nader uitgediept door achtereenvolgens stil te staan bij de functionele overeenkomsten tussen de lexeemklassen (H3.4.4), de classificatie van gebonden lexemen (H3.4.5) en de classificatie van stammen (H3.4.6). Hierdoor zal duidelijk worden dat het syntactische classificatiesysteem fundamenteel tekort schiet, maar dat deze problemen eenvoudig zijn op te lossen als men overstapt op een morfologisch classificatiesysteem. Dit nieuwe systeem, dat een paradigmatisch georganiseerd lexicon veronderstelt,[128] classificeert de morfemen op basis van

[128] Deze organisatievorm is reeds door Schultink (1961) voorgesteld, maar nauwelijks uitgewerkt; zijn netwerkidee is later opgepakt door Van Marle en Koefoed (1980), die het een veel concretere vorm hebben gegeven. Ook hun voorstel is vrijwel onopgemerkt gebleven. Maar het netwerkconcept als morfologisch ordeningsprincipe is inmiddels wijdverbreid dankzij het morfologiemodel van Bybee (1985; 1988).

hun morfologische distributieklassen. Hierbij corresponderen de syntactische functies van de verschillende lexeemklassen met optionele toepassingen van een morfologische stam, d.w.z. een woordinterne, mogelijk gelede kenniseenheid die drager is van een paradigma met alle morfologische en syntactische functies die beschikbaar zijn. Dit paradigma kent een gelaagde opbouw, in de zin dat elke door de basisstam geselecteerde functor (te weten, een affix of een onhoorbare operator) weer drager kan zijn van een subparadigma van nieuwe, meer uitwendige functors.

Het hier geschetste systeem maakt het mogelijk om de syntactische functies (niet te verwarren met syntactische categorieën) een morfologische basis te geven: voor elke functie kan namelijk een aparte operator worden geïntroduceerd, waarbij elke operator drager kan zijn van een subparadigma met de aan deze functie verbonden inflectiekenmerken. Zo correspondeert een voltooid deelwoord (bijv. *verbeterd*) met een morfologische stam die onder meer in staat is om een V-functie (zonder inflectiemogelijkheden, bijv. *de theorie is verbeterd*) en een A-functie (met inflectiemogelijkheden, bijv. *de verbeterde theorie*) te vervullen. De hierin ingebedde stam (namelijk *verbeter*) geeft niet alleen toegang tot de voltooid deelwoord-toepassing, maar ook tot andere functies uit het traditionele V-paradigma (zoals de kale stam, de infinitief, het tegenwoordig deelwoord en de indicatief-modus) alsmede andere affixatie-mogelijkheden (zoals aanhechting van het affix -ING of het affx -BAAR). Dankzij deze analyse is het geen probleem meer dat er overlap bestaat in de syntactische functies van de traditionele lexeemklassen of dat een lexeemvorm meerdere functies kan vervullen.

3.4.4 De morfologische classificatie van lexemen

3.4.4.1 Introductie

In deze subsectie bespreek ik een aantal fundamentele afbakeningsproblemen met betrekking tot de traditionele, op syntactische criteria gebaseerde lexeemklassen. Voor elk afbakenings-probleem wordt een alternatieve analyse gegeven in termen van morfologische distributie-klassen; op deze wijze wordt stap voor stap een morfologisch classificatiesysteem opge-bouwd. Ik laat eerst zien dat het inflectieparadigma van de categorie V, dat een grote hoe-veelheid functies omvat, systematische overlap vertoont met de inflectievormen van andere lexeemcategorieën, namelijk N (§2) en A (§3). Vervolgens bespreek ik de functie-overlap tussen N-, A- en T-lexemen (§4). Tot slot (§5) wordt uiteengezet welke functieovereen-komsten er bestaan tussen stammen uit lexeemklassen zonder inflectieparadigma, in het bijzonder tussen B, P en C; hierbij wordt aannemelijk gemaakt dat er een overkoepelende stamklasse bestaat, namelijk de klasse van partikels.

3.4.4.2 Overeenkomsten tussen V- en N-lexemen

De infinitiefvorm van een V-lexeem is niet alleen bruikbaar voor V-toepassingen, maar ook voor N-toepassingen. Omgekeerd kunnen N-lexemen die het resultaat zijn van nominalisatie met -ING (en in enkele gevallen met -ST of -T), door middel van conversie (loop(V) > loop(N), verzet(V) > verzet(N)) of door stamaanpassing (bijv. *draag*(V) > *dracht*(N), *spreek*(V) > *spraak*(N)) semantisch gezien sterk op een V-toepassing lijken. Hieruit volgt dat het onder-scheid tussen V-lexemen en N-lexemen niet zo scherp is als men zou wensen. Het basis-probleem is dat de categorie N zowel temporele als niet-temporele stammen omvat. In mijn optiek is dit geen toeval, want de temporele N-lexemen berusten bijna altijd op een stam die (etymologisch of synchroon) ook V-toepassingen kent (al kan de klankvorm wat verschillen vertonen). Semantisch gezien vertonen dergelijke lexeemkoppels slechts één systematisch verschil: het N-lexeem dient namelijk een eigen tijdsvariabele te introduceren, terwijl het V-lexeem aan de tijd op zinsniveau kan worden gekoppeld, dus temporeel gezien in de zin is verankerd. Dit is blijkbaar zo'n belangrijk contrast dat relatief veel temporele stammen zowel een N-functie als een V-functie kunnen vervullen. Een bijkomend voordeel van de N-modus

ten opzichte van de V-modus is dat deze kan inzoomen op een deelaspect van het bijbehorende proces, zoals de handeling waarmee het proces in gang wordt gezet (*duw*, *gooi*, *draai*), het proces zelf (*draai*, *loop*, *kap*) en het resultaat van dit proces (*gooi*, *zet*, *gok*).

Het hierboven besproken betekeniscontrast tussen een V-lexeem en het bijbehorende N-correlaat is ook waarneembaar bij de infinitiefvorm; want hoewel infinitieven van oudsher als de neutrale modus van een werkwoord worden geclassificeerd (zoals de infinitief *duwen* in *het dorp zag hem een wagen duwen*) kan de infinitief ook als een N-lexeem worden gebruikt, zoals in *het duwen van de wagen kostte een kwartier*; hierbij behoudt het N-lexeem alle eigenschappen van de V-stam, maar verschilt het van deze V-stam doordat de handeling van het N-lexeem een eigen tijdsvariabele nodig heeft en dus niet in het tijdpad van de zin kan worden verankerd: *het duwen van de wagen* verwijst immers naar een gebeurtenis die vooraf gaat aan de berekening van de tijdsduur, en staat dus los van het tijdspad van het werkwoord (cf. *het duwen van de wagen trok veel bekijks*). De betekenis van de infinitiefvorm is sterk verwant aan die van N-lexemen met het suffix -ING;[129] zo is het N-lexeem *draaiing* in *de draaiing* meestal equivalent aan de constructie *het draaien*, waarin de infinitiefvorm *draaien* zichtbaar is. Hoewel er wel degelijk betekenisverschillen zijn aan te wijzen, gaat het me nu alleen om de observatie dat N-lexemen met het suffix -ING in wezen hetzelfde concept uitdrukken als de infinitiefvorm van een V-lexeem. Ook in dit geval bestaat er een systematisch contrast met betrekking tot de temporele verankering van het concept als geheel. Deze eigenschap correspondeert dus niet met de stam, maar met die van de syntactische functie, dus met de N-operator of de V-operator.

Volgens Don (1993) zijn temporele N-lexemen het resultaat van een conversie-operatie op het V-lexeem. In zijn visie kan dit worden aangetoond door na te gaan of de selectie-eigenschappen van het N-lexeem overeenstemmen met die van een geconverteerde V-stam (die altijd het lidwoord *de* en een meervoud op -EN zou kiezen): zo ja, dan zou het N-lexeem van het V-lexeem zijn afgeleid; zo nee, dan is de derivatierichting onbepaalbaar. In mijn optiek kan deze test echter alleen aantonen dat er sprake is van een morfologische operator. Hij geeft dus geen antwoord op de vraag of deze operator op een V-lexeem wordt toegepast of op een ander type eenheid, bijvoorbeeld een morfologische stam waarvan de betekenis dynamisch aspect vertoont (zoals TOCHT in *de optocht*). Die laatste optie lijkt me echter veel waarschijnlijker, want er zijn allerlei problemen met het idee dat een (syntactisch geclassificeerd) lexeem als basis kan dienen voor de afleiding van een ander lexeem.

Om te beginnen acht ik het conceptueel onzuiver om aan te nemen dat een lexeem een operatie kan ondergaan waarmee opnieuw een lexeem wordt afgeleid, aangezien de notie lexeem met een eindstadium van een morfologisch derivatieproces correspondeert. Bovendien impliceert een dergelijke analyse dat men de eigenschappen van dit basislexeem terug moet kunnen vinden in het hiervan afgeleide, dus morfologisch complexe lexeem, waaronder de klankvorm, de betekenis, de inflectiecategorie en de subcategorisatie-eigenschappen. In de praktijk blijken veel stammen echter meerdere klankvormen en betekenissen te kunnen aannemen, terwijl de oorspronkelijke inflectiecategorie en de subcategorisatie-eigenschappen volledig zijn verdwenen. In mijn morfologische classificatiemodel kunnen deze problemen effectief worden opgelost door morfologische operators als functies van stammen naar stammen te definiëren, en pas in laatste instantie lexeemstatus toe te kennen. Zo kent de stam DRAAI een derivatieparadigma met de derivatie-opties DRAAI+0 (N/V), draai+en (N/V), DRAAI+ING (N), DRAAI+ER (N), DRAAI+BAAR (A) en DRAAI+ERIG (A). Hierbij corresponderen de eerste twee opties met een ondergespecificeerd affix in de zin dat dit affix toegang geeft tot

[129]Dit geldt ook voor N-lexemen met het suffix -ATIE; dergelijke derivaties berusten op een uitheemse stam die tevens de basis vormt van een werkwoord met de uitgang -EER, bijv. de N *inspir+atie* naast de V *inspir+eer*.

twee syntactische functies, namelijk N en V. De overige derivatie-opties zijn op dit punt eenduidig. Wel geldt voor alle affixen dat ze niet automatisch met een lexeemtoepassing corresponderen, maar ook als basis kunnen dienen voor een volgend affix of voor de opbouw van een samenstelling.

3.4.4.3 Overeenkomsten tussen V- en A-lexemen

Het tegenwoordige deelwoord (TD) en het voltooide deelwoord (VD) van een V-lexeem zijn systematisch ambigu tussen een toepassing als V-lexeem en een toepassing als A-lexeem. Alleen in de laatste toepassing kunnen dergelijke lexemen inflectie ondergaan. Voorts kunnen beide deelwoordtypes in een N-lexeem worden omgezet door het suffix -E toe te voegen (wat mijns inziens niet op één lijn mag worden gesteld met de toekenning van een buigings-e); dit is overigens een standaardmogelijkheid van A-lexemen (cf. *de slimme*; *het goede*). Een en ander blijkt uit de onderstaande voorbeelden:

1. (a) Het meisje kwam zingend (TD-V) binnen
   (b) Een zingend (TD-A) meisje / Het zingende (TD-A) meisje
2. (a) Hij heeft een boek gekocht (VD-V)
   (b) Een gekocht (VD-A) boek / Het gekochte (VD-A) boek
3. (a) We vroegen ons af wie de zingende (TD-N) was
   (b) We vroegen ons af wat het gekochte (VD-N) was

Hoewel de A-toepassingen vaak een volledig transparante betekenis kennen (gegeven de betekenis van het grondwoord, namelijk de V-stam), kunnen ze ook een gelexicaliseerde (dus minder transparante) betekenis aannemen. Bij de TD-functie gaat het om lexemen als *razend*, *woedend*, *kokend* (alle drie met de betekenis "boos"), *spannend* ("meeslepend"), *innemend* ("sympathiek"), *sprekend* ("precies gelijk") en *stralend* ("vrolijk"); bij de VD-functie gaat het om lexemen als *versteld* ("verbaasd"), *bezopen* (dronken"), *aangeschoten* ("halfdronken"), *aangeslagen* ("bedrukt"), *ontdaan* ("boos"), *gebroken* ("mentaal kapot") en *gelaten* ("emotieloos"). Omgekeerd zijn er ook A-lexemen die qua vorm op een voltooid deelwoord lijken, maar waarvoor geen V-correlaat bestaat, waaronder *gelaarsd*, *gekapt*, *geharnast*, *betoeterd*, *behuisd* en *gebekt*. Voorbeelden van lexemen met TD-functie zijn *innemend* en *ontsierend*.

Binnen het traditionele classificatiesysteem kunnnen dit soort dwarsverbanden niet goed worden verklaard. Ook hier biedt mijn morfologische classificatiemodel uitkomst, want in dit model kunnen aparte distributieklassen worden gedefinieerd voor TD-lexemen, VD-lexemen en VD-achtige lexemen (zonder V-gerelateerde stam), en per klasse vastleggen welke syntactische functies er beschikbaar zijn.

3.4.4.4 Overeenkomsten tussen N-, A- en T-lexemen

Er zijn tal van N-lexemen die morfologisch gezien op een A-stam of een T-stam zijn gebaseerd. Hiernaast kennen T-lexemen ook een toepassing waarin ze veel weg hebben van een A-lexeem. Ik zal nu eerst ingaan op de A-gebaseerde nomina, om vervolgens aandacht te besteden aan de N- en A-functies van T-lexemen.

Bij sommige nomina is de vorm identiek aan die van een A-lexeem, bijv. in *de gek* (van de A *gek*) en *de dood* (van de A *dood*). Meestal is echter een buigings-*e* vereist, bijv, in *de zotte* (van *zot*), *de slimme* (van *slim*) en *de rode* (van *rood*); deze optie is overigens (zoals ik al eerder meldde) ook beschikbaar voor V-stammen in de toepassing van tegenwoordig of voltooid deelwoord. Hoewel de N-interpretatie meestal in termen van conversie wordt verantwoord, acht ik het inzichtelijker om de N-toepassing als een standaardfunctie van de genoemde stamtypes te beschouwen; in termen van distributieklassen betekent dit dat de N-toepassing tot hetzelfde subparadigma behoort als de A-toepassing of de N-toepassing.

Bij de bespreking van de T-lexemen werd gemeld dat de bepaalde telwoorden uit twee subklassen bestaan, namelijk de hoofdtelwoorden en de rangtelwoorden. De hoofdtelwoorden bleken zowel een nominale functie (bijv. *de twee*, *met z'n vieren*) als een adjectivische functie (bijv. *in vier dagen*) te kunnen vervullen, terwijl de rangtelwoorden in beginsel een adjectivische functie vervullen (bijv. *de derde gooi*), al kunnen ze (net als andere adjectieven) ook als elliptisch substituut dienen voor een nomen (bijv. *deze maat bestaat uit derden*). In mijn morfologische classificatiemodel kunnen deze observaties eenvoudig worden verantwoord door een aparte stamklasse voor bepaalde telwoorden te introduceren, en vervolgens twee gebruiksmodi te onderscheiden, namelijk de hoofdtelwoorden en de rangtwelwoorden en bij beide modi aan te geven dat ze zowel een N-functie als een A-functie kunnen vervullen.[130]

3.4.4.5 Overeenkomsten tussen B-, P- en C-lexemen

Hoewel de ANS aparte secties wijdt aan bijwoorden (B), adposities (P) en connectieven (C), vertonen deze lexeemklassen nogal wat overeenkomsten. Daarom ligt het voor de hand om een overkoepelende categorie te introduceren. De ANS spreekt in dit verband van *partikels*:

> "Formeel zijn bijwoorden een erg heterogene categorie met als enig gezamenlijk kenmerk dat ze onveranderlijk zijn. Deze eigenschap hebben ze gemeen met de voorzetsels en de voegwoorden. Dergelijke onverbuigbare en onvervoegbare woorden vat men ook wel samen onder de term partikels." (ANS, 8.1)

De hier verwoorde classificatiemogelijkheid werd reeds door Jespersen (1928) bepleit:[131]

> "I use the comprehensive term 'particles' for adverbs, prepositions, and conjunctions, as these three "parts of speech" have so much in common that they are best treated together. A preposition may be called a transitive adverb having a noun or pronoun as its object; those conjunctions which serve to introduce a subordinate clause are adverbs (prepositions) having a clause as their object; other so-called conjunctions (e.g. *and*) are simply particles used to join words or clauses."

Hieronder zal ik enkele observaties bespreken waaruit blijkt dat de bijwoorden, adposities en connectieven van het Nederlands zoveel functieoverlap vertonen dat het handiger is om ze (in overeenstemming met de voorgaande citaten) als subfuncties van een overkoepelende klasse te beschrijven, namelijk de partikelklasse. Ten eerste komen de stammen van (niet-adjectivische) bijwoorden, adposities en connectieven overeen in het feit dat ze geen inflectie vertonen (d.w.z. geen contextueel bepaalde vormverandering ondergaan, zoals een buigings-*e*). Ten tweede laten de stammen van bijwoorden, adposities en connectieven zich makkelijk samenvoegen tot complexe lexemen, zoals de bijwoorden *omhoog*, *bovenop*, *buitenlangs, eroverheen* en *voorwaarts* en de connectieven *hierdoor*, *omdat*, *opdat*, *voordat* en *daarom*. Ten derde kunnen partikelstammen vaak meerdere syntactische functies vervullen, namelijk precies de functies die hier centraal staan. Ik zal dit toelichten aan de hand van enkele voorbeelden.

In het algemeen geldt dat adpositiestammen zeer flexibele gebruiksmogelijken kennen. Zo is het partikel *voor* niet alleen bruikbaar als prepositie, waarbij men onderscheid kan maken tussen een ruimtelijke betekenis (*voor het huis*), een temporele betekenis (bijv. *voor zijn komst*), en een intentionele betekenis (bijv. *voor dit doel*), maar ook als (verkort) connectief, namelijk als temporele modificator van een bijzin (bijv. *voor hij binnenkomt*); verder kent *over* ook adverbiale toepassingen (bijv. *zijn ziekte is over, die leerling is over*), al valt moeilijk uit te maken of het hier nu een prepositie of een adverbium betreft. Bij het partikel *op* treft

[130] Dit idee wordt ondersteund door V-constructies als *vierendelen* (N-modus) en *honderuit praten* (A-modus).
[131] Dit idee kan men ook terugvinden in Jesperen's *Modern English Grammar* (1949-1958), Vol. II, 1.15.

men een soortgelijke ambiguïteit aan: naast toepassingen als adpositie van plaats (bijv. *de vaas staat op de tafel*) en richting (bijv. *hij liep de trap op*, *hij pakte zijn tas op*) zijn namelijk ook adverbiale toepassingen mogelijk (bijv. *het eten is op*, *hij belde een vriend op*). Het onderscheid tussen de adpositiefunctie en de bijwoordfunctie berust op twee factoren, namelijk of het partikel met een ruimtelijke eigenschap correspondeert (namelijk een locatie of richting) en of het partikel intransitief is (d.w.z. met een zelfstandig modificerende eenheid correspondeert) of transitief (d.w.z. naast het te modificeren predicaat ook een complementair argument vereist). Indien een partikel transitief is en een ruimtelijke relatie uitdrukt, is sprake van een adpositiefunctie (bijv. *op* in (2a)), anders van een bijwoordfunctie (bijv. *boven* in (2b)).

4. (a) Hij klom het dak op F(*op*) = R(*klimmen*, *dak*)
   (b) Hij is eindelijk boven F(*boven*) = R(*zijn*, -)

Op soortgelijke wijze kan ook een parallel worden getrokken tussen adposities en connectieven. Want in beide gevallen gaat het om partikels die syntactisch transitief zijn in de zin dat ze een relatie (kunnen) leggen tussen een (N-, V- of A-gebaseerd) predicaat en een tweede argument. Bij adposities correspondeert dit tweede argument met een nominale constituent, bij connectieven met een deelzin.

De verwantschap van partikels en bijwoorden blijkt ook uit het feit dat de modificatorpositie van samenkoppelingen zowel met adpositiestammen als met adverbiale stammen kan corresponderen (zonder dat de betekenis van de kernconstituent wezenlijk verandert), mits deze stammen qua betekenis met een richtingaanduiding corresponderen. Voor adposities heeft deze eis relatief weinig gevolgen (aangezien de meeste preposities van nature een ruimtelijke relatie uitdrukken). Voor bijwoorden daarentegen leidt het tot uitsluiting van een groot aantal subklassen; wel beschikbaar zijn partikels als *weg*, *heen*, *voort*, *terug*, *binnen* en *samen*. De onderstaande tabel geeft voor beide partikelklassen een aantal constructievoorbeelden.

partikelklasse	voorbeeldlexemen met de structuur [partikel + kern + uitgang]
A: adpositie	aan+geef+en, voor+stel+baar, voor+loop+er, over+plaats+ing
B: bijwoord	weg+geef+en, samen+stel+ing, binnen+kom+st, terug+plaats+ing

Behalve in samenkoppelingen kan men adposities en bijwoorden ook aantreffen in samenstellingen met een statisch nomen (d.w.z. een nomen waarvan de betekenis met een concept zonder temporeel pad correspondeert, zoals een persoon, object of eigenschap). In dergelijke samenstellingen gaat het echter niet om partikels van richting, maar om partikels van locatie, namelijk locatie-adposities als *voor*, *onder, tussen, bij* (bijv. in woorden als *voor+kant*, *onder+laag*, *tussen+deur*, *bij+keuken*) en locatie-bijwoorden als *binnen*, *buiten*, *boven* (bijv. in woorden als *binnen+deur*, *buiten+kant*, *boven+kamer*).

Deze observaties tezamen wijzen erop dat bijwoorden, adposities en connectieven het beste als subfuncties van een overkoepelende stamklasse kunnen worden geanalyseerd, namelijk de klasse van partikels. Deze analyse heeft als voordeel dat hij een rechtstreekse verklaring biedt voor de observatie dat de semantische eigenschappen van een partikel veel invloed hebben op de syntactische en semantische functies van de hiermee geconstrueerde stam.[132] In het traditionele classificatiemodel daarentegen is men gedwongen om per stam vast te leggen in welke lexeemklassen (en eventuele subklassen) deze kan worden aangetroffen, zonder dat wordt verklaard waarom de stam juist deze lexeemklassen kiest of waarom er zoveel overeenkomsten bestaan tussen deze lexeemklassen.

[132] Dit komt ook duidelijk naar voren uit Blom's gedetailleerde studie van partikelwerkwoorden (Blom, 2005).

Beard (1995) gaat nog een stap verder door te betogen dat elk van de lexeemklassen die Jespersen onder de koepelterm *adverbium* schaarde, in feite met functionele subklassen van het adjectief corresponderen, met uitzondering van een kleine subset van adposities, namelijk alle adposities die met een casusmarkering corresponderen. In Beard's universeel bedoelde theorie corresponderen traditionele woordklassen als bijwoorden, partikels en adposities met voorspelbare inperkingen op de adjectivale gebruiksmogelijkheden. Het betreft echter niet meer dan een globale aanduiding, want elk van deze woordklassen kan nog verder in subklassen worden onderverdeeld; zo valt de klasse van bijwoorden uiteen in deadjectivale, denominale en deverbale bijwoorden, die elk gespecialiseerde functies vervullen. Hierbij hangt het van de taal af of dergelijke subklassen overt of covert worden gecodeerd. Dit voorstel verdient zeker nadere bestudering. In aanvulling hierop zou ook moeten worden nagegaan in hoeverre determiners en telwoorden zich als adjectieven of bijwoorden gedragen.

3.4.5 De morfologische classificatie van gebonden lexemen (woorddelen)

In de voorgaande sectie heb ik alleen aandacht besteed aan identificatieproblemen bij de categoriale classificatie van zelfstandige lexeemvormen c.q. woorden. Maar veel lexemen kunnen ook als constituent in een samenstelling fungeren. In de morfologische theorievorming gaat men er meestal vanuit dat deze gebonden lexemen dezelfde categorie bezitten als hun vrije varianten (cf. MHB, ANS, Booij & Van Santen (1998)). Verder neemt men aan dat de categorie van de meest rechtse constituent normaal gesproken bepalend is voor de categorie (en dus de inflectiekenmerken) van de samenstelling als geheel; deze generalisatie staat bekend als de rechterhoofdregel voor samenstellingen (zie ook H3.6). Voor de overige constituent(en) wordt aangenomen dat deze de functie van modificator hebben. Zo bestaat de samenstelling *dorpsgezicht* uit de constituenten *dorps* en *gezicht*, waarbij *dorps* als een toepassing van het lexeem *dorp* geldt, en *gezicht* als een toepassing van het lexeem *gezicht*. Dit laatste lexeem draagt categorie N (met een meervoudsvorm op -EN), wat impliceert dat de samenstelling als geheel ook categorie N bezit (met dezelfde meervoudsvorm).

Ten aanzien van de categorie van de linkerconstituent stelt de standaardtheorie dat deze identiek is aan de categorie van het onderliggende lexeem, in dit geval het N-lexeem *dorp*. Dit kan echter niet rechtstreeks worden geverifieerd, want de categorie van een modificator heeft geen zichtbaar effect op zijn distributiemogelijkheden of zijn inflectiegedrag (al wordt soms een bindmorfeem geselecteerd, zoals de *-s* in *dorps*). Indien de linkerconstituent met een lexeemvorm correspondeert die meerdere categorieën kan aannemen (zoals een V-lexeem dat N-conversie kan ondergaan), is het in een deel van de gevallen moeilijk uit te maken welke lexeemcategorie de voorkeur verdient; want hoewel conversie soms samengaat met betekenisaanpassing, is meestal sprake van dezelfde grondbetekenis. Zo kan de betekenis van *feestdag* op twee manieren worden geparafraseerd, namelijk als een dag met een feest (op basis van de N-betekenis van *feest*) en als een dag waarop gefeest wordt (op basis van de V-betekenis van *feest*); op dezelfde manier kan *werkdag* worden geïnterpreteerd als een dag waarop men naar het werk gaat, of een dag waarop men werkt. Terwijl in deze voorbeelden nog wel een nuanceverschil waarneembaar is tussen de V-betekenis (waarbij het accent ligt op de activiteit) en de N-betekenis (waarbij het accent ligt op de afbakening van deze activiteit in ruimte en/of tijd), is dit contrast nagenoeg onzichtbaar bij een conversiepaar als *lopen*(V) ↔ *loop*(N), zodat de categorie van de constituent *loop* in *looppad* niet langs semantische weg is vast te stellen. Ook als dit wel kan, is het feit dat veel linkerconstituenten semantisch ambigu zijn een probleem voor het uitgangspunt dat linkerconstituenten normale lexemen zijn.

Een tweede complicatie voor de categoriale identificatie van de samenstellende constituenten is dat sommige gelede constituenten uitsluitend bruikbaar zijn binnen een samenstelling. Hierbij gaat het meestal om rechterconstituenten, waarbij het complement van zo'n niet-zelfstan-

dig lexeem (namelijk de linkerconstituent) meestal vrij gevarieerd kan worden. Het gebruik van een niet-zelfstandig bruikbaar lexeem heeft als gevolg dat de betreffende constituent niet langs lexicale weg van een syntactische categorie kan worden voorzien, maar dat deze categorie uit de eigenschappen van de samenstelling als geheel moet worden afgeleid. Hierdoor is niet zeker of deze eigenschappen rechtstreeks aan de categorie van de linkerconstituent zijn toe te schrijven, aan de uitgang van deze constituent of aan de constructie als geheel.

Indien de rechterconstituent met een niet-zelfstandig lexeem correspondeert, gaat het meestal om gelede lexemen, bijvoorbeeld lexemen met de uitgang -IG, -ER, -ING of -JE. Doordat dit algemeen bruikbare suffixen zijn, kan een deel van de gebonden lexemen qua vorm ook zelfstandig voorkomen, maar dit geldt niet voor de bijbehorende betekenis. Voorbeelden van samenstellingen met een gebonden lexeem op -IG zijn *geleedpotig*, *vierhandig* en *armlastig*. Voorbeelden van samenstellingen met een gebonden lexeem op -ER zijn *driewieler*, *dubbeldekker*, *viermaster* of (uit een fundamenteel andere categorie) *waterkoker* en *schoenlapper*. Voorbeelden met de uitgang -JE zijn *vierkantje*, *theekransje* en *breiwerkje*. Voorbeelden met de uitgang -ING zijn *dankzegging*, *boekhouding* en *kennismaking*.

Indien sprake is van een niet-zelfstandig lexeem in de linkerconstituent (zoals *schoor* in *schoorsteen*), gaat het meestal om een incidenteel geval, en is dus geen systematische variatie van de rechterconstituent mogelijk, op één uitzondering na, namelijk linkerconstituenten met een kale V-stam. Want kale V-stammen zijn in beginsel niet zelfstandig bruikbaar, maar kunnen zeer systematisch als modificator van een samenstelling worden ingezet, onafhankelijk van de vraag of ze een nominale conversievorm bezitten. Enkele voorbeelden zijn *schaatswedstrijd* (met de V-stam *schaats*), *aanlegsteiger* (met de V-stam *aanleg*) en *uitklapbed* (met de V-stam *uitklap*). Los van de hier besproken uitzondering corresponderen linkerconstituenten relatief weinig met onzelfstandige lexemen (waarbij ik voor het moment voorbij ga aan bindmorfemen, stamallomorfie en morfologische substitutie; zie verderop). Enkele voorbeelden zijn *vliegensvlug* (*vliegens), *huishoudboekje* (tenzij het linkerdeel met de stam van het V-lexeem *huishouden* zou corresponderen) en *roerbakgroente* (want er bestaat geen lexeem **roerbak*, alleen een infinitief *roerbakken*).

Nauw verwant aan het voorkomen van gebonden stammen is het fenomeen dat zowel linker- als rechterconstituenten soms met een allomorfische vorm van de vrije stam corresponderen; deze allomorfen worden verder alleen in derivaties aangetroffen en kennen vaak meerdere interpretatiemogelijkheden. Zo bezitten nomina met een meervoud op *-eren* (bijv. *kind-kinderen* en *rad-raderen*) een allomorfstam op *-er* die ook als linkerconstituent voorkomt, bijv. *rader* in *raderwerk* en *kinder* in *kinderboek*. Voorts corresponderen linkerconstituenten die uitgaan van een éénlettergrepig werkwoord met een infinitief op *-n* (waarvan er slechts zes zijn, te weten *gaan*, *staan*, *slaan*, *doen*, *zien*, *zijn*) vaak met de stamallomorf van de nominale vorm (zoals *slag* bij *slaan*, *zicht* bij *zien*, *gang* bij *gaan*, *daad* bij *doen*), al dan niet in geprefigeerde vorm, terwijl de V-betekenis toch beschikbaar blijft (of zelfs primair is). Dit is zichtbaar in voorbeelden als *slagkracht*, *aanslagbiljet*, *zichtas*, *uitzichtpunt*, *ingangscontrole*, *voortgangsgesprek* en *daadkracht*. Bij inheemse werkwoorden komt ook een meer systematisch substitutiepatroon voor, blijkens het contrast tussen *draagvermogen* (vs. **dragingsvermogen*) en *belastingsvermogen* (vs. **belastvermogen*). Dit substitutiefenomeen is in een wat andere vorm ook zichtbaar bij uitheemse werkwoorden, namelijk de verdringing van de V-stam door een vorm op *-ie*, zoals in *communicatielijn* (i.p.v. *communiceerlijn*), *constructiemethode* (i.p.v. *construeermethode*) en *delegatievermogen* (i.p.v. *delegeervermogen*).

Uit de hier besproken fenomenen blijkt dat er tal van samenstellingen zijn waarbij moeilijk valt te bepalen wat de syntactische categorie is van de samenstellende constituenten, vooral bij de linkerconstituent. In mijn optiek zijn deze problemen zo ernstig dat het de vraag is of de constituenten van een samenstelling wel op dezelfde manier geclassificeerd kunnen worden

als vrije lexemen. Vooral het feit dat linkerconstituenten doorgaans een meerduidige interpretatie hebben, wijst er op dat er een fundamenteel verschil bestaat tussen concrete lexeemvormen en de constituenten van een samenstelling.

Binnen mijn morfologische classificatiemodel zijn deze problemen eenvoudig op te lossen: hiertoe dienen vrije lexemen en lexeeminterne constituenten als verschillende verschijningsvormen van een morfologische stam te worden gedefinieerd, wat impliceert dat stammen met een eenheid corresponderen die nog niet gespecificeerd is met betrekking tot zijn syntactische en semantische functies. Hierdoor zijn deze eenheden niet alleen geschikt voor de derivatie van zelfstandige lexemen, maar ook voor de constructie van lexeeminterne constituenten. Door vrije en gebonden eenheden op dezelfde stam te baseren, ontstaat een basis voor de verklaring van hun gemeenschappelijke eigenschappen.

3.4.6 De morfologische classificatie van gebonden stammen (wortels)

In het syntactische categorieën classificatiesysteem is geen plaats voor gebonden stammen, d.w.z. morfologische stammen die niet met een zelfstandig lexeem corresponderen en die meestal ook niet aan de eisen van een fonologisch (c.q. potentieel) woord voldoen. Het MHB duidt dergelijke stammen als wortels aan (met categorie X). De introductie van wortels wordt gemotiveerd door de overweging dat hun lexeeminterne complement qua vorm en functie met een herkenbaar affix correspondeert. Hierdoor kan worden verantwoord dat de eigenschappen van werkwoorden als *bedriegen*, *beginnen*, *begeren* en *bemoeien* deels zijn terug te voeren op het V-vormende prefix BE- (met als complement een X-stam uit het rijtje √DRIEG, √GIN, √GEER en √MOEI) en dat de eigenschappen van adjectieven als *deftig*, *zuinig* en *slordig* deels zijn terug te voeren op het A-vormende suffix -IG (met als complement een X-stam uit het rijtje √DEFT, √ZUIN en √SLORD). Het MHB spreekt in dit verband van formeel gelede lexemen. Bij inheemse affixen betreft het meestal versteende afleidingen waarvan de stam niet langer als lexeem in gebruik is. Uitheemse afleidingen daarentegen corresponderen bijna altijd met een gebonden stam c.q. X-stam. Dit blijkt bijvoorbeeld bij inspectie van het derivationele lexeemparadigma van de X-stam √DUC(T),[133] die bij al deze lexemen vooraf wordt gegaan door een uitheems prefix; dit kan formeel worden verantwoord door een gegeneraliseerde stam (c.q. g-stam) te definiëren, namelijk [<P>+DUC(T)], waarbij P met een stamspecifieke selectie van uitheemse prefixen correspondeert. Tabel 3-3 toont alle mogelijke P-specificaties van deze g-stam (in de verticale kolom) en geeft voor elke P-specificatie aan welke suffixderivaties er beschikbaar zijn (beperkt tot de zes meest voorkomende suffixen).

Uit deze inventarisatie blijkt dat de g-stam [<P>+DUC(T)] voor de meeste uitheemse prefixen P in staat is om een combinatie aan te gaan met de suffixen -EER, -IE, een agentief suffix (namelijk -ENT of -OR, die een complementaire distributie vertonen) en het adjectiverende suffix -IEF (dat qua semantiek verwant is met het inheemse suffix -END), maar dat deze slechts sporadisch als zelfstandig lexeem voorkomt, namelijk als [<P>+DUCT] (een optie die met de klankloze N-operator 0_N correspondeert).[134] Afgezien van deze kleine subklasse kent de g-stam [<P>+DUC(T)] dus geen toepassing als zelfstandig lexeem, zodat er ook geen syntactische categorie aan kan worden toegekend. Toch lijkt deze g-stam qua morfologisch gedrag veel op inheemse prefix-stam-combinaties met V-toepassing, zoals [<P_1>+LEID], waarbij de psoitie P_1 met een V-vormende operator (c.q. V-functor) correspondeert. Deze operator correspondeert ofwel met een gebonden prefix, zoals BE-, VER-, ONT- of GE- (zoals in de stammen 0/GE+LEID, VER+LEID en GE+LEID), of een onbeklemtoond partikel (maar niet bij

[133] De notatie "duct(t) geeft aan dat deze X-stam met de stamvormen *duc* en *duct* kan corresponderen.
[134] Ook het lexeem *viaduct* kan tot deze klasse worden gerekend (evenals de varianten *ecoduct* en *aquaduct*).

deze wortel), ofwel met de neutrale V-operator [0/GE], zoals in de V-stam [0/GE]+LEID in het werkwoord *leiden*; deze operator wordt aalleen zichtbaar in de voltooide tijd.

stam	**stam+*eer***	**stam+*ie***	**stam+*ent***	**stam+*or***	**stam+*ief***	**stam+0_N**
ab+duc(t)	abduc+eer	abduct+ie	-	abduct+or	abduct+ief	-
ad+duc(t)	adduc+eer	adduct+ie	-	adduct+or	-	ad+duct
con+duc(t)	-	conduct+ie	-	conduct+or	conduct+ief	con+duct
de+duc(t)	deduc+eer	deduct+ie	-	-	deduct+ief	-
in+duc(t)	induc+eer	induct+ie	-	induct+or	induct+ief	-
intro+duc(t)	introduc+eer	introduct+ie	introduc+ent	introduct+or	introduct+ief	-
ob+duc(t)	obduc+eer	obduct+ie	obduc+ent	-	-	-
pro+duc(t)	produc+eer	product+ie	produc+ent	-	product+ief	pro+duct
re+duc(t)	reduc+eer	reduct+ie	-	reduct+or	reduct+ief	-
se+duc(t)	-	seduct+ie	-	-	-	-
trans+duc(t)	-	transduct+ie	transduc+ent	-	-	-
11	8	11	4	6	7	3

Tabel 3-3: Lexeemderivaties van de X-stam √DUC(T); de tabel toont voor elke g-stam met de structuur [P+DUC(T)] *welke suffixcombinaties er bestaan (op basis van de GWNT).*

De neutrale V-stam [0/GE+LEID] (verder aan te duiden als LEID') kan zelf weer benut worden voor een volgende constructiestap, namelijk de combinatie van een partikel met de complexe stam LEID' op basis van het schema V' = [<P_2>+LEID']. Dit leidt tot complexe V-stammen als [AAN+LEID'], [IN+LEID'], [OP+LEID'], [UIT+LEID'] en de neutrale optie [0_p+LEID'], met de coverte functor 0_P. Hierbij kunnen de (zwakke of sterke) inflectie-eigenschappen van de oorspronkelijke stam integraal worden overgeërfd. Door deze analysewijze kan verklaard worden waarom V'-lexemen met een overt partikel dezelfde inflectie-eigenschappen bezitten als die van de V'-lexemen met een (onhoorbaar) 0-partikel: deze eigenschappen zijn gewoon op het niveau van de stam gecodeerd. Als deze stam een sterke vervoeging kent, geldt dit ook voor alle V'-toepassingen; en als deze stam een overt prefix omvat en daarom geen voltooide tijd met GE- toestaat, geldt deze eigenschap ook voor de hiermee afgeleide V'-lexemen. Een bijkomend voordeel is dat de stamvormen niet alleen beschikbaar zijn voor een toepassing als V-vorm, maar ook voor andere derivatieklassen (zoals V-gerelateerde nomina en adjectieven).

Tabel 3-4 biedt een op de GWNT gebaseerde inventarisatie van bestaande lexeemderivaties; dit hoeft overigens niet te betekenen dat de overige derivaties onmogelijk zijn en ook niet dat deze nooit voorkomen. De weergegeven affix-inventarisatie loopt grotendeels parallel aan die in tabel 3-3; zo correspondeert het V-vormende -EER met het infinitiefsuffix -EN, het procesvormende N-suffix -IE met -ING, de agentieve suffixen met -ER en het A-suffix -IEF met -END. De optie 0_N daarentegen heb ik weggelaten (aangezien deze niet toepasbaar is op de g-stam [<P>+LEID]); in plaats daarvan heb ik het patiens-specificerende N-suffix -E opgenomen (dat uitgaat van een voltooid deelwoord); verder heb ik een extra A-suffix opgenomen, namelijk het potentie-aanduidende suffix -BAAR.

Vergelijking van deze tabellen leert dat de inheemse g-stam [<P_2>+LEID'] dezelfde hoofdfuncties kan vervullen als de uitheemse g-stam [<P>+DUC], maar dat bij de inheemse g-stam geen apart V-modus-suffix nodig is (zoals het uitheemse suffix -EER, waarvan de functie m.i. vergelijkbaar is met de P_1-functor in [<P_1>+LEID]) om een V-lexeem te construeren. Dit contrast lijkt er de oorzaak van te zijn dat de inheemse morfologie doorgaans in termen van syntagmatische c.q. lexeemgebaseerde regels wordt gedefinieerd, terwijl de uitheemse morfologie in termen van paradigmatische substitutieregels wordt beschreven (dus met een stamgebaseerd regelsysteem). Maar mijn analyse wijst uit dat de inheemse morfologie net zo paradigmatisch georganiseerd is als de uitheemse morfologie. Deze analyse heeft als bij-

komend voordeel dat alle stammen de status van X-stam krijgen, dus dat het onderscheid tussen vrije stammen en gebonden stammen (c.q. X-stammen) overbodig wordt.

stam	**stam+*en***	**stam+*ing***	**stam+*er***	**stam+*baar***	**stam+*end***	***ge*+stam+*e***
[0/ge]+leid	leiden	leiding	leider	leidbaar	leidend	-
ver+leid	verleiden	verleiding	verleider	verleidbaar	verleidend	-
ge+leid	geleiden	geleiding	geleider	geleidbaar	geleidend	geleide
be+ge+leid	begeleiden	begeleiding	begeleider	begeleidbaar	begeleidend	-
her+leid	herleiden	herleiding	-	herleidbaar	-	-
aan+leid'	aanleiden	aanleiding	-	-	aanleidend	-
af+leid'	afleiden	afleiding	afleider	afleidbaar	afleidend	afgeleide
in+leid'	inleiden	inleiding	inleider	-	inleidend	-
om+leid'	omleiden	omleiding	-	-	-	-
op+leid'	opleiden	opleiding	opleider	-	-	-
over+leid'	overleiden	-	-	-	-	-
rond+leid'	rondleiden	rondleiding	rondleider	-	-	-
voor+leid'	voorleiden	voorleiding	-	-	-	-
uit+leid'	uitleiden	uitleiding	-	-	-	uitgeleide
14	14	13	8	6	7	3

Tabel 3-4: Lexeemderivaties van de vrije stam LEID*; de tabel toont voor elke g-stam met de structuur* [P+LEID] *welke suffixcombinaties er bestaan (op basis van de GWNT).*

De hier geformuleerde conclusie strookt goed met mijn eerdere observaties met betrekking tot de classificatie van lexemen. Want uit mijn analyse van het syntactische classificatiesysteem bleek dat de bijbehorende categorieën slecht zijn gefundeerd, waardoor de onderscheiden lexeemklassen veel functie-overlap vertonen.

Zoals ik al eerder uitlegde berust mijn alternatieve classificatiesysteem op het idee dat morfologische stammen in termen van morfologische distributieklassen kunnen worden getypeerd en dat syntactische functies geen stamdefiniërende status hebben: in mijn model zijn het niet meer dan mogelijke toepassingen van een gegeven stamvorm. In deze benadering hebben affixen (en onzichtbare operatoren) de functie om nadere informatie te geven over de distributieklasse van de stam, met als gevolg dat de stam een hoger complexiteitsniveau bereikt. Deze distributieklassen geven primair informatie over de morfologische derivatiemogelijkheden van een gegeven stamniveau, maar ze kunnen ook toegang geven tot een of meer syntactische functies. Zo kan het suffix -EER worden benut om een willekeurige wortel in een stam uit de V-klasse om te zetten (als drager van V-gerelateerde functies) , maar men kan de met -EER afgeleide stam ook als basis nemen voor volgende derivatiestappen, zoals de aanhechting van -ING of -BAAR. Het hier beschreven classificatiesysteem lijkt dan ook goed verenigbaar met een paradigmatische benadering van woordrelaties.

3.4.7 Conclusie

In de voorgaande secties is aangetoond dat een op syntactische categorieën gebaseerd classificatiesysteem ontoereikend is voor de verantwoording van morfologisch gelede lexemen. Dit systeem is namelijk erg star en kent veel functionele overlap tussen de traditioneel onderscheiden lexeemklassen. Daarom heb ik een alternatief classificatiesysteem uitgewerkt, namelijk een classificatiesysteem op morfologische grondslag. In dit classificatiesysteem kan voor elke morfologische stam een distributieparadigma worden gespecificeerd met gedetailleerde informatie over de morfologische derivatiemogelijkheden alsmede over de beschikbare syntactische functies (c.q. lexeemklassen) en het hieraan gekoppelde inflectiepatroon. Dit analysemodel lijkt een goede basis te bieden voor de verantwoording van de paradigmatische

samenhang binnen inheemse en uitheemse derivatieparadigma's. In hoofdstuk 4 wordt dit morfologische classificatiesysteem formeel uitgewerkt en op concrete voorbeelden toegepast.

3.5 Lexicale structuurrelaties

3.5.1 Introductie

In het morfologiemodel van het MHB (zie H3.2) geeft het lexicon uitsluitend informatie over lexemen en affixen, d.w.z. over niet verder analyseerbare morfemen. In dit model biedt het lexicon geen plaats voor de specificatie van gelexicaliseerde relaties tussen morfemen, of het nu morfologisch gelede lexemen betreft of vaste affixcombinaties. Dit heeft als gevolg dat niet kan worden vastgelegd dat er relatief veel woorden bestaan waarin het A-vormende suffix -IEF door het V-vormende suffix -EER wordt gevolgd, zoals in het werkwoord *intensiveren* (= $[\text{INTENS}]_A$+IEF_A+EER_V). Dit probleem zou men kunnen omzeilen door een *synaffix*[135] te postuleren, namelijk, -IVEER_V (= -IV + -EER_V), maar deze analyse is alleen wenselijk indien de bijbehorende affixsequentie een idiosyncratische betekenis heeft ontwikkeld. In de komende subsecties leg ik uit waarom de combinatorische mogelijkheden van Nederlandse morfemen niet goed beschreven kunnen worden indien geen gebruik mag worden gemaakt van lexicaal vastgelegde morfeemrelaties. Hierbij ga ik achtereenvolgens in op allomorfie, affixpotentiatie en paradigmatische samenhang.

3.5.2 Allomorfie

3.5.2.1 Stamallomorfie

Stamallomorfie kan worden gedefinieerd als morfologisch gemotiveerde variatie in de verschijningsvorm van een lexeem. Dit fenomeen dient goed onderscheiden te worden van fonologisch gemotiveerde stamvariatie, die wordt aangeduid als stamallofonie. Zo correspondeert de ie/o-alternantie van de stam SCHIET in het lexeempaar *schieten-schot* met allomorfie, maar is de s/z-alternantie van de stam HUIS in het lexeempaar *huis-huizen* een vorm van allofonie. Het verschil is dat er geen morfofonologische context kan worden gedefinieerd waarin altijd ie/o-alternantie is vereist, terwijl dit wel mogelijk lijkt voor de s/z-alternantie. Toch zijn er vele pogingen ondernomen om stamallomorfie als een morfofonologisch verschijnsel te behandelen.[136] Het bestaan van stamallomorfie is namelijk fundamenteel strijdig met het syntagmatische uitgangspunt dat grammaticale bouwstenen een contextonafhankelijke typering kennen. Als een stam vormalternanties vertoont die niet voorspelbaar zijn uit de fonologische context, kan deze variatie alleen verantwoord worden door per lexeem vast te leggen wat de beschikbare stamvormen zijn en door zonodig de bijbehorende morfeemcontexten te specificeren. Deze laatste benadering ligt ten grondslag aan de autonome morfologietheorie van Booij (1997). De door Booij bijeengebrachte structuurobservaties m.b.t. allomorfische variatie bieden mijns inziens sterke evidentie voor het bestaan van lexicale morfeemrelaties (d.w.z. lexicaal vastgelegde relaties tussen concrete morfemen) en daarom ook voor een netwerkgebaseerd lexiconmodel. Ik zal dit toelichten aan de hand van drie concrete problemen voor een analysemodel zonder lexicale morfeemrelaties.

Probleem 1: Er zijn veel inheemse nomina waarvan de pluralis-stam (pl-stam) om historische redenen een andere vorm bezit dan de singularis-stam (sg-stam). Zo zijn er veel nomina waarvan de pl-stam klinkerverlenging of klinkerverandering vertoont ten opzichte van de sg-stam, blijkens sg/pl-paren als *p[a]d* / *p[aa]d+en* en *st[a]d* / *st[ee]d+en*.[137] Deze alternanties zijn

[135] Een synaffix is een formeel geleed, maar semantisch ongeleed affix; zie Booij (2002).
[136] Deze analyserichting gaat terug op Chomsky & Halle (1968) en staat bekend als lexicale morfologie.
[137] Ten behoeve van de leesbaarheid vermijd ik de fonetische notatiewijze; ik geef er de voorkeur aan om klankalternanties via de normale spellingconventies weer te geven, waarbij ik deze klankweergaves steeds tussen vierkante haken zal zetten. Onderstreepte klinkers markeren de hoofdklemtoon.

etymologisch verklaarbaar uit Prokosch' Law (die stelt dat er vocaalrekking optreedt in een beklemtoonde open lettergreep) of ablaut (een fenomeen dat zich beperkt tot de sterke vervoeging van werkwoorden). Maar in het hedendaagse Nederlands zijn deze fonologische klankwetten niet meer van kracht. Om de hedendaagse vormverbanden toch te kunnen verantwoorden, stelt Booij (2002) een theorie voor waarin het lexicon meerdere stamvormen per lexeem kan opslaan. Deze theorie wordt ondersteund door het feit dat deze pl-vormen de basis kunnen vormen voor affix-aanhechting, blijkens de voorbeelden in onderstaande tabel:

N(SG)	N(PL)	N(STAM)	N(DIM)
sch[i]p	sch[ee]p+en	sch[ee]p+vaart	sch[ee]p+je
sm[i]d	sm[ee]d	sm[ee]d+erij	sm[i]d+je
l[o]t	l[oo]t+en	l[oo]t+erij	l[o/oo]t+je
gr[a]f	gr[aa]v+en	be+gr[aa]f]+enis	gr[a]f+je
p[a]d	p[aa]d+en	p[aa]d+je	p[aa]d+je
st[a]d	st[ee]d+en	st[ee]d+elijk	st[a]d+je

Een andere categorie van klinkeralternanties betreft nomina waarvan het pl-suffix de vorm *eren* lijkt aan te nemen, zoals blijkt uit het sg/pl-paar *kind-kinderen*. Volgens Booij is echter sprake van de structuur [pl-stam + *en*], waarbij de pl-stam zich kenmerkt door de vorm [sg-stam + *er*]. Dit volgt uit het feit dat de pl-stam ook vaak de basis vormt voor de aanhechting van derivationele affixen, blijkens *kinderlijk*, *kindertjes*, *kinderachtig* en *kinderloos*, of voor de constructie van samenstellingen, bijv. *kinderwagen* en *kinderkamer*. Deze klasse van nomina telt niet meer dan 15 lexeemstammen.

Probleem 2: Uitheemse N-stammen met de uitgang *or* of *on*, waaronder *demon*, *elektron*, *motor* en *doctor*, kennen doorgaans twee verschillende stamvormen, namelijk een vrije stamvorm, met de korte klinker *[o]* en staminitiële klemtoon (bijv. *dem[o]n*), en een gebonden stamvorm, met de lange klinker *[oo]* en stamfinale klemtoon (bijv. *dem[oo]n*). Deze stamvormen blijken verschillend derivatiegedrag te vertonen. Zo selecteert de vrije stamvorm een meervoud op *s* (bijv. *dem[o]n + s*), maar de gebonden stamvorm een meervoud op *en* (bijv. *dem[oo]n + en*). Vanuit MHB-perspectief is het opmerkelijk dat een nomen als *demon* een meervoudsvorm op *en* accepteert, want indien men aanneemt dat de pl-vorm het resultaat is van suffix-aanhechting aan de sg-stam, kan de stamfinale klemtoon van de pl-vorm op *en* alleen verklaard worden indien men aanneemt dat dit suffix soms klemtoonverschuiving teweeg kan brengen, dus dat het klemtoongedrag van *en* contextafhankelijk is.

Volgens Booij (1997) kan deze vorm van stamalternantie beter langs lexicale weg worden verantwoord, namelijk door beide stamvormen in het lexicon op te slaan en per gebruikscontext (d.w.z. per aan te hechten affix) te bepalen welke stamvorm het meest geschikt is. Maar indien er geen eenduidige keuze mogelijk is, moet men de betreffende morfeemcombinatie langs lexicale weg verantwoorden. Voor de vrije stam geldt dat hij met de sg-vorm van het nomen correspondeert, dat hij een pl-vorm op *s* selecteert en dat hij kan opduiken in samenstellingen (bijv. *dem[o]nmasker*) en in afleidingen met inheemse suffixen (bijv. *dem[o]n+achtig*). Voor de gebonden stam geldt dat hij alleen in combinatie met een suffix mag worden gebruikt, namelijk de inheemse pl-vorm *en* (bijv. *dem[oo]n+en*), of met een uitheems suffix (bijv. *dem[oo]n+isch*, *dem[oo]n+iseren* en *dem[oo]n+ie*).

Booij lijkt ervan uit te gaan dat bijna al deze voorkeuren voorspelbaar zijn uit de selectierestricties van de aangehechte affixen. Uitheemse suffixen hebben immers een duidelijke voorkeur voor gebonden, klemtoonfinale stammen, terwijl inheemse suffixen een voorkeur hebben voor vrije, klemtooninitiële stammen. Volgens Booij is er slechts één suffix dat langs lexicale weg wordt geselecteerd, namelijk het pl-suffix *-en*, want dit suffix zou normaliter een inheemse stam vereisen. Er zijn echter meer suffixen waarvoor de stamvorm gelexicaliseerd

lijkt; ander zou de "gebonden" stamvorm (bijv. *dem[oo]n*) een goede kandidaat zijn voor de singularis, en zou de "vrije" stamvorm (bijv. *dem[o]n*) "uitheems" genoeg zijn om in aanmerking te komen voor uitheemse afleidingen, te meer omdat er ook uitheemse derivaties zijn die een stamvorm zonder klemtoon vereisen (bijv. *demoniseren* en *demonie*). Het lijkt me daarom aannemelijker dat het lexicon alle bestaande morfeemcombinaties vastlegt, ook indien er geen sprake is van stamallomorfie. Gegeven dit uitgangspunt zou de lexicale ingang van het lexeem *demon* er uit kunnen zien als in (5):

(5) lexeem: *demon*
betekenis: ...
synt. categorie: N
stam 1: *dem[o]n* (spelling, uitspraak etc.)
inflectie-affixen: N-sg = [0]; N-pl = *-s*
derivatie-affixen: *-achtig*, *-schap*
stam 2: *dem[oo]n* (spelling, uitspraak etc.)
stam 2a: *dem[oo]n*
inflectie-affixen: --
derivatie-affixen: *-iseer*, *-ie*, *-isme*, *-ologie*, ...
stam 2b: *dem[oo]n*
inflectie-affixen: N-pl = *-en*
derivatie-affixen: *-isch*, *-tje*

Dit representatieschema geeft aan dat het lexeem DEMON (met nader te specificeren betekenis) de syntactische categorie N bezit, en dat dit lexeem twee verschillende stamvormen kent, namelijk *dem[o]n* en *dem[oo]n*, waarbij de laatste stamvorm twee subvarianten kent, te weten een subvariant (2a) met (bij)klemtoon op de eerste syllabe (*dem[oo]n*) en een subvariant (2b) met klemtoon op de tweede syllabe (*dem[oo]n*). Tot slot wordt voor elke stamvariant aangegeven wat de bijbehorende derivatiemogelijkheden zijn (door opsomming van de affixen).

Probleem 3: Het Nederlands kent vijf V-lexemen waarvan de infinitiefvorm niet met de structuur [stam + *en*] maar met de structuur [stam + *n*] correspondeert. Voor deze V-lexemen geldt dat de stamvorm van de bijbehorende N-lexemen een onvoorspelbare klankvorm bezit. Het gaat om de volgende woordkoppels (V/N): *doen* / *daad*, *gaan* / *gang*, *slaan* / *slag*, *staan* / *stand*, *zien* / *zicht*. Volgens Booij (1997) kenmerkt deze klasse van V-lexemen zich door de eigenschap dat de stamvorm van de hierop gebaseerde derivaties niet met de V-stam, maar hetzij met de infinitiefvorm (te weten /doen/, /gaan/, /slaan/, /staan/ en /zien/), hetzij met de nominalisatievorm (te weten /daad/, /gang/, /slag/, /stand/ en /zicht/) correspondeert. Zo correspondeert de eerste stam met derivaties als *aandoening* en *voorziening*, en de tweede stam met derivaties als *gangbaar* en *zichtbaar*, ook al lijken deze lexemen semantisch gezien een afleiding van het V-lexeem. Hieruit volgt dat de keuze van de stamvorm via lexicale morfeemrelaties moet worden verantwoord. Booij spreekt in dit verband van *paradigmatisch bepaalde allomorfie*. Dit type allomorfie komt algemeen voor. Voor veel V/N-paren die stamallomorfie vertonen geldt dat deze stamallomorfie behouden blijft als het betreffende stampaar met een partikel (P) wordt gecombineerd. Dit blijkt bijvoorbeeld uit de reeks *gaan* / *gang*, *afgaan* / *afgang*, *doorgaan* / *doorgang*, *ingaan* / *ingang*, *overgaan* / *overgang, uitgaan* / *uitgang* etc. Booij (1997) spreekt in dit verband van een paradigmatisch constructieschema; zo'n constructieschema kenmerkt zich door de volgende logica:

(6) (lexeem X : lexeem Y) = (prefix P + lexeem X : prefix P + lexeem Y).

Indien lexeem X bijvoorbeeld met *gaan* correspondeert en lexeem Y met *gang*, en indien er ook een lexeem bestaat met de structuur prefix + X (bijv. *uitgaan*), dan kan ditzelfde prefix (te weten *uit*) ook met Y worden gecombineerd (namelijk *uitgang*). Hoewel zulke constructie-

schema's productieve woordvormingsmogelijkheden definiëren, worden deze lang niet altijd ten volle benut; zo bestaat naast het werkwoord *aangaan* (nog) geen correlaat *aangang*.

3.5.2.2 Affixallomorfie

Net als stammen kunnen ook affixen morfologisch geconditioneerde klankvormvariatie vertonen, d.w.z. variatie die noch semantisch noch fonologisch gemotiveerd is, maar die uitsluitend kan worden verantwoord door de bijbehorende morfologische context te specificeren. Dit geldt onder meer voor de vormvariatie in het diminutiefsuffix. Hoewel het MHB ervan uitgaat dat deze variatie volledig uit fonologische principes kan worden verklaard, betoogt Booij (2002) dat de zogenaamde basisvorm *tje* ook uitspreekbaar is in contexten waarin een allomorf (te weten *je*, *pje*, *kje* of *etje*) moet worden gekozen; dit impliceert dat er sprake is van morfologische conditionering. Hieronder volgt een overzicht van alle in Booij (2002; sectie 5.3) vermelde voorbeelden van suffixallomorfie:

inheemse suffixvarianten

cat	**allomorfen**	**voorbeelden**
N	-er / -der	*schrijver / bestuurder*
A	-er / -der	*groter / raarder*
A	-erig / -derig	*vreterig / zeurderig*
N	-erij / -derij	*stomerij / boerderij*
N	-tje / -je / -pje / -kje / -etje	*traantje / huisje / riempje/ koninkje / ringetje*
Adv	-tjes/ -jes / -pjes / -etjes	*gewoontjes / stilletjes / warmpjes / zachtjes*

uitheemse suffixvarianten

cat	**allomorfen**	**voorbeelden**
A	-eel / -aal	*fundamenteel / fundamentalist*
N/A	-air / -aar	*militair / militarist*
N	-eur / -oor	*directeur / directoraat*
A	-eus / -oos	*nerveus / nervositeit*
A	-iek / -ic	*katholiek / katholicisme*
N	-eur / -eus / -ric	*monteur / monteuse, ambassadeur / ambassadrice*

Deze inventarisatie is beperkt tot suffixvormen die historisch gezien van dezelfde basisvorm zijn afgeleid. Vanuit synchroon perspectief is het echter moeilijk om een principiële grens te trekken tussen affixallomorfie en affixconcurrentie (zie ook H3.5.5). Zo vertonen de agentieve suffixen -ER en -AAR, evenals de hiervan afgeleide synsuffixen -ERIJ en -ARIJ, sterke vormverwantschap en een nagenoeg complementaire distributie. Volgens Booij is er daarom niets op tegen om deze suffixen als synchrone allomorfen van een gemeenschappelijk grondsuffix te beschouwen. Dit geldt in principe ook voor de uitheemse, eveneens agentieve suffixen -OR en -EUR, resp. -ATOR en -ATEUR. Maar bij de keuze tussen -TE en -HEID of tussen -ERD en -ERIK kan beter van affixconcurrentie worden gesproken, want in deze gevallen is geen sprake van vormverwantschap of complementaire distributie.

3.5.2.3 Afbakeningsproblemen

In de voorgaande secties heb ik een aantal duidelijke gevallen van stam- en affixallomorfie besproken; het is echter niet altijd even makkelijk om te beslissen of een woordpaar stamallomorfie of affixallomorfie vertoont en of de overgang van stam naar affix met een morfologisch of een fonologisch gemotiveerd klanksegment correspondeert. Ik licht dit toe aan de hand van drie concrete analyseproblemen, waarvan de eerste twee op observaties van Booij (2002) zijn gebaseerd.

Probleem 1: Zoals in H3.3.5 aan de orde kwam, kennen veel samenstellingen een structuur waarbij het linkerwoorddeel en het rechterwoorddeel door een betekenisloze tussenklank worden verbonden; indien het linkerdeel met een inheemse N-stam correspondeert, heeft deze tussenklank, die ik verder als bindmorfeem zal aanduiden, meestal de vorm van het aan dit nomen gerelateerde pl-suffix, te weten *-s* of *-en*, maar het zou ook om een segment van de pl-stam kunnen gaan (bijv. het segment *-er-* van de pl-stam *kinder* in *kinderwagen*). Hieruit volgt dat het niet op voorhand duidelijk is of bindmorfemen met een affix corresponderen of met een vorm van stamallomorfie. Het kan echter niet om een fonologisch verbindingselement gaan (analoog aan het /j/-foneem tussen de segmenten *slee* en *en* in *sleeën*), want de aanwezigheid van een bindmorfeem in de bijbehorende vorm is doorgaans niet op fonologische gronden te voorspellen. Volgens Booij (2002) kunnen bindmorfemen het beste als een vorm van stamallomorfie worden geanalyseerd.

Probleem 2: Er bestaan tal van morfologische derivaties waarbij de formele basis van het laatst aangehechte suffix zelf ook weer uiteenvalt in een stam en een suffix, maar waarbij de betekenis uitsluitend door de hierin ingebedde stam wordt bepaald. Zo is de geografische aanduiding *Amerikaans* formeel gezien een afleiding van de inwonersnaam *Amerikaan*, maar semantisch gezien heeft dit adjectief betrekking op het land dat wordt aangeduid door het segment *amerik*, namelijk Amerika. Hieruit zou men kunnen afleiden dat de N *Amerika* ten minste drie stamvormen kent, namelijk *amerika*, *amerik* en *amerikaan*. Een andere optie is dat de stam van het lexeem *Amerika* met de vorm *amerik* correspondeert, dat er een suffix -A bestaat voor de vorming van geografische namen (vgl. *Afrika*, *Europa*) en dat het vrouwelijke persoonsnaamsuffix met de vorm *-s* een variant met de vorm *-aans* kent. In beide gevallen wordt echter afstand genomen van het idee dat derivaties lexeemgebaseerd zijn en dat de stam van dit lexeem met de onverbogen vorm overeen moet komen. Andere door Booij (2002) genoemde voorbeelden zijn:

landsnaam	inwoner (m.)	adjectief	inwoner (vr.)
Denemarken	Deen	Deens	Deense
Griekenland	Griek	Grieks	Griekse
Zweden	Zweed	Zweeds	Zweedse
Israel	Israëliet	Israëlisch	Israëlische
Rusland	Rus	Russisch	Russische

Het hier besproken alternantiepatroon is niet beperkt tot het domein van de eigennamen. Zo kan men zich afvragen of de woordgroep *een humoristische opmerking* naar een opmerking met humor verwijst (HUMOR+ISTISCH) of een opmerking die door een *humorist* wordt gemaakt (HUMORIST+ISCH). Omdat er vele stammen zijn waarvoor geen nomen met de uitgang *ist* bestaat, maar wel een adjectief met de uitgang *istisch* (bijv. *amateur* - **amateurist* - *amateuristisch*), lijkt de eerste analyse beter gemotiveerd te zijn; want als het synaffix -[IST+ISCH] onafhankelijk nodig is, waarom zou men dan nog gebruik maken van de omweg -IST + -ISCH? Deze redenering geldt ook voor woordparen als *filosoof* - *filosofisch*, *morfoloog* - *morfologisch* en *lexicograaf* - *lexicografisch*.

Probleem 3: Er zijn veel uitheemse afleidingen waarbij de stam niet direct door het affix wordt gevolgd, maar waarbij een extra foneem is ingevoegd of waarbij juist een foneem is verwijderd. Zo geldt voor bijna alle Griekse wortelsuffixen (zoals -GRAAF, -LOOG, -SOOF, -METER, -THEEK etc.) dat ze vooraf moeten worden gegaan door het segment *-o-*; maar het valt moeilijk uit te maken of dit foneem onderdeel van het "suffix" is of dat het suffix alleen eist dat er een segment *o* aan vooraf gaat, hetzij als (optioneel) segment van de stam (of het suffix), hetzij als vrij segment. Dit impliceert dat er waarschijnlijk geen vaste structuur bestaat, maar dat per geval moet worden nagegaan welke analyse de voorkeur verdient (wat af-

hangt van de interactie tussen stamkenmerken en suffixkenmerken). Ik zal dit toelichten aan de hand van de woordvormanalyses in de onderstaande tabel (BF staat voor bindfoneem):

	stam + suffix-paradigma	**stam + BF + suffix**	**woordvorm**
a)	lexic + {on, aal, grafie}	lexic + o + grafie	*lexicografie*
b)	techn + {iek, eut, craat}	techn + o + craat	*technocraat*
c)	psych + {e, isch, oot, paat}	psych + o + paat	*psychopaat*
d)	spectr + {um, aal, meter}	spectr + o + meter	*spectrometer*
e)	bacteri + {-,eel, cide, fagie}	bacteri + o + fagie	*bacteriofagie*
f)	radi + {o, aal, ent, loog}	radi + o + loog	*radioloog*
g)	disco + {-, grafie, theek}	disco + - + theek	*discotheek*
h)	bio + {toop, grafie, loog}	bio + - + loog	*bioloog*
i)	stereo + {toren, type, scopie}	stereo + - + scopie	*stereoscopie*

In deze tabel correspondeert de laatste kolom met de geanalyseerde woordvorm, terwijl de tweede kolom laat zien welke structuuranalyse het meest waarschijnlijk is, gegeven het suffix-paradigma van de stam (dat in de eerste kolom wordt gespecificeerd). Deze analyses berusten op het uitgangspunt dat men de grens tussen stam en suffix kan bepalen door na te gaan wat het laatste stamfoneem is dat door meerdere suffixen wordt gedeeld, gegeven de informatie in de eerste kolom. Voor de voorbeelden (a)-(e) is meteen duidelijk dat de *-o* er dan niet bijhoort; hierbij is voorbeeld (e) bijzonder omdat de stamvorm *bacteri* ook toepasbaar is als lexeem (met de vorm *bacterie*). Voor de voorbeelden (f)-(i) is de stamgrens minder duidelijk, aangezien elk van deze stammen een lexeemtoepassing kent met de eindletter *o*. Maar dit *o*-lexeem heeft niet altijd de gewenste betekenis. Zo is er geen semantische relatie tussen het woord *radio* en het woord *radioloog*, maar bij de voorbeelden (g), (h) en (i) lijkt het *o*-lexeem wel bruikbaar, al blijft onzeker of de *o* van de stam komt of van de suffixen. Dergelijke afbakeningsproblemen komen ook bij andere suffixen voor, zoals men zelf kan vaststellen voor onderstaande voorbeelden met het suffix -EEL (BF staat voor bindfoneem):

woordvorm	**stam + suffix-paradigma**	**stam + BF + eel**
substantieel	substantie + (-, eel}	substantie + [-] + eel
relationeel	relatie + {-, ief, eel}	relatie + on + eel
redactioneel	redact + {ie, eur, eel}	redact + ion + eel
rationeel	ratio + {-, eel}	ratio + n + eel
devotioneel	devoot + {-, ie, eel}	devoot + ion + eel
controversieel	controvers +{e, ist, eel}	controvers + i + eel

3.5.3 Affixpotentiatie

3.5.3.1 Problemen voor de niveau-ordeningstheorie

Het MHB-model veronderstelt dat het morfologisch en fonologisch relevant is om onderscheid te maken tussen inheemse en uitheemse morfemen. Zo merkt het MHB op dat uitheemse affixen zich alleen aan een uitheemse basis kunnen hechten, terwijl inheemse affixen vaak een voorkeur vertonen voor een inheemse basis. Verder zouden uitheemse affixen meer invloed hebben op de fonologische eigenschappen van de basis, wat tot uitdrukking komt in klemtoonverschuiving en allomorfie.

Don & al. (1994) stellen dat het contrast tussen inheemse en uitheemse suffixen eenvoudig te verklaren is indien men uitgaat van de theorie van lexicale niveauordening (cf. Siegel, 1974; Kiparsky, 1982). Deze theorie, die voortbouwt op het door Chomsky & Halle (1968) geïntroduceerde contrast tussen klemtoonverschuivende +-grens-affixen en klemtoonneutrale #-grens-affixen, gaat ervan uit dat het lexicon verschillende strata kent, die elk met een specifieke verzameling morfemen en grammaticaregels corresponderen. Deze strata zouden een

linaire ordening c.q. niveau-ordening vertonen, waardoor er beperkingen ontstaan op de volgorde waarin de bijbehorende affixen aan een stam kunnen worden gehecht. Want de grammatica kan pas affixen van niveau 2 selecteren als hij niveau 1 heeft doorlopen, en hetzelfde geldt voor de hogere lexiconniveaus. Voor talen als het Engels en het Nederlands wordt betoogd dat niveau 1 overeenkomt met klemtoonverschuivende (veelal uitheemse) affixen, niveau 2 met klemtoonneutrale (veelal inheemse) affixen en niveau 3 met inflectie. Dit impliceert dat inflectie altijd na derivatie komt en dat inheemse affixen niet vooraf kunnen gaan aan uitheemse affixen.

Hoewel de niveau-ordeningstheorie op het eerste gezicht een aantrekkelijke generalisatie is, zijn er tal van studies waaruit blijkt dat deze theorie empirisch onhoudbaar is. Zo heeft Fabb (1988) voor 43 Engelse suffixen uitgezocht wat de voorspelde combinatiemogelijkheden zijn en hoe deze verzameling zich verhoudt tot de daadwerkelijk voorkomende combinatiemogelijkheden (op basis van lexicografische bewijsplaatsen). Uit dit onderzoek blijkt dat Siegel's niveau-conditie slechts een beperkte reductie van combinatiemogelijkheden oplevert: van 1849 (op basis van categoriale selectierestricties) naar 459. In de praktijk zouden echter niet meer dan 50 van deze suffixparen daadwerkelijk voorkomen, terwijl er ook affixcombinaties bestaan die ten onrechte worden verboden. De niveau-ordeningstheorie is dus niet in overeenstemming met empirische observaties aan het Engels. Volgens Fabb zijn veel betere voorspellingen mogelijk als men de suffixen onderverdeelt op basis van hun structurele distributiemogelijkheden (die per suffix moeten worden vastgelegd). In de praktijk zouden er (in aanvulling op de categoriale restricties) slechts vier distributieklassen nodig zijn, namelijk:

1) suffixen die een suffixloze stam vereisen;
2) suffixen die ook achter een specifiek suffix kunnen voorkomen;
3) suffixen die zich "vrij" aanhechten;
4) probleemgevallen.

Volgens Plag (1996) is de door Fabb (1988) voorgestelde classificatie net zo problematisch als de door hem aangevallen niveau-ordeningstheorie. Bij nadere beschouwing is het namelijk een non-theorie, want verreweg de meeste suffixen blijken tot klasse 1 te behoren, terwijl de andere klassen erg willekeurig gekozen zijn. Het ogenschijnlijke succes van Fabb's theorie is dan ook grotendeels te danken aan het feit dat er maar weinig suffixen zijn die achter een ander suffix kunnen voorkomen; bovendien blijkt Fabb geen rekening te hebben gehouden met laagfrequente suffixcombinaties. Plag (1996) daarentegen heeft de complete Oxford Dictionary of English (OED) geanalyseerd en concludeert hieruit dat er een fijnmazig systeem van lexicale selectierestricties nodig is om empirisch adequate voorspellingen te doen over de vraag wat mogelijke en onmogelijke morfeemcombinaties zijn. Verder stelt Plag dat veel van deze combinatiemogelijkheden het beste via stamgebaseerde ("base-driven") selectierestricties kunnen worden verantwoord; dit fenomeen is eerder beschreven als affixpotentiatie (naar een voorstel van Williams (1981)).

Ik zal een en ander toelichten aan de hand van Plag's analyse van het distributiepatroon van de deverbale nomen-vormende suffixen -AGE, -AL, -ANCE, -MENT en -Y; deze suffixen hebben met elkaar gemeen dat ze volgens Fabb uitsluitend aan suffixloze lexemen mogen worden aangehecht. Volgens Plag is dit echter empirisch onjuist; volgens hem kan de observatie van Fabb beter verklaard worden uit het gegeven dat alle V-vormende suffixen in het Engels, te weten -IFY, -IZE en -ATE uitsluitend nominalisatie met -(AT)ION toestaan. Als gevolg van deze stamgebaseerde restrictie worden alle andere nominalisatie-suffixen geblokkeerd, blijkens het onderstaande overzicht:

MAGNIFY+CATION VERBALIZE+ATION CONCENTRATE+ION

*MAGNIFY+AGE	*VERBALIZE+AGE	*CONCENTRATE+AGE
*MAGNIFY+ANCE	*VERBALIZE+ANCE	*CONCENTRATE+ANCE
*MAGNIFY+AL	*VERBALIZE+AL	*CONCENTRATE+AL
*MAGNIFY+Y	*VERBALIZE+Y	*CONCENTRATE+Y
*MAGNIFY+MENT	*VERBALIZE+MENT	*CONCENTRATE+MENT

Plag's voorstel heeft als voordeel dat het enerzijds aangeeft welke morfeemcombinaties potentieel beschikbaar zijn, terwijl het anderzijds een streng filter definieert voor niet-toegestane suffixen. Dit lijkt niet mogelijk met suffixgebaseerde selectierestricties, want in dat geval zou men uitsluitend kunnen vastleggen welke stamsuffixen vaak als aanhechtingsbasis dienen, maar niet welke stamsuffixen verboden zijn; anders zou men enorme lijsten van verboden stamsuffixen moeten specificeren. Plag's voorstel vormt in minstens twee opzichten een breuk met het syntactische derivatiemodel: ten eerste gaat Plag er expliciet vanuit dat suffixen in staat zijn om andere suffixen te selecteren; ten tweede gaat Plag ervan uit dat een deel van de suffixgerelateerde selectierestricties stamgebaseerd is.

3.5.3.2 Popma's inventarisatie van suffixparen

Voor het Nederlands leidt de analysemethode van Fabb tot vergelijkbare conclusies als voor het Engels. Dit blijk uit onderzoek van Popma (1992). Doordat Popma's classificatie van Nederlandse morfemen op het analysemodel van Fabb (1988) is gebaseerd, kent Popma's voorstel dezelfde beperkingen als Fabb's classificatie van Engelse morfemen. Het verdient daarom de voorkeur om de observaties van Popma te heranalyseren op basis van het lexiconmodel van Plag (1997). Dit impliceert dat per stam of stamsuffix moet worden vastgelegd welke suffixen erop kunnen volgen, waarbij alleen suffixen mogen worden geselecteerd waarvan de selectierestricties compatibel zijn met de stam.

3.5.4 Paradigmatische woordvorming

In het morfologische onderzoek naar de woordvorming kunnen twee soorten constructiedimensies worden onderscheiden, namelijk een syntagmatische dimensie en een paradigmatische dimensie. Hierbij heeft de syntagmatische dimensie betrekking op relaties tussen opeenvolgende morfemen (zoals A + X en X + Y2 in het onderstaande schema), terwijl de paradigmatische dimensie naar relaties tussen parallel selecteerbare morfemen verwijst (zo kan X door drie verschillende morfemen worden gevolgd, namelijk Y1, Y2 en Y3):

$$A + X + \left\{ \begin{matrix} Y1 \\ Y2 \\ Y3 \end{matrix} \right\}$$

Hoewel het Morfologisch Handboek zich beperkt tot de beschrijving van syntagmatische constructieregels, geeft Booij (2002) tal van voorbeelden waaruit blijkt dat het lexicon ook kan worden uitgebreid door middel van affixsubstitutie of zelfs lexeemsubstitutie (in het geval van samenstellingen); hij spreekt in dit verband van *paradigmatische woordvorming*. Dit type woordvorming is alleen nodig voor situaties waarin het te vormen lexeem niet langs syntagmatische weg van een ander lexeem kan worden afgeleid. Dit komt het meest voor in het uitheemse deel van de woordenschat, want de stam van uitheemse lexemen is meestal niet beschikbaar als zelfstandig lexeem, maar deze stam vormt vaak de kern van twee of meer gelede lexemen; hieruit volgt dat deze lexemen een paradigmatische relatie onderhouden. Er is sprake van paradigmatische woordvorming indien deze relatie ook benut kan worden voor de constructie van nieuwe lexemen, dus indien deze relatie gegeneraliseerd wordt naar lexemen waarvoor nog geen paradigmatisch correlaat bestaat.

Hierbij kan men denken aan lexemen op *loog*, *logie* en *logisch*, lexemen op *graaf*, *grafie* en *grafisch* en soortgelijke clusters van Griekse suffixen. In deze analyse is *morfologisch* geen syntagmatische afleiding van *morfoloog*, maar een paradigmatische variant van *morfologie*, evenals *morfologisch*, waarbij elk van deze vormen als basis kan dienen voor de constructie van de andere vormen; dit staat bekend als de paradigmatische afleiding van lexemen. Deze constructiemethode biedt ook mogelijkheden voor de aanmaak van nieuwe syntagmatische regels door heranalyse van bestaande derivaties. Zo kan men woorden als *beschilderd* en *geisoleerd* heranalyseren als combinaties van een verbale lexeemstam en een discontinu affix, namelijk [SCHILDER$_V$ + [BE..D]] en [ISOL$_V$ + [GE..EERD]]. Vervolgens kan dit constructiepatroon worden gegeneraliseerd door het bereik uit te breiden tot A-stammen (bijv. van droef naar *bedroefd*) en N-stammen (bijv. van *talent* naar *getalenteerd*).

3.5.5 Affixconcurrentie

De paradigmatische dimensie is ook van belang bij de analyse van affixconcurrentie, dus in omstandigheden waarbij een keuze mogelijk is tussen verschillende affixen met dezelfde functie. Van Marle (1986) probeert dergelijke keuzes te verantwoorden door middel van een woordvormingsconditie die hij aanduidt als de Domein Hypothese. Deze conditie zegt het volgende: indien er verschillende suffixen zijn die dezelfde functie uitdrukken, kennen deze suffixen een complementair toepassingsdomein, waarbij onderscheid kan worden gemaakt tussen een standaardsuffix (met een onbegrensd toepassingsdomein) en één of meer specifieke suffixen (met een beperkter toepassingsdomein). Het gevolg is dat elke stam die voor deze suffixfunctie in aanmerking komt slechts één suffix kan selecteren, want indien een stam niet in aanmerking komt voor een specifiek suffix komt hij altijd uit bij het standaardsuffix. Het Nederlands kent bijvoorbeeld een hele reeks suffixen die een vrouwelijke persoonsnaam uitdrukken, te weten -E, -ES, -ESSE, -EUSE, -ICA, -IÈRE, -IN, -IX, -RICE en -STER. Volgens Van Marle zouden deze suffixen (die met elkaar gemeen hebben dat ze allemaal een nominale stam vereisen) dus een complementair toepassingsdomein moeten bezitten. Volgens zijn analyse correspondeert het suffix -E met het standaardsuffix, wat impliceert dat alle andere suffixen met een uniek toepassingsdomein moeten corresponderen: het suffix -ES zou bijvoorbeeld een lexeembasis met het suffix -AAR of -ER vereisen, het suffix -ESSE een basis met het suffix -ARIS en het suffix -EUSE een basis met het suffix -EUR. Maar indien geen van deze suffixen in aanmerking komt, zou altijd het suffix -E moeten worden gekozen, wat resulteert in vormen als *docente*, *echtgenote*, *gidse* en *typiste*.

Uit Booij (2002) blijkt dat de door Van Marle voorgestelde analyse vele empirisch problemen ontmoet. Zo heeft het suffix -E ondanks zijn status als standaardsuffix een duidelijke voorkeur voor aanhechting aan stammen die op een suffix eindigen; deze tendens is zo sterk dat men zich kan afvragen of de suffixen -ESSE, -EUSE en -RICE geen gelexaliseerde suffixcombinaties (c.q. synaffixen) zijn, namelijk suffixcombinaties met de structuur -[ES+E], -[EUS+E] en -[RIC+E]. Een tweede probleem is dat er tal van persoonsnamen bestaan waarvoor geen vrouwelijke vorm beschikbaar is, ook niet het standaardsuffix -E. Dit is het geval voor woorden als *auteur*, *ingenieur* en *minister*. Ten derde blijkt het niet altijd mogelijk om complementaire domeincondities te formuleren; zo lijkt het tamelijk willekeurig bepaald te zijn of een persoonsnaam op -EUR een vrouwelijke vorm op -EUSE of op -RICE kiest. Ten vierde is het de vraag of vrouwelijke persoonsnamen altijd langs syntagmatische weg worden geconstrueerd; zo betoogt Booij dat het suffix -STER met het mannelijke suffix -ER alterneert, dus dat er in dit geval sprake is van affixsubstitutie. Al met al wijzen Booij's observaties erop dat affixselectie primair een lexicale basis heeft en dat analyses die uitgaan van een standaardaffix (of defaultregel) gedoemd zijn om te falen. Dit blijkt ook uit het feit dat bij sommige derivatiefuncties meerdere standaardsuffixen lijken te bestaan; zo kent het Nederlands twee

standaardsuffixen voor meervoudsvorming, namelijk -S en -EN.[138] Uit deze overwegingen volgt dat het dualistische lexiconmodel van Pinker en Prince (1994) en Clahsen (1999), waarin het mentale lexicon wordt onderverdeeld in een component voor productieve woordvorming (op basis van defaultregels) en een component voor improductieve woordvorming (op basis van lexicale redundantieregels), op een fundamenteel verkeerd uitgangspunt berust.

Het vrije selectiegedrag van standaardaffixen kan worden verantwoord door hun toepassingsomein langs inductieve weg te analyseren en de hierbij aangetroffen subklassen beschikbaar te maken voor gerichte domeinselectie. Zo kan men het toepassingsdomein van het vrouwelijke persoonsnaamsuffix -E achterhalen door voor de bestaande persoonsnamen na te gaan welke stamkenmerken vaak voorkomen; zoals reeds aan de orde kwam blijken deze stammen meestal op een suffix te eindigen, zodat men alle suffixen kan opsommen die regelmatig aan het suffix -E voorafgaan, bijvoorbeeld -ANT, -ENT en -IST. Het toepassingsdomein van -EN en -S laat zich op dezelfde manier analyseren: (ORANJE+S, KOEMAN+EN, NEDERLAND+EN), afkortingen (AIO+S, BMW+S, DVD+S, P.S.+EN) en nominale woordgroepen (zoals VERGEET-ME-NIET-JE+S, POOTJE-OVER+S); hierbij kunnen desgewenst ook fonologische subklassen te worden aangebracht. Op deze manier kan een zeer fijnmazig systeem van selectierestricties worden opgebouwd, dat een krachtig alternatief biedt voor de defaultregels van Van Marle, Pinker en anderen. Zo'n kennissysteem heeft bijvoorbeeld geen moeite met de verantwoording van niet-systematische vormblokkades, zoals **steler* / *dief*, *schieter* / *schutter,* **goeder* / *beter*, ?*meer logisch* / *logischer* en **spreekte* / *sprak*, ook als beide opties zijn toegestaan.

3.5.6 Stamconcurrentie

Net als lexicaal verwante affixen kunnen lexicaal verwante stammen onderlinge concurrentie vertonen bij de opbouw van een nieuwe lexeemtoepassing. Indien de te gebruiken stam slechts één vorm kent, is er geen probleem, maar indien meerdere vormen beschikbaar zijn, zal een keuze moeten worden gemaakt. Zo kent het V-lexeem *spreken* de stamvormen *spreek*, *sprak*, *spraak* en *sprook*; desgewenst kan men hier nog enkele nominale stamvormen aan toevoegen, namelijk *sprek* en *spreuk*. Het is dus niet op voorhand duidelijk welke stamvorm het meest geschikt is voor de opbouw van een nieuwe lexeemtoepassing. In de praktijk gaat echter vaak de voorkeur uit naar de stamvorm van de tegenwoordige tijd, in dit geval de stamvorm *spreek*. Gegeven deze voorkeur zou men het selectieprobleem dus eenvoudig kunnen oplossen door (in het geval van sterke werkwoorden) altijd de stamvorm van de tegenwoordige tijd te selecteren. Deze voorkeur zou gemotiveerd kunnen worden door te stellen dat afleidingen die uitgaan van een andere stamvorm per definitie onregelmatig zijn en daarom niet morfologisch hoeven te worden verantwoord. Onder deze aanname moeten onregelmatige inflectievormen (zoals *sprak* in plaats van *spreekte* en *gesproken* in plaats van *gespreekt*) en afleidingen (zoals *spraak* en *spraakzaam*) dus in ongelede vorm in het lexicon worden opgeslagen.

Hoewel dit op het eerste gezicht een adequate analyse lijkt, zijn er een aantal fundamentele problemen aan verbonden. Allereerst leidt deze analyse ertoe dat alle woordvormen die niet volstrekt regelmatig gevormd zijn genegeerd worden, terwijl er toch herkenbare structuurelementen aanwezig zijn. Zo kan de inflectievorm *wonnen* als een regelmatig meervoud worden gezien van de inflectievorm *won*, die met de verleden tijd enkelvoud van het werkwoord *winnen* correspondeert. Maar omdat de vorm *won* geen stamstatus heeft, kan geen recht worden gedaan aan dit regelmatige verband. Iets soortgelijks geldt voor het lexeem

[138] Zie Van Wijk (2002) voor experimenteel onderzoek op dit terrein.

spraakzaam ten opzichte van de stamvorm *spraak*. De systematische uitsluiting van stamvormen leidt ook tot flinke gaten in het inflectieparadigma van sterke werkwoorden.

Een tweede probleem is dat de keuze voor een stamvorm die met de tegenwoordige tijd correspondeert, impliceert dat de betekenis van een geleed woord als *spreker* geparafraseerd moet worden als "iemand die in de tegenwoordige tijd spreekt", wat natuurlijk onzin is. In feite is deze complicatie karakteristiek voor het lexeemgebaseerde derivatiemodel: bij alle afleidingen die op een V-stam zijn gebaseerd, zal de betekenis immers naar het onderliggende werkwoord moeten verwijzen. Het woord *spreker* correspondeert dan met "iemand die een voordracht houdt", het woord *bespreekbaar* met "situatie waarin een nader aan te duiden onderwerp vrij besproken kan worden" en het woord *bespreking* met "een gebeurtenis waarbij iets besproken wordt". Hoewel dit voor de hand liggende definities lijken, zijn ze niet triviaal af te leiden uit de eigenschappen van het werkwoord (in combinatie met het suffix). Daar komt bij dat de werkwoordstam temporeel aspect zou moeten opleggen, maar er is niets dat erop wijst dat een *spreker* of *bespreking* standaard in heden, verleden of toekomst moet worden gesitueerd, terwijl het evenmin mogelijk is om zo'n temporele specificatie toe te voegen, bijvoorbeeld door stamverbuiging: *spreker-spraker-sproker*, of door samenstelling: *heden-spreker*, *gisteren-spreker*, *morgen-spreker*.

Een derde, nog fundamenteler probleem is dat er geen duidelijke grond bestaat voor een voorkeursbehandeling van de werkwoordstam; toch is het alleen deze arbitraire keuze die het onderscheid tussen regelmatige en onregelmatige afleidingen mogelijk maakt. Het enige argument dat deze keuze zou kunnen motiveren is dat de kale stam zelfstandig gebruik toestaat als eerste persoon enkelvoud van een werkwoord in de tegenwoordige tijd; maar om een inflectievorm als bewijs te nemen voor de V-status van de stam is niet erg overtuigend. Zelfs als het de infinitiefvorm is (zoals in het Engels)[139], zou het toch vooral een syntactisch argument zijn, want de infinitief kan in het Nederlands zowel als deel van een werkwoordelijke cluster als in nominale constructies worden gebruikt, en qua betekenis lijkt de infinitief (door het ontbreken van temporeel aspect) zelfs meer op een nomen dan een werkwoord.

Uit de voorgaande beschouwing blijkt dat er geen syntactische legitimatie kan worden gevonden voor de hypothese dat de stamvorm van een werkwoord in de tegenwoordige tijd standaard als basis dient voor de constructie van nieuwe lexemen. In mijn optiek kan de bestaande voorkeur beter worden verklaard uit het feit dat de basisvorm van het werkwoord tevens de meest gebruikte vorm is, zodat de geobserveerde voorkeur een statistische verklaring kan krijgen.

3.5.7 Conclusie

In deze sectie is empirische evidentie bijeengebracht voor de stelling dat een adequate theorie van de Nederlandse woordvorming niet kan volstaan met een lexicon van atomaire morfemen; zo'n theorie zal namelijk ook kennis over vaste, "gelexicaliseerde" morfeemcombinaties moeten kunnen vastleggen. De onderbouuwing van deze stelling heb ik grotendeels op bestaande literatuur gebaseerd.

In H3.5.2 is gedemonstreerd dat veel morfemen vormalternanties (c.q. allomorfie) vertonen die uitsluitend langs lexicale weg zijn te verantwoorden, d.w.z. door de beschikbare vormvarianten op te sommen en per vormvariant aan te geven wat de bijbehorende morfeemcontexten zijn; deze strategie biedt ook een oplossing voor de analyse van bindfonemen op de grens van twee morfemen: deze morfeemspecifieke fonemen kunnen namelijk als een speciale vorm van allomorfie worden beschouwd.

[139] Deze taal kent zo weinig inflectie dat dit mogelijk de populariteit van het categoriale regelmodel verklaart.

In H3.5.3 is aangetoond dat er veel minder suffixcombinaties voorkomen dan door de regelgebaseerde benadering wordt voorspeld, wat samenhangt met het feit dat suffixen zich niet in algemene morfologische klassen (c.q. strata) laten indelen; het is daarom efficiënter om de beschikbare suffixcombinaties rechtstreeks in het lexicon op te slaan (en dus af te zien van een abstract regelsysteem).

In H3.5.4 is getoond dat lexemen langs paradigmatische weg van andere lexemen kunnen worden afgeleid, dat het lexicon daarom ook kennis dient op te slaan over affixparadigma's, d.w.z. bundels van affixen die dezelfde stammen kunnen selecteren. Het lijkt echter niet nodig om paradigma's te introduceren voor suffixen met een identieke functie (c.q. betekenis), aangezien de selectie van suffixen grotendeels langs lexicale weg kan worden verantwoord (in de vorm van "positieve" selectierestricties).

In de secties 3.5.5 en 3.5.6 is aandacht besteed aan paradigmatische concurrentie-effecten bij de selectie van affixen en bij de selectie van stamvormen. Deze effecten bieden aanvullende evidentie voor de hypothese dat het lexicon een paradigmatische ordening kent.

In mijn optiek wijzen de hier besproken fenomenen erop dat het lexicon met een complex netwerk van morfeemcombinaties correspondeert. In hoofdstuk 4 wordt dit idee formeel uitgewerkt. Hiertoe wordt een lexiconmodel voorgesteld waarin paradigma's een cruciale rol spelen bij de identificatie van morfemen en van grote invloed zijn op de organisatie van het lexicon als geheel.

3.6 De hiërarchische structuurdimensie

3.6.1 Introductie

Het MHB gaat er (net als Don & al. (1994)) vanuit dat alle morfologisch gelede lexemen (behalve enkele uitzonderingsklassen) aan de Rechterhand Hoofd Regel (RHR) voldoen. Met Booij (2002, 2005a) ben ik van mening dat dit een onhoudbare hypothese is. Om dit aan te tonen bespreek ik eerst een aantal conceptuele problemen (H3.6.2) en vervolgens een aantal empirische problemen (H3.6.3). Tot slot (in H3.6.4) leg ik uit dat de RHR overbodig is indien men uitgaat van een compositioneel representatiesysteem.

3.6.2 Conceptuele problemen met de RHR

3.6.2.1 Definitievragen

Williams (1981) is een van de eerste taalkundigen die een poging heeft gedaan om een grammaticaprincipe te formuleren dat een verklaring biedt voor de empirische observatie dat talen als het Nederlands en het Engels veel lexemen kennen waarvan de grammaticale eigenschappen volledig door het meest rechtse suffix of woorddeel worden bepaald. Zijn voorstel staat bekend als de Righthand Head Rule, en is aanleiding geweest voor een hele stroom aan vervolgpublicaties met varianten op de RHR. Volgens Don & al. (1994) verdient een model dat uitgaat van de RHR de voorkeur boven een model dat uitgaat van woordformatieregels (WFR's). Dit wordt toegelicht aan de hand van een concreet voorbeeld, namelijk de morfologische analyse van het N-lexeem *speler*. In de WFR-benadering is dit lexeem het resultaat van een woordformatieregel die aangeeft dat het segment *-er* een N kan vormen door zich aan een V-stam te hechten, in dit geval de stam SPEEL:

(7) [V] + -er $\rightarrow$ [V + -er]$_N$ ("uitvoerder van handeling V")

In de RHR-benadering daarentegen wordt aangenomen dat het segment *-er* met een morfeem van categorie N correspondeert (namelijk het suffix -ER) en dat dit morfeem uitsluitend kan voorkomen in combinatie met een lexeem van categorie V; hierbij voorspelt de RHR dat het resulterende lexeem dezelfde categorie heeft als het hoofd, namelijk de categorie N. Deze

analyse veronderstelt dat voor elk suffix een lexicale ingang bestaat die informatie geeft over zijn categorie, zijn morfologische subcategorisatiematrix, zijn klankvorm en zijn betekenis; voor het suffix -ER zou men bijvoorbeeld de volgende specificaties kunnen aantreffen:

(8) -ER: N, [V_], *-er*, "uitvoerder van handeling V"

Hoewel (7) en (8) op het eerste gezicht notationele varianten zijn, is er een subtiel, maar verstrekkend verschil: de representatie in (7) geeft namelijk geen rechtstreekse informatie over de eigenschappen van het afgeleide lexeem. Volgens Don & al. is dit ook niet nodig, aangezien de RHR voorspelt dat de woordeigenschappen identiek zijn aan de eigenschappen van het hoofd. Deze analyse zou als voordeel hebben dat men slechts één morfologische grammaticaregel hoeft te postuleren, namelijk de RHR. Alle andere informatie kan uit het lexicon worden gehaald. Naar mijn mening is dit echter geen sterk argument, want in beide benaderingen moet per suffix worden gepostuleerd wat de categorie van het hiermee afgeleide woord is, terwijl het qua representatieruimte niet uitmaakt of men de combinatiemogelijkheden van een suffix langs lexicale weg of door middel van een regel verantwoordt.

Ernstiger is dat de RHR bij nadere beschouwing niet toetsbaar is. De RHR stelt namelijk dat het hoofd van een lexeem per definitie met het meest rechtse morfeem correspondeert en dat de eigenschappen van dit hoofd per definitie identiek zijn aan de grammaticale eigenschappen van het hiermee gevormde lexeem. Verder wordt aangenomen dat er onafhankelijke criteria zijn om lexemen morfologisch te ontleden. Zonder deze aannames zou het niet mogelijk zijn om affixen te onderscheiden en van een categorie te voorzien. Op dit laatste punt verschillen affixen namelijk cruciaal van stammen, want in tegenstelling tot stammen kunnen affixen niet als zelfstandig lexeem worden gebruikt. Hieruit volgt dat er geen enkele empirische basis is voor de hoofdgebaseerde analyse. De hier gevolgde redenering impliceert ook dat prefixen nooit als hoofd kunnen optreden, dus niet als categoriebepalend morfeem kunnen fungeren. Volgens de RHR correspondeert het hoofd immers altijd met het meest rechtse morfeem, dus nooit met een suffix of een stam. Don & al. leiden hieruit af dat het onmogelijk is om prefixen van een categorie te voorzien, een hypothese die bevestigd zou worden door voorbeelden als *disharmonie* (dat als $[\text{DIS+HARMONIE}_N]_N$ wordt geanalyseerd). Dit verklaart misschien waarom ze het geen probleem vinden dat het Nederlands vele duizenden lexemen bezit waarin het prefix wel als categoriebepalend element lijkt op te treden (namelijk werkwoorden waarin het prefix met een N-stam of een A-stam is gecombineerd, zoals *beplanten* resp. *versterken*).

Ook Trommelen & Zonneveld (1986), die overtuigende evidentie aandragen voor de stelling dat het Nederlands een groot aantal woordvormingsregels kent waarvoor geldt dat de eigenschappen van het hiermee afgeleide woord rechtstreeks van de eigenschappen van het door deze regel toegevoegde affix kunnen worden afgeleid, zien in dergelijke problemen geen aanleiding om de RHR (in de formulering van Williams) af te zwakken: zij beschouwen de door de RHR opgelegde beperkingen juist als een conceptueel voordeel, aangezien het taalverwervende kinderen zou helpen bij het analyseren van morfologisch complexe woorden. Hoewel deze morfologen het bestaan van tegenvoorbeelden erkennen, stellen ze voor om die dan maar via lexicale redundantieregels te verantwoorden. Blijkbaar realiseren ze zich niet dat de RHR op deze manier onfalsifieerbaar wordt.

Toch is het niet moeilijk om het hoofdcriterium zo aan te passen dat het wel empirische betekenis krijgt, namelijk door het hoofd te definiëren als de morfologische constituent die het meest bepalend is voor de eigenschappen van het lexeem als geheel.[140] Deze formulering doet niet alleen recht aan de observatie dat er vele prefixen zijn die zich als (locaal of globaal)

[140] Jack Hoeksema (p.c.) heeft mij erop geattendeerd dat deze definitie van *hoofd* niet nieuw is, maar de standaarddefinitie is in de GPSG-literatuur over syntactische structuuranalyse (cf. Gazdar & al., 1985).

hoofd gedragen, maar biedt ook de mogelijkheid om het hoofdcriterium als ontleedprincipe toe te passen. Dit principe vormt een belangrijk uitgangspunt van mijn lexicale representatietheorie (zie hoofdstuk 4). Voor zover mij bekend gaat het om een nieuw voorstel, al zie ik enige gelijkenis met de percolatietheorie van Lieber (1980). In deze theorie wordt een fundamenteel onderscheid gemaakt tussen het percolatiegedrag van affixgebaseerde derivaties en samenstellingen: Lieber gaat er namelijk vanuit dat de woordkenmerken van samenstellingen altijd door de meest rechtse constituent worden bepaald (conform de RHR). In het geval van affixgebaseerde derivaties stelt Lieber echter dat de kenmerken van de afgeleide eenheid rechtstreeks bepaald worden door de kenmerken van het laatst geïntroduceerde affix (of het nu een prefix of een suffix is), tenzij er sprake is van een categorieloos affix: in dat geval dienen de ontbrekende kenmerken namelijk aan het complement te worden ontleend (namelijk de morfologische basis, die minimaal uit een stam bestaat). Maar ook deze theorie veronderstelt dat er onafhankelijke criteria bestaan om lexemen van morfologische structuur te voorzien en om de categorie van affixen vast te stellen, zodat de door Lieber geformuleerde principes niet onafhankelijk toetsbaar zijn. Volgens mij kan dit probleem alleen worden opgelost door de vraagstelling om te keren: men moet niet op zoek gaan naar een principe waarmee men een reeds gegeven morfeemstructuur kan interpreteren, maar naar een principe waarmee die morfeemstructuur langs inductieve weg gegenereerd kan worden. Dit idee vormt de basis van mijn lexicale representatietheorie.

3.6.2.2 Het domein van de RHR

Los van de reeds besproken definitieproblemen roept de RHR ook vragen op met betrekking tot de begrenzing van het toepassingsdomein. Want men kan de RHR (d.w.z. de claim dat het morfeem met de grootste invloed op de woordexterne kenmerken altijd met de meest rechtse lexeempositie correspondeert) pas toetsen, indien men aangeeft op welk type lexemen deze claim van toepassing is en hoe men deze lexemen moet afbakenen. Meestal wordt stilzwijgend aangenomen dat de RHR alleen van toepassing is op endocentrisch gevormde lexemen, d.w.z. lexemen waarvan de eigenschappen door een overt, lexeemintern morfeem worden bepaald. Hieruit volgt dat alle lexemen die het resultaat zijn van conversie buiten het bereik van de RHR vallen, al was het maar omdat onzichtbare hoofden niet gelocaliseerd kunnen worden. Maar indien men conversie in termen van een 0-affix analyseert (met een reconstrueerbare positie binnen het lexeem), vervalt een belangrijke reden om zulke derivaties uit te zonderen. Deze onzekerheid bemoeilijkt de evaluatie van de RHR.

Gegeven de bovenstaande domeindefinitie is de volgende vraag hoe men lexemen kan afbakenen. Zo is onduidelijk waar precies de grens ligt tussen inflectie en derivatie. Deze grens is echter van grote invloed op de evaluatie van de RHR, want derivationele affixen zouden van inflectie-affixen verschillen doordat derivatie-affixen potentieel categorieveranderend zijn. Maar hieruit mag niet worden afgeleid dat categorieneutrale affixen per definitie inflectie-affixen zijn, want een deel van deze affixen voldoet niet aan andere inflectiecriteria, zoals syntactische afhankelijkheid, paradigmatische organisatie en perifere positie. Bovendien kan de RHR alleen objectief getoetst worden indien men een onafhankelijk criterium hanteert voor de identificatie van de lexeemgrens. Dit criterium zal bijvoorbeeld antwoord moeten geven op de vraag welke categorieneutrale affixen met inflectie-affixen corresponderen en welke niet. Zo hecht het nomaliserende diminutiefsuffix -JE (en zijn vormvarianten) zich normaal gesproken aan een nominale basis. Ondanks dit categorieneutrale gedrag gaat men er meestal vanuit dat -JE geen inflectioneel, maar een derivationeel suffix is. Want bij nominaal gebruik is het suffix -JE bepalend voor woordgeslacht (onzijdig) en meervoudsvorm (altijd -S), terwijl het ook een voorspelbaar betekeniseffect heeft In de RHR-literatuur is de eerste eigenschap al voldoende om dit morfeem een categoriebepalend hoofd te noemen.

Een ander afbakeningsprobleem betreft de vraag of het partikeldeel van scheidbaar samengestelde werkwoorden als lexeeminterne constituent moet worden geanalyseerd. Indien men een werkwoord als *uithollen* als een samenstelling met de structuur [UIT + HOL + EN] analyseert (waarbij UIT en HOL met twee verschillende lexemen corresponderen), zou dit een uitzondering opleveren voor de regel dat het hoofd van een samenstelling altijd met het rechterwoorddeel correspondeert. Gegeven deze analyse rijst ook de vraag of niet-scheidbare woorden als *uitholling* en *uitholbaarheid* dan eveneens als een samenstelling moeten worden geanalyseerd, en of deze samenstellingen dan eveneens een linkerhoofd bezitten. Maar men zou natuurlijk ook een analyse als samenstellende afleiding kunnen overwegen (analoog aan woorden als *driewieler*), die mogelijk weer een ander hoofdgedrag vertonen. Dergelijke voorbeelden laten zien dat de RHR een zwak fundament kent en dat dit principe nauwelijks valt te toetsen.

3.6.2.3 Structuurvragen

Om de RHR te kunnen evalueren dient niet alleen bekend te zijn hoe men lexemen moet afbakenen, maar ook hoe ze intern gestructureerd zijn. Ook op dit punt zijn nog tal van vragen te beantwoorden. Om te beginnen is het onduidelijk hoe de RHR kan weten of een aan te hechten morfeem met het meest rechtse morfeem in de klankvorm correspondeert. In het algemeen bestaat namelijk geen direct verband tussen de affixatievolgorde en de fonologische affixpositie. Bovendien lijkt evaluatie achteraf onmogelijk, aangezien syntagmatische woordvormingsmodellen er meestal van uitgaan dat de grammaticaregels geen toegang hebben tot de interne structuur van reeds afgeleide constituenten. Ook kan men zich afvragen of een formeel geleed lexeem zijn interne structuur behoudt indien het een gelexicaliseerde betekenis krijgt en of deze morfeemstructuur zichtbaar is voor de RHR.

Een hiermee verwant probleem is dat er tal van vaste affixcombinaties bestaan waarvan niet duidelijk is of ze als een synaffix moeten worden geclassificeerd of als twee autonome suffixen; soortgelijke vragen zijn mogelijk met betrekking tot de analyse van samenstellingen. De hierbij gekozen analyse is uiteraard van invloed op de vraag welk morfeem of woorddeel als het hoofd van het lexeemdomein moet worden aangemerkt. Tot slot is onduidelijk of er enige relatie bestaat tussen de wijze waarop de woordkenmerken worden opgebouwd en de inbreng van het morfeem dat als morfologisch hoofd wordt geïdentificeerd. De hier naar voren gebrachte structuurvragen tonen opnieuw aan dat de RHR zwak is gefundeerd.

3.6.3 Empirische problemen met de RHR

3.6.3.1 Introductie

Deze sectie biedt een overzicht van de empirische problemen met de RHR. Ik besteed achtereenvolgens aandacht aan de problemen die samenhangen met categoriebepalende prefixen (§2), categorieneutrale suffixen (§3), discontinue affixen (§4), coverte affixen (§5), partiële hoofden (§6) en samenstellingen (§7). Ter introductie bespreek ik eerst een reeks lexemen met de N-stam LUCHT. Hierbij zijn drie verschillende structuurklasses te onderscheiden, te weten derivaties van de ongelede stam LUCHT, derivaties van de gelede stam LUCHTIG, en samenstellingen met een LUCHT-constituent. Deze structuurklassen corresponderen met aparte deeltabellen, te weten tabel 3.5a, 3.5b en 3.5c. In deze tabellen wordt voor elk lexeem aangegeven wat zijn grammaticale eigenschappen zijn. Verder geef ik voor elke lexeem een mogelijke structuuranalyse (uitgaande van de RHR-literatuur); deze representaties laten precies zien in welke volgorde de aanwezige affixen en woorddelen met de stam zijn gecombineerd. Deze derivatievolgorde is gemarkeerd door de wortel en elk hieropvolgend tussenproduct tussen vierkante haken te plaatsen en een deel van deze tussenproducten van een categorie te voorzien. Verder is het morfeem dat het meest bepalend is voor de functionele

eigenschappen (het *functionele hoofd*) steeds in vette letters weergegeven. Hierdoor kan snel worden nagegaan of het functionele hoofd en het RHR-hoofd overeenkomen.

vb	morfeemstructuur	categorie	inflectie	functiewoorden
a)	[**lucht**]	N	N-pl = -EN	*de*
b)	[lucht]+**je**	N	N-pl = -S	*het*
c)	[lucht]+**[0]**$_V$	V	zwakke vervoeging	*heeft*
d)	**ont**+[lucht]	V	zwakke vervoeging	*heeft*
e)	[ont+[lucht]]$_V$+**ing**	N	(alleen N-sg)	*de*

Tabel 3.5a: De relatie tussen lexeeminterne structuur en functionele eigenschappen.

Voorbeeld a) toont een ongeleed lexeem, namelijk de N *lucht*. In dit soort gevallen doet de RHR de triviale voorspelling dat het hoofd samenvalt met het hele lexeem en dat dit lexeem bepalend is voor de woordcategorie. Voorbeeld b) toont het gelede N-lexeem *luchtje*, dat als een stam-suffix-combinatie kan worden geanalyseerd. Alleen is onduidelijk of het suffix -JE een derivationele of inflectionele functie heeft. Als het om inflectie gaat, voorspelt de RHR dat de categorie van de inflectievorm als geheel gelijk is aan die van het basislexeem, wat (zoals gewenst) de categorie N oplevert. Als het om een derivatie gaat, voorspelt de RHR dat het suffix -JE met het hoofd correspondeert en dus bepalend is voor de woordcategorie. Dit is echter moeillijk te controleren, want -JE heeft hier dezelfde categorie als de stam.

Voorbeeld c) correspondeert met een V-toepassing van de stamvorm LUCHT. In de hier weergegeven analyse is sprake van een intern gelede lexeemvorm, waarbij de V-categorie aan een 0-affix met suffix-status wordt ontleend. Deze analyse is in overeenstemming met de RHR, en verklaart waarom het V-lexeem LUCHT een betekenis heeft die gebruik maakt van het N-lexeem LUCHT, namelijk "frisse lucht laten binnenstromen". Maar omdat een 0-affix geen zichtbare positie heeft, is onafhankelijke evidentie nodig om vast te stellen of sprake is van een prefix of een suffix. Pas dan kan worden bepaald of aan de RHR wordt voldaan. In Trommelen & Zonneveld (1986) worden argumenten pro en contra de prefixbenadering besproken, waarna een lichte voorkeur wordt uitgesproken voor de suffixbenadering (in wat andere termen). Dit standpunt wordt gedeeld door Don (1990) en Neeleman & Schippers (1992). Er bestaat echter ook steun voor de prefixbenadering (bijv. Lieber & Baayen (1993) en Plag (1997)). Het voordeel van de prefixbenadering is dat deze een verklaring biedt voor de analogie met V-derivaties op basis van prefixen als BE-, VER- of ONT-, zoals wordt geïllustreerd door het V-lexeem *ontlucht* in voorbeeld d). In de alternatieve analyse zou het V-lexeem *ontlucht* zijn categorie rechtstreeks aan de V-stam LUCHT ontlenen, waarbij het prefix ONT- een categorie-neutrale toepassing krijgt toebedeeld, net als in *ontlopen* en *ontstaan*. De onderliggende logica is echter problematisch in het licht van derivaties als *ontluchting* en *ontluchter*. Want deze lexemen kunnen niet worden geanalyseerd als een combinatie van het prefix ONT- met hypothetische N-lexemen als *luchting* of *luchter*, maar uitsluitend als een afleiding van de V-stam ONT+LUCHT. Deze analyse, die onder e) wordt uitgewerkt, is volledig in lijn met de RHR, maar leidt impliciet tot een voorkeur voor de prefix-gebaseerde conversie-analyse van voorbeeld c). Er ontstaat dus een paradox.

Beschouw nu tabel 3.5b, waarin voorbeelden staan met het A-lexeem *luchtig*:

vb	morfeemstructuur	categorie	inflectie	functiewoorden
f)	[lucht]$_N$+**ig**	A	[0/-ER/-ST](-E)	-
g)	**ver**+[[lucht]+ig]$_A$	V	zwakke vervoeging	*zijn*
h)	[[lucht]+ig]$_A$+**[e]**$_N$	N	N-pl = -N	*het*
i)	[[lucht]+ig]$_A$+**heid**	N	N-pl = -EN	*de*

Tabel 3.5b: De relatie tussen lexeeminterne structuur en functionele eigenschappen.

Voorbeeld f) toont de interne structuur van het A-lexeem *luchtig*. Dit lexeem berust op een derivatie met het A-vormende suffix -IG, dat zich hier aan de nominale stam LUCHT heeft gehecht en zich conform de RHR gedraagt. In voorbeeld g) dient het op deze wijze gevormde lexeem als basis van een volgende derivatiestap, namelijk de afleiding van het V-lexeem *verluchtig* door aanhechting van het V-vormende prefix VER-. Deze afleiding roept dezelfde vragen op als de V-afleiding in voorbeeld d), want volgens de RHR kunnen prefixen niet als morfologisch hoofd functioneren; maar het in voorbeeld d) voorgestelde alternatief, namelijk het vooraf converteren van de stam, is hier evenmin aantrekkelijk, want dit zou betekenen dat de A-invloed van het suffix -IG moet worden onderdrukt door een onzichtbaar V-vormend 0-affix. Dit lijkt me een uiterst onwenselijke analyse, die ook op geen enkele manier valt te toetsen. Voorbeeld h) correspondeert met het N-lexeem *luchtige*, dat men aantreft in de zin *Bij deze voorstelling worden het luchtige en het serieuze goed afgewisseld*. Qua vorm is dit enigszins gemarkeerde lexeemgebruik equivalent aan de geïnflecteerde vorm van het A-lexeem *luchtig*, zodat het voor de hand ligt om het N-lexeem als een conversieproduct te analyseren. Maar dit roept onmiddellijk de vraag op hoe men dan het domein van de RHR moet bepalen; want als geïnflecteerde lexemen onderdeel kunnen zijn van een lexeem is de inflectiegrens niet langer bruikbaar als afbakeningscriterium. Verder rijst opnieuw de vraag hoe men moet vaststellen of het 0-affix in hoofdpositie staat. Indien men aanneemt dat dit lexeem dezelfde structuur bezit als het enigszins verwante N-lexeem *luchtigheid* (zie voorbeeld i), zou het 0-affix nu als een suffix kunnen worden aangemerkt; maar in voorbeeld c) leidde een zelfde soort redenering tot de tegenovergestelde conclusie, zodat er geen enkele zekerheid aan kan worden ontleend.

De laatste vijf voorbeelden (wederom gebaseerd op het lexeem *lucht*) laten zien dat de RHR ook problemen ondervindt bij de analyse van samenstellingen.

vb	morfeemstructuur	categorie	inflectie	functiewoorden
j)	$[[\text{buiten}]_P+[\mathbf{lucht}]_N]$	N	(alleen N-sg)	*de*
k)	$[[\text{lucht}]_V+[\mathbf{rooster}]_N]$	N	N-pl = -s	*het*
l)	$[[\text{lucht}]_N+[\mathbf{fiets}]_V]$	V	alleen infinitief (-EN)	*(te)*
m)	$[[\text{lucht}]_N+[[\text{fiets}]_V+\mathbf{er}]_N]$	N	N-pl = -S	*de*
n)	$[[\text{lucht}]_N+[\text{hart}]_N+\mathbf{ig}]$	A	[0/-ER/-ST](-E)	-

Tabel 3.5c: De relatie tussen lexeeminterne structuur en functionele eigenschappen.

Volgens de RHR is het rechterdeel van een samenstelling bepalend voor zijn categorie en zijn externe selectie-eigenschappen. Voorbeeld j) toont een samenstelling waarvan het rechterdeel, te weten het N-lexeem *lucht*, inderdaad als hoofd fungeert, want dit lexeem is bepalend voor de categorie en de selectie-eigenschappen van de hele samenstelling (in tegenstelling tot het P-lexeem *buiten*); maar in tegenstelling tot het zelfstandige N-lexeem *lucht* kent deze samenstelling geen meervoudsvorm. Bovendien blijkt uit voorbeeld k) dat het lexeem *lucht* ook met een V zou kunnen corresponderen, dus dat er geen blinde categorietoekenning mogelijk is. Dit is ook van belang voor voorbeeld l), want het V-lexeem *luchtfiets* (dat voornamelijk in infinitiefvorm wordt gebruikt, namelijk *luchtfietsen*) kan op twee manieren worden geconstrueerd, namelijk als een samenstelling met het V-lexeem *fiets* (dat zelf weer als een V-toepassing van het N-lexeem *fiets* kan worden geanalyseerd) of als een V-toepassing van een samenstelling met het N-lexeem *fiets*; deze laatste optie is echter minder waarschijnlijk, omdat er geen objecten bestaan die men *luchtfiets* noemt.

Gegeven het V-lexeem LUCHTFIETS zou men voorbeeld m), te weten *luchtfietser*, als een derivatie met het suffix -ER kunnen analyseren. Er is echter ook een andere analyse mogelijk, namelijk als een samenstelling met de woorddelen LUCHT en FIETSER, waarbij het woorddeel FIETSER weer kan worden onderverdeeld in de morfemen FIETS en -ER. Maar als gevolg van

deze recursieve structuur is niet duidelijk welk element als hoofd moet worden geïdentificeerd. Er zijn immers twee constituenten die in de meest rechtse positie staan, namelijk het woorddeel FIETSER, dat hoofd is van de samenstelling, en het suffix -ER, dat hoofd is van het woorddeel FIETSER. Voorbeeld n) is nog problematischer, want het A-lexeem *luchthartig* kan niet van een hierin ingebed lexeem worden afgeleid, aangezien noch HART+IG noch LUCHT+HART zelfstandig voorkomen (althans niet in deze betekenis). Daarom kan dit lexeem uitsluitend als een samenkoppeling van twee lexemen en een suffix worden geanalyseerd, namelijk LUCHT+HART+IG. Toch is dit voorbeeld eenvoudiger te analyseren dan de woordvorm in m), want onder de hier gegeven analyse is er slechts één element dat aan de RHR-definitie van hoofd voldoet, namelijk het suffix -IG.

3.6.3.2 Categoriebepalende prefixen

Volgens de RHR zijn prefixen niet in staat om de categorie van een woord te bepalen, aangezien prefixen per definitie links van de woordstam staan. Hier valt echter wel wat op af te dingen, want er zijn vele werkwoorden die zijn opgebouwd uit een prefix en een niet-werkwoordelijke stam; hierbij gaat het meestal om een N-stam (bijv. BE+KROON$_N$+EN, ONT+ZADEL$_N$+EN) of een A-stam (bijv. VER+STERK$_A$+EN, BE+SPOEDIG$_A$+EN).[141] Dit probleem zou kunnen worden ondervangen door de RHR af te zwakken tot de claim dat bij gelede (maar niet-samengestelde) woorden het laatst aangehechte affix (hetzij een prefix, hetzij een suffix) als morfologisch hoofd fungeert. Onder deze aanpassing kan bijvoorbeeld verklaard worden waarom het lexeem *verluchtig* niet de categorie A, maar de categorie V draagt: hiertoe dient het lexeem *verluchtig* als een afleiding van de gelede stam *luchtig* te worden opgevat, waarbij het prefix VER- net als bij andere afleidingen aangeeft dat het om een transitieve V gaat. Gegeven de hier besproken aanpassing lijkt de RHR een bruikbare generalisatie te zijn over de relatie tussen woordstructuur en functionele eigenschappen.

3.6.3.3 Categorieneutrale suffixen

Het Nederlands kent enkele suffixen die geen eigen categorie introduceren, maar de categorie van hun morfologische basis overnemen; wat de overige selecties (zoals de inflectievormen) betreft kunnen deze suffixen overigens wel bepalend zijn. Bovendien is er meestal sprake van toevallige neutraliteit, d.w.z. van neutraliteit die voortkomt uit het feit dat de basis toevallig dezelfde categorie heeft als de doelcategorie. Dit laatste is het geval bij de suffixen -JE (N>N), -IG (A>A) en -SCHAP (N>N).[142] Structurele neutraliteit treft men aan bij de trappen van vergelijking (-ER en -ST) en bij suffixen van de vrouwelijke vorm, zoals -STER, -ES en -E.

3.6.3.4 Discontinue affixen

Zoals reeds bij de samenvatting van het MHB aan de orde kwam, kent het Nederlands niet alleen prefixen en suffixen, maar ook discontinue affixen (zie ook H3.3.4). Discontinue affixen corresponderen met een gelexicaliseerde combinatie van een prefix (PRE) en een suffix (SUF); het zijn dus affixen met de structuur PRE-[...]-SUF. Dergelijke affixen vormen een fundamenteel probleem voor de RHR, want ze corresponderen met twee structuurposities tegelijk, zodat niet zeker is of het affix zich in de meest rechtse positie bevindt.

3.6.3.5 Coverte affixen

Coverte affixen (c.q. 0-affixen) zijn een middel om een morfologische verklaring te geven voor de observatie dat bepaalde lexeemklassen systematisch in staat zijn om een hiervan afgeleide functie te vervullen, zonder dat deze afgeleide functie via een morfeem tot uitdrukking wordt gebracht. Door een 0-affix te postuleren kan namelijk eenvoudig worden verant-

[141] In *verdonkeremanen* vindt men zelfs een N+A-stam-combinatie: ver+donker-e$_A$+maan$_N$+en.
[142] Dat -SCHAP categorieneutraal zou zijn, wordt mogelijk weerlegd door de deadjectivsche vorm *zwangerschap*.

woord dat het "afgeleide" lexeem een aantal voorspelbare eigenschappen bezit. Zo correspondeert de 0-nominalisatie van de V GOOI met de N GOOI-0 (met vorm *gooi*). Deze afgeleide eenheid kiest het lidwoord *de* en meervoud -EN; deze eigenschappen vindt men ook bij andere 0-nominalisaties op basis van een V-lexeem, wat als rechtvaardiging kan dienen voor de postulatie van het 0-affix. Voor de RHR vormen zulke 0-affixen echter een probleem, want hoewel ze bepalend zijn voor de categorie en de selectie-eigenschappen van het gemodificeerde lexeem, hebben ze geen zichtbare positie, zodat ze niet aan de RHR voldoen. Dit kan alleen worden opgelost door te stellen dat 0-affixen een morfologische markering bezitten die aangeeft of het om een prefix of een suffix gaat. Dit lijkt me echter een zeer dubieuze stap.

3.6.3.6 Partiële hoofden

Geprefigeerde werkwoorden met een sterke stam hebben geen eenduidig hoofd: want terwijl de V-stam bepalend is voor de inflectievormen van de verleden en voltooide tijd, is het prefix verantwoordelijk voor de morfologische en syntactische selectiemogelijkheden. Hoeksema (1984)) heeft voorgesteld om dit fenomeen te verantwoorden door morfologische hoofdoperaties te introduceren. Met deze operatie kan men voor elke gewenste structuurdimensie (zoals de semantische dimensie en de morfologische dimensie) het prefix (ONT-) uit de complexe stam halen, en het suffix (in dit geval de operator M_{imp}) er in plaatsen. Hierdoor gaat de V zich morfologisch gezien als een ongelede V-stam gedragen, met als gevolg dat hij zijn sterke inflectiegedrag vertoont. Maar semantisch gezien blijft de stam gewoon met het prefix verbonden, hetgeen in een gelexicaliseerde betekenis resulteert.

(9) semantische structuur: [ont + loop] + M_{imp} vorm: **ontloopte*
morfologische structuur: ont + [loop + M_{imp}] vorm: *ontliep*

3.6.3.7 Samenstellingen

De meeste RHR-definities maken een principieel onderscheid tussen de analyse van derivaties en de analyse van samenstellingen. Deze RHR-definities gaan echter voorbij aan het feit dat het samenstellingsniveau compositioneel moet kunnen worden afgeleid uit het morfeemgebaseerde representatieniveau. Doordat niet over deze relatie is nagedacht, ontstaan analyseproblemen bij samenstellingen van het type [A B] (waarbij A en B voor interne constituenten c.q. lexemen staan) waarvan de meest rechtse constituent met een eenheid met de structuur $[\text{stam} + \text{suffix}]_B$ correspondeert. Want in deze structuur kan zowel het suffix als constituent B als hoofd worden aangemerkt (gegeven een RHR-definitie waarin beide niveaus afzonderlijk worden geanalyseerd). Er is dus een aanvullende regel nodig die voorspelt dat de categorie en de selectie-eigenschappen van de samenstelling als geheel zijn terug te voeren op de eigenschappen van het laatste suffix in de dominante constituent; volgens de bestaande RHR-definitie(s) is dit normaliter de meest rechtse constiuent, maar in het geval van *directeur-generaal* correspondeert de dominante constituent met *directeur*, zodat men niet bij het suffix -AAL, maar bij het suffix -EUR uitkomt. Verder zou ook een oplossing moeten worden bedacht voor samenstellingen waarvan de samenstellende constituenten niet zelfstandig bestaan of andere eigenschappen bezitten dan bij zelfstandig gebruik. Zo'n mechanisme is van belang met het oog op samenstellingen als *touwslager*, *fijnschrijver* en *vierhandig* (wegens de speciale betekenis van de lexemen *slager*, *schrijver* en *handig*) of *waarzegger*, *zingeving en gelijkzijdig* (wegens het niet-zelfstandig voorkomen van de lexemen *zegger*, *geving* en *zijdig*). Om dit goed te kunnen verantwoorden is een compositioneel woordvormingsmodel nodig, d.w.z. een model dat laat zien hoe een gegeven woord stap voor stap uit de samenstellende morfemen kan worden geconstrueerd. Dit wordt nader toegelicht in H3.6.4.

3.6.4 Een compositioneel alternatief

In de voorgaande secties heb ik laten zien dat het overervingsprincipe dat bekend staat als de RHR zowel conceptueel als empirisch op grote problemen stuit. Indien men uitgaat van een compositioneel grammaticamodel is de RHR echter overbodig,[143] want in een compositioneel model is informatie-overerving inherent aan structuuropbouw. Dankzij deze eigenschap biedt de compositionele grammaticabenadering een krachtig alternatief voor grammaticamodellen die hun toevlucht nemen tot overervingsprincipes als de RHR.

De compositionele grammaticabenadering berust op het idee dat de opbouw van morfologische en syntactische structuur gelijk op dient te gaan met de opbouw van de vormrepresentatie en de betekenisrepresentatie.[144] Dit compositionaliteitsbeginsel vindt zijn oorsprong in de ideeën van Frege (1892). Montague (1974) was de eerste die dit beginsel in een formeel grammaticamodel implementeerde.[145] Hoewel Montague's grammaticamodel primair was bedoeld voor de beschrijving van syntactische derivatieprocessen, heeft Dowty (1979) laten zien dat dit model een uitstekende basis biedt voor de definitie van de Engelse woordvormingsregels, en hebben Moortgat (1981; 1987) en Hoeksema (1984) deze representatiemethode met succes op de Nederlandse morfologie toegepast. Het Montague-model biedt tevens een goede basis voor de computationele analyse en synthese van natuurlijke taal. Voor het Nederlands werd deze mogelijkheid voor het eerst uitgewerkt door Van der Hulst en Moortgat (1980), die (in opdracht van het INL) onderzoek deden naar de vraag hoe het Nederlandse lexicon langs automatische weg van morfologische structuur kan worden voorzien; hiertoe definieerden zij een aantal basisprincipes voor het analyseprogramma ALEX (zie verder H5.3). Deze studie leidde later tot de ontwikkeling van een automatische morfeemparser, te weten KASIMIR (zie Moortgat (1985)), die een centrale rol speelde bij de analyse van het morfologisch geannoteerde corpus CELEX (zie Van der Wouden (1988)). Een soortgelijke aanpak ligt ten grondslag aan de morfologische parser MORPA, die deel uitmaakt van een automatisch tekst-naar spraak-systeem (zie Heemskerk & Van Heuven (1993)).

In een compositioneel grammaticamodel (cf. Moortgat, 1987) zijn morfologisch complexe woorden het resultaat van een gefaseerd concatenatieproces. In dit proces worden de lexicale basiseenheden stap voor stap tot grotere eenheden samengevoegd, waarbij elke combinatiestap met de toepassing van een functor op een argument correspondeert (indien sprake is van een monadische functor), of op meerdere argumenten (indien sprake is van een polyadische functor). Indien dit model wordt toegepast op het morfologische domein, corresponderen de basiseenheden met morfemen; deze vallen uiteen in stammen en affixen, waarbij stammen gedefinieerd zijn als eenheden met (morfologische) argumentstatus en affixen als eenheden met (morfologische) functorstatus. Dit wordt typelogisch verantwoord door stammen gelijk te stellen aan eenheden met een lexicale basiscategorie (zoals N, V, A, B of P), en affixen aan eenheden die een functie van broncategorie naar doelcategorie representeren. Zo'n functie correspondeert met een complexe categorie, d.w.z. een categorie die langs recursieve weg uit lexicale basiscategorieën kan worden opgebouwd; zo correspondeert de categorie [N/A] met een functie van A naar N en de categorie [(N/A)\V] met een functie van [N/A] (wederom een complexe categorie!) naar V. [146]

[143] Deze conclusie deel ik met Booij (2002, 2005a).
[144] Dit beginsel staat ook wel bekend als het Curry-Howard-De Bruin-isomorfisme.
[145] Frege (1892) geldt als grondlegger van het taalkundige compositionaliteitsprincipe. Zie Verkuyl (1996b) voor nadere informatie over de receptie van dit principe.
[146] In werkelijkheid kent het door Moortgat (1987) gehanteerde model een wat subtieler classificatiesysteem, want in dit systeem weerspiegelt de aan een stam toegekende categorie het op syntactisch niveau in te vullen argumentgrid. Hierbij correspondeert een transitieve V-stam bijvoorbeeld met de categorie (NP\S)/NP; dit leidt

In het categoriale morfologiemodel geldt voor elk morfologisch geleed lexeem dat de laatst toegevoegde functor bepalend is voor de grammaticale eigenschappen van de nieuw geconstrueerde eenheid c.q. lexeem. Voor de identificatie van het hoofd hoeft dus geen beroep te worden gedaan op een aanvullend grammaticaprincipe, zoals de RHR, want welk morfeem als hoofd fungeert volgt rechtstreeks uit de derivationele lexeemstructuur. Het morfologische hoofd correspondeert namelijk altijd met de laatst toegevoegde functor - ongeacht de positie van deze functor (dus niet per se uiterst rechts) en ongeacht de vraag of hij een waarneembare klankvorm heeft (zodat ook recht kan worden gedaan aan "exocentrische" lexemen).[147] Het categoriale morfologiemodel doet dus geen voorspellingen ten aanzien van de positie van het hoofd, maar gezien de onbetrouwbaarheid van de RHR is dit eerder een voordeel dan een nadeel. In de Nederlandse morfologie blijken functors immers zowel rechts als links van hun argument te kunnen voorkomen, zodat het handiger is om deze positie langs lexicale weg te verantwoorden. Toch is het best denkbaar dat er andere talen zijn die wel duidelijk een links- of rechtshoofdige morfologie hebben. Dit zou men dan kunnen verantwoorden door alle morfologische regels uit die taal langs compositionele weg uit een lexicaal (dus niet universeel) basisschema voor morfologische regels op te bouwen.

Het categoriale morfologiemodel is ook in staat om onderscheid te maken tussen categoriebepalende affixen (die functietype <X,Y> bezitten) en categorieneutrale affixen c.q. modificators (die functietype <X,X> bezitten). Indien de functor met een modificator correspondeert, is de gemodificeerde basis indirect (namelijk via de modificator) bepalend voor de grammaticale eigenschappen van de afgeleide eenheid. In dat geval fungeert de modificator alleen als (evt. partieel[148]) doorgeefluik van deze aan de basis gekoppelde eigenschappen. In alle andere gevallen is ten minste een deel van de doorgegeven eigenschappen van de functor zelf afkomstig. Om een en ander te verduidelijken bespreek ik nu enkele voorbeeldanalyses.

Beschouw allereerst de derivationele opbouw van het N-lexeem WERK+ING:

(10) **Lexicon**
[werk] ↔ ⟨ V, *werk*, "werk" ⟩
[-ing] ↔ ⟨ V\N, *-ing*, "-ing" ⟩
Derivatie
1. $[werk]_V + [-ing]_{V\N} = [-ing]_{V\N}([werk]_V) = [werk+ing]_N$
↔ ⟨ [werk + ing], ["werk" + "-ing"] ⟩

Volgens deze analyse bestaat *werking* uit twee morfemen, namelijk de stam WERK en het suffix -ING. De bijbehorende derivatie omvat daarom slechts één combinatiestap (met nummer 1). Alvorens deze combinatiestap kan worden uitgevoerd, moet eerst worden achterhaald wat de eigenschappen zijn van de te combineren morfemen. Voor dit doel dient het lexicon te worden geraadpleegd. Hierin wordt voor elk morfeem informatie gegeven over de categorie, de klankvorm en de betekenis (via de correspondentierelatie ↔). Op basis van de categorie kan worden vastgesteld dat -ING een functor is die een werkwoord (V) in een naamwoord (N) omzet. Bij toepassing op de V-stam WERK leidt dit tot een lexeem met de morfeemstructuur WERK+ING, de categorie N, de klankvorm *werk+ing* (= *werking*) en de betekenis "werk"+"-ing" (= "mechanisme"). De derivationele opbouw van het lexeem *bewerking* verloopt ana-

echter tot een aanzienlijke toename van de complexiteit binnen het morfologische derivatiedomein. Ik laat deze structuurdimensie (die slechts een deel van de stammen beïnvloedt) verder buiten beschouwing.

[147] Bij exocentrische lexemen (waaronder lexemen die het resultaat zijn van een reguliere conversie-operatie) correspondeert de laatst toegevoegde functor met een 0-functor c.q. klankloze functor. Dit type functor onttrekt zich aan de RHR, maar geldt in de compositionele benadering als een volwaardige functor.

[148] Zie Hoeksema (1984) voor de noodzaak van partiële hoofden.

loog, maar in dit geval is sprake van een complexe stam, te weten de stam BE+WERK, die een aparte derivatiestap vereist. Dit wordt gedemonstreerd in analyse (11).

(11) **Lexicon**

[werk] ↔ ⟨ V, *werk*, "werk" ⟩
[be-] ↔ ⟨ V/{V,N,A}, *be-*, "be-" ⟩
[-ing] ↔ ⟨ V\N, *-ing*, "-ing" ⟩

Derivatie

1. [be-] + [werk] = [be-] ([werk]) = $[\text{be-} + \text{werk}]_V$
2. [be- + werk] + [-ing] = [-ing] ([be- + werk]) = $[[\text{be-} + \text{werk}] + \text{-ing}]_N$

Nog een graadje complexer is de constructie van de stam van het werkwoord *uitwerken*, te weten UIT+[[0/GE]+WERK] (zoals gemotiveerd in H3.4.6). Omdat deze stam een tweede functor heeft, is een extra derivatiestap nodig om zijn morfotactische representatie te verantwoorden; hiernaast is er een mechanisme nodig om de vormalternantie van de functor [0/ge] te beregelen. Hiertoe moet de buitenste functor doorgeven wat de modus is van het V-lexeem; als dit de voltooide tijd ([+vt]) blijkt te zijn, dient de functor [0/ge] de vorm *ge-* te krijgen, anders de 0-vorm. Derivatie (12) laat zien hoe dit technisch kan worden opgelost.

(12) **Lexicon**

[werk] ↔ ⟨ V, *werk*, "werk" ⟩
[0/ge-] ↔ λC.⟨ V/{V,N,A}, {*0*, *ge-*}, [+V] ⟩:
if C = [+vt] **then** R_{fon}([0/ge-]) = /ge-/ **else** R_{fon}([0/ge-]) = /0/
[uit-] ↔ ⟨ V\N, uit-, "uit-" ⟩

Derivatie

1. [0/ge-] + [werk] = [0/ge-]([werk]) = [[0/ge-] + werk]
2. [[0/ge-] + werk] + $[\text{uit-}]_{\pm vt}$ = [uit-] ([[0/ge-] + werk]([±vt]))
= $[\text{uit-} + [[0/ge]_{\pm vt} + \text{werk}]]_V$
if C = [+vt] **then** R_{fon}([0/ge-]) = /ge-/ **else** R_{fon}([0/ge-]) = /0/

Beschouw tot slot de afleiding van de lexeemvorm *besproken*. Dit lexeem correspondeert met een sterk werkwoord, zodat de stam (met de structuur 0_2+[BE+SPREEK]) meerdere klankvormen kan aannemen. Dit kan technisch worden opgelost (zie (13)) door de wortel SPREEK een ondergespecificeerde representatie te geven, met een keuze tussen de vormen *spreek*, *spraak* en *sproken* (voor het gemak analyseer ik het voltooide tijd coderende suffix *-en* hier als onderdeel van de stamallomorf *sproken*). Verder moet er weer een contextvariabele worden geïntroduceerd om informatie over de V-modus op te vragen. In dit geval moet deze variabele bovendien langs twee functors, namelijk het prefix BE- (als C_1) en langs de coverte functor 0_2 (als C_2). Zo krijgt de ingebedde wortel uiteindelijk een [vt]-modus-representatie.

(13) **Lexicon**

[spreek] ↔ λC_1.⟨ V, {*spreek*, *spraak*, *sproken*}, "spreek" ⟩
if C_1 = [+spec] **then** R_{fon}(spreek) = R_{fon}(C_1)
[be-] ↔ λC_2.⟨ V/{V,N,A}, *be-*, "be-"⟩
if C_2 = [+spec] **then** C_1= C_2
[0_2] ↔ ⟨ V/V, [-], "0_2" ⟩

Derivatie

1. $[\text{be-}]_{c2}$ + [spreek] = $[\text{be-}]_{c2}$ ($[\text{spreek}]_{c1}$)
↔ ⟨ $[\text{be-} + \text{spreek}]_{c2}$, [*be-* + {*spreek*, *spraak*, *sproken*}], ["be-" + "spreek"] ⟩
2. $[\text{be-} + \text{spreek}]_{c2}$ ($[0_2]_{vt}$) =$[0_2]$ ($[\text{be-} + \text{spreek}]_{c2}$([vt]))
↔ ⟨ $[0_2 + [\text{be-} + \text{spreek}_{vt}]_V]$], [[*be-* + *sproken*], [["0_2" + be-" + "spreek"]]] ⟩

Hoewel het compositionele representatiemodel een goed alternatief biedt voor de RHR-benadering, levert dit model geen kant-en-klare oplossing voor de in sectie H3.4 besproken problemen met categorietoekenning. Zoals ik al eerder betoogde, kan men deze problemen oplossen indien men afstapt van het idee dat affixen relaties moeten leggen tussen lexemen c.q. syntactische categorieën. Hiervoor is een model nodig dat de combinatorische mogelijkheden van morfemen en lexemen in termen van morfologische distributieklassen analyseert.

3.7 Conclusie

In dit hoofdstuk heb ik betoogd dat het syntactische morfologiemodel (dat onder meer ten grondslag ligt aan de morfeemclassificatie van het Morfologisch Handboek) niet toereikend is als basis voor een integrale beschrijving van de (Nederlandse) woordbouw. Dit modeltype is namelijk niet in staat om lexicale kennis te verantwoorden en kan daardoor geen recht doen aan de morfologische relaties tussen de hierin opgeslagen eenheden. Daarom heb ik een alternatief model voorgesteld, dat uitgaat van paradigmatische distributieklassen. Deze distributieklassen geven informatie over de combinatorische mogelijkheden van het lexeem of morfeem waar ze betrekking op hebben. Hierdoor wordt een nieuwe benadering van syntactische categorieën mogelijk: deze kunnen worden geheranalyseerd als een syntactische functie binnen de distributieklasse van de lexeemvorm. Hierdoor kan eenvoudig worden verantwoord dat een lexeem uit een specifieke morfologische klasse voorspelbare syntactische en semantische eigenschappen bezit; deze volgen namelijk rechtstreeks uit de distributiecategorie van dit type lexemen. Dit idee is voor het eerst uitgewerkt in Koornwinder & Verkuyl (2000).[149] Het voorgestelde systeem biedt een uitstekende basis voor de beschrijving van paradigmatische relaties en voor de verantwoording van de lexicale selectiefenomenen uit H3.5 (waaronder allomorfie en affixpotentiatie). De hiërarchische classificatie-effecten uit H3.6 zijn eveneens goed te verantwoorden, want doordat mijn morfologische classificatiesysteem uitgaat van hiërarchisch geordende stammen die elk hun eigen derivatieparadigma bezitten, is het systeem inherent compositioneel. Al met al kan worden geconcludeerd dat mijn netwerkgebaseerde morfologiemodel aanzienlijk krachtiger is dan de syntactische benadering die ten grondslag ligt aan de affixinventarisatie in het Morfologisch Handboek en veel van de onderliggende literatuur.

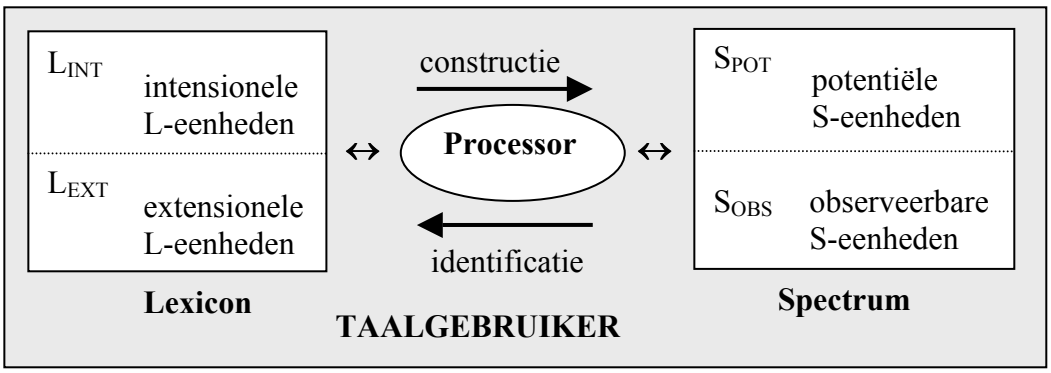
[149] Een soortgelijke gedachte ligt ten grondslag aan de aspectuele theorie van Anna Młnarczyk (2004), met wie ik ook een tijdje heb samengewerkt.

4 De L-KRING-theorie: lexicale kennis-representatie door inductieve naamgeving

4.1 Introductie

In dit hoofdstuk introduceer ik de basisprincipes van mijn formele lexicontheorie. Deze theorie berust op Lexicale KennisRepresentatie door Inductieve NaamGeving, en heet daarom de L-KRING-theorie. Hij biedt een integrale verklaring voor de identificatie, opslag en activatie van de morfologische bouwstenen die de basis vormen van de mentale kennis over de woordenschat. Hierbij geldt morfologische structuur als een bijverschijnsel van de wijze waarop het mentale lexicon woordkennis opslaat. De theorie kent een aantal formele structuurcriteria die het mogelijk maken om een willekeurige verzameling woorden langs inductieve weg van morfologische structuur te voorzien. De langs deze weg toegekende structuur vormt de theoretische basis voor de opbouw van het grammaticale regelsysteem, d.w.z. voor het cognitieve vermogen om de reeds bekende bouwstenen aan te wenden voor de systematische ("regelgebaseerde") aanmaak van nieuwe woorden.

De L-KRING-theorie is voortgekomen uit mijn streven om een representatiesysteem uit te werken dat aan de eisen van een Integraal Dynamisch Lexiconsysteem kan voldoen. Figuur 4-1 toont de algemene structuur van zo'n systeem. Het bestaat uit een processor, een spectrum en een lexicon. Zoals ik in H2.5 uiteen heb gezet, dient het lexicon van een integraal lexiconsysteem een complete dekking te bieden van het observationele woordspectrum van de gemodelleerde taalgebruiker (d.w.z. van diens complete woordenschat) en moet het deze kennis zo gecomprimeerd mogelijk opslaan. Dit hoofdstuk heeft als doel om de representatieprincipes van het lexicon te beschrijven en om een analysemethode uit te werken die identificatie van morfologische structuurkenmerken mogelijk maakt.



Figuur 4-1: De structuur van een Integraal Dynamisch Lexiconsysteem.

Zoals ik in hoofdstuk 3 heb onderbouwd, is het niet erg aannemelijk dat het mentale lexicon van Nederlandse taalgebruikers met een simpele woordenlijst of een lijst van morfologische bouwstenen (lees: morfologisch ongelede woorden en affixen) correspondeert. Het is veel waarschijnlijker dat het om een paradigmatisch georganiseerd kennisnetwerk gaat waarin een enorme hoeveelheid taaleenheden kan worden opgeslagen, of het nu morfemen, morfeemcombinaties, lexemen, samenstellingen, inflectievormen of woordgroepen zijn, en waarbij elke intern gelede eenheid zijn structuur lijkt te behouden. Dit laatste blijkt onder meer uit het feit dat mensen heel goed in staat zijn om bestaande woorden in morfologische families in te delen (d.w.z. om aan te geven welke woorden een gemeenschappelijke stam bezitten), om woorden van morfologische structuur te voorzien en om een verband te leggen tussen de interne woordstructuur en de combinatorische eigenschappen van een woord. Bovendien blijkt uit psychologisch onderzoek dat de interne woordstructuur grote invloed heeft op de

snelheid waarmee zo'n woord wordt herkend of geproduceerd. Dergelijke observaties wijzen erop dat het traditionele onderscheid tussen lexicon en productieve woordvormingsregels achterhaald is. Tot nu toe was echter onduidelijk hoe men de hier genoemde structuur-observaties formeel zou moeten verantwoorden.

Het volstaat duidelijk niet om analoog aan de opzet van een woordenboek een representatie-model te postuleren waarin het mentale lexicon in staat is om alle taalkundige eenheden die in gebruik zijn rechtstreeks op te slaan, en om hierbij te veronderstellen dat voor elk opgeslagen woord behalve de woordvorm en de woordbetekenis ook informatie over zijn morfologische structuur en zijn grammaticale eigenschappen wordt opgeslagen. Want zo'n lexiconmodel zou enorm redundant zijn, en onverklaard laten waarom woorden interne structuur bezitten. Daar komt bij dat het simpel opslaan van morfologische structuurrepresentaties niet toereikend is als verklaring voor de observatie dat woorden sneller met elkaar in verband worden gebracht naarmate ze meer structuurovereenkomsten vertonen, want in zo'n model correspondeert elke morfologische representatie met een reeks unieke representaties (c.q. instanties) van de samenstellende morfemen. Hierdoor zou er geen enkele grond zou zijn om te stellen dat de structuurrepresentaties [[ken]+baar] en [[ken]+merk] meer overeenkomst vertonen dan de woordvormen [kenbaar] en [kenmerk] (waarbij het niet uitmaakt of men de klankvorm van de woorden of hun samenstellende morfemen in spelvorm, spraakvorm of door middel van een onderliggende representatie weergeeft). Dit probleem treedt niet op indien gebruik wordt gemaakt van een productieregel, want in dat geval berusten alle morfologische afleidingen op dezelfde bronmorfemen, zodat ze in die zin formeel verwant zijn. Alleen is deze bronstructuur niet meer zichtbaar op het moment dat deze woorden geuit worden. Hierdoor zijn de meeste taalgebruikers zich er niet (of niet voortdurend) van bewust dat woorden uit morfologische eenheden bestaan. Mijn L-KRING-theorie biedt een oplossing voor deze paradox.

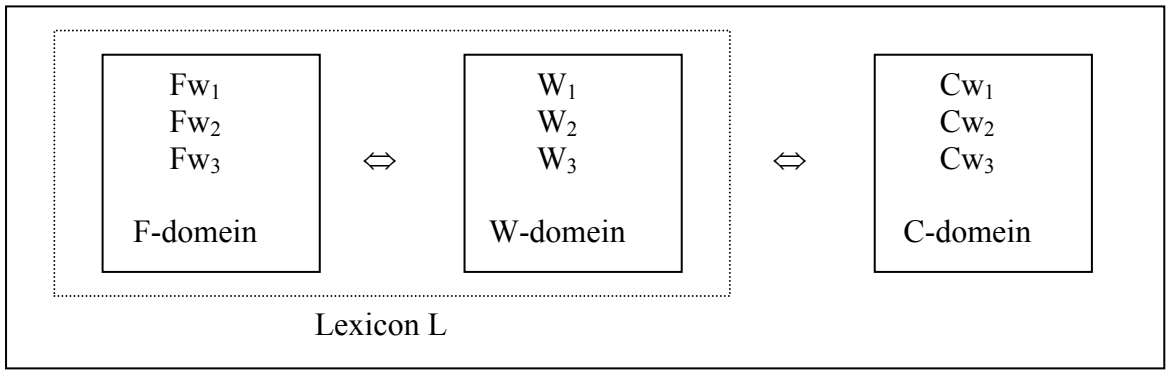
Dit hoofdstuk is als volgt opgebouwd. In H4.2 behandel ik het semantische overervingsmodel van Verkuyl (1978; 2000), te weten het L-model. In tegenstelling tot de gangbare lexiconmodellen kenmerkt het L-model zich door een lexicon dat uit een complete inventarisatie van bestaande woorden bestaat, waarbij deze woorden niet met structuurloze taaleenheden corresponderen, maar deel uitmaken van een netwerk van semantisch overervingsrelaties. Gegeven dit uitgangspunt is het een relatief kleine stap naar het idee dat de woordinterne eenheden c.q. de morfemen (en morfeemcombinaties) eveneens een lexiconstructurerende rol vervullen (wat de basisaanname is van de L-KRING-theorie). Het L-model heeft als bijkomende voordelen dat het inzicht geeft in de relatie tussen grammaticale en psychologische aspecten van woordbetekenis en dat het een formele parallel trekt tussen lexicografische en linguïstische kennismodellen.

Dankzij deze eigenschappen biedt het L-model een geschikte basis voor de opzet van mijn L-KRING-theorie. Daarom zal ik het L-model ook in detail bespreken. Hierbij zal duidelijk worden dat het L-model in veel opzichten slechts schetsmatig is opgezet en daarom nog de nodige technische en conceptuele problemen kent. Daarom zal ik enkele technische verbeteringen voorstellen die mijns inziens bijdragen aan de realisatie van de semantische doelstelling van het L-model (en die dus ook ten goede komen aan de L-KRING-theorie). Hierbij zullen structuurprincipes worden besproken die ook van belang zijn bij de uitwerking van de lexicale representatieprincipes die ten grondslag liggen aan de morfologische dimensie van het lexicon. De L-KRING-theorie zelf staat centraal in H4.3.

4.2 Het lexicale basismodel

4.2.1 Het L-model: een semantisch netwerkmodel

In morfeemgebaseerde grammaticamodellen (zie H2.3) correspondeert het lexicon meestal met een simpele lijst van morfemen en dienen alle morfologisch complexe woorden via regels uit deze kleinste bouwstenen te worden geconstrueerd. In de visie van Verkuyl (1978; 2000) kent het lexicon echter een veel complexere structuur, want semantisch gezien maken woorden deel uit van een netwerk van overervingsrelaties. Bij de beschrijving van deze netwerkstructuur gaat Verkuyl ervan uit dat woorden niet met autonome taalkundige eenheden corresponderen, maar dat er sprake is van een intermodulair fenomeen. In Verkuyl's L-model (zie figuur 4-2) corresponderen woorden namelijk met een equivalentierelatie tussen (taalkundige) woordvormen F_W (c.q. eenheden uit het fonologische domein, aan te duiden als F-domein) en (psychologische) concepten C_W (c.q. eenheden uit het conceptuele domein, aan te duiden als C-domein). Met andere woorden, in dit model behoren woorden tot de interface van het F-domein en het C-domein. Voor de duidelijkheid zal ik deze interface als W-domein aanduiden (namelijk het domein van W-functies) en de combinatie van F-domein en W-domein als het Lexicon L; want in de praktijk stelt Verkuyl woorden meestal gelijk aan een lexicale combinatie van een klankvorm met een W-functie, terwijl de betekenis van deze eenheid in het conceptuele domein c.q. C-domein moet worden opgezocht.



Figuur 4-2: Schematische weergave van Verkuyl's L-model.

Met het hier gepresenteerde model probeert Verkuyl primair antwoord te geven op de vraag hoe woordenboekdefinities zich tot encyclopedische informatie verhouden. Volgens Verkuyl is deze vraag niet alleen relevant voor de makers van woordenboeken en encyclopedieën, maar betreft het een fundamenteel probleem met betrekking tot de afbakening van linguïstische resp. lexicografische en conceptuele resp. encyclopedische kennis. In Verkuyl (2000) wordt dit L-model verder uitgewerkt.

Het L-model berust op de volgende aannames:

1) Het M-lexicon omvat alle woorden (d.w.z. relaties tussen woordvormen en concepten) die daadwerkelijk in gebruik zijn, een aanname die ook ten grondslag ligt aan de morfologiemodellen van Vennemann (1974), Jackendoff (1975) en Aronoff (1976). Maar Verkuyl geeft niet aan hoe het M-lexicon zich tot het morfologische regelsysteem verhoudt.

2) Elk woord W correspondeert met een equivalentierelatie (dus een 1:1-relatie)[150] tussen een woordvorm F_w en een lexicaal concept C_w. Verkuyl (1978) definieert deze lexicale relatie als

[150] Dergelijke 1:1-relaties zijn niet toereikend voor de representatie van namen met meerdere betekenissen of concepten met meerdere namen. Verkuyl geeft niet aan hoe dit fundamentele probleem kan worden opgelost, al zijn er twee voor de hand liggende opties, namelijk het toestaan van 1:n-relaties (of zelfs n:n-relaties) of de bundeling van woordvormen of betekenissen tot een abstractere eenheid.

volgt: $\forall x[C_w(x) \leftrightarrow$ x heet $F_w(x)]$. Hier staat dat elk element x dat kan worden aangeduid met de vorm F_w ook tot de denotatie van concept C_w behoort en vice versa. Deze analyse bouwt voort op een idee van Kripke (1972), die als eerste heeft ingezien dat woorden als eigennamen van concepten kunnen worden geïdentificeerd. Indien men dit idee omkeert, zijn persoonsnamen niets anders dan nomina die slechts op één persoon van toepassing zijn. Gewone nomina zijn immers van type <e,t>, terwijl persoonsnamen van type e zijn.

3) Lexicale concepten maken deel uit van een netwerk van semantische *overervingsrelaties* (c.q. inclusierelaties), zodat bijna elk concept als een *hyponiem* van een algemener concept (het *hyperoniem*) kan worden gedefinieerd. In de meeste gevallen zal dit volstaan, want indien men voldoende referentiële kennis heeft over de hyperoniemen, kan men ook de belangrijkste eigenschappen van het hyponiem afleiden. Voor een nadere invulling van deze concepten (zoals de modeltheoretische extensie) dient men echter het cognitieve kennisdomein te raadplegen, want het lexicon zelf bevat geen referentiële informatie, alleen overervingsrelaties.

4) De conceptrepresentaties in het C-domein kennen zowel een symbolische als een subsymbolische component. De symbolische component is nodig om de modeltheoretische extensie (bijvoorbeeld van een <e,t>-predicaat) te verantwoorden. Voor meer inhoudelijke betekenisaspecten (zoals de waarheidscondities) dient men echter de subsymbolische component te raadplegen, want deze is verantwoordelijk voor de specificatie van "prototypische" kenmerken (inclusief visuele en auditieve deelrepresentaties). In deze voorstelling van zaken zijn woorden niets anders dan wegwijzers naar referentiële informatie: ze geven toegang tot een unieke conceptrepresentatie in het cognitieve domein (via W $\Leftrightarrow$ C) en verbinden dit concept door middel van overervingsrelaties met hyponiemen en hyperoniemen.[151]

5) Lexicale relaties hebben een dynamisch karakter (in tegenstelling tot betekenispostulaten): het zijn in feite hypothetische relaties tussen woordvormen en concepten, die op elk moment kunnen worden aangepast aan de nieuwste empirische observaties. Hierdoor kan een kind stapsgewijs de betekenis van de gememoriseerde woordvormen achterhalen, namelijk door inductieve generalisatie over concrete gebruikscontexten van deze woordvormen. Verkuyl (2003) suggereert in dit verband dat het lexicon voor elke gebruikscontext van een gegeven woordvorm een index kan aanmaken die informatie geeft over de aangetroffen betekenis (p. 14). Op basis van deze indices kunnen vervolgens relevante betekeniskenmerken worden geselecteerd, die tezamen de basis vormen voor een lexicale betekenisdefinitie. Indien dergelijke vorm-betekenis-relaties weinig worden gebruikt, kunnen ze ook weer verdwijnen.

De introductie van overervingsrelaties correspondeert met een fundamentele aanpassing in het klassieke analysemodel voor woordbetekenis: want terwijl het klassieke model vereist dat elk concept in noodzakelijke en voldoende voorwaarden wordt ontleed,[152] is het in Verkuyl's model voldoende om alleen de noodzakelijke voorwaarden te identificeren, te weten de inherente (c.q. overerfbare) kenmerken. Volgens Verkuyl dient een concept C namelijk niet op basis van equivalentie ($\Leftrightarrow$), maar op basis van implicatie ($\Rightarrow$) te worden ontleed, dus als volgt: C $\Rightarrow$ X_1, X_2, X_3, etc. Hierdoor kan een concept ook in meerdere overervingsklassen tegelijk vallen. Men kan deze klassen achterhalen door woorden aan een *componentiële betekenisanalyse* te onderwerpen. Zo kan het woord *hengst* als een lexicaal hyponiem van *mannelijk* en *paard* worden ontleed (hengst $\Rightarrow$ paard, mannelijk), terwijl het woord *paard* een concept aanduidt dat minimaal de eigenschappen *dier*, *vierpotig* en *eenhoevig* moet omvatten

[151] Volgens Verkuyl (2000) is echter nog erg onduidelijk hoe deze overervingsrelaties zich tot de cognitieve representaties verhouden. Dit blijkt uit een opmerking op pag. 50: "Hoe de verbinding tussen deze neurale structuur en de pijl-informatie in de boxen tot stand komt, is op dit ogenblik voor iedereen onbekend."

[152] In een beroemd artikel over de conceptuele analyse van het concept *spel* betoogt Wittgenstein (1953) echter dat deze definitiemethode fundamenteel strijdig is met de essentie van concepten.

(paard ⇒ dier, vierpotig, eenhoevig). Een woord als *vliegenmepper* is echter veel lastiger te classificeren: het is een voorwerp met een aantal optionele eigenschappen.[153]

Ter verduidelijking van het L-model zal ik nu een concreet voorbeeld bespreken, te weten de lexicale representatie van het woord *roodborstje* (of kortweg *roodborst*). De woordvorm *roodborst* verwijst naar een concept dat kan worden omschreven als een zangvogel met een roodoranje borst. Anderzijds is de *roodborst* een ondersoort van de *zangvogels*, die zelf weer tot de *vogels* behoren. De *vogel* kan worden getypeerd als warmbloedig dier met snavel en vleugels; hij behoort zelf weer tot de *dieren*, *dieren* tot *levende organismen* enz. Aan het eind van deze reeks staan de *dingen* of nog abstracter, de *concepten*. Verkuyl (2000) spreekt in dit verband van een A-reeks; deze heeft als kenmerkende eigenschap dat de eigenschappen van het hoogste begrip automatisch "overerven" naar de ondersoorten van dit begrip. Als het waar is dat vogels altijd vleugels hebben, geldt dit bijvoorbeeld ook voor roodborstjes. Maar uit het feit dat *roodborstjes* in staat zijn om met deze vleugels te vliegen mag men niet concluderen dat vogels altijd kunnen vliegen. Zo vormen pinguïns een bekend tegenvoorbeeld.[154] Een ander kenmerk van de A-reeks is dat een begrip meer onderscheidende kenmerken bezit naarmate het verder is ingebed:

(1)

object	= vaste onderscheidbare want telbare *substantie*
organisme	= zelfstandig levend *object*
dier	= zelfstandig bewegend *organisme*
warmbloedig dier	= *dier* dat zijn eigen temperatuur regelt
vogel	= *warmbloedig dier* met vleugels
zangvogel	= *vogel* die melodieus geluid kan maken
roodborstje	= *zangvogel* met een roodoranje borst

Elke categorie in de A-reeks onderhoudt dus een inclusie-relatie (⊆) met de overkoepelende categorieën: roodborstje ⊆ vogel en vogel ⊆ dier etc. Naast de A-reeks onderscheidt Verkuyl ook een B-reeks. Deze correspondeert met hiërarchische relaties die de overervingseigenschap normaal gesproken missen (gespecificeerd als ≤), zoals *deel-geheel*-relaties. Zo bezit een fiets meestal twee wielen, maar als de fiets stuk is hoeft dit niet voor de wielen te gelden.[155]

Verkuyl (2000) argumenteert dat de A-informatie essentieel lexicaal genoemd kan worden doordat via overerving toegang wordt verschaft tot alle relevante informatie, terwijl de B-informatie encyclopedisch (c.q. cognitief) van aard is.[156] Voor de definitie van *roodborstje* is het bijvoorbeeld gebruikelijk om te zeggen: "Een roodborstje is een vogel die ..." De relatie R ⊆ V (met R = *roodborst* en V = *vogel*) garandeert nu dat R-kenmerken worden doorgekoppeld naar V-kenmerken. Op de stippels komt typisch B-informatie te staan, bijvoorbeeld dat het een vogel met een roodoranje borst betreft (dit kenmerk is namelijk niet essentieel voor vogels). In die zin analyseert Verkuyl het klassieke definitiemodel van *genus et differentiae specificae* (dat teruggaat op Aristoteles) als bestaande uit A-informatie (het *genus*) en B-informatie (de *differentiae specificae*). Het nieuwe element is dat Verkuyl deze traditionele componenten van de woorddefinitie met verschillende mathematische modelleringen in verband

[153] Zie Verkuyl (2000), p.33 e.v.
[154] Pinguïns kunnen overigens wel door het water vliegen!
[155] Anderzijds kan men wel concluderen dat als een fiets groen is geverfd, dit ook voor de meeste onderdelen zal gelden. Dit hangt samen met het feit dat er bij deze inbeddingsrelaties sprake is van een extra structuurlaag waar men soms bewust van af kan zien. Als men het over de kleur van de fiets heeft, is het *fietswiel* gewoon een (materieel) deel van de fiets, maar als men het over de werking van een fiets heeft, verandert het wiel plotseling in een onderdeel, waardoor de overervingsrelatie niet meer geldt.
[156] In dit verband suggereert Verkuyl (2000) dat woordenboeken en encyclopedieën ook met tegenovergestelde zoekstrategieën corresponderen: woordenboeken leiden de gebruiker van items naar betekenisklasse, terwijl encyclopedieën er juist op gericht zijn om een klasse tot zijn samenstellende items te herleiden.

brengt, namelijk met $\subseteq$ (inclusie) en $\leq$ (inbedding). Volgens mij bestaat er overigens een structurele relatie tussen deze twee componenten. Deze kan als volgt worden getypeerd: indien concept X een (A)-hyperoniem is van Y, dienen de B-eigenschappen (P_B) van X een subset te zijn van die van Y, dus: $Y \subset X \Rightarrow P_B(X) \subset P_B(Y)$.

4.2.2 Van L-model naar L-KRING-theorie

4.2.2.1 Introductie

In mijn optiek biedt het L-model een bruikbare basis voor lexicale kennisrepresentatie, al is het huidige model tamelijk schetsmatig opgezet. In deze sectie zal ik dit model nader proberen uit te werken door enkele fundamentele problemen aan de orde te stellen en per probleem aan te geven hoe het kan worden opgelost. Deze voorstellen zijn een eerste stap naar de ontwikkeling van mijn compositionele morfologiemodel, te weten de L-KRING-theorie. In deze sectie beperk ik me echter tot de analyse van niet-compositionele aspecten van overerving.

4.2.2.2 De representatie van n:n-relaties

Het L-model veronderstelt dat er een 1:1-relatie bestaat tussen woordvorm en concept. Empirisch gezien is dit echter geen aantrekkelijk uitgangspunt; er zijn immers vele vormen die meerdere betekenissen toestaan (= betekenisambiguïteit c.q. polysemie); zo kan de vorm *blik* zowel naar "metaal" als naar "ogen" verwijzen.[157] Omgekeerd zijn er ook vele betekenissen die door meerdere vormen kunnen worden uitgedrukt (= vormambiguïteit); zo hebben de vormen *gek* en *gestoord* dezelfde betekenis.[158] Verder zijn er woorden die meerdere stamvormen kennen, waardoor ze tegelijk vormambiguïteit en betekenisambiguïteit kunnen vertonen; zo correspondeert het woord *schieten* onder meer met de betekenissen 'snel bewegen' en 'een projectiel 'afvuren'; omgekeerd kunnen deze betekenissen zowel door de stamvorm *schiet* (tegenwoordige tijd) als door de stamvorm *schoot* (verleden tijd) worden uitgedrukt.

Men kan dit probleem oplossen door woorden rechtstreeks als een n:n-relatie tussen woordvormen en concepten te definiëren[159], namelijk als een verzameling van één of meer <C,F>-paren: $\blacklozenge(W) = \{<C,F> | <C,F> \in W\}$. In deze benadering kunnen woordvormen en concepten meerdere relaties aangaan, die ook met verschillende woorden kunnen corresponderen. De hier gegeven definitie maakt gebruik van een door ♦ gemarkeerde structuuroperator, die in dit geval de interne samenstelling van een woord W zichtbaar maakt. Deze woorddefinitie vertelt overigens niet hoe men kan vaststellen welke <C,F>-paren tot hetzelfde woord behoren; zoals ik later in dit hoofdstuk zal uitleggen, kan dit probeem alleen worden opgelost indien een paradigmatisch structuurcriterium wordt gedefinieerd (zie H4.3 en verder). Indien men onderscheid wil maken tussen bestaande en mogelijke woorden dient per woord een lexiconparameter te worden gespecificeerd, namelijk de parameter [±L]. Gegeven dit kenmerk kan het lexicon L bijna triviaal worden gedefinieerd als de verzameling woorden met het kenmerk [+L]: $\blacklozenge(L) = \{W | W \in L\}$. Wat deze definitie toevoegt is het gegeven dat er een entiteit L bestaat die een equivalentieklasse definieert met betrekking tot de in L aanwezige woorden, wat vergezeld kan gaan met de specificatie van metakenmerken als de taalnaam, de bijbehorende taalgebruikers en de geïnventariseerde periode.

4.2.2.3 De representatie van woordvormen

Het L-model introduceert woordvormen als onderdeel van een lexicale relatie met een conceptpointer. Indien men echter meerdere betekenissen per woordvorm toelaat, is het efficiën-

[157] In woordenboeken worden deze betekenisvarianten typisch met subindexen aangeduid, bijv. ^{1}blik en ^{2}blik.
[158] In woordenboeken treft men vaak kruisverwijzingen aan bij dergelijke synoniemen, bijv. *gek* = *gestoord*.
[159] Dit heeft als bijkomend voordeel dat men niet langer afhankelijk is van de problematische veronderstelling dat het element x tegelijk argument van een vormpredicaat ("heet F_w") en een betekenispredicaat (C_w) kan zijn.

ter om de woordvormen elders te introduceren, d.w.z. los van de bijbehorende woordrelaties, en deze woordvormen te activeren door gebruik te maken van een woordvormpointer. Deze analyse heeft als bijkomend voordeel dat woordvormen analoog aan concepten kunnen worden behandeld, namelijk als cognitieve entiteiten die op een subsymbolische wijze zijn opgeslagen, inclusief contextspecifieke realisatiekenmerken; hierdoor kan makkelijker worden verantwoord dat er bij lees- en schrijfprocessen allerlei vormen van patroonherkenning meespelen.

Deze analyse biedt ook uitkomst voor een hieraan gerelateerd probleem. Het L-model gaat er namelijk van uit dat woorden slechts één representatievorm kennen: de spelvorm. Maar naast een spelvorm kennen woorden ook een uitspraakrepresentatie, en mogelijk ook nog tussenvormen, zoals een syllabische representatie. Tenzij men aanneemt dat al deze representaties van een onderliggende vorm (zoals een abstracte uitspraakrepresentatie) kunnen worden afgeleid (wat volgens mij een zeer problematische aanname is; zie hoofdstuk [2]), dient het lexicon dus meerdere representaties per woordvorm op te slaan. Het L-model beperkt zich echter tot equivalentierelaties tussen spelvormen en conceptpointers. Dit probleem kan worden opgelost door per woordvorm (F_i) een bundel van vormrepresentaties te introduceren, waarbij elke bundel onder meer uit een ortografische representatie (F_i,orth) en een fonologische representatie (F_i,fon) dient te bestaan. Elk van deze representaties zou via een vormpointer moeten worden geactiveerd, want veel spelvormen en uitspraakrepresentaties kunnen (los van elkaar) door meerdere woorden worden gebruikt, zodat het efficiënter is om deze representatievormen onafhankelijk van elkaar te kunnen aanroepen en per vorm nadere selecties te maken (zoals de selectie van het juiste klemtoonpatroon). Dit idee wordt hieronder gedemonstreerd voor de woorden *gek* en *gestoord*, die allebei één (hoofd)vorm (in feite vormbundel) en twee (hoofd)betekenissen omvatten:

(2)

cognitieve vormrepresentaties	lexicale relaties		cognitieve conceptrepresentaties
[F1,orth] = |gek|	W_1 = {<F1,C3>,	→	C3 = "raar, vreemd"
[F1,fon] = /gek/	<F1,C5>}	→	C5 = "geestesziek"
[F2,orth] = |gestoord|	W_2 = {<F2,C5>,	→	
[F2,fon] = /gestoord/	<F2,C8>}	→	C8 = "afgeleid"

4.2.2.4 De lexicale representatie van concepten

In tegenstelling tot het klassieke *genus et differentiae specificae* model gaat Verkuyl (2000) ervan uit dat er meerdere hyperoniemen per concept kunnen worden onderscheiden, dus dat A-reeksen naar boven kunnen vertakken.[160] Verkuyl laat echter niet zien hoe dit idee technisch kan worden uitgewerkt. Ik zal nader op deze kwestie ingaan aan de hand van het concept *roodborstje*. Indien men hier een componentiële analyse op loslaat, krijgt men onder meer de volgende B-eigenschappen te zien (ten opzichte van het hoofdgenus *dier*):

(3)
- B1: het roodborstje is warmbloedig
- B2: het roodborstje heeft vleugels
- B3: het roodborstje kan vliegen
- B4: het roodborstje legt eieren (eig. eitjes)
- B5: het roodborstje kan melodieus fluiten
- B6: het roodborstje vertoeft vaak in tuinen

[160] Elke B-eigenschap van een gegeven A-concept voldoet namelijk triviaal aan de overervingseis dat zijn B-eigenschappen een subset zijn van de B-eigenschappen van dit A-concept.

B7: het roodborstje is een overwinteraar
B8: het roodborstje heeft de grootte van een vuist(je)
B9: het roodborstje heeft een roodoranje borst
B10: het roodborstje leeft onder meer in Nederland

Elk van deze eigenschappen correspondeert met een hyperoniem van *roodborstje*, namelijk met een predicaat van het type P = {X| X is een dier met eigenschap(cluster) Y}, waarbij X onder meer voor *roodborstje* kan staan en Y voor een of meer van de hierboven opgesomde eigenschappen. Volgens Verkuyl dient *roodborstje* echter als ondersoort van de *zangvogel* te worden beschouwd, want dit concept omvat een groot deel van de hierboven opgesomde eigenschappen, namelijk eigenschap B1-B5; de resterende eigenschappen zijn dan automatisch soortspecifiek. In deze analyse kan *roodborstje* dus als 'zangvogel met de eigenschappen B6-B10' worden gedefinieerd. Nu is dit op zichzelf geen verkeerde analyse, maar ten aanzien van het genus *dier* zijn ook andere (niet-genetische) indelingen denkbaar die dwars door de klasse van vogels heen lopen, bijvoorbeeld *waterdieren* (zoals meeuwen en eenden) versus *landdieren* (zoals roodborstjes en adelaars) of *dagdieren* (zoals nachtegalen) versus *nachtdieren* (zoals uilen).

Deze flexibiliteit impliceert dat een soortaanduiding als *vogel* op een tamelijk arbitraire selectie van kenmerken berust (iets wat ook blijkt uit het bestaan van gemengde diersoorten, zoals *loopvogels*, *vleermuizen* en *walvissen*). Bovendien zijn er andere, niet dier-gebonden perspectieven denkbaar, zoals een classificatie op fysieke omvang, populatiegrootte (wel/ niet bedreigd), leefgebied of eetbaarheid. Niet-biologische concepten zijn nog moeilijker te classificeren, want in dit geval is geen natuurlijk (genetisch) ordeningsprincipe beschikbaar. Zo is onduidelijk of men een *lamp* als een gebruiksvoorwerp, een meubelstuk of een lichtbron moet classificeren; dit lijkt ook mede af te hangen van de soort lamp (*zaklamp*, *hanglamp* of *toneellamp*). En dan heb ik het nog niet eens over abstractere concepten als *liefde*, *muziek, lotsverbondenheid en wil*.[161, 162]

Ik zal nu aangeven hoe het hier gesignaleerde probleem kan worden opgelost. Deze oplossing berust op het idee dat een concept als *dier* langs meerdere structuurdimensies (d_i) in subklassen kan worden onderverdeeld, bijvoorbeeld:

(4) d1: 'biologische familie' (bijv. *zoogdieren*, *vogels*, *vissen* en *insecten*),
d2: *landdier* / *waterdier*
d3: *dagdier* / *nachtdier*
d4: *trekdier* / *standdier*
d5: 'leefgebied' (bijv. *Europa*, *Amerika*, *Afrika*, *Azië*)

Het is ook mogelijk om nieuwe subklassen te construeren door structuurdimensies te combineren: hoe meer structuurdimensies men combineert, hoe specifieker de subklassen. Zo levert combinatie van d1 en d2 onder meer de subsoorten *landvogel* en *watervogel* op. Deze kan men vervolgens weer onderverdelen in *dagdieren* versus *nachtdieren,* de resulterende subsoorten in *standdieren* versus *trekdieren* enz. Indien gewenst kan men hier net zolang mee doorgaan totdat men bij de ondersoort *roodborstje* is aangekomen. Maar dit laatste concept kan (evenals zijn hyperoniemen) ook langs andere weg worden geconstrueerd, bijvoorbeeld als een ondersoort van de Nederlandse landdieren. Volgens mij dienen individuele exemplaren

[161] Zie Verkuyl (2003) voor een nadere beschouwing over de concepten *lotsverbondenheid* en *onwil*.
[162] Dit is niet alleen een theoretisch probleem. Bij VDL is namelijk een semantisch classificatiesysteem ingevoerd waarbij voor elke woordvorm is nagegaan wat zijn directe hyperoniem is. Deze doelstelling leverde in de praktijk talloze dilemma's op, onder meer door ambiguïteit van de klassenamen en door de eis dat slechts één hyperoniem per woordvorm (i.p.v. betekenis) mocht worden aangewezen (zie bijv. NRC-H, W&O, 3 mei 2003).

op soortgelijke wijze te worden gedefinieerd:[163] roodborstje r_{23} kan bijvoorbeeld als een unieke ondersoort van de soort *roodborstje* worden beschouwd. Zo'n exemplaar kan op zijn beurt weer in tijdstokens worden onderverdeeld (d.w.z. unieke subinstanties met betrekking tot de tijdsindex of de verrichte handeling). Meer in het algemeen geldt: hoe groter het aantal onderscheiden kenmerken, hoe groter de soortresolutie.

De keuze van het constructiepad komt doorgaans ook tot uitdrukking in de wijze waarop men naar een concept verwijst (inclusief de bijbehorende inferentie-effecten): zo kan roodborstje r_{23} niet alleen als als *dit roodborstje* worden aangeduid, maar ook als *deze zangvogel, dit Nederlandse landdier* of gewoon *dit dier* (vgl. Verkuyl, 1984). Maar deze aanduidingen kunnen ook op soortniveau worden geïnterpreteerd: zo kan een predicaat als *deze zangvogel* zonder enig probleem naar de soort *roodborstje* verwijzen. Om dezelfde reden is een vraag van het type 'hoeveel vogels heb je vandaag gezien?' niet goed te beantwoorden zonder dat de vragensteller aangeeft op welk niveau er gekwantificeerd moet worden en welk classificatiesysteem hierbij moet worden gehanteerd.

Deze observaties onderstrepen het belang van een conceptueel representatiesysteem waarin exemplaren en soorten deel uitmaken van een (multidimensionaal) continuüm. Want in een dergelijk representatiesysteem zijn per concept evenveel representatiemogelijkheden beschikbaar als er constructiepaden zijn, en dus evenzoveel omschrijvingsmogelijkheden. Toegepast op het concept *roodborstje* (gedefinieerd als een dier met de kenmerken B1-B10; zie boven) kan men bijvoorbeeld de volgende representaties construeren:

(5) [ROODBORST, c1] = dier (met kenmerken B1-B10)
[ROODBORST, c2] = vogel (met kenmerken B5-B10)
[ROODBORST, c3] = zangvogel (met kenmerken B6-B10)
[ROODBORST, c4] = tuindier (met kenmerken B1-B5, B7-B10)
[ROODBORST, c5] = Nederlands tuindier (met kenmerken B1-B5, B7-B9)

De hier gepresenteerde analyse is niet (zonder meer) compatibel met het modeltheoretische onderscheid tussen predicaten (c.q. soorten) en elementen (c.q. instanties). Ik zal dit uitleggen aan de hand van het schema in figuur 4-3. Dit schema laat twee mogelijke analyses zien van de wijze waarop roodborstje r_{23} (en het hierin ingebedde tijdstoken t_5) zich tot het soortniveau verhoudt, namelijk mijn eigen (conceptuele) analyse en de modeltheoretische (MT) standaardanalyse (bijv. Gamut, 1991), die ook ten grondslag aan het model van Verkuyl (2000).

conceptuele benadering	klasseniveau	modeltheorie
gewerveld dier	(hoofdsoort)	$\{x \mid P_1(x)\}$
Nederlandse zangvogel	(familie)	$\{x \mid P_2(x)\}$
roodborstje	(basissoort)	$\{x \mid P_3(x)\}$
roodborstje II	(ondersoort)	$\{x \mid P_4(x)\}$
..., r_{23}, ...	(exemplaar)	x_{23}
..., t_5, ...	(tijdstoken)	$x_{23:t5}$

Figuur 4-3: De conceptuele relatie tussen exemplaren en soorten.

[163] Deze visie berust op de kwantificatietheorie die door mij uiteen is gezet in Koornwinder (1997, ms.).

In de MT-benadering, die op hogere-orde predicatenlogica berust, verwijzen concepten (c.q. predicaten) naar typelogische verzamelingen in een domein D, namelijk verzamelingen die kunnen worden getypeerd in termen van elementen e en waarheidswaardes t, bijvoorbeeld <et>, <et,t> of <et,<et,t>>. Hierbij corresponderen zelfstandige naamwoorden doorgaans met een eenvoudige elementverzameling met type <et> (namelijk de verzameling {x| P(x)}, waarbij P met een predicaat correspondeert). Voor de elementen zelf wordt echter geen modeltheoretisch predicaat gereserveerd; in plaats daarvan wordt aangenomen dat men elementen uniek kan definiëren door alle predicaten te specificeren die op het element kunnen worden toegepast (door ze systematisch in verzamelingen in te delen).

In deze benadering bestaat dus een fundamenteel contrast tussen soorten en exemplaren, want soorten staan wel conceptuele analyse toe, maar elementen kunnen alleen indirect worden gedefinieerd, namelijk door alle predicaten op te sporen die erop van toepassing zijn. Omgekeerd kunnen predicaten alleen op elementen (en elementverzamelingen) worden toegepast; hierdoor kunnen predicaten slechts indirect op soortniveau worden toegepast, namelijk door de bijbehorende verzameling te specificeren.[164]

In mijn optiek biedt een puur conceptuele analyse (zoals weergegeven in de linkerkolom), verder aan te duiden als het C-model, grote voordelen boven de modeltheoretische benadering. Allereerst heeft het C-model geen kunstmatig onderscheid nodig tussen een soortniveau (boven de stippellijn) en een elementniveau (beneden de stippellijn). In plaats daarvan is er sprake van een multidimensionaal continuüm waarbij elk concept als een nadere specificatie van een algemener concept kan worden gedefinieerd. Hierbij correspondeert het exemplaarniveau met concepten waarvan de exemplaar-dimensie is geactiveerd (bijvoorbeeld op basis van een tijd-positie-criterium).

Dankzij deze eigenschappen biedt het C-model een fundamentele verklaring voor het feit dat predicaten zowel op "exemplaren" (inclusief tijdstokens) als op "soorten" kunnen worden toegepast. Ten tweede leidt het C-model automatisch tot een hiërarchische ordening van predicaten; in de MT-benadering is deze ordening slechts indirect zichtbaar, namelijk door de bijbehorende verzamelingen te inspecteren (die inclusie zullen vertonen). Ten derde is geen apart representatiedomein nodig voor modeltheoretische extensies, mits voor elk predicaat onderscheid kan worden gemaakt tussen verschillende predicatiedimensies, bijvoorbeeld P^{D1} voor soorten, P^{D2} voor subsoorten, P^{D3} voor exemplaren enz. Ten vierde is de multidimensionale opzet van het C-model noodzakelijk om recht te doen aan de rijke structuur van concepten (zie de voorgaande uitleg).

4.2.2.5 De cognitieve representatie van concepten

In het L-model bestaat geen structureel verband tussen de overervingsstructuur van de lexicale eenheden en de cognitieve representatie van de concepten. Dit impliceert dat de overervingsstructuur van het L-model slechts een beperkte functie heeft, namelijk het signaleren van een inclusieverhouding tussen onafhankelijk (want subsymbolisch) gerepresenteerde concepten.[165] Het zou echter logischer zijn als de overervingsrelaties tussen lexicale concepten een structurele bijdrage leveren aan hun cognitieve representatie. Indien er slechts één hyperoniem per concept bestaat is dit ook heel eenvoudig te formaliseren. Men dient dan

[164] Verkuyl (1993) presenteert bijvoorbeeld een modeltheoretische analyse van de ambiguïteit in zinnen als '*Rembrandt verkocht drie etsen*, waarin het zowel om token-etsen als om type-etsen kan gaan.

[165] De hierbedoelde overervingsrelaties kunnen niet rechtstreeks op de modeltheoretische extensie van de predicaten worden gebaseerd, want er zijn vele predicaten waarvan de extensies bij toeval een inclusieverhouding vertonen, maar waarvan iedereen weet dat bij uitbreiding van het observatiedomein toch tegenvoorbeelden kunnen worden aangetroffen; het negeren van deze kennis leidt meestal tot stereotypering (vgl. Verkuyl, 2000a).

het hyperoniem als functor (F) te beschouwen en het te definiëren concept als argument (X). Het enige wat X dan hoeft te doen is een nieuw (door X gedefinieerd) kenmerk toevoegen aan de eigenschappen die reeds door F worden geïntroduceerd. Desgewenst kan deze procedure recursief worden toegepast. Zo kent het concept "warmbloedig dier" vaste dimensies als uiterlijk, bewegingsmogelijkheden en voortplantingsgedrag. Gegeven dit concept kan men soorten construeren door een of meer dimensies van een specifieke eigenschap te voorzien.

(6)
warmbloedig dier ⇔ concept Y1: D1 = ?, D2 = ?
↓ + "met vleugels"
vogel ⇔ concept Y1: D1 = met vleugels, D2 = ?
↓ + "melodieus fluitend"
zangvogel ⇔ concept Y1: D1 = met vleugels, D2 = melodieus fluitend

Men kan bijvoorbeeld de diersoort *vogel* construeren door het als functor (F) op de eigenschap "met vleugels" (X) toe te passen.[166] Het resulterende concept kan vervolgens als basis dienen voor de constructie van de subsoort *zangvogel* door het als functor op de eigenschap "melodieus fluitend" toe te passen. Dit wordt hierboven gedemonstreerd.

Deze analyse is echter te simpel, want zoals ik onder punt d) heb betoogd kan elke combinatie van conceptgerelateerde eigenschappen als hyperoniem van dit concept gelden, zodat er in de praktijk tientallen of zelfs honderden hyperoniemen per concept mogelijk zijn. Ik ga er daarom van uit dat concepten ook meerdere conceptdefinities kunnen krijgen, namelijk evenveel als er nodig zijn om alle concepttoepassingen te kunnen verantwoorden: concepten corresponderen dus met een equivalentieklasse van conceptconstructies. Dit kan formeel worden verantwoord door equivalente conceptconstructies een identieke index te geven, bijvoorbeeld de index RB (van roodborst). Deze index kan bijvoorbeeld een verband leggen tussen conceptconstructies op basis van de hyperoniemen *zangvogel*, *tuinvogel* en *standvogel* (die zelf ook weer met een geïndexeerde reeks concepten corresponderen):

(7) ROODBORST ⇔ {$ZANGVOGEL_{RB}$, $TUINVOGEL_{RB}$, $STANDVOGEL_{RB}$}

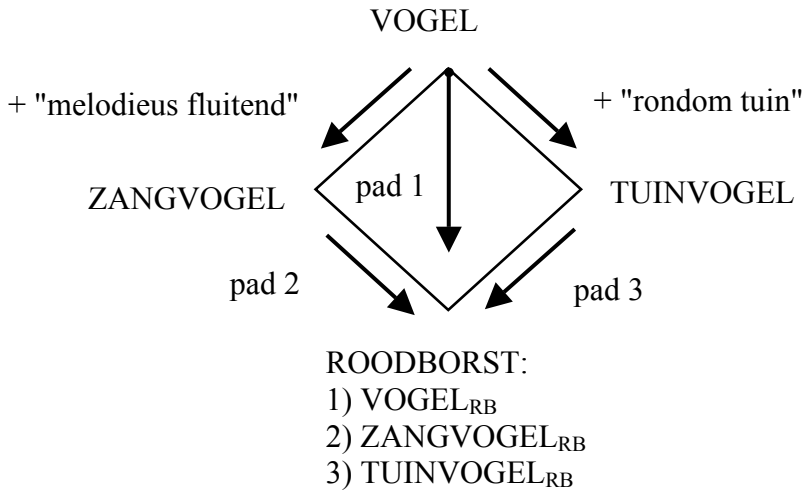
In deze analyse corresponderen de conceptconstructies $ZANGVOGEL_{RB}$, $TUINVOGEL_{RB}$ en $STANDVOGEL_{RB}$ met verschillende perspectieven op het concept ROODBORST; ze hebben echter dezelfde extensie, want de index RB zorgt ervoor dat elk van de onderliggende hyperoniemen precies tot de verzameling van roodborstjes wordt beperkt.[167] Ik zal deze index-gebaseerde analyse nader toelichten aan de hand van de representatie in figuur 4-4.

Deze representatie toont een aantal constructiepaden van het concept ROODBORST (weergegeven door pijlen), namelijk de constructiepaden vanuit VOGEL en vanuit de VOGEL-hyponiemen ZANGVOGEL en TUINVOGEL. Elk van deze constructiepaden resulteert in een conceptconstructie van ROODBORST, namelijk de onder ROODBORST opgesomde RB-concepten. Deze RB-concepten bestaan uit dezelfde kenmerken, namelijk VOGEL +

[166] Hier zou nog de eigenschap "eierleggend" aan toe moeten worden gevoegd, ter uitsluiting van vleermuizen.

[167] Deze index-gebaseerde analyse dient niet verward te worden met een predicaatlogische analyse waarin het predicaat ROODBORST als de doorsnede van zijn hyperoniemen wordt gedefinieerd. Maar in zekere zin zijn het notationele varianten, want indien men de intensie van een concept C als de vereniging van al zijn conceptconstructies (en de daarin vervatte eigenschappen) kan definiëren, zal de extensie van dit concept identiek zijn aan de doorsnede van de extensies van de onderliggende hyperoniemen; indien deze doorsnede desondanks elementen bevat die niet onder concept C vallen, dan zijn er waarschijnlijk ook conceptconstructies te vinden die nog geen deel uitmaken van C's intensionele definitie.

"melodieus fluitend" (= F1) + "rondom tuin" (= F2), maar deze kenmerken zijn op verschillende manieren geclusterd. Pad 1 leidt het concept ROODBORST namelijk direct van het concept VOGEL af, terwijl pad 2 en pad 3 van een hyponiem van VOGEL uitgaan, te weten ZANGVOGEL en TUINVOGEL, waardoor ze één eigenschap minder hoeven toe te voegen dan pad 1. Doordat elke eigenschap met een aparte constructiedimensie correspondeert, kan elke combinatie van eigenschappen als een aparte hyponiem worden gedefinieerd, zonder dat dit extra representatiekosten met zich meebrengt (afgezien van de toevoeging van een clusterindex). Hoewel in dit voorbeeld slechts twee eigenschappen zijn gebruikt kan het model eindeloos met nieuwe eigenschappen worden uitgebreid. Dit leidt tot een proliferatie van clusteringsmogelijkheden, maar slechts een klein deel van deze mogelijkheden zal daadwerkelijk als concept worden gebruikt. Per concept kan bovendien exact worden bijgehouden hoe vaak het als uitgangspunt dient voor een constructiepad, wat verklaart hoe men intuïties kan hebben over de meest waarschijnlijke conceptconstructie van een concept als ROODBORST.



Figuur 4-4: De cognitieve constructie van het concept ROODBORST.

Volgens mij biedt de hier voorgestelde analyse een fundamentele verklaring voor het feit dat het lexicon een overervingsstructuur vertoont: deze speelt immers een cruciale rol in de opbouw van de cognitieve representaties. Ik geloof dan ook dat mijn representatiesysteem een aantrekkelijk alternatief biedt voor het integrale opslagmodel van Verkuyl.

4.2.2.6 De representatie van morfologische structuur

Verkuyl (2000) geeft niet aan hoe het overvingsprincipe zich tot compositionele woordvorming verhoudt. Hoewel hij kort aandacht besteedt aan de interpretatie van samenstellingen (p. 79 e.v.), wordt geen systeem gepresenteerd waarin de semantische kenmerken van samenstellingen langs compositionele weg uit hun interne constituenten kunnen worden afgeleid. In plaats daarvan betoogt Verkuyl dat vrijwel alle samenstellingen een gelexicaliseerde betekenis hebben, en dat de hierin aanwezige redundantiepatronen slechts een marginale rol spelen bij de productie en interpretatie van nieuwe samenstellingen. Toch valt hier heel wat meer over te zeggen. Meer in het algemeen roept het L-model allerlei vragen op met betrekking tot de interactie tussen morfologie en betekenisopbouw en de cognitieve structuur van concepten.[168] Zo is onduidelijk hoe woordvormen als *spreken*, *spreker* en *spraak* zich tot elkaar verhouden in termen van overervingsrelaties. In de volgende sectie presenteer ik een nieuw representatiemodel (het L-KRING-model) dat antwoord kan geven op dit soort vragen.

[168]In andere publicaties heeft Verkuyl echter interessante voorstellen gedaan met betrekking tot de morfologische aspecten van woordbetekenis (cf. Verkuyl (1993)) en de structuur van concepten (cf. Zwarts & Verkuyl (1991)).

4.3 De representatieprincipes van de L-KRING-theorie

4.3.1 Introductie

In deze sectie zet ik de centrale principes van mijn L-KRING-theorie uiteen. Met deze theorie beoog ik concreet invulling te geven aan mijn bouwplan voor een Integraal Dynamisch Lexiconsysteem.[169] Hiertoe wordt de in hoofdstuk 3 ontwikkelde visie op de Nederlandse woordbouw verder uitgebouwd en geformaliseerd. De resulterende theorie geeft inzicht in de psychologische functie van morfologische structuur, verklaart hoe lexicaal opgeslagen woorden aan hun morfologische structuur komen en legt uit hoe deze informatie benut wordt voor het genereren en analyseren van nieuwe woorden.

In mijn visie is morfologische structuur een epifenomeen van het streven om bestaande woorden zo gecomprimeerd mogelijk in het lexicon op te slaan, met als extra conditie dat er geen informatie verloren mag gaan. Om dit idee te onderbouwen heb ik een algoritme ontworpen dat automatische identificatie van gemeenschappelijke bouwstenen mogelijk moet maken (zoals morfemen).[170] Zulke bouwstenen worden door lexicale indexen gerepresenteerd. De relatie tussen indexen en bouwstenen is vergelijkbaar met de verhouding tussen letters en hun onderliggende klanken: want net als letters hebben indexen een verwijsfunctie. En net zoals letters tot woordvormen kunnen worden samengevoegd, kunnen indexen worden samengevoegd tot indexcombinaties. Hierdoor wordt het mogelijk om morfologisch complexe woorden als een lexicale combinatie van morfeemindexen te definiëren.

De hier beschreven representatiemethode leidt tot een hiërarchisch gestructureerd lexicon, d.w.z. een op overerving gebaseerd representatiesysteem waarbij de eigenschappen van complexe kenniseenheden (zoals gelede lexemen) altijd in termen van minder complexe kenniseenheden (zoals morfemen) zijn gedefinieerd. In een dergelijk lexicon zijn complexe taaleenheden op twee manieren toegankelijk, namelijk als zelfstandige lexicale eenheid met niet-afleidbare informatie over frequentie en idiosyncratische eigenschappen, en als (semi)-compositioneel product van twee of meer basiseenheden. Hierbij is de gebruiksfrequentie van de basiseenheden medebepalend voor de snelheid waarmee de complexe eenheden worden herkend, conform recente inzichten uit psycholinguïstisch onderzoek (bijv. Baayen, Dijkstra & Schreuder (1997) en Taft (1994)).

In vergelijking met bestaande modellen brengt de indexgebaseerde representatiewijze van de L-KRING-theorie grote voordelen met zich mee. Ik zal dit toelichten door drie alternatieve benaderingen te bespreken. In tegenstelling tot deze alternatieve modellen kenmerkt de L-KRING-theorie zich door een optimale balans tussen complete kennisverantwoording, efficiënt ruimtegebruik en snelle toegankelijkheid. Bovendien kent dit model een inductief analysemechanisme, waardoor de morfologische structuurdimensie inherent leerbaar is.

Het eerste alternatief is om het lexicon alleen te gebruiken voor de opslag van (ongelede) morfemen, een benadering die reeds door Bloomfield (1933) werd voorgesteld en veel navolging heeft gekregen in generatieve en categoriale morfologiemodellen (zie ook H2.2). Hoewel deze opzet tot een economischer gebruik van de opslagruimte leidt, zijn er ernstige nadelen aan verbonden. Zo vermindert de toegankelijkheid van morfologisch complexe woorden, want deze moeten steeds opnieuw worden afgeleid. Bovendien brengt deze opzet veel informatieverlies met zich mee, want het is niet meer mogelijk om na te gaan welke

[169] Zoals in hoofdstuk 2, sectie 5, uiteen werd gezet, kenmerkt een IDL-systeem zich door een lexicon met een morfologisch gestructureerd netwerk van bestaande woorden. Hierbij worden de morfologische structuurklassen niet aan een abstract regelsysteem ontleend, maar langs inductieve weg uit de opgeslagen taaleenheden afgeleid.

[170] In deze studie beperk ik me tot een conceptuele beschrijving van dit principe. Het is mijn bedoeling om dit principe in een computationeel (en psycholinguïstisch) leeralgoritme te vertalen.

woordafleidingen reeds eerder zijn voorgekomen, wat hun gebruiksfrequentie is, in welke contexten ze zijn gebruikt, wat voor uitspraakvarianten er bestaan en welke betekenissen eraan kunnen worden toegekend. In de L-KRING-theorie kunnen deze kenmerken via het lexicon worden verantwoord.

Het tweede alternatief is een model waarbij het lexicon alle bestaande woorden integraal opslaat (evt. inclusief uitspraak- en betekenisvarianten), zoals is voorgesteld door Jackendoff (1975) en Aronoff (1976). Maar hoewel deze aanpak een maximale kennisdekking oplevert, kost deze ongestructureerde opslagwijze relatief veel ruimte,[171] terwijl het (opnieuw) tot een verminderde toegankelijkheid leidt. Want door de grotere omvang van het lexicon en het ontbreken van interne woordstructuur kost het meer tijd om de opgeslagen woorden terug te vinden. De L-KRING-theorie vermijdt dit probleem door de lexicale representaties zo veel mogelijk intern te structureren.

Het derde alternatief bestaat uit een hybride combinatie van de voorgaande twee modellen. In het concurrentiemodel van Baayen, Dijkstra & Schreuder (1997) wordt bijvoorbeeld aangenomen dat morfologisch complexe woorden langs twee verschillende wegen geactiveerd kunnen worden, namelijk rechtstreeks via activatie van een lexicaal opgeslagen woordvorm of indirect door activatie van de samenstellende morfemen. In deze benadering bestaat er geen compositioneel verband tussen de lexicale representatie van het morfologisch complexe woord en de bijbehorende morfeemcombinatie, wat impliceert dat de morfeemrepresentatie overbodig is. Men zou dit kunnen oplossen door morfologisch complexe woorden als een vaste morfeemcombinatie op te slaan. Maar in dat geval zou elke morfeemtoepassing met een nieuw morfeem corresponderen, wat impliceert dat het vastleggen van de morfeemstructuur geen enkel representatie- of activatievoordeel zou opleveren. In de L-KRING-theorie wordt dit probleem omzeild door woordinterne bouwstenen als indexen te representeren.

De L-KRING-theorie biedt dus belangrijke voordelen (voor zover de beoogde eigenschappen waargemaakt kunnen worden). Om te beginnen kan het alle kennis over de bestaande woordenschat verantwoorden. Ten tweede wordt deze kennis efficiënter opgeslagen dan bij woordgebaseerde lexiconmodellen, want door het gebruik van indexen hoeven de onderliggende representaties slechts eenmaal te worden gedefinieerd. Ten derde verklaart deze representatiewijze de invloed van morfeemfrequentie op de activatiesnelheid van morfologisch complexe woorden. Ten vierde biedt deze opslagwijze een formele basis voor de identificatie van patronen die ten grondslag liggen aan de productie en interpretatie van nieuwe woorden. Ten vijfde garandeert de indexbenadering consistentie bij de toekenning van woordkenmerken, wat vooral van belang is met het oog op taaltechnologische toepassingen. Tot slot werpt de L-KRING-theorie nieuw licht op de fundamentele mechanismes van kindertaalverwerving, want dankzij het inductieve algoritme voor woordanalyse is het lexicon in staat tot zelfstandige ("unsupervised") identificatie van morfologische structuurkenmerken.

[171] Indien de woordgebaseerde representaties geen interne bouwstenen kennen, is elk opgeslagen woord een soort blackbox die niet intern kan worden geanalyseerd. Aan de andere kant zijn redundantieregels wel in staat om fonologische en semantische kenmerken te identificeren. Dit wijst erop dat Jackendoff's woordrepresentaties wel degelijk interne bouwstenen hebben, namelijk semantische en fonologische eenheden. Deze eenheden kunnen gelijk worden gesteld aan lexicale indexen, aangezien het slechts symbolen (bijvoorbeeld grafemen) zijn waarvan de eigenschappen elders worden gespecificeerd, namelijk in een impliciet lexicon met semantische resp. fonologische eenheden. Hieruit volgt dat woordgebaseerde modellen uitgaan van een lexicon dat wel klankeenheden en betekeniseenheden kan specificeren, maar geen morfemen; ik acht deze discontinuïteit ongewenst (vgl. de discussie in Margolis & Laurence (1999)).

4.3.2 De architectuur van het lexicon

Het lexicon van de L-KRING-theorie kenemrkt zich door een hiërarchisch netwerk van intern gestructureerde kenniseenheden. Dankzij deze netwerkstructuur kan elke lexicale eenheid in beginsel uit kleinere eenheden worden opgebouwd zonder dat deze bouwstenen elke keer opnieuw hoeven te worden gedefinieerd. Deze informatie kan namelijk worden overgeërfd van de lexicale ingang waar de betreffende bouwsteen zelf wordt gedefinieerd. Dit is mogelijk door de aanname dat lexicale eenheden, waaronder woorden en woordstammen, uit indexen, c.q. (bundels van) namen, zijn opgebouwd. Hierbij correspondeert elke index met een kopie van een kenniseenheid die elders in het lexicon wordt gedefinieerd. Bij de modellering van dit systeem kan elke index worden genoteerd als een bundel van één of meer kenmerkende eigenschappen.[172] Zo kan het segment *tover* in *betovering* eenduidig worden getypeerd via de indexdefinitie [<citatievorm> tover; <klasse> morfeem, wortel; <fon-kenmerken> inheems, twee segmenten, el-uitgang; <betekenis> "langs onverklaarbare weg transformeren"]. Deze indexkenmerken kunnen ook worden gebruikt om selectierestricties te formuleren. De door mij voorgestelde indexeringsmethode is enigszins vergelijkbaar met het gebruik van padnamen in de lexicale representatietaal DATR (Evans & Gazdar, 1996).

Binnen het L-KRING-lexicon kunnen drie fundamenteel verschillende kennismodaliteiten (c.q. representatielagen of tiers) worden onderscheiden, namelijk een fonologische modaliteit (onder te verdelen in een akoestische en een orthografische submodaliteit), een semantische modaliteit en een morfosyntactische modaliteit. In deze studie zal ik me concentreren op de morfosyntactische modaliteit. Deze heeft als functie om fonologische ene semantische (index)informatie met elkaar te verbinden. De bijbehorende bouwstenen (waaronder morfemen, lexemen, samenstellingen en vaste woordverbindingen) noem ik *taxemen*. De morfotactische modaliteit bestaat zelf weer uit drie submodaliteiten, te weten een morfofonologische kennislaag, een morfosemantische kennislaag en (wederom) een overkoepelend representatieniveau, te weten de morfotactische kennislaag (met informatie over de relatie tussen de morfofonologische en de morfosemantische bouwstenen en over hun categoriale en combinatorische eigenschappen).[173]

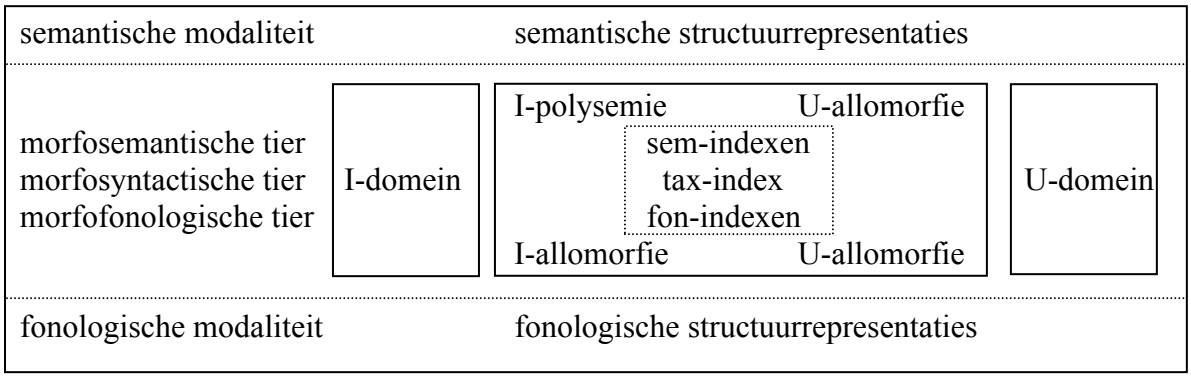
Al deze submodaliteiten hebben betrekking op morfosyntactische aspecten van de lexicaal opgeslagen kenniseenheden (waaronder hun categoriale typering). Hierdoor wordt het mogelijk om onderscheid te maken tussen formeel gelede woorden (zoals *deftig*, een formele afleiding met het affix -IG, of *druiloor*, een formele afleiding met het lexeem OOR), semantisch gelede woorden (zoals *beter* (= GOED+ER) en *best* (= GOED+ST), of *piloot* (= VLIEG+ER)) en transparant gelede woorden (zoals *moedig* (= MOED+IG) en *strijder* (= STRIJD+ER). Meer specifiek kan onderscheid worden gemaakt tussen morfofonologische bouwstenen (die wel morfosyntactische eigenschappen dragen, maar geen betekenis toevoegen, zoals de morfemen in *deftig*), morfosemantische bouwstenen (die wel betekenis dragen, maar niet aan een vormkenmerk zijn te koppelen, zoals de betekenis "goed" in *beter*) en morfotactische bouwstenen (die een verband leggen tussen een morfofonologische en een morfosemantische bouwsteen; zo legt het morfeem STRIJD een relatie tussen de vorm *strijd* en de betekenis "strijd" en het affix er tussen de vorm *-er* (of *-der*) en de functie van "agens".

Zowel op morfofonologisch niveau als op morfosemantisch niveau kunnen meerdere representaties per taxeem worden gespecificeerd. Dit is nodig om te kunnen verantwoorden dat taxemen vaak *allomorfie* (meerdere vormen per eenheid) of *polysemie* (meerdere betekenissen per bouwsteen) vertonen. Bij de specificatie van deze 1:n-relaties kan een nader onderscheid worden gemaakt tussen twee variatiedomeinen, namelijk variatie onder invloed

[172] Op cognitief niveau corresponderen deze indexen echter met abstracte knooppunten van lexicale relaties.
[173] Maar in de praktijk gebruik ik bijna altijd de conventionele termen *semantisch* en *fonologisch*.

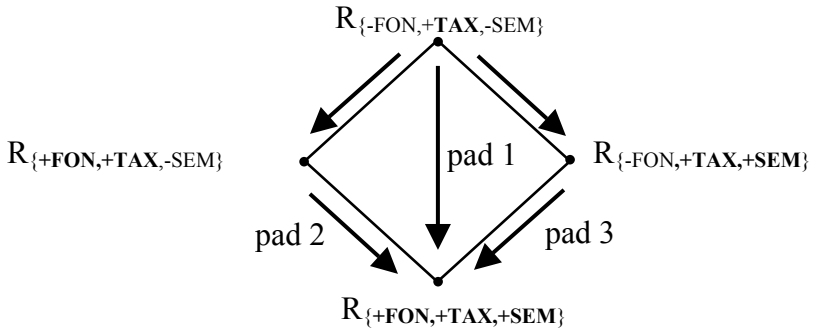
van stamkenmerken (c.q. het inwendige "I-domein") en variatie onder invloed van functorkenmerken (c.q. het uitwendige "U-domein"); deze selectiedomeinen zullen nog in detail worden behandeld (zie H4.3.4). Wat betreft de morfofonologische representatielaag kan de alternantie tussen de affixvormen *-er* en *-aar* van het morfeem -[ER/AAR] met indexdefinitie [<klasse> morfeem, affix <citatievorm> er/aar, <fon-kenmerken> inheems, <sem-kenmerken> agens/instrument] duidelijk in verband worden gebracht met de stamkenmerken: gegeven een inheemse stam (= selectieconditie) wordt de affixvorm *-aar* gekozen als de stam minstens twee segmenten telt en een *el*-uitgang bezit, zoals het geval is bij stam TOVER in *toveraar*. Een voorbeeld van functorgeconditioneerde vormvariatie is de alternantie tussen *-aal* en *-eel* bij de stam van het lexeempaar *intensioneel* / *intensionaliteit*: de suffixvorm *-eel* verandert hier in *-aal* onder invloed van de functor -ITEIT.

Bij wijze van samenvatting biedt het schema in figuur 4-5 een overzicht van de hierboven besproken kennisdimensies van een taxeem.



Figuur 4-5: De lexicale kennisdimensies van een morfotactische bouwsteen.

Het schema in figuur 4-6 laat zien hoe een basistaxeem (namelijk een lexicale representatie R met de representatiekenmerken $R_{\{-FON,+TAX,-SEM\}}$) kan worden verrijkt met informatie over zijn fonologische representatie ($R_{\{+FON,+TAX,-SEM\}}$) en zijn semantische representatie ($R_{\{-FON,+TAX,+SEM\}}$). Hiertoe dient men het pad te activeren dat het taxeem met de betreffende tierrepresentatie verbindt (resp. pad 2 en pad 3). Indien beide representatiekenmerken worden geactiveerd, ontstaat een complete taxeemrepresentatie (namelijk $R_{\{+FON,+TAX,+SEM\}}$) Deze representatie kan overigens ook rechtstreeks worden geactiveerd (namelijk via pad 1).



Figuur 4-6: De overerving van fonologische en semantische taxeemkenmerken.

Voor elke taxeemrepresentatie geldt dat zijn I-domein met de doorsnede van de I-domeinen van zijn samenstellende indexen correspondeert (althans op het niveau van de instanties; op het niveau van de selectiekenmerken dient juist de vereniging te worden genomen). Zo geldt voor de suffixvorm *-aar* van de functor -[ER/AAR] dat het normaal gesproken alleen stammen kan selecteren waarvan de fonologische representatie op *-el*, *-er* of *-en* eindigt (een fono-

logische restrictie), en die bijvoorbeeld minimaal een agentieve component moeten bezitten (een semantische restrictie). De vereniging van deze restricties levert een taxeem op waarvan het I-domein onder meer de stammen HANDEL, WANDEL en BAGGER omvat (blijkens *handelaar*, *wandelaar* en *baggeraar*), maar waarvan stammen als VERKOOP, TREK en REINIG om fonologische redenen zijn uitgesloten, en stammen als KRONKEL, HOBBEL en FLAKKER om semantische redenen.

De hier beschreven representatiestructuur maakt het mogelijk om partieel gelede woorden (namelijk woorden die alleen fonologisch of semantisch geleed zijn) als een deelrepresentatie van het integrale geleed woord te analyseren. Als een woord bijvoorbeeld betekenisspecialisatie ondergaat (zoals het woord *handelaar*), is alleen het fonologische overervingspad actief (zodat sprake is van een fonologisch geleed woord), en als het transparant gelede woord een niet-voorspelbare uitspraak aanneemt (bijv. /rinkeler/ i.p.v. /rinkelaar/ en /opener/ i..p.v. /openaar/)[174] alleen het semantische overervingspad. Dit leidt automatisch tot de aanmaak van een nieuwe taxeemindex, maar dankzij het gedeelde overervingspad is deze oplossing "goedkoper" dan het definiëren van een geheel onafhankelijk taxeem. Bovendien wordt zo verklaard dat taalgebruikers zowel bewust (meta-talig) als onbewust (via priming) een verband kunnen leggen tussen een transparante en een niet-transparante woordtoepassing.

4.3.3 Compositionele structuurprincipes

In de L-KRING-theorie wordt, conform de modeltheoretische traditie,[175] een fundamenteel onderscheid gemaakt tussen functors (representaties met één of meer interne variabelen) en stammen (representaties zonder variabelen). Deze stammen kunnen zelf eveneens het product zijn van een functor-stam-toepassing. Ik zal dit toelichten aan de hand van de lexicale functierepresentaties (F_{LEX}) voor de taxeemindexen in (8):

(8a) F_{LEX} (werk+baar) = [werk] ⊕ -baar
(8b) F_{LEX} (ver+werk+baar) = [ver+werk] ⊕ -baar = [ver ⊕ [werk] ⊕ -baar

Deze functierepresentaties bestaan uit een combinatie-operator + voor indexen, een compositie-operator ⊕ voor de combinatie van taxemen, een stam S (tussen vierkante haken) en een functor F (het complement van de stam). Het onderscheid tussen + en ⊕ hangt samen met het uitgangspunt dat het samenvoegen (c.q. combineren) van twee indexen (die niet meer dan lexicale verwijzers zijn) iets anders is dan het functioneel afleiden (c.q. componeren) van een nieuwe taxeemrepresentatie (door de taxeemrepresentaties te integreren).

Uit functiepresentatie (8a) blijkt dat het lexeem *werkbaar* is opgebouwd uit een stamindex WERK (die zelfstandig kan voorkomen) en een functorindex -BAAR (die altijd met een stam moet worden gecombineerd). Uit indexrepresentatie (8b) blijkt dat het lexeem *verwerkbaar* valt onder te verdelen in een functorindex -BAAR en een complexe stamindex VER+WERK; deze kan zelf weer worden geanalyseerd als een combinatie van de functorindex VER- en de stamindex WERK. Hoe men de functor identificeert, komt later aan de orde. Hier volstaat de kennis dat de functor overeenkomt met het segment dat de grootste invloed heeft op de (locale) kenmerken (zoals inflectie) van de afgeleide eenheid.

Zoals ik in hoofdstuk 3 uiteen heb gezet, gaan zowel generatieve als categoriale morfologiemodellen ervan uit dat productieve affixen met een functie van lexemen naar lexemen corresponderen en dat deze functie in termen van lexicale basiscategorieën (zoals N, V of A) kan worden gedefinieerd. Zo kenmerkt een affix met de functiespecificatie $F_{N>V}$ (zoals een sub-

[174]Dit zijn de enige GWNT-voorbeelden. Maar in de spreektaal zijn er denk ik tal van woorden waar de verwachte uitgang -aar als -er wordt gerealiseerd, met name bij geprefigeerde stammen (bijv. /beoordeler/).
[175]Meer specifiek doel ik op het Montague-framework. Zie H3.6.4 voor een korte introductie.

optie van het prefix BE-) zich door de eigenschap dat dit affix een N-lexeem in een V-lexeem omzet. Deze analyse veronderstelt dat elke morfologische stam (of deze nu geleed is of ongeleed) een lexicale basiscategorie bezit, ook als dit lexeem als stam dient voor een volgende derivatiestap. Zo is het prefix BE- in staat om zich aan het N-lexeem *plant* te hechten, wat een V-lexeem *beplant* zou opleveren; dit V-lexeem kan vervolgens weer als stam dienen voor een derivatie met het $F_{N>V}$-suffix -ING, wat een N-lexeem *beplanting* zou opleveren.

In de L-KRING-theorie corresponderen affixen niet met relaties tussen lexemen (c.q. lexeemklassen), maar met relaties tussen stammen c.q. stamklassen. In deze benadering hechten affixen zich dus aan stammen, terwijl het product van zo'n combinatie eveneens een stam is. Hierbij introduceert elke stam een paradigma met derivatiemogelijkheden (c.q. functors), waarbij de toepassing als zelfstandig lexeem (dus als de drager van een specifiek soort inflectie, zoals V-inflectie) slechts één van de mogelijke opties is. Daarom is het niet wenselijk om morfemen in termen van categoriale functies te definiëren, want deze beperken het functorparadigma tot de meest waarschijnlijke (dus ongemarkeerde) lexeemtoepassing. Een dergelijke typering gaat voorbij aan het feit dat veel morfemen meerdere functors kunnen kiezen. Bovendien is er lang niet altijd een ongemarkeerde lexeemtoepassing beschikbaar (zoals bij X-stammen), terwijl een gegeven vorm ook meerdere lexeemtoepassingen kan toestaan (wat meestal in termen van conversie wordt verantwoord). Tot slot valt moeilijk in te zien hoe een stam na categorietoekenning als basis kan dienen voor volgende affixatiestappen: zo is nergens aan te zien dat het N-lexeem *werking* van een V-lexeem *werk* is afgeleid, want binnen het lexeem *werking* staat de stam WERK geen V-interpretatie of V-inflectie toe, terwijl deze stam qua vorm en betekenis net zo goed met het N-lexeem *werk* kan corresponderen.

Het hier verwoorde inzicht kan formeel worden verantwoord door verschillende klassen van morfotactische bouwstenen (c.q. taxemen) te onderscheiden, waaronder morfemen (M), lexemen (L), woorden (W) en phrases (P). Hierbij correspondeert elk taxeemdomein met een domeinspecifieke begrenzer B. Deze begrenzer (die tevens een lexicale subklasse moet toekennen) heeft de functie om aan te geven dat de eenheden uit het bijbehorende structuurdomein tezamen een basiseenheid vormen in het volgende structuurdomein. Zo kan men een reeks van één of meer morfemen ($[M_1\ M_2\ ...\ M_n]$) in een lexeem L omzetten door er een lexeembegrenzer (B_L) op toe te passen (zie de voorbeelden in (9)). Op dezelfde wijze kan men een reeks lexemen ($[L_1\ L_2\ ...\ L_n]$) in een woord W omzetten door er een woordbegrenzer B_W op toe te passen (zie de voorbeelden in (10)).

(9) $[M_1\ M_2\ ...\ M_n]_M \oplus B_L$
(9a) $[M_1(\text{WERK})] \oplus B_L = [\textit{werk}]_L$
(9b) $[M_1(\text{WERK})] \oplus [M_2(\text{BE-})] + B_L = [\textit{bewerk}]_L$
(9c) $[M_1(\text{WERK})] \oplus [M_2(\text{BE-})] + [M_3(\text{-ER})] + B_L = [\textit{bewerker}]_L$

(10) $[L_1\ L_2\ ...\ L_n]_L \oplus B_W$
(10a) $[L_1(\textit{werk})] \oplus B_W = [\textit{werk}]_W$
(10b) $[[L_1(\textit{werk})] \oplus [L_2(\textit{woord})]]_L \oplus B_W = [\textit{werkwoord}]_W$
(10b) $[[L_1(\textit{werk})] \oplus [L_2(\textit{woord})]]_L \oplus [L_3(\textit{functie})] \oplus B_W = [\textit{werkwoordfunctie}]_W$

De begrenzer maakt het mogelijk om elk niveau zijn eigen subcategorieën te geven. Zo dienen de woordgerelateerde subcategorieën informatie te geven over het inflectieparadigma en het functiewoordenparadigma; voor deze functie bieden de traditionele woordcategorieën (zoals N, V en A) een geschikt uitgangspunt (al zullen wel subspecificaties nodig zijn). Op soortgelijke wijze dienen lexeemgerelateerde categorieën informatie te geven over de derivationele mogelijkheden op lexeemniveau (dus over de samenstelling van de lexeemparadigma's), terwijl de morfeemgerelateerde categorieën informatie dienen te geven over de combinatorische mogelijkheden op morfeemniveau (dus over de samenstelling van de morfeemparadigma's).

Voor elk structuurniveau kan dan ook onderscheid worden gemaakt tussen vrije c.q. onbegrensde eenheden en gebonden c.q. begrensde eenheden. Bij de gebonden eenheden kan een nader onderscheid worden gemaakt tussen linkerstammen, waarvan het paradigma met rechtshechtende eenheden correspondeert, rechterstammen, waarvan het paradigma met linkshechtende eenheden correspondeert en positieneutrale stammen, die (nog) niet gespecificeerd zijn voor de keuze tussen een linkerparadigma en een rechterparadigma.

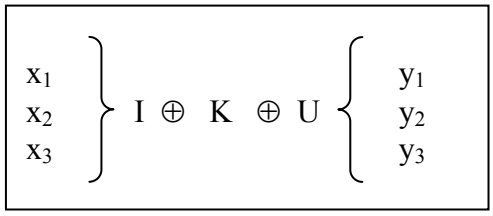
Elk morfotactisch structuurniveau kent zijn eigen categorietypes, al zijn er wel relaties tussen deze niveaus. Om de representaties zo leesbaar mogelijk te houden maak ik gebruik van de gangbare lexeemklassen Z (met Z ∈ {N, V, A, etc.}), maar op morfeemnivau zal ik deze als #z noteren en op lexeemniveau als $z:[176]

lexicaal toepassingsniveau	notatie	voorbeelden
categorie op phraseniveau	ZP	NP, VP, AP, PP, ...
categorie op woordniveau	Z	N, V, A, P, ...
categorie op lexeemniveau	$z	$n, $v, $a, $p, ...
categorie op morfeemniveau	#z	#n, #v, #a, #p, ...

De onderstaande tabel geeft voor elk toepassingsniveau enkele voorbeelden van vrije en gebonden taxemen. Merk op dat zowel op het niveau van de morfemen als op het niveau van de lexemen een syntactische oriëntatie wordt aangegeven (linkerdelen versus rechterdelen). Dit verschijnsel is ook relevant op hogere structuurniveaus, maar krijgt hier een steeds complexere (syntactisch gedifferentieerd) karakter.

taxeemtype	vrije taxemen	gebonden taxemen
morfemen	BE-, -ER, -IEK, -ISCH	[#v: BE-], [#n: -ER], [#n: -IEK], [#a: -ISCH]
lexemen	*lezers*, *handig*	[$n: *lezers*-], [$a: -*handig*]
woorden	*lezersvraag*, *vierhandig*	[N: *lezersvraag*], [A: *vierhandige*]
phrasen	*vierhandige pianostukken*	[NP: *vierhandige pianostukken*]

Ik ga ervan uit dat de combinatorische eigenschappen van de per domein onderscheiden distributieklassen grotendeels zijn terug te voeren op de semantische kenmerken (zoals gebeurtenis versus object) en de fonologische kenmerken (zoals inheems versus uitheems) van de geclassificeerde eenheden. Om meer inzicht te krijgen in deze verbanden is empirisch onderzoek nodig naar de vraag wat de morfologische distributiecategorieën van het Nederlands zijn en in hoeverre deze daadwerkelijk op fonologische en semantische kenmerken zijn terug te voeren.



Figuur 4-7: Het lexicale analysevenster van een kern K

De morfologische bouwstenen uit de L-KRING-theorie kunnen systematisch worden gedefinieerd door gebruik te maken van een lexicaal analysevenster. Zo'n analysevenster (zie figuur 4-7) bestaat uit drie domeinen, te weten een kern (K), die met het te identificeren segment correspondeert, een inwendig (stamgeoriënteerd) selectiedomein (I) en een uitwendig (functorgeoriënteerd) selectiedomein (U). Indien een selectiedomein leeg is, krijgt het de specificatie '[-]' (ten teken dat de kern hier begrensd wordt); alle eenheden tezamen vormen een locaal selectieparadigma van kern K (die als centrale eenheid fungeert). Om de localiteit

[176] Zie H4.3.5 voor een meer gedetailleerde bespreking.

van deze domeinen te benadrukken spreek ik meestal van *inwaarts* en *uitwaarts* selectiedomein (in plaats van *inwendig* en *uitwendig* selectiedomein). In figuur 4-15 omvatten het inwaartse (I) en het uitwaartse (U) domein allebei drie eenheden, resp. x_1, x_2 en x_3 en y_1, y_2 en y_3, maar elk ander aantal is mogelijk.

Op alle structuurniveaus geldt dat de functor in principe uit meerdere stammen kan kiezen. Al deze stammen tezamen vormen het inwendige domein I van de functor F. Hiernaast hebben functors ook een uitwendig domein U; dit bestaat uit alle functors die toepasbaar zijn op de met F afgeleide eenheden; indien nodig kunnen deze in subklassen worden onderverdeeld. Dit wordt uitgewerkt in (11) en (12):

(11) D_I(-BAAR) = { WERK, VER+WERK, HOOR, ZICHT, VER+STAAN, VER+PLAATS, ... }

BAAR- + D_I(-BAAR) = { [WERK]+BAAR, [VER+WERK]+BAAR, [HOOR]+BAAR, ... }

D_U(-BAAR, $D_{I,1}$) = { $O_{\$A}$, -HEID, -DER/ST }

D_U(-BAAR, $D_{I,2}$) = { $O_{\$A}$, -HEID }

(12) D_I(VER-) = { WERK, PLAATS, DENK, GROOT, TEL, WACHT, MIS, HUUR... }

VER- + D_I(-VER) = { VER+[WERK], VER+[PLAATS], VER+[GROOT], VER+[TEL] }

D_U(VER-, $D_{I,1}$) = { $O_{\$V}$, -ING, -BAAR }

D_U(VER-, $D_{I,2}$) = { $O_{\$V}$, -ING }

Uit (11) blijkt dat het I-domein van het suffix -BAAR onder meer de stammen WERK, VER+WERK, HOOR, ZICHT, VER+STAAN en VER+PLAATS omvat. Deze kunnen allemaal de basis vormen voor de adjectiefvormende operator $O_{\$A}$, voor het suffix -HEID en in het geval van stammen uit subdomein $D_{I,1}$ voor de A-modificerende suffixen -DER en -ST. Uit (12) blijkt dat het prefix VER- een I-domein bezit met stammen als WERK, PLAATS, DENK, GROOT, TEL en DEDIG. Toepassing van de functor VER- leidt hier tot complexe eenheden als VER+PLAATS, VER+DENK en VER+GROOT. Deze hebben met VER+WERK gemeen dat ze stam kunnen zijn van de functor -ING en van de operator $O_{\$V}$ voor werkwoordsvorming; de laatste twee eenheden verschillen echter van de andere twee doordat ze geen stam kunnen zijn van de functor -BAAR. Dit impliceert dat het U-domein van de functor VER- eveneens in twee subklassen uiteenvalt, te weten subklasse $D_{I,1}$ met de functors $O_{\$V}$, -ING en -BAAR, en subklasse $D_{I,2}$ met alleen de functors $O_{\$V}$ en -ING..

De hier geschetste classificatiemethode kan een fijnmazig netwerk opleveren van morfologische, deels semantisch of fonologisch gemotiveerde equivalentieklassen. Dit netwerk kan bovendien een empirische basis vormen voor de identificatie van morfosyntactische categorieën. Zoals al in hoofdstuk 3 aan de orde kwam, mag worden verwacht dat dit een veel inzichtelijker model oplevert dan de gangbare indeling op basis van syntactische hoofdcategorieën. In de volgende subsecties wordt de hier geïntroduceerde analysemethode nader uitgewerkt en op concrete voorbeelden toegepast.

4.3.4 Inductieve lexiconanalyse

Deze subsectie heeft als doel om een inductieve analysemethode te presenteren voor het construeren van morfotactische bouwstenen. Hierbij ga ik ervan uit dat het mentale lexicon van een willekeurige taalgebruiker alle woorden omvat die hij recent gebruikt heeft of die met enige regelmaat voorkomen, ongeacht de vraag of deze woorden morfologisch geleed zijn. Gegeven dit mentale woordspectrum kan langs inductieve weg worden vastgesteld welke lexeeminterne segmenten morfeemstatus verdienen. Dit zijn segmenten die een vaste relatie

vertonen tussen hun vorm, hun betekenis (of betekenisklasse) en hun morfologische combinatiemogelijkheden, d.w.z. hun inwaartse en hun uitwaartse selectiedomein.

Hieronder demonstreer ik mijn analysemethode voor het morfologisch complexe lexeem *uitspreekbaar*. Dit lexeem heeft de taxeemstructuur [UIT ⊕ [SPREEK]] ⊕ BAAR], die uit volgende kernen bestaat: UIT-, SPREEK, [UIT ⊕ SPREEK] en -BAAR. Voor elk van deze kernen kan een apart analysevenster worden gespecificeerd. In onderstaande voorbeeldvensters beperk ik me tot de specificatie van de kern (K), het eerste I-domein (I_1) en het eerste U-domein (U_1) bij elk van de genoemde kernen in de morfologische structuurrepresentatie van het lexeem *uitspreekbaar*:

(13)

Venster	**I_1**	**K**	**U_1**
1	[-]	SPREEK	UIT-
2	SPREEK	UIT-	-BAAR
3	[-]	UITSPREEK	-BAAR
4	UIT-	-BAAR	$A

Indien men dit analysevenster niet beperkt tot informatie over de structuurdomeinen van één lexeem, maar ook lexicale informatie over andere lexemen verwerkt, zijn meerdere eenheden per structuurdomein mogelijk. Gegeven een lexicon met de $A-lexemen *uitspreekbaar*, *uitneembaar* en *uitklapbaar* kan de kern UIT- bijvoorbeeld de volgende specificatie van het I-domein en het U-domein krijgen:

(14)

Venster	**I_1**	**K**	**U_1**	**U_1**
2	SPREEK	UIT-	-BAAR	$A
	NEEM			
	KLAP			

En indien het lexicon ook nog de infinitieven *uitspreken*, *uitnemen* en *uitklappen* bevat (met lexeemcategorie $V), alsmede de participia *uitsprekend*, *uitnemend* en *uitklappend* (met lexeemcategorie $A), kan het analysevenster als volgt gespecificeerd worden:

(15)

Venster	**I_1**	**K**	**U_1**	**U_1**
2	SPREEK	UIT-	-BAAR	$A
	NEEM		-EN	$V
	KLAP		-END	$A

Volgens dit analysevenster correspondeert het inwaartse selectiedomein van de kern UIT- met een (locaal) paradigma dat uit drie eenheden bestaat, te weten de wortels SPREEK, NEEM en KLAP. Voor elk van deze eenheden geldt dat de kern met drie uitwaartse eenheden kan samengaan, te weten de suffixen -BAAR ($A), -EN ($V) en -END ($A). Alle wortels uit het I-domein vertonen dus hetzelfde U-paradigma als ze deel uitmaken van een stam met de kern UIT. Stel nu dat de wortel KLAP als enige in staat is om na combinatie met de kern UIT een lexeem op te bouwen met het suffix -ER, namelijk het lexeem *uitklapper*. In dat geval zal het lexicon een extra U-eenheid moeten specificeren voor het U-domein bij de stam UIT + KLAP; dit kan als volgt tot uitdrukking worden gebracht in het lexciale analysevenster:

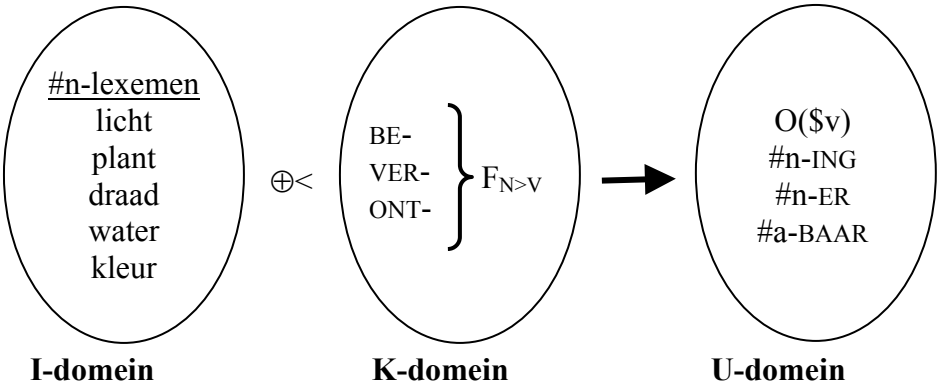
(16)

Venster	**I_1**	**K**	**U_1**	**U_1**
2	SPREEK	UIT-	-ER	
	NEEM		-BAAR	$A
	KLAP		-EN	$V
			-END	$A

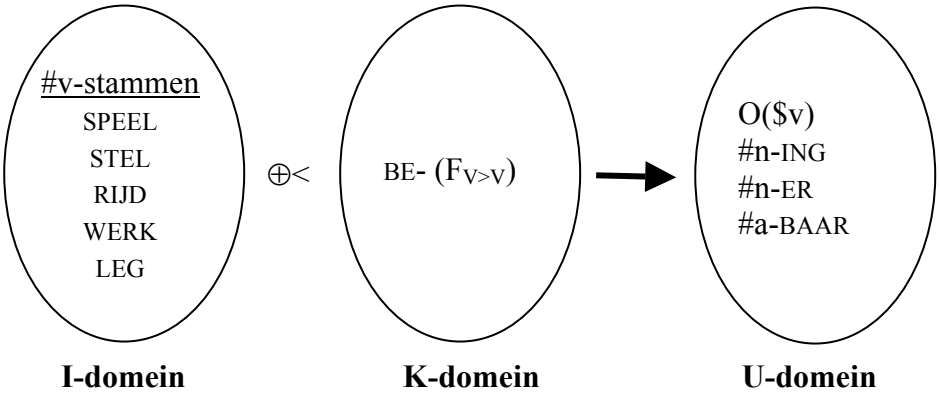
Beschouw nu de lexemen *werken* (V), *werk* (N) en *werking* (N). Deze kunnen alledrie op een #v-stam WERK worden gebaseerd, waarbij het U-domein van WERK minimaal de volgende opties specificeert: U(WERK) = {$v, #n-0, #n-ING}. Indien dit U-domein vaak genoeg

voorkomt,[177] kan het tevens de basis vormen voor de introductie van een distributieklasse X, waarbij X een lexicale index is die alle stammen omvat die minimaal de hier gespecificeerde U-opties bezitten. Elk affix dat tot het U-domein van deze X-stammen behoort, bezit per definitie een I-domein met het selectiekenmerk X.

Ook affixen kunnen een distributieklasse krijgen; maar in dat geval dient de definitie van het U-domein aan de invulling van het I-domein te worden gerelateerd, wat impliceert dat per affix meerdere distributieklassen kunnen voorkomen. Ik zal dit idee toelichten aan de hand van enkele voorbeelddiagrammen. Beschouw eerst de diagrammen in figuur 4-8 en 4-9.



Figuur 4-8: Distributiediagram voor de $F_{N,V}$-toepassing van het prefix BE-.



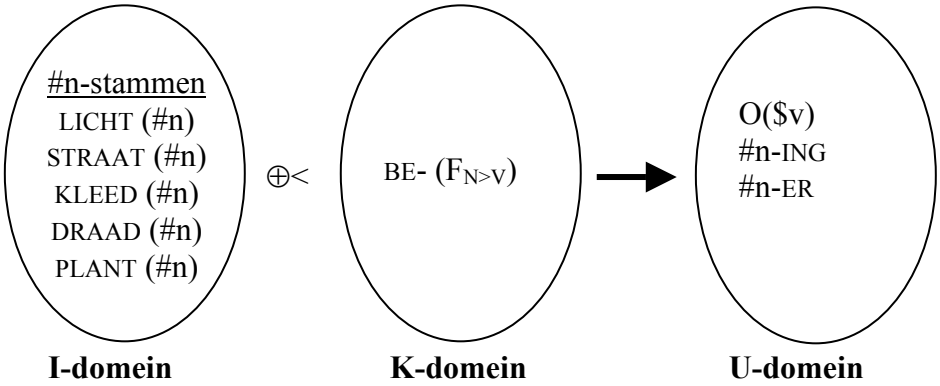
Figuur 4-9: Distributiediagram voor de $F_{V,V}$-toepassing van het prefix BE-.

Het eerste distributiediagram toont de $F_{N>V}$-toepassing van het prefix BE-. Hierbij correspondeert het I-domein met #n-stammen en geldt voor elke [BE ⊕ #n-stam]-combinatie dat deze een V-lexeem oplevert waarvan het U-domein uit de functors bestaat: de V-vormende operator O($v) en de nominaliserende suffixen #n-ING en #n-ER. Het tweede distributiediagram, toont het prefix BE- in zijn $F_{V>V}$-toepassing. In dit geval correspondeert het I-domein met #v-stammen. Ook hier geldt voor elke [BE + #v-stam]-combinatie dat deze een V-lexeem oplevert, maar het U-domein omvat nu niet alleen de functors O($v), #n-ING en #n-ER, maar ook de functor #a-BAAR.[178] Het bestaansrecht van deze distributieklassen blijkt ook uit het feit dat de prefixen VER-, ONT- en HER- zich qua inwaartse en uitwaartse selectiecondities door-

[177] In principe kan voor elk U-domein een aparte distributieklasse worden gedefinieerd, maar dit zou weinig informatief zijn; daarom is het raadzaam om een drempelwaarde te hanteren (bijv. 5 types); hoe zwaarder de drempel, hoe informatiever de distributieklasse.

[178] Dit contrast zou kunnen samenhangen met het feit dat de betekenis van [BE + N] bijna altijd neerkomt op "een object van N voorzien"; dit is een handeling die zo algemeen van aard is dat hij voor elke N uitvoerbaar is, zodat de constructie met -BAAR overbodig is: zo is een weg in beginsel altijd "bestraatbaar".

gaans analoog aan het prefix BE- gedragen, wat impliceert dat ze tot dezelfde distributie-



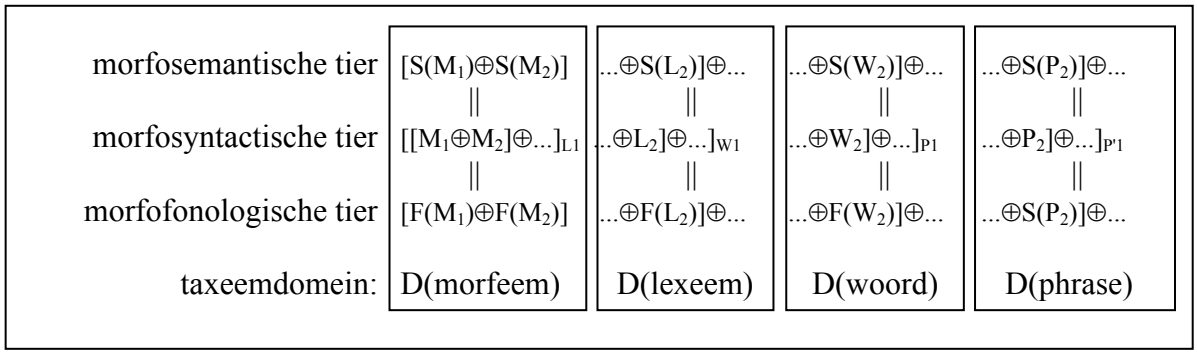
klasse(n) behoren. Dit blijkt bijvoorbeeld uit het distributiediagram in figuur 4-10.

Figuur 4-10: Distributiediagram met de $F_{N,V}$-toepassing van de prefixen BE-, VER- *en* ONT-.

Dit diagram toont het I-domein en het U-domein van de prefixen VER-, BE- en ONT- in hun toepassing als <#n,#v>-functor ($F_{N>V}$), d.w.z. in hun toepassing als denominale, verbaliserende functors. Voor alle stammen uit het I-domein geldt dat ze voor elk van deze drie prefixen het hetzelfde U-domein selecteren, namelijk O($v), #n-ING en #n-ER. Hoewel dit patroon niet algemeen geldig is, illustreert dit diagram de mogelijkheid om algemene functorklassen te introdcueren die over meerdere affixtypes kunnen generaliseren, zowel wat betreft het I-domein (dat alleen op intensioneel niveau overeen hoeft te komen) als wat betreft het U-domein (dat voor alle affixtypes samen minimaal één identieke functor moet omvatten). De L-KRING-theorie biedt daarom de mogelijkheid om individuele taxemen in klassen onder te brengen. Het idee is dat taxeeminstanties t systematisch aan een overkoepelende taxeemklasse kunnen worden gerelateerd door op zoek te gaan naar overeenkomsten in hun combinatorische eigenschappen. Gegeven zo'n taxeemverzameling kan een overkoepelende klasse T worden gedefinieerd die alle gemeenschappelijke eigenschappen van de instanties introduceert. Hierna hoeft per taxeeminstantie uitsluitend informatie te worden gegeven over de taxeemspecifieke eigenschappen, want de rest van de eigenschappen kan van de klasse T worden overgeërfd.

4.3.5 Hiërarchische structuuraspecten

Deze subsectie belicht de hiërarchische structuur van het L-KRING-lexicon. Het diagram in figuur 4-11 laat zien hoe de morfotactische structuurrepresentaties uit de vier lexicale hoofddomeinen formeel aan elkaar zijn te relateren. Het diagram toont de lexicale ruimte die ontstaat door combinatie van de morfotactische domeindimensie (de horizontale as) en de tierdimensie (de verticale as). Hieruit blijkt dat de opbouw van de morfologische representatie gelijk oploopt met die van de semantische (S) en de fonologische (F) representaties. Zo correspondeert de morfeemindex M_1 met de vormindex $F(M_i)$ en met de betekenisindex $S(M_i)$. Zowel de fonologische als de semantische representatie van M_i kunnen worden onderverdeeld in tierspecifieke bouwstenen.



Figuur 4-11: De compositionele structuur van de lexicale ruimte.

In de onderstaande representatie correspondeert het morfeem $[\text{WERK}]_M$ bijvoorbeeld met F- en S-representaties die allebei als een compositioneel product van kleinere eenheden zijn gedefinieerd:

$M_i = [\text{WERK}]_M$
$F(M_i) = [/w/\oplus/e/\oplus/r/\oplus/k/]_F = [/\text{werk}/]$
$S(M_i) = [\text{inspanning} \oplus \text{doelgericht}]_S = [\text{"werk"}]$

Het resulterende morfeem ($M_{1\oplus2}$) kent uiteraard ook weer een fon-representatie en een sem-tier-representatie, maar deze corresponderen nu met een combinatie van de representaties van de samenstellende morfemen, bijvoorbeeld $F(M_{1\oplus2}) = F(M_1) \oplus F(M_2)$:

$M_{1\oplus2} = \lambda X.[\text{BE} + X]_M \oplus [\text{WERK}]_M = [\text{BEWERK}]_M$
$F(M_{1\oplus2}) = \lambda X.[/\text{be}/ + X]_F \oplus [/\text{werk}/]_F = [/\text{bewerk}/]_F$
$S(M_{1\oplus2}) = \lambda X.[\text{"X ergens op richten"}]_S \oplus [\text{"doelgerichte inspanning"}]_S =$
$= \lambda X.[\text{"(een) doelgerichte inspanning ergens op richten"}]_S$

Bij de representatie van de functors maak ik gebruik van de lambda-operator λ, die afkomstig is uit de type-logica (cf. Gamut, 1991); deze operator geeft aan dat de functor waarop hij betrekking heeft een door de lambda-term gespecificeerde variabele nodig heeft om een grotere eenheid te kunnen vormen: zo heeft M_1 een variabele X^M (van type M = morfeem) nodig om een M-eenheid met de structuur be+X^M te kunnen vormen. Er zijn ook constructie-stappen waarbij een eenheid uit domein D_i wordt opgetild naar een eenheid uit domein D_{i+1} door toepassing van een domeinoperator $O^V(D_i>D_{i+1})$. Hierbij specificeert het superscript V de valentie van de door O te construeren eenheid, te weten het aantal stammen dat deze eenheid kan selecteren. De begrenzing van een lexicaal domein gaat in principe altijd samen met de introductie van een functioneel kenmerk (zoals inflectie), maar dit kenmerk kan een onhoorbare klankvorm hebben.

Het morfeem $M_{1\oplus2}$ kan zelf weer de basis vormen voor de opbouw van grotere morfeem-clusters (wat in figuur 4-11 door een reeks puntjes is gemarkeerd), maar deze mogelijkheid verdwijnt zodra er een lexeemgrens wordt bereikt; vanaf dit moment kan de opgebouwde morfeemcluster worden aangeduid als het lexeem L_1. Dit lexeem kan vervolgens met andere lexemen worden gecombineerd tot een samenstelling, zoals het lexeem L_3 in het domein D(lexeem). Dit kan weer net zolang doorgaan totdat er een woordgrens wordt bereikt, waarna de opgebouwde cluster kan worden aangeduid als het woord W_1. Ook voor woorden geldt dat ze kunnen worden geclusterd (wat overeenkomt met modificatie van een N of een V door adjectieven en bijwoorden), wat in een phrase resulteert (= syntactische woordgroep, bijvoorbeeld een NP of een PP), zoals de phrase P_1. Tot slot kunnen de eenheden op phrase-niveau tot grotere phrasen worden gecombineerd, zoals P_3 en P_5. Ik zal een en ander demonstreren aan

de hand van een voorbeeld, namelijk de opbouw van de nominale woordgroep *programma voor automatische gegevensbewerking*:

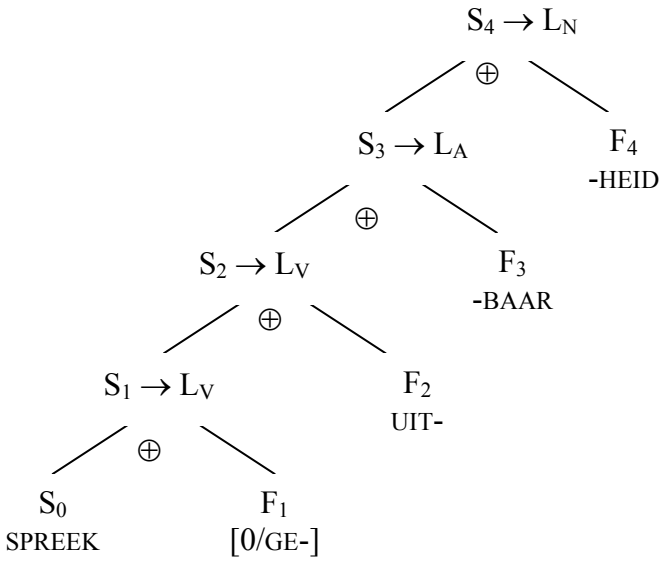
$M_1 = \lambda X^M.\ [be \oplus X^M]_M$
$M_2 = [werk]_M$
$M_3 = M_1 \oplus M_2 = [bewerk]_M$
$M_4 = \lambda X^M.\ [ing \oplus X^M]_M$
$M_5 = M_4 \oplus M_3 = [bewerking]_M$
$O^0{}_{M>L} = \lambda X^M.\ [[0/s] \oplus [X^3]_M]_L$
$O^1{}_{M>L} = \lambda X^M\ \lambda X^L.\ [[[0/s] \oplus [X^M]_M]_L \oplus X^L]$
$L_1 = O^1{}_{M>L} \oplus M_5 = \lambda X^L.\ [[[0/s] \oplus M_5]_M \oplus X^L] = \lambda X^L.\ [[bewerking\_0]_L \oplus X^L]_L$
$L_2 = O^0{}_{M>L} \oplus [gegeven]_M = [gegeven\_s]_L = [gegevens]_L$
$L_3 = L_1 \oplus L_2 = [gegeven\_s]_L \oplus \lambda X^L.\ [[bewerking\_0]_L \oplus X^L]_L = [gegevensbewerking]_L$
$O^0{}_{L>W} = \lambda X^L.\ [[X^L]_M]_L$
$W_1 = [O^0{}_{L>W} \oplus L_3]_W = [gegevensbewerking]_W$
$W_2 = \lambda X^W.\ [automatische \oplus X^W]_W$
$W_3 = [W_2 \oplus W_1]_W = [automatische\ gegevensbewerking]_W$
$O^0{}_{W>P} = \lambda X^W.\ [[X^W]_W]_P$
$P_1 = O^0{}_{W>P} \oplus W_3 = [automatische\ gegevensbewerking]_P$
$P_2 = \lambda X^P.\ [voor \oplus X^P]_P$
$P_3 = [P_2 \oplus P_1]_P = [voor\ automatische\ gegevensbewerking]_P$
$P_4 = \lambda X^P.\ [programma \oplus X^P]_P$
$P_5 = [P_4 \oplus P_3]_P = [\ [\textit{automatische}\ ]_W\ \oplus\ [\ [\textit{gegevens}]_L \oplus [\textit{bewerking}]_L\ ]_W\ ]_P$
$= [programma\ voor\ automatische\ gegevensbewerking]_P$

De L-KRING-theorie kent dus een opzet waarbij alle morfotactische domeinen op dezelfde structuurprincipes berusten. Het lexicon als geheel kent echter een asymmetrische structuur in de zin dat elk element uit een gegeven domein het bestaan van elementen uit de lagere domeinen veronderstelt: zo is het morfeemdomein structureel ingebed in het lexeemdomein, het lexeemdomein in het woorddomein en het woorddomein in het woordgroepdomein. Hoe dieper een domein in het lexicon is ingebed, hoe hoger de gemiddelde gebruiksfrequentie van de bijbehorende eenheden, en hoe sterker de interne samenhang van deze eenheden. Deze eigenschappen kunnen helpen bij de identificatie van vaste lexicale eenheden, zoals woorden. Dit zou een belangrijke voorwaarde kunnen zijn voor de verwerving van een taal. Want een kind kan pas op zoek gaan naar combinatorische regelmaat als het inzicht heeft gekregen in de vraag welke fonologische patronen als basiseenheid kunnen worden aangemerkt.

Om meer inzicht te krijgen in de hiërarchische structuuraspecten van het L-KRING-lexicon is het handig om gebruik te maken van een boomdiagram. Zo toont figuur 4-12 het boomdiagram van het lexeem *uitspreekbaarheid*. Hierbij staat S_i voor de i^e stam, F_i voor de i^e functor en i zelf met de volgorde waarin de stammen en functors met elkaar gecombineerd worden (door toepassing van de compositie-operator $\oplus$); combinatie van een stam S_i met een functor leidt tot een gelede stam S_{i+1}. De notatie $S_i \rightarrow L_C$ geeft aan dat stam S_i de basis kan vormen voor een lexeem L met categorie C ($C \in \{N,V,A\}$) (door toepassing van een domeinbegrenzer).

Het boomdiagram laat zien dat het lexeem *uitspreekbaarheid* een recursieve stamstructuur heeft, in de zin dat elke stam kan worden onderverdeeld in een functor (die het meest bepalend is voor de eigenschappen van de stam als geheel) en een dieper ingebedde stam. Hierbij kan elk stamniveau S aan een I-K-U-analyse worden onderworpen in de zin dat men elke intern gelede S (neem bijv. S_2) in een I-segment (c.q. substam, hier S_1) en een K-segment (c.q. subfunctor, hier F_2) kan onderverdelen, terwijl het U-segment met de functor correspondeert

die het complementaire deel vormt van de eerste overkoepelende eenheid (hier F_3). Zo bestaat de stam S_2 (te weten [SPREEK ⊕ [0/GE]]$_{S1}$ ⊕ UIT]$_{S2}$, die onder meer de vormen *uitspreek*, *uitspraak* en *uitgesproken* kan aannemen, uit de stam [SPREEK ⊕ [0/GE]] en de functor UIT-, terwijl het U-segment van deze eenheid met de functor -BAAR correspondeert.



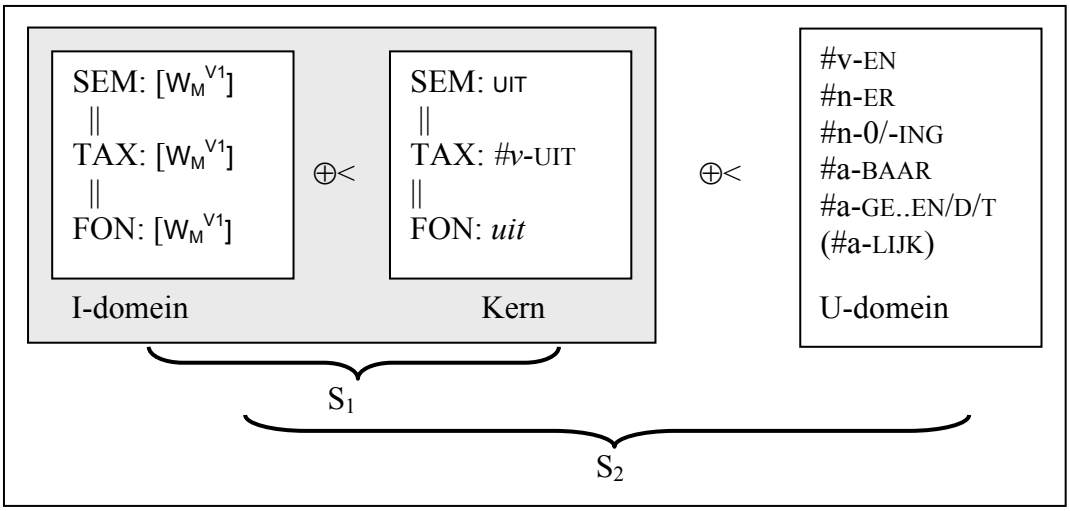
Figuur 4-12: De morfologische analyse van het lexeem uitspreekbaarheid.

Beschouw nu de kern [#v-UIT] (d.w.z. het V-stam-vormende partikel UIT).[179] Men treft dit partikel niet alleen aan in V-lexemen als *uitspreken*, *uitdragen*, *uitleggen*, *uitdrukken*, *uitroepen* en *uitwerken*, maar ook in morfologisch verwante lexemen (al dan niet met gelexicaliseerde betekenis) als N-lexemen met het (onhoorbare) suffix [#n-0] (zoals *uitspraak*, *uitleg* en *uitroep*), N-lexemen met het suffix [#n-ING] (zoals *uitdrukking* en *uitwerking*), N-lexemen met het suffix [#n-ER] (zoals *uitdrager* en *uitlegger*), A-lexemen met het suffix [#a-BAAR] (zoals *uitspreekbaar* en *uitlegbaar*), A-lexemen met het suffix [#a-LIJK] (zoals *onuitsprekelijk* en *uitdrukkelijk*) en A-lexemen met het (discontinue) affix [#a-GE..EN/D/T] (zoals *uitgesproken, uitgedoofd* en *uitgelokt*). Al deze lexemen hebben met elkaar gemeen dat ze op het stampatroon S_1 = [[W_M^{V1}] ⊕< [#v-UIT]] zijn gebaseerd; hierbij geeft de notatie 'x ⊕< y' aan dat y functor is ten opzichte van x, d.w.z. dat functor y een compositionele relatie (c.q. ⊕-relatie) kan aangaan met stam x, onder afleiding van een eenheid y' = [y⊕x] (met primair door y bepaalde eigenschappen). De eenheid [W_M^{V1}] correspondeert met morfemen uit de ad hoc gedefinieerde taxeemklasse [M:V_1], d.w.z. met wortelstammen die minimaal in staat zijn om het V-stam-vormende partikel [#v, UIT]$_M$ te selecteren; meestal kunnen zulke wortels ook andere partikels selecteren, en ook één of meer gebonden prefixen uit de reeks 0/GE-, BE-, VER-, ONT- en HER-.

Figuur 4-13 laat zien hoe de [W_M^{V1}]-eenheden zich tot kern [#v-UIT] verhouden en wat het U-domein (c.q. uitwaarts selectiedomein) is van de resulterende (door combinatie verkregen) S1-stammen. Elke combinatie van S_1 met een U-functor levert een S_2-stam op; deze S_2-stam kan vervolgens lexeemstatus en zelfs woordstatus krijgen, maar kan ook de basis vormen voor verdere afleidingen. Uit het diagram blijkt dat [#v-UIT] tal van functors (voornamelijk suffixen) kan selecteren, want het U-domein van deze kern omvat de suffixen [#v-EN], [#n-ER], [#n-0], [#n-ING], [#a-BAAR], [#a-GE..EN/D/T] en (optioneel) [#a-LIJK]. Al deze functoraanduidingen hebben de structuur [#x-y], waarbij #x de morfologische hoofdklasse specificeert en y een

[179] Zie H4.3.3 voor een toelichting op het onderscheid tussen #x, $x en X.

specifieke morfeemindex (die in enkele gevallen met een klankloze vorm correspondeert, aangeduid als '0'). Dit betekent echter niet dat elke S1-stam met de kern [#v-UIT] verplicht is om al deze suffixen te selecteren. Voor elke U-functor van een gegeven kern K kan namelijk apart worden gecodeerd of het om een verplichte of om een optionele functor gaat. In het eerste geval moet voor alle eenheden uit het I-domein van K een lexeemtoepassing bekend zijn die met de betreffende functor is gevormd; in het tweede geval hoeft dit slechts voor enkele I-eenheden te gelden; in figuur 4-13 geldt dit laatste alleen voor het suffix [#a-LIJK]. Op grond van dit soort distributiekenmerken kan een kern worden onderverdeeld in meer gespecialiseerde subkernen k(1)..k(n) die elk met een uniek I-domein en U-domein corresponderen.



Figuur 4-13: Distributieschema dat informatie geeft over het inwaartse (I) domein en het uitwaartse (U) domein van kern K, te weten het morfeem [*#v*-UIT] *in de stam* S_1.

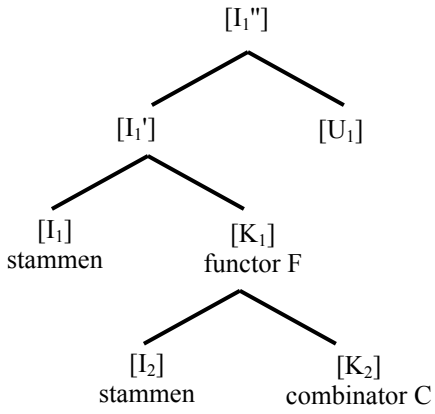
Het hierboven besproken schema kan als volgt worden geformaliseerd:

K = [#v, UIT]
S_1 = I(K) ⊕< K = [W_M^{V1}] ⊕< [#v-UIT]
S_2 = S_1 ⊕< U(K) = [[W_M^{V1}] ⊕< [#v-UIT]] ⊕< U(K)
I(K) = [W_M^{V1}] = {SPREEK, DRAAG, LEG, DRUK, ROEP, WERK, ...}
U(K) = {#v-EN, #n-ER, #n-0, #n-ING, #a-BAAR, #a-GE..EN/D/T, (#a-LIJK) ...}

Elke stam S_0 die aan de hier gespecificeerde condities voldoet, behoort tot het inwaartse domein I(K) van de kern K. Eén van die stammen is SPREEK (zonder stamprefix), want deze stam kan als basis dienen voor de lexemen *uitspreken*, *uitspreker, uitspraak*, *uitspreekbaar* en *uitgesproken*; alleen de (niet-verplichte) lexeemtoepassing *uitspreek(e)lijk* is ongebruikelijk.

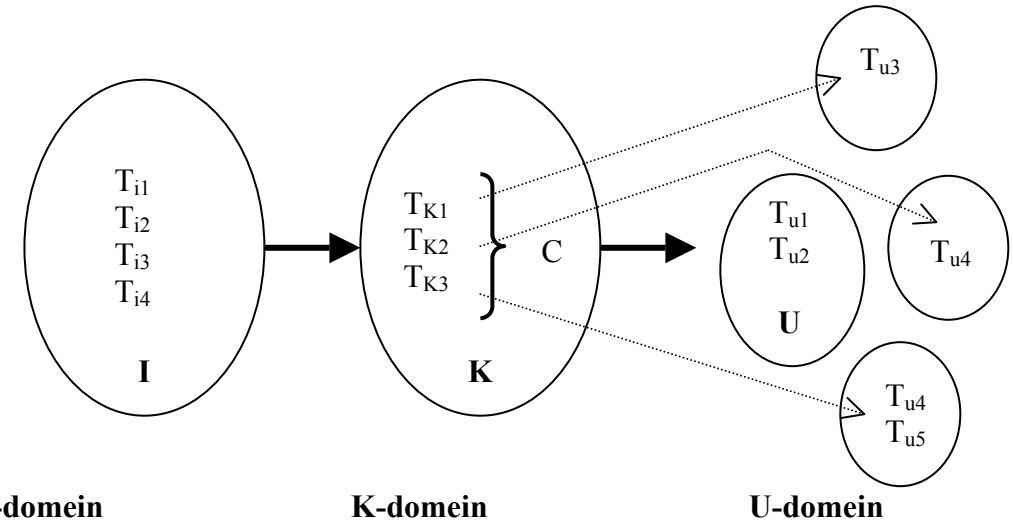
Samenstellingen verschillen van gewone derivaties doordat er een speciale constructiestap nodig is om twee eenheden met stamstatus (dus zonder interne variabele) in een grotere structuur in te bedden. In traditionele samenstellingen gaat het altijd om een combinatie van twee zelfstandig bruikbare lexemen, zoals de lexemen *bloemetjes* en *gordijn* in de samenstelling *bloemetjesgordijn*. Maar het is ook mogelijk om een samenstelling op morfeemniveau te vormen; zo is het lexeem *vierhandig* een afleiding van een lexeem dat is samengesteld uit de morfeemstammen VIER en HAND, en het lexeem *pianobouwer* van een lexeem met de morfeemstammen PIANO en BOUW. Dit type samenstelling komt ook veel voor in de uitheemse woordenschat, blijkens lexemen als *morfofonologisch* (van [MORFO + FOON]), en *psychoanalyticus* (van [PSYCHO + ANALY{S/T}]). Dit type structuur kan formeel worden verantwoord door gebruik te maken van een combinator.

Combinators kenmerken zich door de eigenschap dat ze niet één, maar twee stammen nodig hebben om een nieuwe eenheid te vormen; hierbij correspondeert de volgorde waarin deze stammen geslecteerd worden met de volgorde waarin deze stammen invloed kunnen uitoefenen op de eigenschappen van de resulterende samenstelling. Hierdoor kan eenvoudig worden verantwoord dat Nederlandse lexeemgebaseerde samenstellingen normaal gesproken aan de rechterhoofdregel voldoen: dit betekent namelijk dat de combinator voor lexeemgebaseerde samenstellingen eerst de rechterstam selecteert (via het I_2-domein) en dan pas de linkerstam (via het I_1-domein). Het boomdiagram in figuur 4-14 illustreert dit idee.



Figuur 4-14: Boomdiagram van een functortaxeem.

De eerste selectiestap leidt tot de constructie van een functorversie van het rechterlexeem (bijv. *gordijn*), waarna deze functor (F) op een linkerlexeem (bijv. *bloemetjes*) kan worden toegepast, namelijk via het I_1-domein van deze door C gevormde functor F. Dit resulteert in een samenstelling *bloemetjesgordijn* met dezelfde syntactische en semantische eigenschappen als het lexeem *gordijn*. Een andere optie is dat combinator C aangeeft dat het linkerlexeem medebepalend is voor de eigenschappen van de samenstelling als geheel, zoals het geval is in het *luchtbelwaterpas* (dat een ander lidwoord selecteert dan het rechterlexeem *waterpas*), *waterloop* (dat uitsluitend als nomen voorkomt), *zweefvliegen* (dat uitsluitend als infinitief voorkomt), *druiloor* (geen speciaal soort *oor*) en *vederlicht* (minder verbuigbaar dan *licht*).



Figuur 4-15: De relatie tussen het I-domein, het K-domein en het U-domein.

Het hier bedoelde effect, dat de stam uit het secundaire domein mede bepaalt hoe het U-domein van de kern is samengesteld, komt ook voor bij prefixgebaseerde stammen. Dit wordt inzichtelijk gemaakt door figuur 4-15. Alle taxemen uit het K-domein (te weten T_{K1}, T_{K2}, T_{K3}) hebben hier toegang tot een U-domein met de taxemen T_{U1} en T_{U2}; daarnaast is elk

afzonderlijk K-taxeem via een stippellijn met een uniek subdomein van U verbonden (resp. T_{U3}, T_{U4}, T_{U5}). Men zou I, K en U bijvoorbeeld als in (17) kunnen specificeren:

(17)

I = actie	K = actiepad	U = actieperspectief
i1: spring	k1: [0/ge]-	u1: #v-en (werkwoord)
i2: schiet	k2: ver-	u2: #n-er (agens nominalisatie)
i3: jaag	k3: be-	u3: #n-ge[..] (iteratief proces)
i4: zoek		u4: #n-ing (proces nominalisatie)
i5: rijd		u5: #a-baar (adjectief van potentie)

De stammen in I kunnen met alle prefixen in K worden gecombineerd, te weten het 0-prefix (dat alterneert met de prefixvorm GE-), het prefix VER- en het prefix BE-. Bovendien bezitten ze een gemeenschappelijk U-domein (te weten U_0), bestaande uit de U-opties V-lexeem (#v) en agensnominalisatie (#n-ER). Hiernaast is (bij wijze van voorbeeld) voor elk prefix nog minstens één aanvullende U-optie gespecificeerd. Blijkens deze informatie is het prefix BE- niet alleen compatibel met de standaard u-functors [#v-EN] en [#n-ER], maar ook met de u-functors [#n-ING] en [#a-BAAR].

4.3.6 Indexgebaseerde kennisopbouw

In deze subsectie wordt uiteengezet hoe de indexgebaseerde opbouw van het lexicon in zijn werk gaat. Paragraaf 1 legt uit wanneer een lexeemintern segment als (morfologische) index kan worden aangemerkt en hoe men langs distributieve weg onderscheid kan maken tussen stammen en functors. Paragraaf 2 demonstreert hoe men stamklassen en functorklassen kan construeren door te generaliseren over indexrepresentaties op een lager structuurniveau.

4.3.6.1 De identificatie van stammen en functors

Het lexicon wordt opgebouwd door alle in het lexicon opgenomen taxemen stap voor stap van (sub)indexen te voorzien, terwijl en passant hiërarchische structuur wordt aangebracht. De basisconfiguratie voor indexintroductie ziet er als volgt uit:

$$\begin{array}{l} t1 = [x1 + y1] \\ t2 = [x1 + y2] \\ t3 = [x1 + y3] \end{array} \quad \Rightarrow \quad x1 = \mathbf{i} \quad \Rightarrow \quad \begin{array}{l} t1 = [\mathbf{i} + y1] \\ t2 = [\mathbf{i} + y2] \\ t3 = [\mathbf{i} + y3] \end{array}$$

Deze representatie dient als volgt te worden geïnterpreteerd: indien er drie of meer taxemen t zijn die zo in componenten kunnen worden opgedeeld dat er één component is die een constante waarde bezit, namelijk x1 (bijv. een klanksegment f, of een vaste relatie R(f,s) tussen een klanksegment en een vormsegment), terwijl de andere component, namelijk y, een variabele waarde vertoont, mag de constante component door een index **i** worden vervangen. Hieronder wordt deze procedure toegepast op taxemen met de #v-stam SPEL:

$$\begin{array}{r} |spellen| = [|spel| + |len|] \\ |speller| = [|spel| + |ler|] \\ |spelbaar| = [|spel| + |baar|] \end{array} \quad \Rightarrow \quad |spel| = \mathbf{i_1} \quad \Rightarrow \quad \begin{array}{r} |spellen| = [\mathbf{i_1} + |len|] \\ |speller| = [\mathbf{i_1} + |ler|] \\ |spelbaar| = [\mathbf{i_1} + |baar|] \end{array}$$

In dit voorbeeld heb ik de analyse rechtstreeks op de spelvorm van de weergegeven taxemen gebaseerd. Hierbij heb ik de te substitueren component, namelijk de stamvorm *spel*, een constante vorm gegeven, met als gevolg dat het effect van de l-verdubbeling in de suffixcomponent is verwerkt. De suffixen -EN en -ER hebben daarom een extra *l* gekregen (resp. *len* en *ler*). Hierdoor is het niet mogelijk om deze segmenten rechtstreeks aan de segmenten *-en* en *-er* te koppelen, dus om ze als vormvarianten van de affixen -EN en -ER te analyseren; hiervoor moet eerst nadere informatie over hun distributiegedrag en/of betekenis beschikbaar komen. Dit probleem speelt ook bij de identificatie van de stam in de onderstaande taxeemvormen met de #v-stam $SPEEL_I$; deze taxeemvormen weerspiegelen hun uitspraak, waardoor

de stamvorm *speel* lichte variatie vertoont in de syllabe-opbouw (zoals uit de positie van het streepje blijkt); maar omdat deze stamvormen qua functie duidelijk bij elkaar horen, kunnen ze toch aan dezelfde index (i_2) worden gekoppeld. (De variatie in de klankvorm zou men kunnen verantwoorden door subindexen te introduceren).

$$\begin{array}{r} \text{/spee-len/} = [\text{/spee-l/} + \text{/en/}] \\ \text{/spe-ler/} = [\text{/spee-l/} + \text{/er/}] \\ \text{/speel-baar/} = [\text{/speel/} + \text{/baar/}] \end{array} \Rightarrow \left\{ \begin{array}{l} \text{/speel/} \\ \text{/spee-l/} \end{array} \right\} = \mathbf{i_2} \Rightarrow \begin{array}{r} \text{/spee-len/} = [\mathbf{i_2} + y1] \\ \text{/spee-ler/} = [\mathbf{i_2} + y2] \\ \text{/speel-baar/} = [\mathbf{i_2} + y3] \end{array}$$

In het derde voorbeeld treedt een nog sterkere vorm van variatie op, namelijk allomorfie: de #v-stam SPREEK neemt hier drie verschillende vormen aan. Maar ook voor deze niet-voorspelbare variatie geldt dat hij geen belemmering hoeft te zijn voor de toekenning van een gemeenschappelijke index (namelijk i_3), want dit kan worden gemotiveerd op basis van hun semantische en distributionele equivalentie.

$$\begin{array}{r} \text{bespreek} = [\text{be} + \text{spreek}] \\ \text{versprak} = [\text{ver} + \text{sprak}] \\ \text{gesproken} = [\text{ge} + \text{sproken}] \end{array} \Rightarrow \left\{ \begin{array}{l} \text{spreek} \\ \text{sprak} \\ \text{sproken} \end{array} \right\} = \mathbf{i_3} \Rightarrow \begin{array}{r} \text{bespreek} = [\text{be} + \mathbf{i_3}] \\ \text{versprak} = [\text{ver} + \mathbf{i_3}] \\ \text{gesproken} = [\text{ge} + \mathbf{i_3}] \end{array}$$

In de hier besproken voorbeelden heb ik me uitsluitend op de indexering van de stam gericht, maar het complement van de op deze wijze geanalyseerde eenheden zou eigenlijk ook een morfeemindex moeten krijgen, want dit complement correspondeerde in alle gevallen met een affix (dus met een morfeem). Voor deze indexeringsstap is echter aanvullende evidentie nodig van andere taxemen met hetzelfde affix. Het is ook denkbaar dat men sommige taxemen eerst met een affixindex verrijkt en pas later met een stamindex, bijvoorbeeld indien het affix veel frequenter is dan de stam. Dit leidt tot de vraag hoe men eigenlijk vaststelt welke component als stam fungeert en welke als affix (c.q. functor).Een lexeem als *lezer* kan bijvoorbeeld worden onderverdeeld in een morfeem LEES en een morfeem -ER. De vraag is nu welk van deze twee morfemen als functor moet worden aangemerkt.[180]

Deze kwestie kan niet worden opgelost door uit te gaan van de Rechterhoofdregel (RHR), want los van het feit dat deze "regel" zowel conceptueel als empirisch op grote problemen stuit,[181] correspondeert de RHR met een morfologiemodel waarin het hoofd (c.q. functor) zowel met een "stam" als met een "affix" kan corresponderen. Maar in de L-KRING-theorie corresponderen affixen per definitie met een functor, terwijl de stam is gedefinieerd als de drager van een of meer externe functortoepassingen, waarbij het zowel basisstammen als morfologisch gelede stammen kan betreffen. Als een morfeem zowel stamfuncties als affix-functies kan vervullen, corresponderen deze functies per definitie met verschillende representaties. In de L-KRING-benadering bestaat dus een duidelijke taakverdeling tussen stam en functor. Hierdoor kan de functor worden gedefinieerd als het element dat de grootste invloed heeft op de U-eigenschappen van de hiermee opgebouwde stam-affix-combinaties; dit impliceert dat de functor zich van de stam onderscheidt doordat deze niet door een andere eenheid kan worden gesubstitueerd zonder dat dit gevolgen heeft voor de U-eigenschappen van de eenheid als geheel, terwijl de stam juist legio substitutiemogelijkheden bezit. Hieruit volgt dat men de functor kan identificeren door een substitutietest uit te voeren:

[180] In de grammaticale lexiconbenadering wordt nauwelijks aandacht besteed aan dit soort kwesties, want meestal wordt aangenomen dat de morfeemregels ofwel aangeboren zijn, ofwel reeds verworven zijn.
[181] Dit wordt uitgebreid toegelicht in hoofdstuk 3, sectie 6.

Substitutietest
Gegeven een lexeem L met de morfologische segmentstructuur [E1+E2] en het uitwaartse paradigma U kan de functor worden geïdentificeerd door na te gaan welk lexeemintern segment E het meest bepalend is voor de samenstelling van het inflectieparadigma in het U-domein; dit is het segment E waarvoor het U-domein de meeste veranderingen ondergaat als het door een segment van een ander lexeem wordt vervangen (wat een lexeem [E1⊕E2'] resp. [E1'⊕E2] oplevert), dus het segment dat de minste U-neutrale substitutiemogelijkheden bezit.

Hieronder demonstreer ik deze test aan de hand van het lexeem *lezer* met de interne structuur [lees⊕er]. Uit de voorbeeldsubstituties in tabel 4-1 blijkt dat het morfeem LEES meer substitutiemogelijkheden biedt dan het morfeem ER. Hieruit volgt dat het statistisch gezien het meest waarschijnlijk is dat de functor van *lezer* met het suffix -ER correspondeert, conform de intuïtie. Bij substitutie van -ER door een ander suffix gaat het inflectieparadigma van *lezer* namelijk altijd verloren, maar er zijn wel tal van substitutiemogelijkheden voor de stam LEES. Omgekeerd geldt dat de kans dat de stam LEES de inflectiesuffixen -S en -TJE selecteert aanzienlijk kleiner is dan voor de functor -ER. Behalve dat dit verband (namelijk het verband LEES...{-S,-TJE}) op niet-locale selectierelaties berust, zoals zichtbaar is aan afleidingen als LEES⊕[ER]⊕S en LEES⊕[ER]⊕TJE (waarin niet het morfeem LEES, maar het mofeem -ER tussen stamhaken is geplaatst), is het meestal ook een ongeldig verband, blijkens de onwelgevormdheid van afleidingen als als *LEES⊕[ING]⊕S en *LEES⊕[BAAR]⊕TJE.

"I"	"K"	lexeem	U-paradigma
LEES	⊕ ER	*lezer*	U = {#n: 0,-S,-TJE}
	↓		
WERK	⊕ ER	*werker*	U = {#n: 0,-S,-TJE }
BREEK	⊕ ER	*breker*	U = {#n: 0,-S,-TJE }
LOOP	⊕ ER	*loper*	U = {#n: 0,-S,-TJE }
DROOM	⊕ ER	*dromer*	U = {#n: 0,-S,-TJE }
			4 identieke U's

"K"	"I"	lexeem	U-paradigma
LEES	⊕ ER	*lezer*	U = {#n: 0,-S,-TJE }
	↓		
LEES	⊕ STER	*lezeres*	U = {#n: 0,-EN,-JE}
LEES	⊕ ING	*lezing*	U = {#n, 0,-EN,-KJE}
LEES	⊕ BAAR	*leesbaar*	U = {#a: 0,-E}
LEES	⊕ EN	*lezen*	U = {#v: 0,-T,-EN}
			geen identieke U's

(↓ = substitutierelatie; "K" *= hypothetische kern;* "I" *= hypothetische I-stam)*
Conclusie: R_{LEX}(*lezer*) = [LEES ⊕< ER], met de stam LEES en de functor -ER

Tabel 4-1: Demonstratie van de substitutietest aan de hand van het lexeem lezer.

4.3.6.2 De introductie van stamindexen en functorindexen

Tabel 4-2 toont de morfotactische representatie en de hieraan verbonden selectiekenmerken van een aantal lexemen met eindsegment *ing*, te weten *herkenning*, *verzending*, *bespeling*, *ontleding*, *vertelling* en *beheersing*. Uit de tabel blijkt dat de lexemen uit klasse Y niet alleen overeenkomst in de vorm vertonen (aangezien ze allemaal op *ing* eindigen), maar ook in de betekenis ('het X-en', d.w.z. proces waarbij het V-concept ten uitvoer wordt gebracht of het resultaat van dit proces), de syntactische categorie (N), het bijbehorende inflectiepatroon (N-pl = -EN, d.w.z. meervoud op *en*) en de keuze van het lidwoord (bijv. *de* bij enkelvoud); men zou deze lijst nog kunnen uitbreiden, bijvoorbeeld met informatie over de door X bepaalde argumentstructuur, bijv. <Agens,Thema> indien X transitief is.

De L-KRING-theorie verantwoordt de overeenkomsten tussen deze lexemen door ze als instanties van dezelfde basisrepresentatie te representeren, namelijk Y = [[M_V X] ⊕< [M_N -*ing*]] (waarbij M voor een morfeem staat en waarbij X⊕<Y aangeeft dat Y functor is van X). Uit deze representatie blijkt dat lexeem Y is opgebouwd uit een V-stam (c.q. wortel) X en een N-functor (c.q. suffix) -ING (met de klankvorm *ing*). Bij de toepassing van dit patroon moet voor

elke instantie van X en Y een nieuwe index worden aangemaakt. In tabel 4-2 zijn deze instanties door een subscript i (namelijk i1, i2, etc.) gemarkeerd.

	lexeemvorm Y	**lexicale functiestructuur** Y = [[M_V X] ⊕< [M_N *ing*]]	**overerfbare lexeemkenmerken** N, N-pl = -EN, "het X-en", sg-lidw.= *de*
1.	*herkenning*	[[X=*herken*$_i$] ⊕< [M_N *ing*]$_{i1}$]	N, N-pl = -EN, "het X-en", sg-lidw.= *de*
2.	*verzending*	[[X=*verleg*$_i$] ⊕< [M_N *ing*]$_{i2}$]	N, N-pl = -EN, "het X-en", sg-lidw.= *de*
3.	*bespeling*	[[X=*bespeel*$_i$] ⊕< [M_N *ing*]$_{i3}$]	N, N-pl = -EN, "het X-en", sg-lidw.= *de*
4.	*ontleding*	[[X=*ontleed*$_i$] ⊕< [M_N *ing*]$_{i4}$]	N, N-pl = -EN, "het X-en", sg-lidw.= *de*
5.	*vertelling*	[[X=*vertel*$_i$] ⊕< [M_N *ing*]$_{i5}$]	N, N-pl = -EN, "het X-en", sg-lidw.= *de*
6.	*beheersing*	[[X=*beheers*$_i$] ⊕< [M_N *ing*]$_{i6}$]	N, N-pl = -EN, "het X-en", sg-lidw.= *de*

Tabel 4-2: Demonstratie van het overervingsprincipe: overerving met het suffix -ING.

De hier geanalyseerde ING-lexemen bezitten allemaal een stamlexeem met een prefix. Er zijn echter vele ING-lexemen waarvan de stam met een niet-geprefigeerd V-lexeem correspondeert (bijvoorbeeld *zitting*, *lezing*, *stalling* en *schutting*). De lexemen in tabel 4-2 vormen dus een subklasse van de lexemen die ING-affixatie kunnen ondergaan, namelijk V-lexemen met de structuur [M_V F ⊕< X'], waarbij F voor een willekeurige functor in prefixpositie staat en X' voor de gemodificeerde wortel (het resterende deel van lexeem X). Deze subklasse specificeert functor F als een overt prefix (F_P). De andere subklasse specificeert de functor F als een onhoorbaar element (F_0). Deze structuuropties kunnen economischer worden gerepresenteerd door ze op hetzelfde basispatroon te baseren, de hieraan verbonden kenmerken over te erven en deze informatie zonodig aan te vullen. Dit wordt in tabel 4-3 gedemonstreerd.

	lexicale functiestructuur	**overerfbare lexeemkenmerken**
patroon 1:	[[M_V F ⊕ X] ⊕< [M_N: *ing*]	patroon 1, [±dyn, ±N_{pl}]
variant 1a:	[[M_V F_P ⊕< X'] ⊕< [M_N: *ing*]	patroon 1a = patroon 1, [+dyn, -N_{pl}]
variant 1b:	[[M_V F_0 ⊕< X'] ⊕< [M_N: *ing*]	patroon 1b = patroon 1, [-dyn, +N_{pl}]

Tabel 4-3: Nadere specificatie van de toepassingscontexten van het suffix -ING.

Hierbij definieert patroon 1 de meest algemene lexeemcontext van het suffix -ING (althans, voorzover dit voorbeeld reikt), terwijl de varianten 1a en 1b deze lexeemcontext en het bijbehorende overervingspatroon specifieker invullen. In dit voorbeeld hebben de lexemen van variant 1a een voorkeur voor een dynamische lezing ([+dyn]), namelijk een interpretatie als proces (waarbij doorgaans geen meervoudsvorming mogelijk is, wat gemarkeerd is als [-N_{pl}]), terwijl de lexemen van variant 1b een voorkeur vertonen voor een statische lezing ([-dyn]), namelijk een interpretatie als situatie of object (waarbij wel meervoudsvorming mogelijk is, wat gemarkeerd is als [+N_{pl}]). Als deze generalisaties alleen opgaan voor een subklasse van de varianten 1a en 1b, dienen deze varianten verder te worden opgesplitst.

	stam S=[M_V X]	**suffixoptie 1** S ⊕< [M_N *ing*]	**suffixoptie 2** S ⊕< [M_{INF} *en*]	**suffixoptie 3** S ⊕< [M_A *baar*]	**suffixoptie 4** S ⊕< [M_N *er*]
1.	*herken*	$f_{1.1}$	$f_{1.2}$	$f_{1.3}$	-
2.	*verzend*	$f_{2.1}$	$f_{2.2}$	$f_{2.3}$	$f_{2.4}$
3.	*bespeel*	$f_{3.1}$	$f_{3.2}$	$f_{3.3}$	$f_{3.4}$
4.	*ontleed*	$f_{4.1}$	$f_{4.2}$	-	-
5.	*vertel*	$f_{5.1}$	$f_{5.2}$	-	$f_{5.3}$
6.	*beheers*	$f_{6.1}$	$f_{6.2}$	$f_{6.3}$	$f_{6.4}$
=	6	6	6	4	4

Tabel 4-4: De lexicale combinatiemogelijkheden van enkele stamlexemen met categorie V.

Veel stamlexemen kunnen met verschillende affixen worden gecombineerd. Zo kunnen de in tabel 4-3 opgenomen stammen naast het suffix -ING allemaal het suffix -EN selecteren, terwijl een deel van deze lexemen ook -BAAR en -ER toestaan. In de L-KRING-theorie wordt dit formeel verantwoord door voor elk affix een index aan te maken, waarbij elke index toegang geeft tot lexeemspecifieke frequentie-informatie. Dit wordt in tabel 4-4 gedemonstreerd. Elke bestaande lexeem-suffix-combinatie correspondeert hier met een index (c.q. frequentieteller) $f_{m.n}$ (waarbij 'm.n' met een unieke nummercombinatie correspondeert en waarbij de index als geheel informatie geeft over de tokenfrequentie van de bijbehorende morfeemcombinatie), terwijl niet-bestaande combinaties met een streepje corresponderen. De tabel wijst uit dat het stamlexeem *herken* drie verschillende suffixen kan selecteren, namelijk [M_N *ing*], [M_{INF} *en*] en [M_A *baar*], maar dat het suffix [M_N *er*] niet beschikbaar is (dus dat de betreffende vorm niet bekend is). Bovendien blijkt *herken* tot dezelfde klasse te behoren als de stamlexemen *verzend*, *bespeel* en *beheers*, want ze kunnen allemaal de suffixen [M_N *ing*], [M_{INF} *en*] en [M_A *baar*] selecteren. In de L-KRING-theorie kan dit formeel worden verantwoord door de betreffende stammen als instanties van een paradigmatische lexeemfamilie U te analyseren, in dit geval [M_V F⊕<X'] ⊕< {U: [M_N *ing*], [M_{INF} *en*], [M_A *baar*]}]; hierbij specificeert de door 'U' gemarkeerde component het affixparadigma. Bij de constructie van paradigmatische lexeemfamilies moet eerst de grootste familie worden gedefinieerd (door selectie van de twee hoogstfrequente affixen), om deze familie vervolgens stap voor stap in subfamilies onder te verdelen door het gespecificeerde U-paradigma uit te breiden met het hoogstfrequente suffix van de resterende verzameling en de selectiekenmerken voor het bijbehorende stamdomein aan te passen.

De hier beschreven analysemethode leidt tot een sterk gecomprimeerde, dus economische opslag van lexicale kennis. Bij consequente toepassing van deze methode zouden de gangbare lexeemklassen vanzelf boven water moeten komen, mits de analyse op een compleet lexicon wordt uitgevoerd. Anders is het niet mogelijk zijn systematisch na te gaan welke correlaties er bestaan tussen stammen en lexicale selectiekenmerken, laat staan welke patronen generaliseerbaar zijn naar nieuwe woorden.

4.3.7 De productieve toepassing van distributiepatronen

Een overervingspatroon leent zich beter voor de constructie van nieuwe lexemen naarmate dit patroon betrouwbaarder is als generalisatie over het combinatorische gedrag van de stammen uit het lexicale toepassingsdomein van dit patroon. Zo zou men de minimumeis kunnen hanteren dat de introductie van een morfologisch patroon pas acceptabel is als minstens 60% van de stammen uit de intensionele karakterisering van het toepassingsdomein eraan voldoet. Gegeven de informatie in tabel 4-4 zou het patroon [[M_V F⊕<X'] ⊕< [M_N *er*]] bijvoorbeeld op minimaal vier van de zes stammen toepasbaar moeten zijn. Aan dit criterium wordt inderdaad voldaan. Dit betekent dat men een morfologische regel kan postuleren die stelt dat alle lexeemstammen met de structuur [M_V F⊕<X'] het suffix -ER kunnen selecteren, dus ook de lexeemstammen HERKEN$_V$ en ONTLEED$_V$ (wat *herkenner* en *ontleder* zou opleveren). Dit soort limietwaardes kunnen niet worden voorspeld, maar vereisen empirisch onderzoek (dat mogelijk ondersteund kan worden door computationele simulaties van het mentale lexicon).

De in het lexicon aanwezige patronen kunnen ook worden benut om nieuwe lexemen te analyseren en om voorspellingen te doen over de combinatorische eigenschappen van hun stammen. Indien men bijvoorbeeld de lexeemvorm *ontvanging* zou tegenkomen (met dezelfde betekenis als de gebruikelijke lexeemvorm *ontvangst*), kan dit lexeem direct als een ING-constructie met de stam ONTVANG worden geanalyseerd, dus als een lexeem met de structuur [[M_V F⊕<X'] ⊕< [M_N *ing*]] (en de bijbehorende overervingskenmerken). Hieruit kan worden afgeleid dat deze stam ook met de affixen [M_N *er*], [M_{INF} *en*] en [M_A *baar*] kan worden

gecombineerd. Op dezelfde manier kan de lexeemvorm *verluking* direct als een ING-constructie met de nog onbekende stam VERLUUK worden geanalyseerd, waarna deze stam op dezelfde wijze met informatie over zijn combinatorische eigenschappen kan worden verrijkt. In de L-KRING-theorie kan elk lexicaal patroon op deze wijze worden gegeneraliseerd, al kunnen er grote verschillen bestaan in de waarschijnlijkheid dat deze patronen daadwerkelijk worden geactiveerd.

Elke functor F kan in beginsel productieve toepassingen krijgen. Hierbij kan men onderscheid maken tussen nieuwe I-toepassingen (waarbij F met een stam wordt gecombineerd) en nieuwe U-toepassingen (waarbij de door F gevormde eenheid de basis vormt voor de toepassing van een functor). Maar niet alle I- en U-combinaties zijn even waarschijnlijk. Dit hangt namelijk af van de vraag in hoeverre de nieuwe combinaties met het reeds bestaande distributiepatroon overeenstemmen. Dit kan men achterhalen door na te gaan hoe de samenstelling van het U-domein afhangt van die van het I-domein. Zo kan men onderscheid maken tussen U-eenheden die tot het inherente U-domein behoren (namelijk U-eenheden waarvoor geldt dat deze met alle stammen uit het I-domein kunnen samengaan) en U-eenheden die tot een potentieel U-domein behoren, bijvoorbeeld het 60^{+}%-domein; dit domein omvat alle U-eenheden die nog niet in een U-domein met een hoger selectie-pecentage zijn opgenomen, en waarvoor geldt dat ze door minimaal 60% van de I-stammen worden geselecteerd. Dit is een maat voor de waarschijnlijkheid dat de overige I-stammen in staat zullen zijn alsnog een combinatie aan te gaan met deze U-eenheden. In beginsel dient voor elke combinatie van I-kenmerken te worden nagegaan wat de bijbehorende distributieschema's zijn, want hoe gedetailleerder deze informatie in kaart wordt gebracht, hoe beter de voorspellingen van het model zijn. In het L-KRING-model is deze informatie automatisch af te leiden.

I-domein	**kern**	**100% U-domein**	**60^{+}% U-domein**
X1	-en$_{V}$	$O_{\$V}$:, $[O]_{\$V}$	$[C: s]_{L2}$
X2	-ing	$O_{\$N}$, $[C]_{L2}$	-$[kje]_{\#N}$
X3	-er	$O_{\$N}$, $[C: s]_{L2}$	-$[es]_{\#N}$, -$[schap]_{\#N}$
X4	-baar	$O_{\$A}$, -$[heid]_{\#N}$	-$[der]_{\#A}$/-$[st]_{\#A}$

Tabel 4-5: Het 100% en 60^{+}% U-domein van de functors -EN$_{INF}$, -ING, -ER en -BAAR.

I-domein	**kern**	**100% U-domein**	**60^{+}%U-domein**
$[lees]_{\#V}$	-en$_{V}$	$[lezen]_{\$V}$	$[lezens(waardig)]_{L2}$
$[lees]_{\#V}$	-ing	$[lezing]_{\$N}$, $[lezingen(reeks)]_{L2}$	$[lezinkje]_{\#N}$
$[lees]_{\#V}$	-er	$[lezer]_{\$N}$, $[lezers(onderzoek)]_{L2}$	$[lezeres]_{\#N}$, $[lezerschap]_{\#N}$
$[lees]_{\#V}$	-baar	$[leesbaar]_{\$A}$, $[leesbaarheid]_{N}$	$[leesbaar\{der/st\}]_{\#A}$

Tabel 4-6: Een voorbeeldtoepassing van de informatie uit tabel 4-5: specificatie van de derivatiemogelijkheden van de I-stam LEES$_{\#V}$.

De tabellen 4-5 en 4-6 geven een mogelijke invulling van het 100% en het 60^{+}% U-domein van de suffixen -EN$_{INF}$, -ING, -ER en -BAAR. Deze functors dienen als kern van de hier uitgewerkte distributieschema's. Tabel 4-5 geeft per suffix een intensionele karakterisering van de selectiemogelijkheden in het U-domein. De functors uit de U-domeinen kunnen tot drie hoofdklassen behoren, namelijk lexeemvormende operators $O_{\$Z}$ (waarbij Z de lexeemklasse specificeert), combinators $[C:Y]_{T}$ (waarbij Y optionele C-markeringen specificeert en waarbij T aangeeft welk type taxemen nodig is; C_{L} correspondeert met lexemen) of stamvormende functors, zoals het #n-stam-vormende suffix -[heid]#n. Het I-domein bestaat uit een stamverzameling X die per functor kan verschillen, resp. X1, X2, X3 of X4. Tabel 4-6 toont een aantal mogelijke toepassingsproducten bij tabel 4-5 (op basis van de I-stam LEES$_{\#V}$).

De U-domeinen van de weergegeven functors omvatten relatief weinig derivationele affixen: eigenlijk kent alleen het suffix -BAAR "productieve" derivatiemogelijkheden, namelijk met het suffix -HEID (tot *baarheid*) en met de A-modificators -DER en -ST. Het inherente U-domein van de andere suffixen (te weten -EN, -ER en -ING) omvat alleen lexeemoperators (en de bijbehorende inflectiesuffixen), namelijk de lexeemoperator voor categorietoekenning (te weten $O_{\$V}$, $O_{\$N}$ en $O_{\$A}$) en de lexeemoperator voor modificatie (O_{mod}). Indien ook het [60^{+}%] U-domein in ogenschouw wordt genomen, zijn er iets meer derivaties mogelijk, want -ING kan in sommige gevallen het verkleinsuffix -KJE selecteren, wat de suffixcombinatie -[INK+JE] oplevert, terwijl -ER in sommige gevallen het suffix -SCHAP kan selecteren, wat de suffix-combinatie -[ER+SCHAP] oplevert.

4.4 Conclusie

De in dit hoofdstuk beschreven L-KRING-theorie is een concrete poging om een formeel representatiesysteem te ontwikkelen dat aan alle eisen van een Integraal Dynamisch Lexicon-systeem (IDL-systeem) kan voldoen. De L-KRING-theorie berust op het uitgangspunt dat taalgebruikers over een cognitief analysesysteem beschikken waarmee alle binnenkomende woorden van een representatie kunnen worden voorzien door ze stap voor stap in reeds opgeslagen lexicale eenheden op te delen, waarna het mentale lexicon deze morfologisch gestructureerde representaties integraal kan opslaan. Dit is technisch mogelijk door aan te nemen dat morfologische bouwstenen in feite indexen zijn die naar gedeelde informatie-eenheden verwijzen (met zowel fonologische als semantische kenmerken), en dat het analyseren van een woord equivalent is aan het substitueren van gemeenschappelijke eenheden door indexen die naar de locatie verwijzen waar de betreffende informatie-eenheden worden gedefinieerd. Deze representatiewijze verklaart niet alleen waarom woorden (en woordgroepen) interne structuur bezitten, maar ook hoe het mogelijk is dat woorden met onvoorspelbare vorm- en betekeniskenmerken toch morfologisch complex zijn in de zin dat ze voorspelbare woordeigenschappen vertonen, en dat hun herkenbaarheid gevoelig is voor de gebruiksfrequentie en de familiegrootte van hun interne bouwstenen. Dankzij deze eigenschappen biedt de L-KRING-theorie een krachtig alternatief voor alle bestaande morfologiemodellen.

5 Ontwerp en aanmaak van de Morfologische Gegevensbank

5.1 Introductie

In dit hoofdstuk bespreek ik het ontwerp en de aanmaak van de Morfologische Gegevensbank voor het Nederlands (MGBN). Zoals in hoofdstuk 1 aan de orde kwam, dient de MGBN een bijdrage te leveren aan de systematisering van de woordkenmerken in VDL's Woord-KenmerkenBank Nederlands (WKB-Ned) door morfologische informatie te verstrekken over de hierin opgenomen woorden. De MGBN is tot stand gekomen door de spelvorm van deze woorden op het niveau van de basislexemen van een morfologische structuurlaag te voorzien. Hierbij heb ik de structuurprincipes van mijn L-KRING-theorie toegepast.

Zoals in hoofdstuk 4 uiteen is gezet, berust de L-KRING-theorie op het idee dat de morfologische structuur van mentale lexeemrepresentaties een bijverschijnsel is van het streven om deze lexemen zo gecomprimeerd mogelijk op te slaan. Hiertoe dienen taalgebruikers voor alle woorden die ze tegenkomen een mentale lexeemrepresentatie aan te maken; het mentale lexicon kan deze representaties vervolgens intern structureren door langs inductieve weg op zoek te gaan naar eenheden met een systematische relatie tussen vorm, betekenis en combinatiemogelijkheden.[182] In deze benadering wordt het morfologische regelsysteem dus niet gemotiveerd door de potentiële woordenschat (in de vorm van productieregels), maar door de waarneembare woordenschat (in de vorm van redundantieregels).[183]

De inductieve analysemethode van de L-KRING-theorie vormt de theoretische grondslag voor de ontwikkeling van de MGBN. Omgekeerd is de ontwikkeling van de MGBN een middel om een L-KRING-model op te bouwen van de Nederlandse woordvormingspatronen.[184] Om dit doel te bereiken heb ik voor alle basislexemen onderzocht welke segmenten naar mijn intuïtieve oordeel (dus niet op basis van "regels") morfologisch relevant zijn, d.w.z. welke vormsegmenten vaak dezelfde betekenis en/of voorspelbare distributiekenmerken bezitten. Ter vergroting van de werksnelheid en de consistentie heb ik een semi-automatische werkwijze gevolgd, wat inhoudt dat ik de lexemen in een cyclisch proces afwisselend "handmatig" en langs automatische weg van structuur heb voorzien en op consistentie heb gecontroleerd. Wegens de grote omvang van de te analyseren lexeeminventarisatie was het niet mogelijk om rekening te houden met semantische transparantie.[185]

De hier beschreven werkwijze heeft een grote inventarisatie van morfologisch geanalyseerde lexemen opgeleverd, waarbij de toegekende morfeemrepresentaties een gesystematiseerde afspiegeling vormen van de morfeemrepresentaties in mijn eigen mentale lexicon. Doordat deze morfeemrepresentaties een formele (niet-semantische) basis hebben, biedt de door mij ontwikkelde gegevensbank unieke mogelijkheden voor onderzoek naar de vraag in hoeverre formele morfemen (d.w.z. de segmenten die potentieel als morfeem kunnen dienen) een

[182] Hierbij moet het aantal toepassingen boven een nader te bepalen minimum uitkomen.
[183] Dit type analyse valt buiten het bereik van de grammaticale standaardtheorie, want in deze theorie is morfologische structuur geen permanent beschikbare woordeigenschap, maar een hulpmiddel om nieuwe woorden te construeren en van betekenis te voorzien. Zodra een woord (c.q. lexeem) eenmaal gevormd is en in het lexicon van een taal is opgenomen, kan zo'n woord de status van autonome kenniseenheid krijgen.
[184] Het lexicon van "mijnheer Van Dale" heeft een omvang van ca. 250.000 trefwoorden, die uit ca. 80.000 basislexemen zijn opgebouwd. Hiermee is dit lexicon (dat vol zit met archaïsche en vakspecfieke woorden) zeker twee keer zo groot als de woordenschat van een goedgeschoolde Nederlander.
[185] Deze eis zou veel complicaties met zich meebrengen, want vrijwel alle woorden die formeel uit morfemen zijn opgebouwd, bezitten naast de compositionele betekenis ook gelexicaliseerde betekenissen; verder bestaat er meestal geen scherpe grens tussen compositionele en niet-compositionele morfeemtoepassingen.

voorspelbaar effect hebben op de lexeemkenmerken. Deze informatie is niet alleen van belang met het oog op de morfologische theorievorming over het Nederlands, maar kan ook ingezet worden voor de verbetering van de MGBN en de systematisering van de lexeemkenmerken in de WKB-Ned.

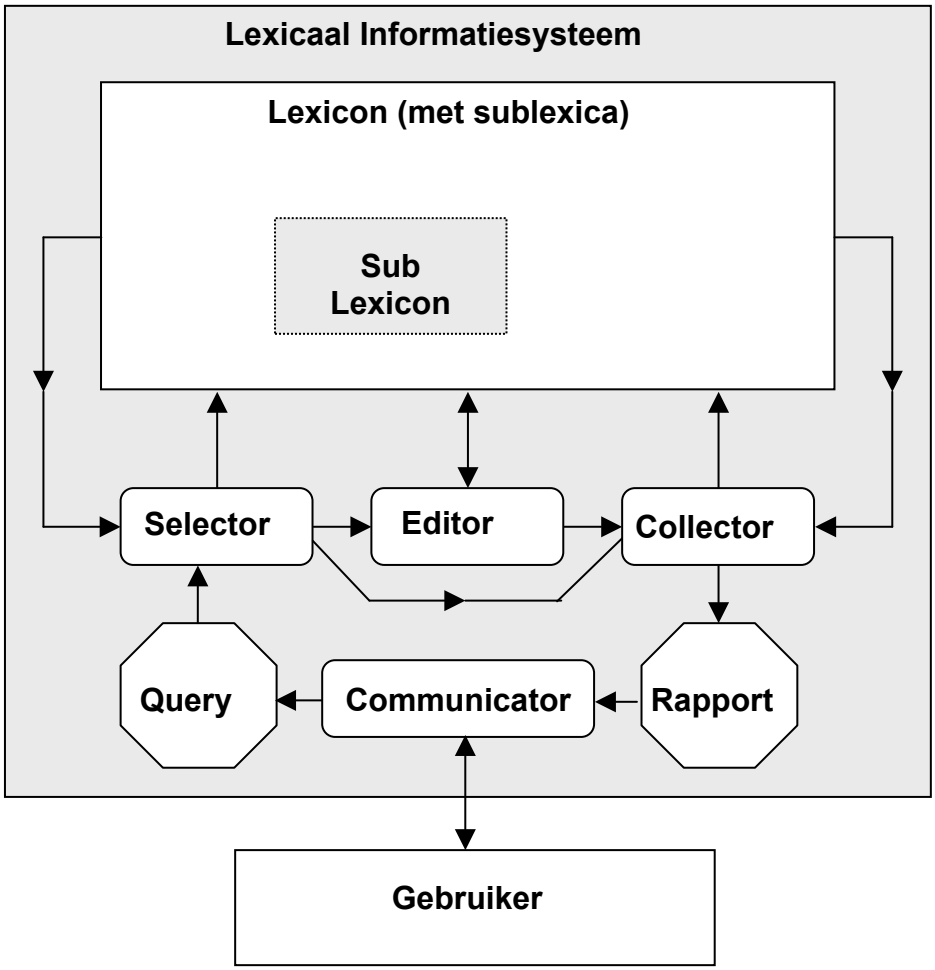
5.2 Het theoretische ontwerp

5.2.1.1 De structuur van het informatiesysteem

In mijn visie op het mentale lexicon bestaat er een fundamenteel verschil tussen de wijze waarop mensen bestaande woorden waarnemen en de wijze waarop deze woorden lexicaal zijn gerepresenteerd. Want terwijl mensen het gevoel hebben dat woorden met zelfstandige (niet-deelbare) taaleenheden corresponderen, stelt mijn theorie dat morfologisch complexe woorden niet rechtstreeks in het lexicon zijn terug te vinden, maar alleen als een hiërarchisch gestructureerde sequentie van indexen, waarbij elke index naar een door meerdere woorden gedeelde structuureenheid (c.q. bouwsteen) verwijst. Zoals reeds in hoofdstuk 4 aan de orde kwam, leidt dit abstracte representatiesysteem tot een aanzienlijke compressie van het lexicon, terwijl een hoge mate van lexicale samenhang ontstaat. Voor de taalgebruiker is deze interne, compositionele structuur niet zomaar toegankelijk: hij kan de lexicaal opgeslagen woorden pas waarnemen als de bouwstenen zijn samengevoegd tot een groter geheel, namelijk de woordvorm (hetzij als spelvorm, hetzij als klankvorm) of de betekenis.

Deze informatiekloof kan worden overbrugd door het lexicon in een lexicaal informatiesysteem in te bedden. Dit informatiesysteem heeft de taak om de zoekwensen van de gebruiker in een zoekprocedure, namelijk de (formele) Query, om te zetten, deze zoekprocedure uit te voeren, de gevraagde kenmerken te verzamelen en desgewenst te wijzigen en de verzamelde informatie, namelijk het (formele) Rapport, vervolgens in een voor de gebruiker begrijpelijk formaat te presenteren. Hiertoe is het informatiesysteem met een Communicator, een Selector, een Editor en een Collector uitgerust. Het schema in figuur 5-1 laat zien hoe deze componenten zich tot het lexicon en tot de gebruiker verhouden.

De Communicator verzorgt de communicatie tussen gebruiker en informatiesysteem. Hierbij kunnen twee hoofdfuncties worden onderscheiden, namelijk de omzetting van gebruikersvragen (c.q. Queries) in systeeminstructies en de omzetting van systeemgegevens in voor de gebruiker toegankelijke Rapporten. De Selector draagt zorg voor de activatie van de lexicale eenheid of eenheden waar de gebruiker informatie over wil opvragen (waarbij wordt uitgegaan van de criteria in de Query) door deze eenheden in een nader te specificeren sublexicon op te zoeken en te activeren. De Collector verzamelt vervolgens alle kenmerken waarover de gebruiker geïnformeerd wil worden (bijv. de betekenis en de woordcategorie), waarna deze informatie in een Rapport wordt verwerkt. Door de indexgebaseerde zoekmethode is voor elke woordinterne eenheid een aparte zoekstap nodig. Indien een zoekstap meerdere indexkandidaten oplevert, moeten deze kandidaten net zolang worden vastgehouden totdat er een keuze kan worden gemaakt. Desgewenst kan de gebruiker ook aangeven dat hij één of meer van de geactiveerde indexen wil wijzigen. In dat geval dient de Editor te worden ingeschakeld. De Editor is een module die wijzigingen kan aanbrengen in de inhoud van het lexicon, zoals de opslag van nieuwe lexicale eenheden, het aanpassen van hun gebruiksfrequentie en het doorvoeren van correcties.



Figuur 5-1: De hoofdcomponenten van het lexicale informatiesysteem.

5.2.1.2 De domeinparameters

Het lexicon kent drie domeinparameters, namelijk [±mentaal], [±idiolect] en [±diachroon]. De parameter [±mentaal] geeft aan of het om een mentaal of een computationeel informatiesysteem gaat. Deze keuze heeft ook gevolgen voor de selectie van de gebruiker. Indien er sprake is van een mentaal lexicon ([+mentaal]), correspondeert de gebruiker van nature met een *persoon*, maar het is ook mogelijk om dit informatiesysteem als *onderzoeker* te benaderen, namelijk in situaties waarin de gebruiker bewust nagaat wat voor kennis er in het mentale lexicon zit (bijvoorbeeld ten behoeve van taalkundig onderzoek). Indien de L-KRING-theorie als lexicografisch systeem wordt gerealiseerd, kan de gebruiker eveneens een *persoon* zijn (namelijk de raadpleger van een elektronisch woordenboek, al dan niet ten behoeve van taalkundig onderzoek), maar ook een *redacteur* (die het lexicon van dit woordenboek bewerkt) of een *applicatie* (zoals een automatische spellingchecker of een voorleesprogramma).

Het basismodel biedt ook de mogelijkheid om het centrale lexicon onder te verdelen in hiërarchisch geclassificeerde sublexica. Allereerst kan het lexicon verschillende talen omvatten, waarbij de semantische informatielaag de mogelijkheid biedt om vertaalrelaties tussen deze talen te leggen.[186] Gegeven een specifieke taal is het ook mogelijk om aan te geven of het lexicon met het taalgebruik van een specifieke persoon correspondeert (= [+idiolect]) of met de taalkennis van de taalgemeenschap als geheel (= [-idiolect] c.q. *sociolect*). In het

[186] Het vertaalsysteem SIMULLDA van Janssen (2002) laat zien hoe dit idee concreet kan worden uitgewerkt

laatste geval kan men een aanvullend onderscheid maken tussen synchrone kennis (over de actuele woordenschat) en diachrone kennis (over historische stadia van de woordenschat, dus over de etymologische samenhang van de woordenschat), namelijk door specificatie van de basisparameter [±diachroon]. Individuele taalgebruikers beschikken per definitie over een idiolect (de taal die ze in eigen kring spreken, bijvoorbeeld een dialect), maar de meeste taalgebruikers zijn ook in staat om over te schakelen op de standaardtaal (c.q. sociolect), in elk geval passief; in feite is hier sprake van een continuüm tussen de polen "formeel" en "informeel" taalgebruik.[187]

In de L-KRING-theorie corresponderen al deze informatiedomeinen met hetzelfde basismodel en dezelfde informatiestructuur, maar de "technische" invulling van het informatiesysteem is natuurlijk sterk afhankelijk van de parameter [±mentaal], terwijl de inhoud van het lexicon sterk afhangt van de parameters [±idiolect] en [±diachroon], en natuurlijk ook van de gespecificeerde taal. In het ideale geval zijn idiolect en sociolect identiek, en omvat het elektronische woordenboek dezelfde kennis als het mentale lexicon.[188]

5.2.1.3 De selector

Indien de gebruiker de betekenis van de Nederlandse woordvorm *bewerking* zoekt, zal de selector de opdracht (c.q. query) ontvangen om de orthografische woordvorm *bewerking* te zoeken. De selector zal deze woordvorm eerst in segmenten onderverdelen (wat tot de representatie [*b e w e r k i n g*] leidt) en dan proberen om de bijbehorende indexen in het lexicon te vinden; zodra alle fonemen zijn gevonden, kan de selector proberen om deze grafeemreeks in morfologische segmenten onder te verdelen, bijvoorbeeld als volgt: [m1: *be*] + [m2: *werk*] + [m3: *ing*]]. Tegelijk dient ook een hiërarchische structuur te worden toegekend, bijvoorbeeld m3 > m1 > m2. Volgens deze hiërarchie staat m3 hoger in de hiërarchie dan m1 en m2, wat betekent dat de eigenschappen van de resulterende constructie sterker door m3 worden bepaald dan door de hieraan ondergeschikte morfemen. Hieruit volgt dat m3 functor is bij de stam (m1+m2); binnen deze stam treedt het morfeem m1 als functor op bij m2, die dus met de kleinste stam correspondeert. Bij het huidige voorbeeld leidt deze hiërarchie tot de volgende structuurrepresentatie van lexeem L: L = [m1: *be*] ⊕> [m2: *werk*] <⊕ [m3: *ing*]. Hierbij correspondeert de structuur a⊕b met de (hiërarchische) *compositie* van segment a en segment b, terwijl > en < de hiërarchische ordening markeren (zo betekent a⊕>b dat a functor is bij b).

Indien het te analyseren woord nog niet in het lexicon is opgenomen, maar wel uit bestaande morfemen is opgebouwd, kan in beginsel dezelfde procedure worden gevolgd; maar het verschil is dat de compositie-operator geen gebruik kan maken van een lexicaal opgeslagen resultaat, maar actief een nieuwe eenheid moet construeren. Hierdoor kost de identificatie van nieuwe woorden meer tijd dan de identificatie van bestaande woorden. Stel bijvoorbeeld dat het lexeem *bewerksel* een nieuwvorming is met de bestaande morfemen BEWERK en -SEL. In dat geval kan de morfeemcombinatie BE+WERK direct uit het lexicon worden gehaald, maar zal de combinatie BEWERK+SEL actief moeten worden aangemaakt. Eenmaal aangemaakt kan deze morfeemcompositie (namelijk BEWERKSEL) echter als kant-en-klare eenheid L in het

[187] In de taalkundige theorievorming werd [+mentaal] lange tijd gelijk gesteld aan [+idiolect] en [-mentaal] aan [-idiolect]. Dat belet taalkundigen overigens niet om hun eigen taalintuïties als basis te nemen voor generalisaties over de hele taalgemeenschap, terwijl woordenboeken in de praktijk heel wat lemma's bevatten met idiolectisch getinte informatie (cf. Verkuyl 1993; zie ook sectie [4.1]). De door mij ontwikkelde morfologische gegevensbank heeft eveneens een hybride status: de gegevensbank berust namelijk op een lexicon met de basiskenmerken [-idiolect], [+diachroon], maar de hieraan toegekende morfologische structuur is noodgedwongen [+idiolect] van aard (in afwachting van verder onderzoek).

[188] Door inherente verschillen tussen deze media kan de "technische" implementatie van het informatiesysteem aanzienlijk verschillen.

lexicon worden opgeslagen. Indien het lexeem *bewerksel* opnieuw wordt aangeboden, zal dus een snellere identificatie mogelijk zijn.

5.2.1.4 De collector

Zodra er een woordindex is geselecteerd, kan de collector nagaan wat de betekenis is van deze woordindex; indien er meerdere betekenissen beschikbaar zijn, kan een voorkeursvolgorde worden aangebracht, bijvoorbeeld door op de gebruiksfrequentie te letten of op basis van contextuele informatie. Tot slot dient de collector een rapport op te stellen waarin de gevonden betekenis(sen) worden vermeld of een door de gebruiker bepaalde selectie, bijvoorbeeld alleen de meest waarschijnlijke betekenis. Maar de gebruiker kan ook vragen om alle betekenismogelijkheden te vermelden, en bovendien per betekenis aan te geven wat voor morfologische en syntactische eigenschappen ermee samengaan. En indien het lexicon meerdere talen bevat, kan de gebruiker vragen om de opgegeven woordvorm te vertalen (via een gemeenschappelijke betekenisdefinitie).

Deze rapportagemogelijkheden bestaan zowel voor een mentaal lexicon als voor een computationeel informatiesysteem: alle zoekopdrachten die men aan de computer geeft kan men immers ook aan zichzelf geven (dus aan het mentale lexicon). Gegeven dit perspectief op het mentale lexicon kan een woordenboek (d.w.z. een lexicon in boekvorm) worden gedefinieerd als het resultaat van een zoekprocedure waarbij alle indexen (c.q. lemma's) zijn geselecteerd waarvan de gebruiksfrequentie boven een bepaald minimum uitkomt, waarbij elk lemma bijvoorbeeld is opgebouwd uit velden met de orthografische vorm, de uitspraak, de syntactische klasse, inflectievormen, collocaties, etymologische informatie en hoofdbetekenissen.

5.2.1.5 De editor

De editor speelt een cruciale rol bij de opbouw en aanpassing van het lexicon. Hij kan op twee manieren worden aangestuurd, namelijk via de collector of door directe manipulatie via de user-interface. In een mentaal systeem correspondeert de eerste route met onbewuste aanpassing van het lexicon, en de tweede route met bewuste aanpassing (zoals het corrigeren van een verkeerde uitspraak).[189] Bij de verwerving van een taal zal de collector regelmatig nieuwe woorden tegenkomen, waarvan een groot aantal een nog onbekend woordvormingspatroon vertoont of gebruik maakt van een nieuwe stam. In al deze gevallen zal de editor de instructie ontvangen om het betreffende woord aan het lexicon toe te voegen, en hetzelfde te doen voor nog onbekende componenten binnen het woord (mits er voldoende andere woorden zijn met dezelfde component). Op deze manier kan stap voor stap een volwaardig lexicon worden opgebouwd; vanaf dat moment zal de editor alleen nog nodig zijn om nieuwe instanties van bestaande woordvormingspatronen toe te voegen en om na voltooiing van een query de gebruiksfrequentie van de geselecteerde indexen op te hogen.

5.2.2 De inhoud van het lexicon

In het door mij beoogde eindstadium dient de MGBN voor elk hierin opgenomen lexeem een multidimensionale, op L-KRING-principes gebaseerde morfeemstructuur te specificeren. Hoewel deze morfeemstructuur zowel een fonologische als een semantische dimensie heeft, beperk ik me hier tot de uitwerking van de fonologische dimensie (die uit een orthografische en een auditieve subdimensie bestaat). Voor een nadere uitwerking van de semantische dimensie is verder onderzoek nodig. Indien sprake is van een morfologisch gelede eenheid, is altijd sprake van een hiërarchische ordening in de zin dat elk segment met een uniek

[189] Ik heb de werking van de mentale editor bewust ervaren toen ik ontdekte dat het Engelse woord *schema* met een /k/ moet worden uitgesproken en niet met een /sj/; het heeft enige weken geduurd voordat dit inzicht volledig in mijn mentale lexicon was geïntegreerd; tot die tijd betrapte ik mijzelf nog vaak op spontane toepassing van het oude (verworpen) klankbeeld.

functorniveau correspondeert: hoe hoger het functorniveau, hoe groter de invloed op de woordkenmerken (en hoe minder substitutiemogelijkheden); de diepst ingebedde stam correspondeert altijd met functor-niveau 0. Indien een woord hiërarchische structuur bezit, dient deze structuur in alle representatiedimensies te worden doorgevoerd. Ik zal de door mij beoogde morfeemstructuur toelichten aan de hand van een reeks voorbeelden. Beschouw om te beginnen de lexicale representatie van het lexeem *gaan* (dat met de infinitiefvorm van een werkwoord correspondeert):

<lexeemindex>	gaan:1
<lexeemcategorie>	$v-inf
<lexeemstatus>	zelfstandig
<lexeempositie>	0
<mstructuur0>	1:[0/GE]$_{\#V}$ ⊕ 0: #w-GAAN
<orth-representatie>	|gaan|
<mstructuur1-orth>	1:|0| ⊕ 0:|gaan|
<mstructuur2-orth>	1:|[0/GE]$_{\#V}$| ⊕ 0:|[#w-GAAN]|
<audi-representatie>	/gaan/
<mstructuur1-audi>	1:/0/ ⊕ 0:/gaan/
<mstructuur2-audi>	1:/[0/GE]$_{\#V}$/ ⊕ 0:/[#w-GAAN]/

Hieronder volgt een toelichting op de in deze tabel opgenomen informatievelden:

<lexeemindex>	citatievorm van het lexeem, gevolgd door een betekenis-index: 'gaan:1' staat voor het lexeem *gaan* met betekenisindex 1
<lexeemcategorie>	distributiecategorie op het lexeemniveau (vóór samenstelling): $v-inf staat voor een lexeem met kenmerken van een V-infinitief
<lexeemstatus>	specificatie van de combinatorische status (±zelfstandig): het lexeem *gaan* heeft hier de status van zelfstandig lexeem
<lexeempositie>	positie binnen samenstelling (bij niet-zelfstandig gebruik): positie 0 correspondeert met zelfstandige lexemen
<mstructuur0>	de morfologische indexstructuur van het lexeem (waarbij van de fonologische representatiedimensies wordt geabstraheerd)
<orth-representatie>	de orthografische lexeemrepresentatie (c.q. spelvorm) van het lexeem *gaan* (waarbij |x| voor de spelvorm van x staat)
<mstructuur1-orth>	morfeemstructuur op het orthografische gebruiksniveau: de orth-vorm van het lexeem *gaan* correspondeert met een ⊕-compositie van de orth-segmenten |0| en |gaan|
<mstructuur2-orth>	overkoepelende morfeemstructuur (bij mstructuur1-orth); deze representatie bestaat uit indexen die generaliseren over alle beschikbare vormvarianten: zo generaliseert de niveau-0-index |[#w-gaan]| (van wortelmorfeem *gaan*) over de vormen |gaan|, |ga|, |gang|, |gank| en |ging| (waarvan de eerste met de orth-representatie van het hier behandelde lemma correspondeert; de niveau-1-index [0/GE]$_{\#V}$ correspondeert met de vormen |0| en |*ge*|.
<audi-representatie>	de auditieve lexeemrepresentatie (c.q. klankvorm) van het lexeem *gaan* (waarbij /x/ voor de klankvorm van x staat)
<mstructuur1-audi>	morfeemstructuur op het auditieve gebruiksniveau: de klankvorm van het lexeem *gaan* correspondeert met een ⊕-compositie van de audi-segmenten /0/ en /*gaan*/
<mstructuur2-audi>	overkoepelende morfeemstructuur (bij mstructuur1-audi); zie verder de toelichting bij mstructuur2-orth

Beschouw nu de lexicale representatie van het lexeem *gang*:

<lexeemindex>	gang:1
<lexeemcategorie>	$n-dyn
<lexeemstatus>	zelfstandig
<lexeempositie>	0
<mstructuur0>	1:[0]$_{\#N}$ ⊕ 0: #w-GAAN
<orth-representatie>	|gang|
<mstructuur1-orth>	1:|0| ⊕ 0:|gang|
<mstructuur2-orth>	1:|[0]$_{\#N}$| ⊕ 0:|[#w-GAAN]|
<audi-representatie>	/gang/
<mstructuur1-audi>	1:/0/ ⊕ 0:/gang/
<mstructuur2-audi>	1:/[0]$_{\#N}$/ ⊕ 0:/[#w-GAAN]/

Wegens de vorm- en betekenisovereenkomsten tussen de lexemen *gang* en *gaan* ga ik ervan uit dat hun orthografische niveau-2-representatie (te weten de mstructuur2-orth) op dezelfde stamindex is gebaseerd (te weten de wortel [#w-GAAN]), maar dat de niveau-1-representatie met verschillende stamvormen correspondeert, namelijk *gaan* resp. *gang*. Verder verschillen de niveau-2-representaties in de keuze van de bijbehorende functor: bij *gaan* is dit de functor [#v-0/GE] (dus een $v-vormende functor), maar bij *gang* is het [#n-0] (dus een $n-vormende functor). Merk op dat de niveau-1-indexen niet geïnterpreteerd kunnen worden zonder de niveau-2-index erbij te betrekken; een en dezelfde morfeemvorm kan immers verschillende functies uitdrukken. Omdat nog geen betekenisinformatie is geïntroduceerd, kunnen de hier gespecificeerde niveau-2-indexen geen rekening houden met betekenisvariatie; dit heeft als gevolg dat het U-domein alle formele affixatiemogelijkheden dient te geven, ongeacht de semantische condities van deze affixatiemogelijkheden. Maar de L-KRING-theorie is zo opgezet dat altijd aanvullende differentiatiedimensies geactiveerd kunnen worden. In de rest van deze sectie zal ik me beperken tot de analyse van de orthografische dimensie; de auditieve dimensie zal dus, net als de semantische dimensie, buiten beschouwing blijven.

Hieronder geef ik de morfeemrepresentatie van het werkwoord *begaan*:

<lexeemindex>	begaan:1
<lexeemcategorie>	$v-inf
<lexeemstatus>	zelfstandig
<lexeempositie>	0
<mstructuur>	1:[BE]$_{\#V}$ ⊕ 0: #w-GAAN
<orth-representatie>	|begaan|
<mstructuur1-orth>	1:|be| ⊕ 0:|gaan|
<mstructuur2-orth>	1:|[BE]$_{\#V}$| ⊕ 0:|[#w-GAAN]|

Volgens deze representatie is *begaan* uit twee morfeemsegmenten opgebouwd; op het eerste orthografische representatieniveau (<mstructuur1-orth>) bestaat dit lexeem uit de orth-segmenten |be| en |gaan|, waarbij |gaan| met de diepst ingebedde stam (namelijk de niveau-2-index [#w-GAAN]) correspondeert en |be| met de eerste functor (namelijk de niveau-2-index [BE]$_{\#V}$). In dit lexeem neemt [BE]$_{\#V}$ dus dezelfde positie in als de functor [0/GE]$_{\#V}$ in het werkwoord *gaan*. De resulterende stam kan weer als basis dienen voor de toepassing van een volgende functor, bijvoorbeeld het suffix -BAAR:

<lexeemindex>	begaanbaar:1
<lexeemcategorie>	#a
<lexeemstatus>	zelfstandig
<lexeempositie>	0

<mstructuur>	[1:[BE]$_{\#V}$ ⊕ 0: #w-GAAN] ⊕ 0: #a-BAAR
<mstructuur1-orth>	[1:|be| ⊕ 0:|gaan|] ⊕ 2:|baar|
<mstructuur2-orth>	[1:|[BE]$_{\#V}$| ⊕ 0:|[#w-GAAN]|] ⊕ 2:[#a-BAAR]

Dit suffix correspondeert met een dyadische functor, wat inhoudt dat deze functor niet één maar twee morfemen selecteert. Want behalve een #v-stam vereist deze functor ook een waarheidsindicator. Indien er sprake is van een positieve indicator, blijft deze meestal ongespecificeerd (al kan hij ook de vorm *wel* aannemen: *welbegaanbaar*), maar indien er sprake is van een negatieve indicator, is een expliciete markering nodig, in dit geval ON-. In de onderstaande representaties komt dit tot uitdrukking door de hoofdstam als morfeem$_A$ te markeren en de waarheidsindicator als morfeem$_B$, en beide morfemen in de accolade-structuur {morfeem$_B$, morfeem$_A$} op te nemen. Hierbij heb ik [#a-BAAR] het functor-nummer 3 gegeven, aangezien deze functor is bij de complexe stamstructuur tussen de accolades.

<lexeemindex>	onbegaanbaar:1
<lexeemcategorie>	$a-neg
<lexeemstatus>	zelfstandig
<lexeempositie>	0
<mstructuur>	{2:[#a-#neg]$_B$, [1:[#v-be] ⊕ 0:[#w-gaan]]$_V$} ⊕ 3:[#a-baar]
<mstructuur1-orth>	{2:|0| ⊕ [1:|be| ⊕ 0:|gaan|]$_V$} ⊕ 3:|baar|
<mstructuur2-orth>	{2:[#a-#neg]$_B$, [1:[#v-be] ⊕ 0:[#w-gaan]]$_V$} ⊕ 3:[#a-baar]
<lexeemindex>	(wel)begaanbaar:1
<lexeemcategorie>	$a-pos
<lexeemstatus>	zelfstandig
<lexeempositie>	0
<mstructuur>	{2:[#a-#pos]$_B$, [1:[#v-be] ⊕ 0:[#w-gaan]]$_V$} ⊕ 3:[#a-baar]
<mstructuur1-orth>	{2:|0| ⊕ [1:|be| ⊕ 0:|gaan|]} ⊕ 3:|baar|
<mstructuur2-orth>	{2:|[#a-#pos]$_B$|, [1:|[#v-be]| ⊕ 0:|[#w-gaan]|]$_V$} ⊕ 3:|[#a-baar]|

Uit deze voorbeelden blijkt dat de L-KRING-analyse van Nederlandse lexemen al gauw tot complexe representaties leidt, waarbij voortdurend vragen ontstaan met betrekking tot de classificatie en benoeming van de waargenomen segmenten. In dit analyseproces dienen steeds nieuwe indexen en indexniveaus te worden geïntroduceerd, totdat alle lexemen een gecomprimeerde representatie hebben gekregen (d.w.z. een representatie waarin zoveel mogelijk materiaal met andere lexemen wordt gedeeld). Deze opzet ligt ook ten grondslag aan de MGBN, al is het hier geschetste ideaal natuurlijk nog lang niet bereikt.

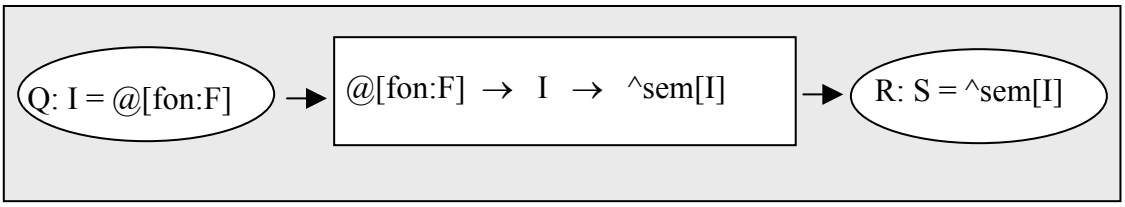
5.2.3 Demonstratie van de querymethode

5.2.3.1 Introductie

Elke lexicale zoekopdracht kan worden onderverdeeld in een index-selectie-opdracht Q (van Query) en een index-collectie-opdracht R (van Rapport). Zo kan Q worden gespecificeerd als de opdracht om een index I te identificeren met fonologische representatie F (via het Selector-commando @), en R als de opdracht om de bijbehorende betekenis (namelijk de semantische representatie S) vast te stellen (via het Collector-commando ^sem). De @-operator is gedefinieerd als een functie die een gegeven representatiekenmerk (zoals de "uitgeschreven" vorm of betekenis) aan een lexicale index probeert te koppelen, in dit geval het fonologische kenmerk F (fon:F). De ^-operator is juist gedefinieerd als een functie die een gegeven index naar een representatie van een van tevoren opgegeven kenmerktype projecteert, waarbij het gewenste kenmerktype als onderdeel van de operator moet worden gespecificeerd (in dit geval het type *sem*: ^sem).

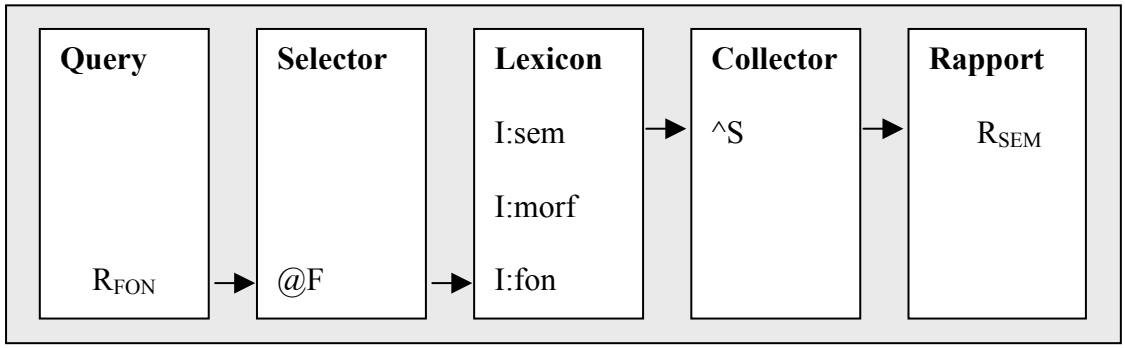
@(fon:x) = een functie f(x) van type Extensie(fon) → I
^sem(x) = een functie f(x) van type I → Extensie(sem)

Voor de overzichtelijkheid heb ik de hier gedefinieerde operators ook gebruikt om de formele zoekprocedure te karakteriseren; het enige alternatief is om de hele procedure uit te schrijven, zoals verderop zal blijken. De zoekprocedure bestaat uit twee stappen: in de eerste stap gaat de Selector op zoek naar de index die het beste aan de fonologische representatie F in Q voldoet; na identificatie van deze index (I) construeert de Collector de semantische representatie S van deze index, die vervolgens aan de R-component wordt doorgegeven. Een en ander wordt schematisch weergegeven in figuur 5-2.



Figuur 5-2: De formele implementatie van een zoekopdracht: deze zoekopdracht bestaat uit een query Q en een rapportinstructie R, die worden verbonden door vertaalrelaties (→).

In figuur 5-3 wordt de hierboven gedefinieerde zoekprocedure verder uitgesplitst door de bijdrage te specificeren van elk van de doorlopen componenten. In de eerste component wordt een Query gedefinieerd, in dit geval de opdracht om lexicale informatie te geven over een eenheid met klankvorm F. In de Selector wordt deze Query omgezet in een algoritme om een lexicale index I te zoeken die aan de eisen uit de Query voldoet, in dit geval een index waarvan de fonologische tier (I:fon) compatibel is met de extensionele representatie F: @[fon:F]. Zodra er een index wordt gevonden die aan deze eis voldoet, kan de Collector aan de slag met de constructie van kenmerken ten behoeve van het Rapport aan de gebruiker, in dit geval de constructie van een extensionele representatie S voor de semantische tier (I:sem) van de gevonden index(en). Deze kenmerken worden tot slot in een Rapport weergegeven.



Figuur 5-3: Schematische weergave van de route die wordt doorlopen bij de uitvoering van een Query die als doel heeft om de representatie R_{FON} in de representatie R_{SEM} om te zetten.

Zoals reeds eerder uiteen werd gezet berust de L-KRING-theorie op het uitgangspunt dat alle eenheden uit het lexicon (van morfemen tot woordgroepen) dezelfde informatiestructuur bezitten. Deze bestaat minimaal uit een kern K, die de lexicale hoofdindex introduceert en twee domeinen met combinatorische informatie (hetzij op extensioneel, hetzij op intensioneel niveau), namelijk het inwaartse domein I (om het toepassingsdomein van een functor te definiëren, al dan niet als paradigma) en het uitwaartse domein U (om de directe en de indirecte gebruikscontext te specificeren, al dan niet als paradigma).

Elk van deze domeinen kan worden onderverdeeld in een morf-tier (waarin de morfotactische hoofdindex van de beschikbare eenheden wordt geïntroduceerd), een fon-tier (voor fonologische of orthografische differentiatie van de hoofdindex) en een sem-tier (voor

semantische differentiatie van de hoofdindex). Voor de kerneenheid kan ook nog een freq-tier worden gespecificeerd (waarin een indicatie van de gebruiksfrequentie kan worden aangetroffen, bijvoorbeeld laag, midden, hoog, zo mogelijk aangevuld met informatie over de *productiviteit* van het morfeem, dus over de kans dat het morfeem in nieuwvormingen kan worden aangetroffen). In de rest van deze sectie zal ik alle indexrepresentaties op het basisformaat <I, K, U>$_D$ baseren, dus op een lexicaal venster waarin behalve het kerntaxeem ook informatie wordt gegeven over het inwaartse en het uitwaartse domein. Dit basisformaat kan als volgt worden uitgewerkt:

i$_n$ = < I , K , U > =
= <[{D-stam: ...}]$_I$, [{**fon**: ...}, {**morf**: ...} ,{**freq**: ...}, {**sem**: ...}]$_K$, [{D-func: ...}]$_U$>$_D$

i$_n$ = index met identificatienummer n
< ... > = afbakening van structuurdomein D van de indexdefinitie
[{D-stam: ...}]$_I$ = specificatie van D-stammen in het inwaartse domein I:
a) extensionele opsomming van beschikbare eenheden (in domein D)
b) opsomming van intensionele kenmerken (evt. op basis van tiers)
[...]$_K$ = specificatie van het kerndomein K:
{**fon**: ...} = kenmerken van de fonologische tier
{**morf**: ...} = kenmerken van de morfotactische tier
{**freq**: ...} = indicatie van de gebruiksfrequentie
{**sem**: ...} = kenmerken van de semantische tier
[{D-func: ...}]$_U$ = specificatie van D-functors in het uitwaartse domein U:
a) extensionele opsomming van beschikbare eenheden (in domein D)
b) opsomming van intensionele kenmerken (evt. op basis van tiers)

Bij de constructie van een nieuwe eenheid zal ik de kenmerken van de stam-index meestal overhevelen naar de morf-tier van de functor-index, zonder aan te geven wat hun oorspronkelijke locatie is (bijv. fon-tier of sem-tier). Dit heeft als voordeel dat de hoofdindex als een reeks stamkenmerken kan worden gedefinieerd. Indien sprake is van een nieuwe eenheid kan ook worden aangegeven of de constructie grammaticaal is ([±gram]), d.w.z. of de samengevoegde eenheden onderling compatibel zijn. In de praktijk zal ik dit alleen aangeven indien een index niet grammaticaal is; in dat geval kan hij uiteraard niet worden opgeslagen.

Ter verduidelijking van de querymethode zal ik nu een concrete query beschrijven. In deze query dient een klankvorm in een betekenis te worden omgezet, namelijk de klankvorm van het lexeem *wikkeling*. Gemakshalve ga ik er van uit dat dit lexeem deel uitmaakt van een gesproken betoog, zodat de klankvorm van dit lexeem als een lineaire reeks fonemen (die ik voor het gemak als grafemen zal blijven weergeven) binnenkomt, dus als /w-i-kk-e-l-i-ng/ .

5.2.3.2 De selectiefase

Het selectiedeel van de opdracht kan als volgt worden geformaliseerd:

$$Q = @(\{R_{FON} = /wikkeling/\}, \{R_{MORF} = lexeem\})$$

Omdat het onmogelijk is om *wikkeling* als integrale vorm in het lexicon terug te vinden, dient Q net zolang in kleinere zoekopdrachten (= Sel-taken) te worden opgesplitst totdat er deeleenheden ontstaan die wel terug kunnen worden gevonden. Bij de identificatie van gesproken woorden is de meest voor de hand liggende strategie om per binnenkomend foneem na te gaan of reeds een herkenbare morfeemvorm is ontstaan; zo ja, dan moet de bijbehorende morfeemindex tijdelijk worden geactiveerd in afwachting van volgende morfemen. Bij de verwerking van nieuwe morfemen moet echter ook worden nagegaan of er nog andere morfemen kunnen worden geconstrueerd met het beschikbare foneemmateriaal. Al deze varianten dienen tijde-

lijk te worden vastgehouden totdat er een morfeemcombinatie ontstaat die al het foneemmateriaal dekt; de alternatieven kunnen dan worden weggegooid. Toegepast op het lexeem *wikkeling* leidt deze strategie tot de volgende deelstappen:

Sel-1a: @(/wik/) =
i_{1a}: <[-]$_I$, [{**fon**: /wik/}, {**morf**: [M-stam][+inh][+temp]}, {**freq**: laag},
{**sem**: 'overwegen', 'wichelen'}]$_K$, [{**M-func**: #v, #n-er, #n-erij}]$_U$>
i_{1b}: <[-]$_I$, [{**fon**: /wik/}, {**morf**: [M-stam][+inh][-temp][-mod]}, {**freq**: zeer laag},
{**sem**: 'vijg'}]$_K$, [{**M-func**: #n-tje,}, {**L-func**: L-Op(N)}]$_U$>

toelichting: de eerste foneemcluster die als morfeem kan worden herkend is /wik/ (al kan men betogen dat de /w/ Nederlandse vraagwoorden markeert). Het morfeem 'wik' is echter ambigu tussen twee mogelijke betekenissen, namelijk de betekenis van de [+temp]-stam van het werkwoord *wikken* (dat 'overwegen' of 'wichelen' betekent), die met index i_{1a} correspondeert, en de betekenis van het segment *wik* in het nomen *paardenwik* (dat 'paardenvijg' betekent), die met index i_{1b} correspondeert. De eerste betekenis zal waarschijnlijk sneller worden geactiveerd dan de tweede, omdat dit (volgens mijn lexicon althans) de meest gebruikte betekenis is van *wik*; dit blijkt uit het feit dat het freq-veld van index i_{1a} de specificatie 'laag' bezit, en dat van index i_{1b} de specificatie 'zeer laag'. Zolang echter geen contextuele en/of combinatorische informatie beschikbaar is, dienen beide betekenissen in overweging te worden genomen. Uit de bijbehorende index-representaties blijkt dat het betekenisverschil ook gevolgen heeft voor het uitwaartse domein: index i_{1a} selecteert namelijk de morfeemfunctors #v, #n-er en #n-erij, terwijl index i_{1b} toegang geeft tot de morfeemfunctor #n-tje en de lexeemfunctor L-Op(N) (waarmee deze stam in een lexeem kan worden omgezet). Er zijn echter ook enkele overeenkomsten tussen beide indexen: volgens de morf-tier van de kern K corresponderen beide indexen met een M-stam die inheems is [+inh]; hun stamstatus blijkt ook uit het feit dat het inwaartse domein als leeg ([-]) is gespecificeerd.

Sel-2: @(/el/) =
i_2: <[**M-stam**: [+inh][+temp]]$_I$, [{**fon**: /_el/},
{**morf**: [M-func][+inh][+temp][+Lsuf] [+iter]},
{**freq**: hoog}, {**sem**: herhaalde uitvoering van proces I-stam}]$_K$,
[{**M-func**: #v, #n-ing, #n-aar, #a-baar}]$_U$ >

toelichting: Het eerste morfeem dat herkend kan worden in de foneemstring die binnenkomt na identificatie van het morfeem *wik* is *el*. Uit de index-representatie blijkt dat dit morfeem een inheemse M-functor is waarvan het inwaartse domein ter linkerzijde een M-stam vereist met de intensionele specificatie [+inh][+temp], dus een inheemse [+temp]-stam met morfeemstatus (zoals *wik*). Volgens de K-representatie van *el* leidt toepassing van deze functor tot een [+temp]-eenheid waarvan de betekenis omschreven kan worden als de herhaalde uitvoering van het door de inwaartse stam (I-stam) geïntroduceerde proces. Op het niveau van de morf-features wordt deze eigenschap weerspiegeld door het formele feature [+iter] (van 'iteratief'). Het kenmerk [+Lsuf] geeft aan dat het suffix *el* een licht suffix is (evenals *er* en *en*), wat voorspelbare gevolgen heeft voor het uitwaartse paradigma: veel stammen met een [+Lsuf]-suffix zijn namelijk in staat om de functors #v, #n-ing, #n-aar en #a-baar te selecteren (gegeven het betekenisfeature [+iter]). Vooral de selectie van #n-aar (in plaats van #n-er) is tekenend. Volgens het freq-veld ten slotte, bezit het suffix *el* een hoge gebruiksfrequentie (wat echter niet betekent dat het ook een 'productief' suffix is).

Sel-3: @(/wikkel/) =
i_{3a}: **i_2(i_{1a})** = [**?gram**], want **i_{3a}** is neologisme met onproductief suffix (*el*):
<[-]$_I$, [{**fon**: [**i_2**:fon]([**i_{1a}**:fon]) = /wikkel/},
{**morf**: [M-stam][+inh][+temp][+Lsuf][+iter]},

{**freq**: nieuw}, {**sem**: [$\mathbf{i_2}$:sem]([$\mathbf{i_{1a}}$:sem]) = "steeds opnieuw overwegen"}]$_K$,
[{**M-func**: #v, #n-ing, #n-aar, #a-baar}]$_U$ >

$\mathbf{i_{3b}}$: $\mathbf{i_2}$($\mathbf{i_{1b}}$) = [**-gram**], want $\mathbf{i_{1b}}$ (*wik* in betekenis van 'vijg') is [-temp]-eenheid (c.q. object), en kan daarom niet de basis vormen voor toepassing van suffix *el*.

$\mathbf{i_{3c}}$: <[-]$_I$, [{**fon**: /wikkel/}, {**morf**: [M-stam][+inh][+temp][+Lsuf][+iter]},
{**freq**: medium}, {**sem**: "windingen maken"}]$_K$,
[{**M-func**: #v, #n-[0], #n-ing, #n-aar, #a-baar}]$_U$ >

toelichting: Na identificatie van de morfemen *wik* en *el* is het ook mogelijk om de foneemstring /wikkel/ als geheel te identificeren. Uit de bovenstaande informatie blijkt dat er zelfs drie verschillende indexrepresentaties mogelijk zijn, waarvan de eerste twee zijn opgebouwd uit de indexen die in de voorgaande selectiestappen zijn geactiveerd, terwijl de laatste op directe (in plaats van getrapte) herkenning berust. Wat de eerste twee opties betreft geldt dat de onderliggende morfemen in beginsel sneller herkend zullen worden dan de samengestelde eenheid, omdat hun gebruiksfrequentie per definitie hoger ligt (aangezien per morfeem ten minste één andere toepassing zal bestaan).

De eerste optie is gebaseerd op de toepassing van het suffix *el* op index i_{1a} (de stam wik in de [+temp]-betekenis); hoewel de resulterende vorm (met de betekenis 'steeds opnieuw overwegen') niet strijdig is met de voor *el* gespecificeerde selectierestricties, is het een onaantrekkelijke (subgrammaticale) optie ([?gram]), want het betreft geen bestaande index, maar een nieuwe toepassing van een onproductief suffix. Dit probeem geldt ook voor de tweede optie, waar het suffix *el* wordt toegepast op de stam *wik* in de [-temp]-betekenis; maar in dit geval is zelfs sprake van een niet-grammaticale ([-gram]) nieuwvorming, want de functor vereist een [+temp]-stam.

De derde optie berust op directe identificatie van de foneemstring *wikkel* als een [+temp]-stam met de betekenis 'windingen maken'. In tegenstelling tot de vorige twee opties gaat het hier om een bestaande lexicale eenheid met de frequentie 'medium'. Bovendien specificeert het uitwaartse domein tal van functors, namelijk #v, #n-ing, #n-aar en #a-baar. Index i_{3c} lijkt dan ook een goede kandidaat voor de opbouw van een grotere representatie. Deze eenheid zou overigens ook het product kunnen zijn van een formele (betekenisloze) toepassing van het pseudosuffix *el* (met [+Lsuf]-status) op de eenheid *wik* (met dezelfde betekenis als de eenheid *wikkel*); het voordeel van zo'n analyse is dat de uitwaartse selectiemogelijkheden van *wikkel* (en alle andere stammen met een Lsuf) kunnen worden overgeërfd van het (formele) suffix *el*.

Sel-4: @(/ing/) =
$\mathbf{i_4}$: <[{**M-stam**: [+inh][+temp]}]$_I$,
[{**fon**: /_ing/}, {**morf**: [M-func][+inh][+temp][#n]}, {**freq**: hoog},
{**sem**: "stadium in proces van I-stam"}]$_K$,
[{**M-func**: #n-tje}, {**L-func**: L-Op (N: -en, -s)]$_U$ >

toelichting: Het eerst herkenbare morfeem na *wikkel* is het hoogfrequente #n-suffix *ing*. Het I-domein van deze functor vereist ter linkerzijde een inheemse [+temp]-stam, wat in overeenstemming is met de i_3-analyse van de foneemstring *wikkel*. De activering van *ing* wordt nog extra bevorderd door het feit dat deze functor deel uitmaakt van het uitwaartse paradigma van de i_{3c}-representatie van *wikkel*. De toepassing van deze functor resulteert in een [+temp][#n]-eenheid (ofwel een nominale stam) waarvan de betekenis kan worden omschreven als een stadium in het proces dat gedefinieerd wordt door de inwaartse stam (in dit geval dus *wikkel*). Uit het U-domein blijkt dat *ing* op morfeemniveau alleen door de M-functor #n-tje gevolgd kan worden; het kan echter ook de basis vormen voor de toepassing van een lexeem-operator (L-Op), die soms de klankvorm *en* (meervoud) of *s* (modifier) aanneemt.

Sel-5: @(/wikkeling/) =

i_{5a}: **i_4(i_{3a})** = [**-gram**], want **i_{3a}** is [?gram]
i_{5b}: **i_4(i_{3b})** = [**-gram**], want **i_{3b}** is [-gram]
i_{5c}: **i_4(i_{3c})** = <[-]$_I$, [{**fon**: [**i_4**:fon]([**i_{3c}**:fon]) = /wikkeling/},
{**morf**: [M-stam][+inh][+temp][+Lsuf][+iter][#n]},{**freq**: medium},
{**sem**: [**i_4**:sem]([**i_{3c}**:sem]) = "stadium in proces van windingen maken"}]$_K$,
[{**M-func**: #n-tje}, {**L-func**: L-Op (N: -en, -s)}]$_U$ >

toelichting: Zodra het morfeem *ing* is herkend, kan ook de foneemstring /wikkeling/ als geheel worden geïdentificeerd, namelijk als een toepassing van het suffix *ing* op de stam *wikkel*. Deze stam kent drie mogelijke indexrepresentaties, zodat er ook minstens drie analyses voor *wikkeling* bestaan. Maar de eerste twee analyses leiden tot een niet-grammaticale ([-gram]) suffixtoepassing, want de stam met index i_{3a} is [-gram] (zodat hij helemaal niet gebruikt kan worden) terwijl de stam met index i_{3b} een niet-bestaande stam is met status [?gram], zodat het erg onwaarschijnlijk is dat deze potentiële stam de basis vormt voor grotere constructies. Dat betekent dat alleen de stam met index i_{3c} in aanmerking komt, dus de [+temp]-stam met de betekenis 'windingen maken'. De toepassing van -ING leidt dan tot een eenheid waarvan de betekenis kan worden omschreven als "stadium in het proces van windingen maken" en waarvan de gebruiksfrequentie als 'medium' kan worden getypeerd; de overige eigenschappen van deze eenheid, waaronder het uitwaartse paradigma, kunnen rechtstreeks worden overgeërfd van het suffix. In theorie is overigens nog een vierde analyse mogelijk. De stam *wikkel* kan namelijk ook geïnterpreteerd worden als een [+TEMP]-stam die een [-TEMP]-stam met de betekenis 'omhulsel' incorporeert (die zelf weer van de [+TEMP]-stam met index i_{3c} is afgeleid), waarbij de resulterende betekenis kan worden omschreven als 'omhulsels aanbrengen' (bijvoorbeeld in *Het wikkelen van de tijdschriften kostte veel tijd*). Het gaat hier echter om een nieuwe (niet-bestaande) analyse van *wikkeling*, zodat de constructie van deze eenheid (c.q. index) waarschijnlijk veel meer tijd zal kosten dan herkenning van de lexicale index i_{5c}; normaal gesproken zal deze nieuwvorming (waar minstens twee extra constructiestappen voor nodig zijn) dan ook weinig kans maken.

Sel-6: @([L-Op-2{N:sg}]) =
i_6: <[{**L-stam**: [modifier]},{**M-stam**: [#n]}]$_I$, [{**fon**: /-/},
{**morf**: [L-functor][+inh][#n][sg]}, {**freq**: zeer hoog},
{**sem**: "1 eenheid van I-stam"}]$_K$, [{**W-func**: ...}]$_U$ >

toelichting: Hoewel de voorgaande zoekprocedure een index heeft opgeleverd die volledig compatibel is met de foneemstring /wikkeling/, heeft deze index geen lexeem-status, zodat nog niet aan alle eisen van de Query wordt voldaan. Dit probleem kan worden opgelost door een (onhoorbare) lexeem-operator op deze index toe te passen, namelijk de operator L-Op-2{N:sg}; deze operator verandert al dan niet gelede morfemen in een lexeem met categorie N, subspecificatie 'sg' c.q. enkelvoud (waarvoor geen suffixmarkering nodig is) en functorstatus, wat impliceert dat dit lexeem een inwaarts selectiedomein bezit waarmee het modificator-lexemen (van type L-stam) kan selecteren. Deze eigenschappen worden ook weerspiegeld door de features in het morf-veld van de indexkern (K), en de specificatie van het sem-veld: "1 eenheid van de I-stam". Omdat lexeem-operatoren een onmisbare schakel vormen in de constructie van lexemen, kennen ze over het algemeen een zeer hoge gebruiksfrequentie (al kunnen er gradatieverschillen zijn: zo zal de pl-operator minder vaak voorkomen dan de sg-operator). Door de algemene toepasbaarheid van lexeem-operatoren valt er weinig algemeens te zeggen over hun uitwaartse selectiekenmerken; in de representatie van L-Op{N-sg} wordt daarom geen informatie gegeven over de invulling van het U-domein (behalve een indicatie van het te specificeren functordomein, namelijk het domein van de woordfunctors). Wegens de hoge gebruiksfrequentie van domeinoperators is het de vraag of er een aparte selectiestap vereist is voor hun selectie, temeer daar deze operators meestal geen expliciete markering

kennen; men kan bijvoorbeeld ook de hypothese verdedigen dat hier sprake is van permanent geactiveerde coercion-operators die op elk gewenst moment op een stam kunnen worden toegepast. Men zou dit kunnen onderzoeken door na te gaan of de selectie van lexeem-operators beïnvloed wordt door verschillen in gebruiksfrequentie.

Sel-7: @(/wikkeling/ + [L-Op-2{N:sg}]) =
$\mathbf{i_7}$: $\mathbf{i_6}(\mathbf{i_{5c}})$ = <[**L-stam**: [modifier]]$_I$, [{**fon**: [L-Op{N:sg}]([$\mathbf{i_{5c}}$:fon]) = /wikkeling/},
{**morf**: [L-functor][+inh][+Lsuf][+iter][+stadium][#n][sg]}, {**freq**: medium},
{**sem**: [$\mathbf{i_6}$:sem]([$\mathbf{i_{5c}}$:sem]) = "1 stadium in proces van windingen maken"}]$_K$,
[{**W-func**: W-Op (NP: de, een, [0], 1])]$_U$ >

toelichting: Selectiestap 7 correspondeert met de activering van index i_7, die het product is van toepassing van lexeem-operator L-Op-2{N:sg} op de morfeemstam *wikkeling*; dit resulteert in een lexeem dat zich kenmerkt door een middelfrequent gebruik (blijkens het freq-veld). De betekenis van dit nomen singularis (N-sg) kan worden geparafraseerd als "1 stadium in een proces van windingen maken" (namelijk het tot nu toe afgelegde windingstraject of één winding in dit traject). Wat betreft de morfotactische features verschilt i_7 van i_{5c} doordat het feature M-stam nu in het feature L-functor is omgezet, waarbij sprake is van de subspecificatie [sg]. De L-functor-status komt tot uitdrukking in het feit dat het een inwaarts domein bezit waarmee het L-stammen kan selecteren; deze lexemen fungeren dan als modifier van materiaal, kern of effect, zoals respectievelijk het geval is in *draadwikkeling*, *ankerwikkeling*, *veldwikkeling*. Ook het uitwaartse domein verschilt, want door de lexeemstatus van i_7 kan *wikkeling* niet langer als basis fungeren voor morfeemfunctors, maar wel voor woordfunctors, zoals de woordoperator W-Op, die kan corresponderen met de lidwoorden *de*, *een* en het onhoorbare [0] of het telwoord 1. Indien geen L-modifier wordt gespecificeerd, dient het I-domein de specificatie [-] te krijgen; in dat geval geldt *wikkeling* als een zelfstandig woord.

5.2.3.3 De collectiefase

Zodra de Selector de foneemreeks /w-i-kk-e-l-i-ng/ als de fonologische extensie van de lexicale index i_7 heeft herkend, kan de Collector aan de slag met het rapportagedeel van de gebruikersopdracht, namelijk de constructie van de semantische extensie van de index die door de Selector is geactiveerd (te weten index i_7): R: S = ^sem($\mathbf{i_7}$). De Collector kan deze extensie langs deductieve weg construeren door op recursieve wijze de lexicale representatie van de onderliggende indexen te activeren; hierbij worden de volgende deelstappen (Col-instructies) doorlopen:

Col-1: ^sem($\mathbf{i_7}$) = $\mathbf{i_7}$:sem
Col-2: $\mathbf{i_7}$:sem = [$\mathbf{i_6}$:sem]([$\mathbf{i_{5c}}$:sem])
Col-3: $\mathbf{i_6}$:sem = "1 eenheid van **I-stam**"
Col-4: $\mathbf{i_{5c}}$:sem = [$\mathbf{i_4}$:sem]([$\mathbf{i_{3c}}$:sem])
Col-5: $\mathbf{i_4}$:sem = "stadium in proces van **I-stam**"
Col-6: $\mathbf{i_{3c}}$:sem = "windingen maken"
Col-7: $\mathbf{i_{5c}}$:sem = [$\mathbf{i_4}$:sem]([$\mathbf{i_{3c}}$:sem]) = "stadium in proces van [$\mathbf{i_{3c}}$:sem]"
= "stadium in proces van windingen maken"
Col-8: $\mathbf{i_7}$:sem = [$\mathbf{i_6}$:sem]([$\mathbf{i_{5c}}$:sem]) =
= "1 eenheid van [$\mathbf{i_{5c}}$:sem]"
= "1 eenheid van stadium in proces van windingen maken"
Col-9: **S** = "1 eenheid van stadium in proces van windingen maken"

Toelichting: De eerste stap in de uitvoering van het commando ^sem(i_7) correspondeert met de activering van de semantische representatie (c.q. sem-tier) van index i_7, namelijk i_7:sem. In Col-stap 2 blijkt dat deze representatie niet zelfstandig geïnterpreteerd kan worden, want i_7:sem is gedefinieerd als [$\mathbf{i_6}$:sem]([$\mathbf{i_{5c}}$:sem]). Dat betekent dat eerst moet worden nagegaan

wat de semantische representatie van de indexen i_6 en i_{5c} is. Dit gebeurt in de Col-stappen 3 en 4. Col-stap 3 wijst uit dat de betekenis van i_6 is gedefinieerd als "1 eenheid van **I-stam**", waarbij de vetgedrukte I-stam een variabele is die met de betekenis van de inwaartse stam correspondeert. Col-stap 4 wijst echter uit dat index i_{5c} wederom in termen van andere indices is gedefinieerd, namelijk als [**i_4**:sem]([**i_{3c}**:sem]). Om i_{5c} volledig te kunnen interpreteren, dienen dus eerst de semantische representaties van i_4 en i_{3c} te worden geactiveerd. Dit gebeurt in de Col-stappen 5 en 6. Vervolgens worden deze representaties in i_{5c} gesubstitueerd (Col-stap 7), waarna i_{5c} zelf weer in de semantische representatie van index i_7 wordt gesubstitueerd (Col-stap 8). Dit resulteert ten slotte in een representatie die als de extensionele betekenis S kan worden gerapporteerd (Col-stap 9).

5.2.3.4 Discussie

Bij de beschrijving van de zoekprocedure voor *wikkeling* ben ik er voor het gemak van uitgegaan dat de stam *wikkel* niet geprefigeerd kan worden. De analyse wordt namelijk een stuk complexer als men ook rekening wil houden met het bestaan van werkwoorden als *verwikkelen*, *ontwikkelen* en *inwikkelen* of hieraan gerelateerde derivaties als *ontwikkeling, onderontwikkeld* en *ingewikkeld*. Het is bijvoorbeeld niet toereikend om het prefix -ONT in *ontwikkeling* als een modifier van het lexeem *wikkeling* te analyseren: [*ont*](*wikkeling*), ondanks het feit dat *wikkeling* een zelfstandig bruikbare eenheid is (wat traditioneel als een belangrijk mofologisch criterium geldt) en *ontwikkel* niet (behalve in de eerste persoon enkelvoud). Want deze analyse doet geen recht aan het feit dat de eenheid *ontwikkel* ook de (semantische) basis vormt van lexemen als *ontwikkelen*, *ontwikkelaar* en *ontwikkeld*, om nog niet te spreken van *onderontwikkeld*. Dit kan alleen verantwoord worden door aan te nemen dat *ontwikkel* als stam in het lexicon is opgenomen, evenals de eenheden *wikkel* en *verwikkel*.

Volgens dezelfde redenering dienen deze formeel en semantisch verwante stammen op dezelfde lexicale eenheid (c.q. wortel) te worden herleid, die dan verantwoordelijk is voor de introductie van hun gemeenschappelijke eigenschappen, zoals de (partiële) klankvorm *wikkel*, en de betekenis "wikkelen". Deze basiseenheid kan in een [+temp]-stam worden omgezet door middel van een [+temp]-functor, die de basisstructuur van het temporele traject specificeert (bijvoorbeeld door aan te geven of er sprake is van locatieve toename (BE-) of locatieve afname (ONT-); zie hoofdstuk 5 voor een nadere uitwerking van dit voorstel). Deze functor manifesteert zich meestal als prefix. De door mij voorgestelde analyse impliceert dat de [+temp]-stam WIKKEL evenveel structuur bezit als de [+temp]-stam ONTWIKKEL, en dat de laatste stam dus niet als een afleiding van de eerste kan worden beschouwd. Beide beschikken immers over een [+temp]-functor. In onderstaande tabel wordt dit alles op een meer overzichtelijke wijze gepresenteerd:

functor	**wortel**	**[+temp]-stam**	**derivaties**
[0/ge]	wikkel$_0$	[0/ge]-wikkel	*wikkelen*, *gewikkeld*, *wikkeling*
ver	wikkel$_0$	verwikkel	*verwikkelen*, *verwikkeld*, *verwikkeling*
ont	wikkel$_0$	ontwikkel	*ontwikkel*, *ontwikkeld*, *ontwikkeling*
be	wikkel$_0$	(bewikkel)	(*bewikkeld*, *bewikkeld*, *bewikeling)*

Deze tabel laat duidelijk zien dat de [+temp]-stammen [0/ge]-WIKKEL, VER+WIKKEL, ONT+WIKKEL en BE+WIKKEL op dezelfde wortel berusten (de potentiële stam *bewikkel* kent nog toepassingen met een GWNT-vermelding). De index [0/GE] heeft betrekking op een functor die ambigu is tussen een onhoorbare variant (namelijk [0]) en een hoorbare variant (namelijk *ge*); deze laatste vorm komt onder meer aan de oppervlakte in het voltooid deelwoord bij de V-stammen WIKKEL en VERWIKKEL, blijkens *gewikkeld* en *ingewikkeld*. Bij de V-stammen VERWIKKEL en ONTWIKKEL correspondeert de voltooide tijd echter met een vorm zonder *ge*, namelijk *verwikkeld* en *ontwikkeld*. Blijkbaar bevindt het prefix GE- zich op

dezelfde structurele positie als VER- en ONT-, maar kan deze positie vooraf worden gegaan door een prepositie, zoals IN (blijkens *ingewikkeld*) of *onder* (blijkens *onderontwikkeld*).

Tot besluit wil ik aangeven wat voor gevolgen de vaste prefixstructuur heeft voor de zoekprocedure voor het lexeem *wikkeling* en voor de variant *ontwikkeling*. Ik neem aan dat het lexeem *wikkeling* minimaal de volgende functorstructuur bezit (waarbij de functors zijn gemarkeerd door een subscript f, gevolgd door een rangnummer):

functor-structuur (WIKKELING) = $[\text{L-Op}]_{f3}$ $[\text{\#-ING}]_{f2}$ $([0/\text{GE}]_{f1}$ $(\text{WIKKEL}_0))$

Uit deze nieuwe analyse volgt dat er een extra stap nodig is in de identificatieprocedure voor de lexicale index van het lexeem *wikkeling*: bij de identificatie van de spelvorm *wikkel* dient namelijk eerst de basisstam WIKKEL_0 te worden herkend, en pas daarna de [+temp]-stam [0/GE-]-WIKKEL (die een meestal onhoorbare [+temp]-functor bezit). Hierbij dient het (direct aangrenzende) uitwaartse domein van $wikkel_0$ als volgt te worden gedefinieerd:

$U(\text{WIKKEL}_0)$ = {M: {[+temp]-functor: [0/GE-], VER-, BE-}}

Volgens deze definitie dient de stam WIKKEL_0 dus eerst een [+temp]-functor te selecteren, en hangt het vervolgens van de gekozen functor af wat voor verdere afleidingen er mogelijk zijn. Bij de identificatie van het lexeem *wikkeling* zal de index van de [+temp]-functor pas na de stam WIKKEL_0 worden geactiveerd, omdat deze functor met het onhoorbare prefix [0] correspondeert. Bij de identificatie van het lexeem *ontwikkeling* daarentegen zal de functor-index juist eerder worden geactiveerd, omdat de functor in dit geval met het hoorbare prefix ONT- correspondeert, waarvan de klankvorm eerder binnenkomt dan die van de stam WIKKEL_0. Maar na combinatie van de beide eenheden zal de rest van de analyse voor beide lexemen hetzelfde verlopen, namelijk op de wijze die reeds in de voorbeeldprocedure werd getoond.

5.2.4 De bewerking van het lexicon

Bij de opbouw van een op de L-KRING-theorie gebaseerd informatiesysteem zal meestal sprake zijn van directe kennisimplementatie door gespecialiseerde redacteuren; in dat geval dient de communicator ervoor te zorgen dat de inhoud van het lexicon zo aan de redacteur wordt gepresenteerd dat deze gemakkelijk kan beoordelen of de gegevens correct zijn en of er gegevens moeten worden toegevoegd; omgekeerd moet de user-interface ervoor zorgen dat de kennis die door de redacteur wordt ingevoerd wordt omgezet in een formaat dat compatibel is met het lexicon. Indien de redacteur bijvoorbeeld de taak heeft om na te gaan of de woorden in het lexicon een correcte spelvorm hebben, zal de user-interface eerst moeten berekenen wat de spelvorm is van de indexen die met een woord corresponderen (door de spelvorm van de samenstellende morfemen te inspecteren), waarna de redacteur deze spelvorm moet kunnen wijzigen zonder dat hij inzicht hoeft te hebben in de morfeemstructuur van het woord; de user-interface moet vervolgens zelf berekenen welke morfeemindex door de correctie wordt beïnvloed, waarna de editor deze correctie daadwerkelijk kan doorvoeren.

Stel bijvoorbeeld dat een redacteur aangeeft dat het lexeem met de spelvorm |koninkje| per ongeluk als |koningje| is gespeld (wat het gevolg zou kunnen zijn van automatische vormgeneratie), en dat de spelvorm in |koninkje| moet worden veranderd. De communicator moet dan zelf kunnen nagaan welk morfeem moet worden aangepast om deze fout te voorkomen, in dit geval het stammorfeem *koning*. Dit stammorfeem zou bijvoorbeeld de volgende lexicale (deel)representatie kunnen bezitten:

M_{12}: cat = [nomen], orth =
orth(1): gebruik |koning| in onderstaande contexten
$[M_{12} \oplus [N_{sg}, \text{-}]$ (N_{sg} = singuliere vorm van het nomen: |koningen|)
$[M_{12} \oplus [N_{pl}, \text{-en}]$ (N_{pl} = plurale vorm van het nomen: |koningen|)

$[M_{12} \oplus [N_{fem}$, -in] (fem = feminienvorm van het nomen: |koningin|)
$[M_{12} \oplus [N_{dim}$, -je] (dim = diminutiefvorm van het nomen: |koningje|)
orth(2): gebruik |konink| in onderstaande contexten
$[M_{12} \oplus [A_{rel}$, -lijk] (A_{rel} = relationeel adjectief: |koninglijk|)

Hierbij staat 'cat' voor (syntactische categorie (in dit geval nomen), en 'orth' voor orthografisch veld, namelijk het veld waarin de spelvormen worden gespecificeerd, waarbij elke spelvorm wordt gevolgd door een opsomming van de contexten waarin deze spelvorm voorkomt (met een korte toelichting). De lexicale representatie van het stammorfeem *koning* kan verbeterd worden door de context $[N_{dim}$, -je] aan de spelvorm |konink| te koppelen (die ook voorkomt in het reeds correct opgeslagen woord *koninklijk*). Deze wijziging resulteert in de volgende stamrepresentatie:

M_{12}: cat = [nomen], orth =
orth(1): gebruik |koning| in onderstaande contexten
$[M_{12} \oplus [N_{sg}$, -] (N_{sg} = singuliere vorm van het nomen: |koningen|)
$[M_{12} \oplus [N_{pl}$, -en] (N_{pl} = plurale vorm van het nomen: |koningen|)
$[M_{12} \oplus [N_{fem}$, -in] (fem = feminienvorm van het nomen: |koningin|)
orth(2): gebruik |konink| in onderstaande contexten
$[M_{12} \oplus [N_{dim}$, -je] (dim = diminutiefvorm van het nomen: |koningje|)
$[M_{12} \oplus [A_{rel}$, -lijk] (A_{rel} = relationeel adjectief: |koninglijk|)

Indien er sprake is van een systematische fout dient het correctiesysteem natuurlijk ook de mogelijkheid te bieden om alle woorden waar deze fout in voorkomt automatisch te corrigeren door te generaliseren over de correcties die door de redacteur zijn voorgedaan. Zo is het geen toeval dat de diminutiefvorm van de stam *koning* de spelvorm /konink/ vereist, want alle Nederlandse stammen op *ing* vertonen dezelfde spellingsalternantie. Zodra de redacteur dit ontdekt, kan hij dus beter een automatische correctie laten doorvoeren op basis van de door hem ingevoerde voorbeelden. Hierbij dient het correctiesysteem zelf op zoek te gaan naar een gemeenschappelijk correctiepatroon in de opgegeven voorbeelden. Gegeven de correcties in *koninkje*, *harinkje* en *woninkje* moet het systeem bijvoorbeeld zelf kunnen concluderen dat alle stammen M die in de context $M \oplus N_{dim}$ met de spelvorm $|X \oplus ing|$ corresponderen zo moeten worden aangepast dat ze in de context $M \oplus N_{dim}$ met de spelvorm $|X \oplus ink|$ corresponderen. Deze opdracht kan vervolgens aan de editor worden doorgegeven.

5.3 Beschikbare analysetools

5.3.1 Introductie

Deze sectie biedt een overzicht van potentieel bruikbare tools voor de automatische analyse van Nederlandse woorden, te weten de parseringssystemen[190] ALEX, MORPA, FAMBL, Linguistica en Word Manager, en het automatisch geannoteerde CELEX-lexicon.[191] Voor al deze tools wordt een korte specificatie gegeven van opzet en toepassingsmogelijkheden. Vervolgens wordt besproken in hoeverre deze tools nuttig kunnen zijn voor de analyse van de MGBN, gegeven de doelstelling om een bijdrage te leveren aan de systematisering van de formele woordkenmerken in VDL's lexicografische gegevensbank.

[190] In het hier gepresenteerde overzicht beperk ik me tot informatie over morfologische parsers. Op syntactisch niveau is echter veel meer parser-onderzoek gedaan. Zie Coppen & Cremers (2002) voor een overzicht.
[191] Er bestaan twee Taalunie-rapporten die nadere informatie verstrekken over taaltechnologische hulpmiddelen voor het Nederlands en hun beschikbaarheid voor publieke toepassingen, te weten het rapport van Bouma & Ineke Schuurman (1998) en het rapport van Daelemans & Strik (2002). Maar dat heb ik pas later ontdekt.

5.3.2 ALEX

ALEX (Van der Hulst & Moortgat, 1980) is een categoriaal model voor automatische woordanalyse. Deze INL-studie had als doel om na te gaan hoe men het beste te werk kan gaan bij de opbouw van een morfologisch gestructureerd woordenboek van het Nederlands. Het project vormde een belangrijke basis voor de ontwikkeling van twee morfologische parsers, namelijk de parser KASIMIR (Moortgat, 1985), die een cruciale rol speelde bij de morfologische annotatie van CELEX (zie H5.3.3), en de parser MORPA, die deel uitmaakt van een tekst-naar-spraak-systeem (zie H5.3.4). Omdat er nog geen computationeel systeem bestaat dat volgens de principes van ALEX werkt, zal ik deze methode buiten beschouwing laten bij de toepasbaarheidsbeoordeling in H5.3.8.

5.3.3 CELEX

De CELEX databank (Baayen, Piepenbrock and Gulikers, 1995) bestaat uit drie morfologisch geannoteerde tekstcorpora, namelijk een Nederlands corpus (met 124.000 lemma's), een Engels corpus (met 52.000 lemma's) en een Duits corpus (met minstens 52.000 lemma's). Voor de annotatie van deze corpora is gebruik gemaakt van de categoriale parser KASIMIR; de resulterende lexica zijn vervolgens aan een beperkte redactionele controle onderworpen. CELEX is het eerste informatiesysteem over de morfologische structuur van het Nederlands. Het biedt mogelijkheden voor statistisch onderzoek naar de gebruiksfrequentie van affixen.[192] Nadelen van CELEX zijn onder meer dat de morfologische annotatie niet volledig betrouwbaar is, dat deze annotatie in principe beperkt is tot inheemse, meestal productieve morfemen en dat men in principe geen informatie kan opvragen over hapax-woorden.

5.3.4 MORPA

MORPA (Heemskerk 1993; Heemskerk & Van Heuven 1993) is een morfologische parser op categoriale grondslag (net als KASIMIR, de parser die ten grondslag ligt aan CELEX) en bezit een morfeemlexicon van ca. 17.000 lexemen. MORPA herkent alleen lexeemgebaseerde afleidingen, met één uitzondering: bij uitheemse werkwoorden (die altijd op het suffix -EER eindigen) is ook de getrunceerde stamvorm opgenomen, zodat de parser een verband kan leggen met afleidingen op -ATIE en soortgelijke suffixen.

MORPA is ontworpen als component van een tekst-naar-spraak-systeem; voor deze toepassing kan worden volstaan met de herkenning van woordvormen die productief zijn afgeleid van de woorden in het basislexicon. Ook dit bleek overigens een tamelijk complexe doelstelling te zijn, want veel woordvormen zijn ambigu: zo kan de stam van de werkwoordsvorm *knikkeren* (namelijk KNIKKER) op minstens vier verschillende manieren worden geanalyseerd, namelijk als de geconverteerde N-stam KNIKKER (wat de meest voor de hand liggende analyse is), als de gelede V-stam KNIK+ER, als de gelede V-stam KNIK+EER en als de samenstelling KNIK+KEER. Om toch een voorkeur te kunnen aangeven, is MORPA uitgebreid met een module die de waarschijnlijkheid van elke structuur berekent; hiervoor maakt de parser gebruik van frequentiegegevens die zijn gebaseerd op CELEX. Dankzij deze kansmodule bereikt MORPA een hoog herkenningspercentage. In een testverzameling die perfect op het morfeemlexicon aansluit weet deze parser namelijk 92% van de woordvormen correct te analyseren; deze score stijgt zelfs tot 96% als de correcte analyse niet per se als eerste voorkeur hoeft te worden aangemerkt (in dat geval geldt BEL+ANGST+ELLENDE als een correcte analyse van het woord *belangstellende*).

[192] Zo heeft Baayen (1991b, 1992) een statistische methode ontwikkeld waarbij de 'productiviteit' van een affix wordt uitgedrukt als het percentage hapaxen in een representatieve sample van een tekstcorpus, zoals CELEX. In Baayen & Lieber (1991) wordt aangetoond dat deze methode correct intuïties kan voorspellen.

5.3.5 FAMBL

FAMBL (Van den Bosch & Daelemans, 1999) is een voorbeeldgestuurde morfeemparser, d.w.z. een computationeel systeem dat over een leeralgoritme beschikt waarmee het zelf morfologische structuurregels kan ontdekken in een trainingsbestand met voorbeeldanalyses (doorgaans 1000 tot 10000 voorbeeldwoorden). Bij proefsessies met een Nederlands lexicon heeft dit analysesysteem een rendement van 84% gehaald. Hoewel FAMBL over interessante generalisatiemogelijkheden beschikt, heeft het systeem als nadeel dat het herkennings-vermogen beperkt is tot de grammaticale regels die ten grondslag liggen aan de voorbeeldanalyses in het trainingsbestand.

5.3.6 Linguistica

Goldsmith (2001) heeft een compressiemethode ontwikkeld die gedreven wordt door het Minimal Desription Length (MLD) criterium (De Marcken, 1995). Met dit MLD-algoritme kan voor elke taal met een Latijns alfabet een morfologische grammatica worden geconstru-eerd door de analyse van een tekstcorpus uit die taal. Hoe groter dit tekstcorpus, hoe beter het resultaat. Deze methode kan men zelf uitproberen door het programma Linguistica te downloaden en op een corpus toe te passen.[193] De methode berust op het idee dat affixen gedefinieerd kunnen worden als vaste lettercombinaties die niet alleen frequent voorkomen, maar die vaak dezelfde stammen selecteren als andere affixen (c.q. frequente letter-combinaties); in dit verband spreekt Goldsmith van signaturen. Zodra een segment deel uitmaakt van een signatuur kan het als morfeem worden aangemerkt. Deze "blinde" identificatiemethode levert verrassend goede resultaten op.[194] Zo zou kunnen blijken dat de woordfinale segmenten *en*, *aar* en *baar* allemaal met de eenheden *wikkel*, *stapel*, *handel* en *verbeter* kunnen worden gecombineerd. Dit kan worden verantwoord door het volgende signatuur te introduceren:

[{wikkel, stapel, handel, verbeter} {er, aar, baar}]

Al deze segmenten kunnen dus als morfeem worden aangemerkt. Bovendien kan men op basis van de rechterhoofdhypothese speculeren dat de segmenten in de linkergroep met stammen corresponderen en de segmenten in de rechtergroep met affixen.

5.3.7 Word Manager

Het lexicografisch hulpprogramma Word Manager (Domenig & Ten Hacken, 1992) biedt redacteuren hulp bij het morfologisch annoteren van een woordenlijst. Hiervoor dienen ze veel voorkomende patronen als taalkundige regels te implementeren, waarna het systeem alle woorden kan zoeken waarop de regel van toepassing zou kunnen zijn. Zo zou een redacteur kunnen aangeven dat Nederlandse werkwoorden die op *el* of *er* eindigen vaak het suffix aar kunnen selecteren. Word Manager zal dan alle woorden die aan dit patroon voldoen als stam+AAR-derivaties analyseren. De redacteur dient vervolgens per woord aan te geven of deze analyse klopt. Inmiddels is de hele Duitse woordenschat op deze wijze van morfo-logische structuur voorzien.[195]

5.3.8 Toepasbaarheid in het MGBN-project

Het oorspronkelijke projectvoorstel ging ervan uit dat de morfologische parser MORPA een goede basis zou bieden voor de opbouw van de MGBN, zeker als deze parser met informatie uit het Morfologisch Handboek zou worden verrijkt. Inderdaad kon MORPA in de beginfase van het project goed worden gebruikt, maar gaandeweg bleek het toch beter om de MGBN

[193] Zie http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000/.
[194] Bij een Engelse en een Franse sample van 1000 woorden bleek 83% van de analyses correct te zijn.
[195] Dit rijk gestructureerde lexicon is raadpleegbaar via het Canoo-Net: http://www.canoo.net/index.html.

van een andere basis te voorzien. De MGBN dient immers primair een bijdrage te leveren aan de systematisering van de woordkenmerken in de WKB-Ned door informatie te geven over de relatie tussen de morfologische structuureenheden en de formele woordkenmerken uit de LGBN (die is afgeleid uit de WKB-Ned). Hiervoor is een compleet beeld nodig van lexeeminterne segmenten met een voorspelbare invloed op de lexeemkenmerken.

MORPA kon tot dan toe alleen lexemen analyseren waarvan de structuur op productieve grammaticaregels berust waardoor improductieve (o.a. uitheemse) afleidingen en gelede lexemen met niet-productieve of onregelmatige structuurkenmerken ongeanalyseerd blijven. Verder is de LGBN zo omvangrijk dat er tal van inheemse woorden in zitten waarvan de stam niet in het morfeemlexicon van MORPA voorkwam, met als gevolg dat deze woorden niet geanalyseerd konden worden. Veel woordvormen hebben bovendien meerdere structureringsmogelijkheden, zodat alle MORPA-analyses door een redacteur geëvalueerd zouden moeten worden. Voorts was er geen garantie dat de door MORPA onderscheiden morfemen (in het bijzonder de stammen) interne morfeemstructuur bezitten. En tenslotte was MORPA niet in staat om morfologische verbanden te identificeren waarbij de stam allomorfie vertoont omdat staminterne vormvariatie meestal niet tot productieve woordvormingsregels kan worden herleid, maar stam voor stam moet worden geleerd. Dit verschijnsel kan alleen langs lexicale weg worden verantwoord.

De hier genoemde beperkingen stelden ons voor de keuze om of MORPA zodanig aan te passen dat het toch een geschikt hulpmiddel voor de analyse van de MGBN zou kunnen zijn of om iets nieuws op te zetten. Een overweging voor het laatste was dat het niet realistisch is om het regelbestand van MORPA aan te vullen met regels die op de morfologische observaties uit het Morfologisch Handboek zijn gebaseerd. De formalisering van deze informatie is namelijk minder eenvoudig dan het lijkt, want de observaties uit het MHB hebben een sterk informeel karakter, terwijl de nieuwe regels op allerlei manieren kunnen interfereren met reeds opgenomen regels. Bovendien leiden improductieve regels al gauw tot overgeneralisatie. Dit betekent dat MORPA na een dergelijke uitbreiding opnieuw had moeten worden gecalibreerd. Voor een dergelijke uitbreiding zou in feite een apart onderzoeksproject nodig zijn, terwijl het slechts een deel van de problemen voor de MGBN oplost. Dit was dan ook niet haalbaar in het kader van mijn eigen onderzoeksproject.

Veel van de problemen met MORPA golden ook voor CELEX. Want hoewel een deel van de hierin opgenomen morfeemanalyses redactioneel gecontroleerd is, berusten ook deze analyses op productieve woordvormingsregels. Bovendien kwam het CELEX-lexicon slechts voor een deel overeen met het lexicon van de LGBN. Hieruit volgt dat CELEX slechts gedeeltelijk kon voorzien in de informatie die nodig is om het MGBN-lexicon van morfologische structuur te voorzien, terwijl voor alle lexemen moet worden gecontroleerd of de aangebrachte structuur compleet en correct is. Daarom bleek CELEX toch niet een handig vertrekpunt voor de opbouw van de MGBN.

De beperkingen van MORPA en CELEX hadden deels kunnen worden opgelost door een voorbeeldgestuurd leersysteem te gebruiken, zoals FAMBL of neurale netwerken. Dergelijke systemen hebben als voordeel dat ze zowel regelmatige als subregelmatige patronen kunnen herkennen. Maar daar staat tegenover dat dergelijke leersystemen tot nu toe slechts een beperkt trainingslexicon aankunnen. Verder blijft hun herkenningsvermogen afhankelijk van de voorbeeldanalyses in het trainingsbestand. Als dit trainingsbestand een grammaticale basis heeft, vertonen de automatisch geleerde patronen nog steeds grote overeenkomst met de oorspronkelijke grammaticaregels.

Het hier gesignaleerde probleem kon omzeild worden door een puur statistisch analysecriterium te hanteren. Linguistica liet zien dat zo'n puur statistische aanpak (op paradig-

matische grondslag) verrassend goede resultaten opleverde. Maar net als andere parsers wordt dit systeem gehinderd door het feit dat het geen toegang heeft tot betekenisinformatie. Dit probleem viel alleen op te lossen door het automatisch bewerkte databestand langs redactionele weg te controleren.

De algemene bezwaren tegen automatische analysesystemen golden mijns inziens niet voor het lexicografische ondersteuningsprogramma Word Manager. De redacteur definieert zelf een woordvormingspatroon en laat vervolgens alle woorden opsporen waar de regel mogelijk op van toepassing is. Hierna gaat de redacteur per woord na of het opgegeven patroon toepasbaar is; zo ja, dan wordt de regel automatisch toegepast, zo nee, dan wordt het woord genegeerd. Een nadeel van deze aanpak is wel dat de redacteur gedwongen wordt om vanuit woordvormingsregels te denken, terwijl het volgens de L-KRING-theorie cruciaal is om vanuit paradigmatische verbanden te denken. Vandaar dat het programma niet is gebruikt.

5.3.9 Conclusie

Geen van de besproken analysemethodes werd geschikt bevonden voor de morfologische analyse van het MGBN-lexicon. Door het ontbreken van semantische informatie zouden al deze methodes overgeneraliseren, zodat de inzet van zo'n analysesysteem altijd door een uitgebreide redactionele controleronde moet worden gevolgd. Bovendien leende geen van de besproken methodes zich voor de toepassing van inductieve (L-KRING-gebaseerde) structuurcriteria. Daarom heb ik afgezien van het gebruik van bestaande applicaties voor automatische of semi-automatische structuuranalyse.

5.4 De L-KRING-methode

5.4.1 Introductie

De morfologische structuurrepresentaties in de MGBN zijn het resultaat van een semi-automatische analysemethode. Hierbij heb ik me laten leiden door de lexicale representatieprincipes van de L-KRING-theorie. Zoals eerder uiteen werd gezet (zie hoofdstuk 4) berust deze theorie op het idee dat de morfologische structuurrepresentaties in het mentale lexicon een bijproduct zijn van het streven om de in dit lexicon opgeslagen woorden zo gecomprimeerd mogelijk op te slaan door hun gemeenschappelijke bouwstenen (c.q. morfemen) door indexen te vervangen (eenheden die naar een lexicale representatie verwijzen). Deze bouwstenen zijn te herkennen aan het feit dat ze een vaste relatie vertonen tussen vorm en (globale) functie (zoals hun categorie) en dat ze een voorspelbaar combinatieparadigma bezitten, d.w.z. een door voldoende woorden gedeeld cluster van inwaartse (stamgerelateerde) en uitwaartse (affixgerelateerde) combinatiemogelijkheden. Deze structuurcriteria liggen ook ten grondslag aan de morfologische structuurrepresentaties in het MGBN-lexicon. Deze representaties zijn het resultaat van een cyclisch proces van structuurtoekenning. Hierbij bestaat elke cyclus uit vier fasen, te weten:

Fase 1: aanmaak van het te bewerken bestand
- (gefaseerde) aanmaak van een bestand met alle basislexemen uit de LGBN en aanvullende kenmerken, zoals afbreekvorm, uitspraak en inflectiecategorie
-toekenning van morfologische structuurkenmerken

Fase 2: analyse van de beschikbare woordrepresentaties
-kwantitatieve gegevens over het lexicon
-kwantitatieve gegevens over de morfeemstructuur
-kwalitatieve gegevens over specifieke morfeemklassen

Fase 3: evaluatie van de geanalyseerde woordrepresentaties
-vergelijking van MGBN-patronen met MHB-patronen

-identificatie van cognitief bepaalde distributiepatronen
Fase 4: correctie van de foutieve woordrepresentaties
-geautomatiseerde correctie van formele structuurfouten
-handmatige correctie van inhoudelijek structuurfouten

5.4.2 Lexicografische randvoorwaarden

De MGBN dient uit te gaan van de lexicografische informatie in de MKB-Ned en moet zo worden opgezet dat de lexicale inhoud aan de uitgangspunten van een Ideaal Woordenboek[196] voldoet. Dit betekent dat het MGBN-lexicon zich minimaal moet kenmerken door een mentale basis, consistentie, compleetheid en correctheid:

Mentale basis De lexemen in de MGBN dienen op dezelfde manier te worden gerepresenteerd als in het mentale lexicon. In de praktijk komt dit neer op de eis dat de morfologische structuurrepresentaties zoveel mogelijk in overeenstemming dienen te zijn met de morfologische intuïtie van Nederlandse taalgebruikers.

Compleetheid De MGBN moet zo worden opgezet dat alle bestaande en mogelijke Nederlandse woorden van morfologische structuur kunnen worden voorzien. Hierbij dient zoveel mogelijk rekening te worden gehouden met aanvullende lexeemkenmerken (zoals spelling, uitspraak, categorie en betekenis).

Consistentie Bij de opbouw van de morfologische structuurrepresentaties dient zo consistent mogelijk te worden gewerkt; dit betekent dat voortdurend moet worden gecontroleerd of lexemen met vergelijkbare vormkenmerken ook op dezelfde manier worden gestructureerd, tenzij er goede redenen zijn om een afwijkende structuur te kiezen.

Correctheid De morfologische structuurrepresentaties dienen een goede benadering te bieden van de mentale kennis van de redacteur; bovendien dienen ze formeel correct te zijn in de zin dat elke representatie aan systeeminterne vormeisen moet voldoen.

De analysemethode dient verder aan de volgende randvoorwaarden te voldoen:

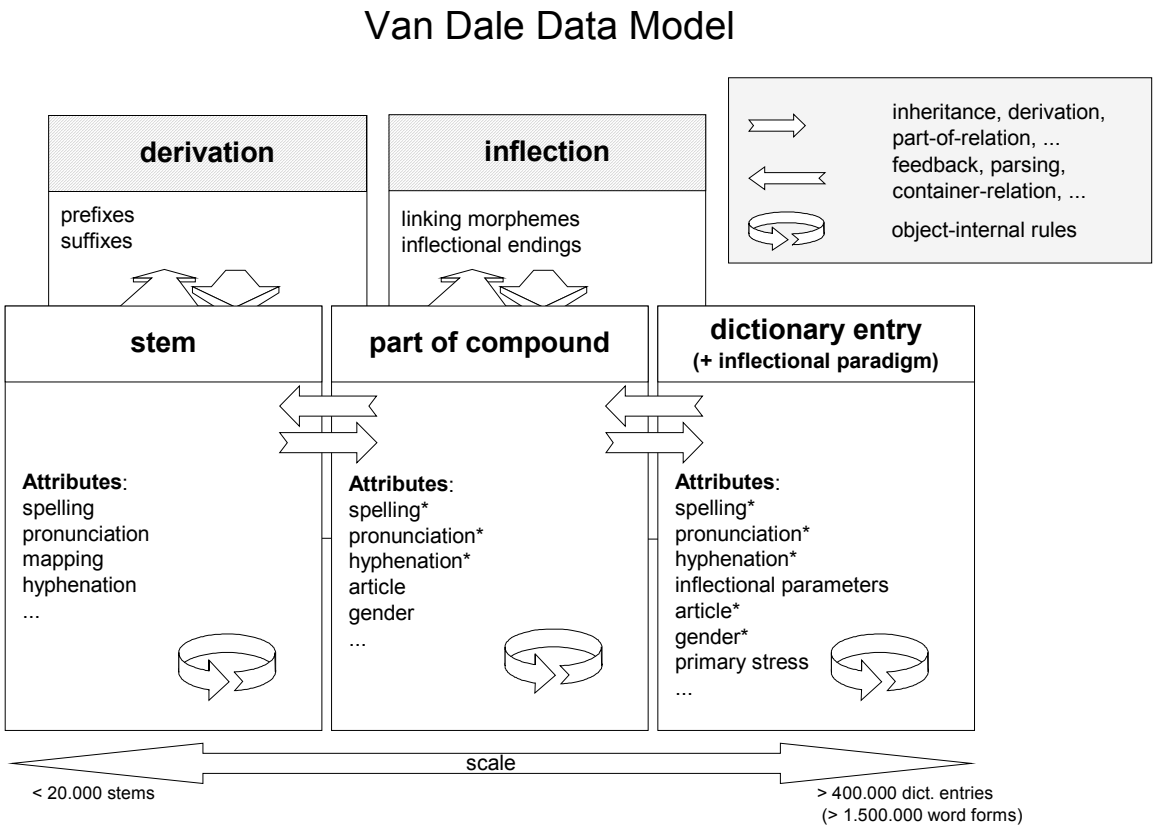
Uitvoerbaarheid De MGBN moet binnen enkele jaren gerealiseerd kunnen worden, want de MGBN vormt de basis voor mijn verdere onderzoek; bovendien zijn er verschillende VDL-projecten die baat hebben bij de MGBN-informatie.

Efficiëntie Bij de structurering van de MGBN dient snelheid voor precisie te gaan; foutjes en inconsistenties zijn namelijk onvermijdelijk, maar zullen doorgaans geen significante invloed hebben op de identificatie van morfologische patronen; omgekeerd kunnen deze patronen wel helpen om foutjes en inconsistenties op te sporen en te corrigeren.

Flexibiliteit Er is een aanpak nodig die flexibel kan omgaan met gegevens die pas in een later stadium beschikbaar komen of die tijdelijk "bevroren" moeten worden; want de onderliggende gegevensbank (de WKB-Ned) heeft een dynamische status in de zin dat de inhoud voortdurend wordt uitgebreid en aangepast.

Leesbaarheid Er dient een leesbaar coderingssysteem te worden gehanteerd, d.w.z. een systeem dat eenvoudig is te coderen en dat zich goed leent voor automatische verwerking.

[196] Deze term verwijst naar het in hoofdstuk 1 besproken lexiconmodel van Verkuyl & al. (1998).



Figuur 5-5: Het lexicografische kennismodel van VDL, onderdeel vormkenmerken.

5.4.3 Van Dale's lexicale kennismodel

De MGBN heeft als doel om Van Dale's WoordKenmerkenBank Nederlands (WKB-Ned; zie H5.5.2 voor nadere informatie) met een morfologische structuurlaag uit te breiden. Het in figuur 5-5 weergegeven kennismodel (het Van Dale Datamodel)[197] laat zien hoe deze morfologische structuurlaag (met stammen en affixen) zich tot de andere informatielagen moet gaan verhouden, te weten de lexeemlaag, die informatie geeft over de woordkenmerken op het niveau van de lexemen (waaronder inflectie en tussenklanken), en de woordlaag, die de citatievorm geeft van de met deze lexemen gevormde woorden en samenstellingen en vaste woordcombinaties (c.q. meervoudige expressies); hierbij wordt voor elk woord informatie verstrekt over woordklasse, samenstellingsgrenzen, inflectiekenmerken, uitspraak, afbreekposities en betekenis. De inflectionele informatie dient automatisch te worden toegevoegd bij de overgang van de lexeemlaag naar de woordlaag (d.m.v. overerving). Elke laag bestaat uit een niveau met basismateriaal en een complex niveau, waar deze bouwstenen al dan niet op basis van regels zijn samengevoegd tot grotere eenheden. Bij de morfeemlaag gaat het om de samenvoeging van stammen en affixen tot lexemen, bij de lexeemlaag om de samenvoeging van lexemen tot (samengestelde) woorden (maar nog zonder inflectiekenmerken).[198]

5.4.4 Morfologische annotatiemethode

Ik zal nu uiteenzetten welke methode ik heb gevolgd bij de aanmaak van de morfologische structuurrepresentaties in de MGBN. Volgens het theoretische ontwerp in H5.2 zou de MGBN op den duur in staat moeten zijn om voor alle hierin opgeslagen woorden informatie te geven over de morfologische structuur van hun fonologische (waaronder orthografische en

[197] Dit model is ontworpen is door Johan Zuidema en Marc du Chatinier, beide werkzaam bij VDL.
[198] Ordelman (2003) benut deze data voor de evaluatie van een automatische spraakherkenningstool.

auditieve) lexeemvormen, waarbij deze informatie op dezelfde structuurprincipes moet berusten als de representaties in het mentale lexicon. Alleen is nog niet bekend hoe die structuurcriteria er precies uitzien, zodat ze geen houvast bieden voor de structurering van de WKB-Ned. Ook als deze principes wel bekend waren, zou het waarschijnlijk lastig zijn om deze op de WKB toe te passen, want als de L-KRING-theorie een correcte beschrijving geeft van het mentale lexicon, bevat het gedetailleerde gebruikskennis over een groot deel van de lexemen en lexeemcontexten die de taalgebruiker ooit is tegengekomen. Meer in het bijzonder geeft het gedetailleerde vorm- en betekenisrepresentaties en fijnmazige frequentiegegevens. In de L-KRING-theorie is deze informatie onmisbaar voor een adequate identificatie van morfemen en andere structuureenheden, al is het een empirische vraag hoe vorm, betekenis en frequentie onderling moeten worden gewogen.

Bij de realisatie van de huidige MGBN heb ik het hier beschreven ideaal afgezwakt tot het doel om voor alle in de MGBN opgeslagen basislexemen een cognitief gemotiveerde structuurrepresentatie aan te maken die inzicht geeft in de kleinste morfologische structuureenheden van hun spelvorm en om deze representaties zo systematisch mogelijk te structureren. Om deze structuur te achterhalen heb ik een semi-automatische analysemethode gehanteerd, wat neerkomt op een cyclisch proces van redactioneel gecontroleerde structuurtoekenning waarbij de redacteur steeds kan afwisselen tussen computationele en redactionele analysetechnieken, bijvoorbeeld door interactief gebruik van automatische zoek- en vervangopdrachten.[199] Deze inductieve (datagestuurde) analysemethode maakt het mogelijk om in relatief korte tijd een groot aantal woorden van gedetailleerde en cognitief gemotiveerde morfeemrepresentaties te voorzien, terwijl de op deze wijze opgebouwde patrooninventarisatie informatie geven van de morfologische structuurkenmerken die een rol spelen bij de mentale representatie van de bestaande woordenschat. Bij deze aanpak ontstaan de structuurcriteria en coderingsconventies tegelijk met het analyseproces. De op deze wijze tot stand gekomen conventies kunnen daarna in het hele bestand worden doorgevoerd. Hierdoor gaan de structuurcriteria steeds beter aansluiten op de geanalyseerde data. De aanpak heeft als nadelen dat hij arbeidsintensief is,[200] dat het onderzoek niet op de gebruikelijke wijze is te reproduceren (doordat de resultaten sterk afhankelijk zijn van de kennis en doelstelling van de bewerker)[201] en dat het resultaat aanvankelijk minder consistent is dan bij een regelgebaseerde parser. Maar het grote voordeel is dat de structuurrepresentaties van begin af aan een directe weerspiegeling vormen van de kennis in het mentale lexicon, dat er gedetailleerdere representaties kunnen worden opgebouwd en dat er veel minder ambiguïteitsproblemen ontstaan.

Ik ga ervan uit dat de MGBN een bruikbare kennisbron is voor statistisch onderzoek naar de morfologische patronen van het Nederlands. Meer specifiek wil ik nagaan welke morfeempatronen potentieel deel uitmaken van het Nederlands, wat hun typefrequentie is en in hoeverre deze patronen taalkundig relevant zijn. Deze vragen kunnen op twee manieren worden onderzocht. De eerste mogelijkheid is om voor alle patronen na te gaan of ze reeds in de taalkundige literatuur zijn beschreven, dus of ze deel uitmaken van de bestaande kennis over de morfologische grammaticaregels van het Nederlands (zoals de regels in het MHB). De tweede mogelijkheid is om op basis van concrete patrooninventarisaties statistische criteria te formuleren die bepalend zijn voor de vraag of een potentieel patroon een significante bijdrage levert aan de compressie en coherentie van het lexicon. Dergelijke criteria maken het mogelijk om de structuurrepresentaties kwalitatief te beoordelen en zonodig aan te passen. Niet alleen kan

[199] Dit houdt in dat de redacteur item voor item beslist over de toepasbaarheid van het opgegeven patroon.
[200] Dit is sowieso een arbeidsintensieve aangelegenheid: het vergde al met al ruim twee jaar.
[201] Maar idealiter zou een andere redacteur tot een vergelijkbaar resultaat moeten komen, anders zouden deze redacteuren structureel verschillende talen spreken.

dit ten goede komen aan de morfologische kwaliteit van de MGBN, maar ook aan het inzicht in de morfologische eigenschappen van het mentale lexicon.

5.5 Aanmaak van het basisbestand

5.5.1 Introductie

Bij de aanmaak van het MGBN-basisbestand ben ik, zoals ik in H5.4.3 aankondigde, uitgegaan van de woordkenmerken in VDL's WoordKenmerkenBank Nederlands (WKB-Ned). Deze kennisbank biedt gesystematiseerde informatie over de vormkenmerken van alle trefwoorden uit VDL's Nederlandstalige woordenboeken. Naast de WKB-Ned kunnen een aantal gespecialiseerde sublexica worden onderscheiden, te weten Van Dale's Groot Woordenboek der Nederlandse Taal (GWNT), c.q. Grote Van Dale, en het Groot Woordenboek Hedendaags Nederlands (WHN). Deze databronnen worden in H5.5.2 besproken.

5.5.2 Databronnen

5.5.2.1 De Woordkenmerkenbank Nederlands (WKB-Ned)

De Woordkenmerkenbank Nederlands (WKB-Ned) omvat alle trefwoorden (c.q. lexemen) uit VDL's Nederlandstalige woordenboeken (onder vermelding van de bronbestanden). In totaal betreft het een kwart miljoen lexemen (dus exclusief inflectievormen), die zijn opgebouwd uit ca. 80.000 samenstellende delen (c.q. basislexemen). Bij elk trefwoord worden de volgende woordkenmerken gespecificeerd:

- citatievorm; deze correspondeert meestal met de onverbogen woordvorm, maar bij niet-scheidbare werkwoorden wordt de infinitiefvorm gebruikt.[202]
- structuurinformatie: voor alle trefwoorden is een representatie beschikbaar waarin de samenstellende delen (c.q. basislexemen) en de afbreekposities zijn gemarkeerd; deze representaties zijn deels langs automatische weg aangemaakt[203]
- syntactische categorie; VDL hanteert een traditioneel classificatiesysteem waarbij per hoofdcategorie tal van functionele subcategorieën worden onderscheiden.
- inflectievormen: per syntactische categorie is een parametrisch analysesysteem ontwikkeld waarmee automatisch woordvormen kunnen worden gegenereerd; uitzonderingen zijn systematisch in kaart gebracht.
- uitspraakrepresentatie: hierbij wordt een codeersysteem gebruikt dat zo is vormgegeven dat de representaties makkelijk interpreteerbaar zijn en tevens bruikbaar zijn voor een automatisch spraaksynthesesysteem
- gebruiksfrequentie: bij elke woordvorm wordt informatie gegeven over de frequentie waarmee de woordvorm voorkomt in een omvangrijk corpus dat voor een groot deel uit Nederlandse krantenartikelen bestaat.[204]
- semantische klasse (via VLIS, VDL's semantische classificatiesysteem)
- specificatie van de bronbestanden: voor elk lexeem wordt aangegeven of het in een woordenboek is opgenomen, en zo ja in welke woordenboeken; deze informatie kan worden benut om toegang te krijgen tot de semantische woorddefinities
- overige kenmerken: etymologische gegevens, registerkenmerken (bijv. standaard/ archaïsch/ volks/ gewestelijk), syntactische collocaties, opmerkingen etc.

[202] De niet-scheidbare V-stam BESPREEK correspondeert bijvoorbeeld met de citatievorm *bespreken*, maar de scheidbare V-stam UITSPREEK met de citatievorm *spreek_uit*.
[203] Ordelman (2003) heeft deze structuurinformatie als uitgangspunt genomen voor de ontwikkeling van een automatische compound-splitter ten behoeve van een spraak-naar-tekst-systeem.
[204] Dit corpus heette destijds Nederlandse Pers Databank (NPD).

5.5.2.2 Groot Woordenboek der Nederlandse Taal (GWNT)

De Grote Van Dale (GWNT) is het meest omvangrijke en gezaghebbende woordenboek van het Nederlands van de afgelopen eeuw. Behalve een complete inventarisatie van de hedendaagse woordenschat biedt dit woordenboek een rijke inventarisatie van zeldzame woorden en bijzondere betekenissen. De GWNT omvat ca. 245.000 lemma's, waarvan ca. 1/3 met een basiswoord en ca. 2/3 met een samenstelling correspondeert. Sinds het jaar 2000 bestaat er naast de driedelige folio-editie ook een elektronische editie (de eGWNT). Alle trefwoorden en een groot deel van de vormkenmerken uit de GWNT zijn ook in de LGBN terug te vinden.

5.5.2.3 Groot Woordenboek Hedendaags Nederlands (WHN)

De WHN geeft informatie over een hedendaagse selectie uit de woorden in de GWNT (in totaal 94.000 lexemen). Bij ca. 10.000 woorden is ook informatie opgenomen over regelmatige afleidingen, zoals argument-nominalisatie (met -ER/-AAR/-OR/-ATOR), procesnominalisatie (met -ING/-ERING/-IE/-ATIE) en vrouwelijke persoonsmarkeringen (-IN/-ES/-ICE). In totaal gaat het om ca. 11.000 extra lexemen. Omdat deze lexemen geen zelfstandige woordingang bezitten, zijn ze niet in de WKB-Ned opgenomen. Maar deze aanvullende informatie is uiteraard zeer interessant met het oog op de inventarisatie van de morfologische derivatiemogelijkheden van de in de MGBN opgenomen woordstammen. Daarom zijn deze derivaties wel in de LGBN opgenomen (en vervolgens morfologisch geanalyseerd).

5.5.3 Opzet van de LGBN

De Lexicale Gegevensbank voor het Nederlands (LGBN) biedt een door de MGBN gemotiveerde selectie uit de woordinformatie in de WKB-Ned (VDL's Woordkenmerkenbank Nederlands). De LGBN specificeert voor alle hierin opgenomen woorden de samenstellingsstructuur en voor elk samenstellend deel (c.q basislexeem) een door de MGBN gemotiveerde selectie uit de beschikbare vormkenmerken (waaronder, klankvorm, afbreekvorm en syntactische klasse). Zo bestaat de afbreekvorm van het lexeem *levensbeschouwing* uit de constituenten *levens* (met bindmorfeem -S) en *beschouwing*. Daarom zijn beide constituenten als basislexemen in de MGBN opgenomen (maar deze zijn niet niet gedesambigueerd voor categorie, betekenis of uitspraak). De beperking tot basislexemen berust op de aanname dat de structuurrepresentatie van een samengesteld lexeem een compositioneel product is van de structuurrepresentaties van de samenstellende basislexemen. Deze basislexemen kunnen zowel met zelfstandige woorden als met samenstellende delen corresponderen.

5.5.4 Aanpassingen

Bij de opzet van de LGBN heb ik de nodige aanpassingen doorgevoerd met betrekking tot de identificatie van woordinterne basislexemen. Zo was het voor mijn morfologische doeleinden handiger om scheidbare prepositie als werkwoord-intern morfeem te analyseren (in plaats van basislexeem). Ik heb ook aanpassingen doorgevoerd in de specificatie van de bijbehorende lexeemkenmerken, onder meer met betrekking tot de inflectiecategorie. Hierbij heb ik de nummersystematiek door een lettersystematiek vervangen (met N, A, V, P etc.). Verder heb ik in een later stadium vele basislexemen die nog geen inflectiecategorie hadden gekregen, hier alsnog van voorzien; dit was mogelijk door gebruik te maken van de morfologische structuurkenmerken. In de appendix zal dit alles nader worden verantwoord.

5.6 Aanmaak van de morfologische representaties

5.6.1 De morfologische structuurkenmerken

Ik zal nu uiteenzetten welke structuurkenmerken ik heb aangemaakt bij de morfologische annotatie van de basislexemen in de MGBN. Als eerste stap heb ik de spelvorm zo systematisch mogelijk in potentiële morfeemsegementen (pm-segmenten) opgedeeld. Bovendien heb ik per representatie minstens één stamsegment gemarkeerd. De onderstaande MGBN-selecties demonstreren dit voor een inheemse stam (ZET) en een uitheemse stam (FORM):

(S1)
```
omzet = om;[zet]
omzetbaar = om;[zet];baar
omzetten = om;[zett];en
omzetting = om;[zett];ing
```

(S2)
```
transformeren = trans;[form];er;en
transformatie = trans;[form];at;ie
transformatief = trans;[form];at;ief
transformationeel = trans;[form];at;ion;eel
```

Bij ca. 5000 lexemen zijn zelfs meerdere stamsegmenten gemarkeerd: hierbij gaat het meestal om niet-transparante samenstellingen. Zo bevat de pseudo-samenstelling *aambeeld* de stammen AAM en BEELD. In de MGBN wordt dit als volgt verantwoord:

(S3)

lexeemvorm	lemmacode	pm1-stam	pm1-structuur
aambeeld	aambeeld.1	aam	[aam];{beeld}
aambeeld	aambeeld.2	beeld	{aam};[beeld]

De onderstaande tabel geeft een overzicht van de mogelijke structuurmarkeringen op het pm1-niveau (inclusief een korte omschrijving en een voorbeeld):

omschrijving	structuur	voorbeeld
morfeemgrens	morfeem_morfeem	be_[sprek]_ing
stamgrenzen	[stam] of {stam}	[ansjo]_{vis}
prefixen	prefix_prefix_[	ver_ge_[lijk]
suffixen	]_suffix_suffix	[treff]_end_heid
middenprefixen	]+_prefix_{	[cito]+_re_{cept}
middensuffixen	]_suffix+_{	[acht]_ens+_{waard}
tussenmorfemen	+_affix+_	[al]+_te+_{met}
bindfonemen	_foneem:suffix	be_[heer]_d:er, [plan]_o:log_ie
woorddeelgrens	deel1+_deel2	{aard}_s+_ge_[zin]_d
MHB-clusters	affix1=_affix2	[amalg]_er=_en

Als tweede stap heb ik voor alle segmenten een onderliggende vormindex (c.q. pm2-index) aangemaakt; op die manier heb ik regelmatige vormvarianten bijeengebracht. Als derde stap heb ik alle pm2-indexen aan een pm3-index gekoppeld; deze pm3-index heeft als functie om etymologisch verwante pm2-indexen bijeen te brengen. Ik zal een en ander verduidelijken door een concreet voorbeeld te geven, namelijk het vormparadigma van de pm3-stam SPREEK (waarvan de vorm op een arbitraire keuze berust):

(S4)

pm3-stam	pm2-stam	pm1-stammen	voorbeeld
spreek	sprEK	spreek, sprek	bespreekbaar, gesprek
spreek	sprAk	spraak, sprak	spraak, sprakeloos
spreek	spreuk	spreuk	spreuk

Bij affixen kan eveneens vormvariatie voorkomen. Dit blijkt uit de onderstaande tabel, die alle vormvarianten van het pm2-affixen ABEL weergeeft (het teken ! geeft aan dat het om een affixtoepassing op lexeemeinde gaat; het teken * markeert een MHB-affix).

(S5)

pm2-vorm	pm1-vorm	voorbeeld
*abel	abel	variabele
*abel	abil	venerabile
*abel	abl	inséparables
*abel	i:abel	justitiabelen
*abel!	abel	venerabel
*abel!	abiel	vegetabiel
*abel!	able	unspeakable
*abel!	e:abel	malleabel
*abel!	i:abel	ministeriabel

Op het pm2-niveau bestaan de volgende structureringsmogelijkheden:

kenmerk	L-KRING-beschrijving
*m	morfeemindex waarvan de spelvorm overeenkomt met een MHB-morfeem
*m1=_*m2	in het MHB vermelde combinatie van twee morfeemindexen (m1 en m2);
_[m]_	morfeemindex met wortelstatus
_m_	morfeemindexen met affixstatus (prefix = 'm_'; suffix = '_m')
L1+L2	combinatie van twee lexeemindexen (waarbij L1 en L2 uit morfemen bestaan)

Speciale markeringen voor pm2-affixen

_#	stamachtig affix van uitheemse herkomst (zoals _#scOp)
_$	stamachtig affix van inheemse herkomst (zoals _$halve)
_&	autonoom affix, bijv. _&te in [al]+_&te+_{met}
_*	affix met vermelding in Handboek, bijv. *be, *ing (of *$kund, *#loog)
=	affix-clusters (slechts incidenteel gecodeerd)
	a) met structuur affix1=affix2, bijv. [anim]_*At=ie
	b) met structuur affix1=_affix2, bijv. [amalg]_*Er=_en
	(structuur b is gemotiveerd door MHB-clusters)

Enkele notatieconventies bij de vorm van de pm2-index:
-hoofdletter A staat voor a/aa-alternantie (idem voor andere klinkers)
-hoofdletter F staat voor f/v-alternantie
-hoofdletter Z staat voor s/z-alternantie
-de pm1-affixen [e:lijk] en [lijk] zijn vormvarianten van de pm2-vorm [lijk]
-de pm1-affixen [baar] en [bar] zijn vormvarianten van de pm2-vorm [bAr]

Hieronder volgt een concreet voorbeeld van de morfologische veldstructuur van de MGBN; het betreft de MGBN-representatie van het lexeem *toegankelijkheid*:

<trefwoord>	toegankelijkheid:1
<categorie>	N
<lexeemstatus>	zelfstandig
<lexeempositie>	0
<pm1-stam>	gank
<pm2-stam>	ganK
<pm3-stam>	gaan:1
<pm1-structuur>	toe_[gank]_e:lijk_heid
<pm2-structuur>	toe_[ganK]_lijk_heid

De MGBN specificeert voor elk lexeem drie morfologische structuurniveaus, namelijk een pm1-structuur (de morfologische structuur op spelvormniveau), een pm2-structuur (de morfologische structuur op het eerste abstractieniveau, met als functie om regelmatige vormvarianten bijeen te brengen) en een pm3-structuur (de morfologische structuur op het tweede abstractieniveau, waar alle etymologisch verwante pm2-eenheden worden gebundeld). Men kan dus ook drie soorten stammen aantreffen, namelijk de pm1-stam ([c.q. "a-stam"]; deze correspondeert met de spelvorm), de pm2-stam ([c.q. "o-stam"]; deze bundelt regelmatige pm1-stam-varianten) en de pm3-stam ([c.q. "s-stam"]; deze bundelt onregelmatige pm1-stam-varianten). Uit de hier gespecificeerde informatie blijkt dat ik ervan uit ben gegaan dat het lexeem *toegankelijkheid* van de stam GAAN is afgeleid (met pm2-vorm GANK en pm1-vorm GANK). Hieronder volgt nadere informatie over de weergegeven velden.

MGBN-term	L-KRING-beschrijving
trefwoord	lexeemindex (die overeenkomt met de lexicografische citatievorm)
categorie	inflectiecategorie bij toepassing als woordfinaal lexeem (bijv. $N, $V of $A)
lexeemstatus	positie-informatie: zelfstandig woord of links/midden/rechts in woord
pm1, pm2	pm = afkorting voor potentieel morfeem binnen een lexeemrepresentatie (in de spelvorm); het nummer markeert het structuurniveau
pm-structuur	potentiële morfeemstructuur van een lexeemrepresentatie (in de spelvorm); deze structuur bestaat uit één of meer pm-segmenten (c.q. morfemen), waaronder minimaal 1 pm-stam; er bestaan meerdere pm-niveaus
pm1-structuur	niveau-1-representatie van de morfologische structuur van een lexeem; deze is opgebouwd uit pm1-morfemen
pm2-structuur	niveau-2-representatie van de morfologische structuur van een lexeem; deze is opgebouwd uit pm2-morfemen
pm1-stam	niveau-1-index van de basisstam; deze index correspondeert met een direct waarneembare spelvorm
pm2-stam	niveau-2-index van de basisstam; deze index bundelt voorspelbare spellingsvarianten
pm3-stam	niveau-3-index van de basisstam op spelvormniveau; deze index bundelt onvoorspelbare spellingsvarianten
pm1-affix	niveau-1-index van de basisstam; deze index correspondeert met een direct waarneembare spelvorm
pm2-affix	niveau-2-index van een affix; deze index bundelt regelmatige vormvarianten
pm2-affix	niveau-2-index van een affix; deze index bundelt niet-regelmatige vormvarianten

5.6.2 Werkwijze

Bij de opbouw van de MGBN ben ik cyclisch te werk gegaan. Hierbij diende het eindstadium van cyclus 1 als beginstadium voor cyclus 2, en zo verder. Deze werkwijze heeft als voordeel dat de reeds opgebouwde structuur bijdraagt aan de mogelijkheden tot verfijning van deze structuur. Elke bewerkingscyclus kende de volgende onderdelen:

1) aanmaak van het werkbestand
- dataselectie vanuit centraal beheersysteem
  a) selectie van bronbestand(en)
  b) selectie en duplicatie van relevante data
  c) herstructurering van de geselecteerde data
- voorbewerking
  - script schrijven voor automatische structuurtoekenning

- uittesten en verbeteren van script
- toepassing van het script

- sortering
  -selectie en ordening van de te sorteren velden
  -keuze tussen alfabetische en getalsmatige sortering
  -keuze tussen oplopende of aflopende sortering
  -keuze tussen lineaire of retrograde sortering
  -wel/niet hoofdlettergevoelig sorteren
  -specificatie van speciale opties, zoals te negeren symbolen
- opmaak
  -keuze van het weergaveformaat: rijen (c.q. tabel) of kolommen
  -ordening van de informatievelden
  -toevoeging van speciale symbolen
  -overheveling naar tekstverwerker
  -eventueel stijlmarkeringen aanbrengen (binnen tekstverwerker)

2) semi-automatische structurering van de data
- redactionele bewerking van het bestand door de aanwezige eenheden zo consistent mogelijk van structuur te voorzien (op basis van intuïtieve structuurcriteria)
- zo mogelijk zoek- en vervangpatroon definiëren en semi-automatisch uitvoeren, zodat de redacteur maximale controle houdt over het wel/niet toepassen van een patroon.
- raadpleging van aanvullende informatiebronnen (indien noodzakelijk)
  - speciale referentielijsten (o.a. veldcodes) en MGBN-hulpbestanden
  - elektronische woordenboeken (vooral de eGWNT en de GWNT-index)
  - naslagwerken (vooral het MHB, de WHN en de EWN)

3) cyclische verfijning van de aangebrachte structuur
- herordening van het werkbestand
- evaluatie van de aangebrachte structuur
- indien nodig: verfijning van deze structuur (met de technieken uit stap 2)
- herhaling van stap 3 (tot het bestand in orde is)

4) terugkoppeling naar centraal beheersysteem
- indien nodig: conversie naar text-formaat
- formele consistentiecontroles (desgewenst interactief)
- optioneel: nabewerking data
  -hulpvelden en hulpsymbolen verwijderen
  -automatisch doorgeneren van structuurkenmerken
- optioneel: aanpassing van de bestandsopmaak
- optioneel: integratie met het moederbestand (c.q. "inritsing")
- opslag in centraal beheersysteem

Het hier weergegeven schema laat zien dat elke cyclus met de aanmaak van een werkbestand begint. Na de aanmaak van dit werkbestand volgt een snelle bewerking van de te analyseren eenheden, om vervolgens de aangebrachte structuur te verfijnen. Dit is efficiënter dan om van begin af aan heel precies te werken. Bij een dergelijke werkwijze is het echter niet mogelijk om allerlei theoretische criteria te hanteren; in plaats daarvan baseerde ik mijn keuzes op inwaartse (stamgerelateerde) analogie en uitwaartse (affixgerelateerde) analogie onder abstractie van semantische transparantie. Bij bekende woorden kon ik dit soort analogie snel vaststellen op basis van mijn intuïties over het Nederlands (en de naslagwerken die mij ter beschikking stonden). Bij onbekende woorden leverde het meer problemen op; hier heb ik waarschijnlijk ook veel meer "fouten" gemaakt (d.w.z. etymologisch onjuiste oordelen).

Bij de structurering van de MGBN heb ik mij op intuïtieve structuuroordelen gebaseerd. Om beter analyseerbare gegevens te verkrijgen heb ik mogelijke variatie in mijn structuuroordelen tot een minimum proberen te reduceren door het databestand tijdens en na elke bewerkingscyclus op consistentie te controleren. Zulke controles zijn in feite een eerste stap naar theorievorming, want zij abstraheren van "toevallige" variatie in de structuuroordelen. Door de ingebouwde controles is de MGBN-methode robuust. Daarom verwacht ik dat de door mij tot stand gebrachte gegevensbank (namelijk de MGBN) een bruikbare basis biedt voor kwalitatief en kwantitatief onderzoek naar de morfologische patronen van het Nederlands. De hier bedoelde onderzoeksmogelijkheden vormen het centrale thema van hoofdstuk 6.

5.6.3 De structuurcriteria

5.6.3.1 Introductie

Zoals ik in hoofdstuk 3 heb toegelicht gaat het Morfologisch Handboek (MHB) ervan uit dat morfemen met de kleinste woordinterne klankeenheden corresponderen die een voorspelbare bijdrage aan de woordbetekenis leveren. Indien niet aan deze transparantie-eis wordt voldaan kan het segment geen morfeemstatus krijgen. In mijn optiek is het dan ook veel belangrijker of woordinterne segmenten een bijdrage kunnen leveren aan de lexicale compressie van de woordkenmerken; hiervoor is het voldoende als dit segment één of meer voorspelbare combinatiemogelijkheden bezit. Er zijn namelijk tal van lexemen die geen compositionele betekenis bezitten, maar wel formeel geleed zijn.

Neem bijvoorbeeld het lexeem *ingewikkeld*: op het niveau van de spelvorm is dit lexeem ambigu tussen een compositionele betekenis (namelijk "de toestand die voortkomt uit het inwikkelen van een voorwerp") en een versteende betekenis, namelijk "moeilijk". Indien men een compositioneel structuurcriterium hanteert, dient het lexeem met de tweede betekenis structuurloos te blijven. Maar indien men alleen in de formele morfeemstructuur is geïnteresseerd, hoeft geen betekenisonderscheid te worden gemaakt.

Zo'n formele morfeemstructuur kan goed worden gemotiveerd indien men aanneemt dat een morfeem primair de functie heeft om een reeks combinatorische eigenschappen te coderen (aangezien dit bijdraagt aan de compressie van lexicale informatie). Hierdoor kan recht worden gedaan aan het feit dat formeel gelede lexemen vaak dezelfde combinatorische eigenschappen bezitten als hun compositioneel interpreteerbare tegenhangers. Hierbij geldt de aanvullende eis dat de te onderscheiden morfemen potentieel dezelfde betekenis kunnen aannemen als in een transparant geleed woord. Om die reden kan het segment *ing* in *koning* niet als morfeem worden aangemerkt, want hoewel dit segment dezelfde inflectiekenmerken met zich meebrengt als *ing* in het transparant gelede *deling*, is het niet potentieel interpreteerbaar. Dit is wel mogelijk voor *ing* in *woning*, want doordat de stam potentieel een werkwoord kan zijn, is het segment *ing* hier ook potentieel interpreteerbaar als een nominaliserend suffix. De hier verwoorde overwegingen berusten op het onderstaande structuurcriterium:

> **Structuurcriterium voor morfeemidentificatie**
> *Men kan een lexeemintern segment als morfeem markeren indien dit segment vaste combinatorische eigenschappen bezit (hetgeen uit analogietesten moet blijken) en als het potentieel in staat is om (binnen de context van het geanalyseerde lexeem) een voorspelbaar betekeniskenmerk toe te voegen.*

Dit structuurcriterium heeft grote gevolgen voor de opdeling van uitheemse woorden: zo gaan de gangbare morfologische theorieën er vanuit dat werkwoorden als *exporteren*, *importeren* en *transporteren* moeten worden afgeleid van de stammen EXPORT, IMPORT en TRANSPORT, hoewel ze een gemeenschappelijke wortel PORT bezitten (die steeds iets betekent in de trant van "dragen" of "verplaatsen"; omdat deze wortel echter nooit zelfstandig wordt gebruikt, en

omdat de gemiddelde Nederlander geen Latijnse prefixen zou kennen, wordt doorgaans aangenomen dat deze structuur alleen etymologisch kan worden gemotiveerd en dus onzichtbaar is voor de grammatica (cf. Don & al. (1994)). In mijn visie is dit onterecht, want de interne structuur van deze woorden is vaak wel degelijk van belang voor het begrip van deze woorden; bovendien vertoont de wortel vaak morfologisch voorspelbare vormalternanties, ongeacht het prefix. Alleen al vanwege lexicografische doeleinden is het dan de moeite waard om deze patronen zichtbaar te maken, en te benutten voor de systematisering van de hieraan gekoppelde woordkenmerken. Mijn op de L-KRING-theorie gebaseerde definitie van morfemen legitimeert deze aanpak, want de prefixen EX-, IN- en TRANS- kunnen in de gegeven gebruikscontext (namelijk voorafgaande aan een wortel) potentieel betekenis dragen.

Gegeven dit criterium kan men zich afvragen in welke contexten het segment *el* als een morfologisch relevante eenheid kan worden aangemerkt. Dit is niet mogelijk in de context van de lexeemvorm *ingewikkeld*, Zo is het niet waarschijnlijk dat de stamvorm *wikkel* semantisch gezien de structuur WIK+EL bezit; maar indien het segment *el* voorspelbare effecten heeft op het derivatiegedrag van de stam WIKKEL (zoals de keuze van het suffix -AAR in *wikkelaar*), kan deze structuur toch motiveerbaar zijn. Tot slot zou men ook nog etymologische overwegingen kunnen hanteren.

Dit ene voorbeeld lijkt me voldoende om aan te tonen dat lexemen vele structuurdimensies kennen, en dat het zonder nadere definitie onmogelijk is om de ideale morfeemstructuur aan te wijzen. Vanuit het mentale lexicon bezien is deze verwarring wel begrijpelijk, want het mentale lexicon kan alle structuurdimensies tegelijk representeren, zonder aan te geven welke dimensie het label "morfologisch" draagt. Dit is namelijk geen mentale categorie, maar een taalkundige categorie. Toch stelt de L-KRING-theorie dat het mentale lexicon een inzichtelijke structuur vertoont, namelijk een multidimensionale indexstructuur, maar deze wordt pas zichtbaar als men lexemen integraal bekijkt.

Indien men zich beperkt tot de analyse van de spelvorm (of de klankvorm), is men gedwongen om de indexstructuur terug te brengen tot een 1-dimensionele projectie. Dit is zo'n onnatuurlijke opgave dat de resulterende morfeemstructuur al gauw willekeur gaat vertonen. Bij de opzet van de MGBN heb ik deze willekeur proberen te bedwingen door al doende morfeemspecifieke analyseconventies te ontwikkelen en deze conventies zo vorm te geven dat de resulterende gegevensbank optimale mogelijkheden biedt voor computationeel onderzoek naar de morfologische dimensie van de Nederlandse woordenschat.

Met betrekking tot het segment *el* heb ik bijvoorbeeld de conventie gehanteerd dat er alleen morfeemstatus mag worden toegekend als er sprake is van een herkenbare wortel, zoals de wortel HAK in *hakkelen* (met structuur HAK+EL+EN); hierdoor wordt gericht onderzoek mogelijk naar de vraag of de aanwezigheid van een herkenbare wortel invloed heeft op de combinatorische mogelijkheden van het segment *el*.

Een andere conventie die ik hier wil noemen betreft de vraag wanneer twee stammen met duidelijke vormovereenkomsten tot hetzelfde basismorfeem kunnen worden herleid; bij inheemse morfemen heb ik dit sterker van de betekenisovereenkomst laten afhangen dan bij uitheemse morfemen, want bij de inheemse stammen wilde ik de selectiecondities preciezer kunnen analyseren dan bij de uitheemse. Het is vrijwel onmogelijk om dit soort conventies systematisch te expliciteren, want dan zal men segment voor segment moeten aangeven welke condities bepaalden of er sprake was van een affix. Maar bij detailanalyses zullen dit soort conventies vanzelf boven water komen.

5.6.3.2 De identificatie van affixen

Bij de morfologische structurering van de MGBN-lexemen heb ik me in eerste instantie op de identificatie van affixen gericht, want affixen zijn meestal hoogfrequent en hebben meer invloed op de lexeemeigenschappen, waardoor ze beter herkenbaar zijn. Zo corresponderen lexemen met het eindsegment *lijk* vrijwel altijd met een modificator (c.q. A-lexeem), d.w.z. met een N-modificerende (adnominale) eenheid (c.q. A_N-lexeem) of een V-modificerende (adverbale) eenheid (c.q. A_V-lexeem). In dergelijke lexemen correspondeert het segment *lijk* duidelijk met een suffix, namelijk het #a-suffix -LIJK. Dergelijke lexemen zijn niet alleen herkenbaar aan hun semantische en syntactische eigenschappen, maar ook aan hun inflectie- en derivatiegedrag. Want *lijk*-lexemen vertonen standaard A-inflectie (namelijk de contextspecifieke selectie van een buigings-*e*), kunnen bijna altijd een vergrotende en een overtreffende trap vormen (door affixatie met -ER of -ST) en staan derivatie toe met het suffix -HEID. Zo vormt het lexeem *aanschouwelijk* de basis voor woordvormen als *aanschouwelijke*, *aanschouwelijker*, *aanschouwelijkst* en *aanschouwelijkheid*.

Ook wat betreft de inwaartse selectiemogelijkheden vertoont het suffix -LIJK voorspelbaar gedrag: want het suffix -LIJK hecht zich bij voorkeur aan inheemse #v-stammen, d.w.z. aan stammen die een inheemse klankvorm bezitten en die direct (zonder affixatie) als V-lexeem kunnen worden toegepast. Dergelijke stammen zijn herkenbaar aan het feit dat ze ook argumentnominalisatie (met de suffixen -ER of -AAR) en procesnominalisatie (met -ING of door conversie) kunnen ondergaan, evenals A-vorming met het modaliserende suffix -BAAR. Dit geldt ook voor de stam van *aanschouwelijk*, te weten de #v-stam AANSCHOUW. Want naast $v en -LIJK staat deze #v-stam ook derivaties toe met -ING, -ER en -BAAR.

Het eindsegment *lijk* van het lexeem *aanschouwelijk* voldoet dus aan alle criteria voor de identificatie van het suffix -LIJK. Meer in het algemeen kan worden gesteld dat er minstens drie soorten criteria zijn op grond waarvan men kan bepalen of een lexeemintern segment affixstatus bezit:

1) semantische en syntactische kenmerken op lexeemniveau

2) uitwaartse selectiekenmerken c.q. inflectie- en derivatiegedrag

3) inwaartse selectiekenmerken c.q. substitutiegedrag

Indien een segment duidelijk als affix herkenbaar is, heb ik het altijd als zodanig gemarkeerd, ook als het lexeem als geheel geen compositionele betekenis lijkt te hebben. Hierbij speelde de productiviteit van het morfeem geen rol, d.w.z. de mate waarin een bouwsteen gebruikt wordt om nieuwe woorden te vormen.

Nu bezit het transparant gelede werkwoord *aanschouwen* een goed herkenbare basisstam, namelijk het #v-morfeem SCHOUW; dit morfeem kan namelijk ook zelfstandig als V-lexeem voorkomen. Maar in het formeel gelede werkwoord *ontginnen* correspondeert de basisstam GIN met een eenheid die niet zelfstandig bruikbaar is en waar dus niet zo makkelijk een zelfstandige betekenis aan kan worden toegekend. Wel is door de combinatie met ONT- onmiddellijk duidelijk dat de afgeleide stam ONTGIN zich als een V-lexeem gedraagt. Om die reden kan het segment ONT- direct als affix worden aangemerkt, met als gevolg dat het complement van ONT- ook morfeemstatus krijgt, ongeacht de eigenschappen van dit segment.

Er zijn ook lexemen waarbij de stam juist beter herkenbaar is dan het affix. Dit is bijvoorbeeld het geval in het lexeem *dievegge*: qua vorm en betekenis is namelijk direct duidelijk dat er sprake is van een lexeem met de stam DIEF, zodat moet worden aangenomen dat het unieke segment -EGGE hier met een affix correspondeert dat als markering van een vrouwelijke persoon dient.

5.6.3.3 Functionele ambiguïteit

Een andere complicatie ontstaat indien een vormsegment meer dan één functie kan aannemen. Zo kan het segment *-er* drie duidelijk herkenbare functies representeren, namelijk *vergrotende trap* (bijv. MOOI+ER) *frequentatief* (bijv. KLAP(P)+ER, MEK(K)+ER) en *argument-nominalisatie* (bijv. WERK+ER); hiernaast kan dit segment ook in minder scherp afgebakende toepassingen opduiken, waaronder toepassingen als persoonsmarkering (bijv. STAK(K)+ER) of toestandsmarkering (WAKK+ER). Dergelijke toepassingen komen bijna alleen in combinatie met een wortelstam (of slecht herkenbare stamallomorf) voor, zodat zelden sprake is van een compositionele betekenisopbouw. Hierdoor zijn dergelijke "pseudo-morfemen" moeilijk van functieloze toepassingen te onderscheiden (die men lijkt aan te treffen in lexemen als *akker* en *snugger*). Bij de ontwikkeling van de MGBN heb ik alleen pseudo-affixen gemarkeerd die met een onafhankelijk gemotiveerde stam corresponderen; anders zou al gauw te veel morfologische ruis zijn ontstaan.

De hier onderscheiden segmentfuncties duiken soms ook binnen één lexeemvorm op. Zo kan het segment -er van de lexeemvorm *lekker* de volgende functies vervullen: agens-nominalisatie bij de #v-stam LEK (persoon die lekt of object dat lekt), vergrotende trap bij de #a-stam LEK en (mogelijk) als pseudomorfeem in de adjectief-functie ('smakelijk') en de bijwoord-functie ('behoorlijk'). Deze segmentfuncties vertonen grote verschillen in hun distributieve gedrag (d.w.z. in de samenstelling van hun inwaartse en uitwaartse selectiemogelijkheden).

Gegeven een langs distributieve weg gemotiveerde hoofdfunctie kan men soms ook semantisch gemotiveerde subfuncties onderscheiden. Zo kent het suffix voor argument-nomninalisatie subfuncties als agens-nominalisatie (bijv. *werker* en *loper*), thema-nominalisatie (bijv. *stijger*, *ontvanger* en *lijder*), instrument-nominalisatie (*wekker* en *knijper*) en effect-nominalisatie (*giller*). Bij dergelijke polysemie is het niet wenselijk om van verschillende affixen te spreken, want vaak zijn meerdere subfuncties per lexeemvorm mogelijk (zo kunnen *wekker*, *knijper* en *giller* ook wel in de agens-functie voorkomen), terwijl de voorkeursfunctie meestal uit de interactie tussen stamconcept, affixconcept en pragmatische toepassingsmogelijkheden valt te voorspellen. Als een stam bijvoorbeeld geen agensfunctie bezit (zoals het geval is bij *stijgen*), zal noodzakelijkerwijs een ander argument moeten worden geactiveerd. Vanwege deze overewegingen ben ik alleen tot functionele onderverdeling van affixen overgegaan indien de onderscheiden affixfuncties een systematisch contrast vertonen met betrekking tot de morfologische selectiemogelijkheden.

5.6.4 Empirische complicaties

Hieronder volgt een overzicht van de complicaties die ik tegenkwam bij de morfologische analyse van de lexemen in de MGBN.

1) De MGBN moest regelmatig met nieuwe lexemen worden uitgebreid.

Op het moment dat ik aan de opbouw van de LGBN en de MGBN begon, was de centrale informatiebron (namelijk de WKB-Ned) nog volop in ontwikkeling. Hierdoor was het niet mogelijk om met een compleet basisbestand te beginnen, want voor een deel van de LGBN-lexemen was aanvankelijk nog geen structuurinformatie beschikbaar (waardoor geen opdeling in basislexemen mogelijk was), en hetzelfde geldt voor andere woordkenmerken (zoals de uitspraak); bovendien onderging een deel van deze kenmerken tussentijds veranderingen. Omgekeerd werden de reeds beschikbare morfeemrepresentaties soms bij andere projecten ingezet, met als gevolg dat deze representaties "bevroren" moesten worden (wat betekent dat er tijdelijk geen aanpassingen in mochten optreden).

2) De lexeeminventarisatie van de MGBN is erg heterogeen, want hij omvat alle basislexemen die ten grondslag liggen aan één of meer woorden uit de LGBN. Dit leidt tot de volgende complicaties:

a) zowel zelfstandige als niet-zelfstandige lexemen

Onder de basislexemen zijn zowel zelfstandige eenheden als niet-zelfstandige eenheden (d.w.z. eenheden die alleen als constituent van een samenstelling voorkomen). De niet-zelfstandige lexemen zijn vaak in het bezit van een bindmorfeem, zoals -S en -EN. Maar doordat de MGBN pas in een laat stadium van informatie over de lexeemstatus (wel/niet zelfstandig) en de lexeemposities (linkerdeel/ middendeel/ rechterdeel) werd voorzien, was het bij de bewerking van de MGBN niet altijd duidelijk wat voor status een lexeem had (aangezien de eindsegmenten *-s* en *-en* behalve de functie van bindmorfeem ook andere functies kunnen aannemen, zoals meervoud of markering van een adverbale betekenis; en soms hebben ze helemaal geen functie). Dit leidde regelmatig tot segmentatieproblemen.

b) zowel gangbare als niet-gangbare lexemen

De LGBN omvat zoveel woorden dat een groot deel (minstens een derde deel) van de onderliggende basislexemen voor mij onbekend waren. Dit leidde soms tot een minder betrouwbare analyse. Maar bij segmenten waarvoor meerdere structuuranalyses mogelijk zijn, zoals *eling*, dat soms met EL+ING (cf. [KRAK]+EL+ING) correspondeert en soms met ELING (cf. [JONG]ELING), heb ik de lexemen vaak één voor één beoordeeld en waar nodig aanvullende betekenisinformatie opgezocht. Hierdoor is een groot deel van de ambigue segmenten toch van een betrouwbare structuur voorzien. Bij niet-ambigue patronen kon deze rigide controle uiteraard achterwege blijven, al bestaat er veel variatie in de mate van semantische transparantie, zoals eenvoudig kan worden aangetoond voor -AGE (wel transparant in *etalage* maar mogelijk niet in *etage*) of -ING (wel transparant in *draaiing*, *woning* maar niet in *haring*, *kling*). Bij de beoordeling van dergelijke patronen ben ik meestal van etymologische structuurcriteria uitgegaan.

c) zowel "Nederlandse" woorden als leenwoorden

De MGBN bevat zeker 1000 lexemen met een on-Nederlandse, niet-geassimileerde klankvorm, zoals *übermensch*, *etablissement*, *economy, peshmerga* en *perestrojka*. Volgens de standaardtheorie bezitten dergelijke leenwoorden geen morfeemstructuur. Maar wie enige kennis van de brontaal heeft, herkent de oorspronkelijke morfeemstructuur. Bij de opbouw van de MGBN ben ik er daarom van uitgegaan dat herkenbare segmenten altijd als morfeem moeten worden gemarkeerd. Zo heb ik *etablissement* als [ETABL]+ISS+EMENT geanalyseerd.

d) zowel gewone stammen als stammen met naamfunctie

De MGBN bevat tal van lexemen waarvan de stam met een persoonsnaam (bijv. in *Platoons* en *Aristoteliaans*) of een locatienaam (bijv. *Zwitserland* of *Australië*) correspondeert. Volgens de standaardtheorie bezitten namen geen morfeemstructuur, naamafleidingen wel. Maar bij alternanties van het type *België, Belgisch, Belg* lijkt de landsnaam *België* op de persoonsnaam *Belg* te zijn gebaseerd in plaats van andersom. Bij de analyse van dergelijke lexeemparadigma's ben ik altijd uitgegaan van de grootste stamvorm die door alle lexemen wordt gedeeld. Zo heeft het lexeem *België* de structuur BELG+I:E gekregen (en dus niet BELGI;E).

e) zowel basislexemen als pseudo-samenstellingen

De MGBN omvat ca. 5000 lexemen die in feite met versteende (of soms ook transparante) samenstellingen corresponderen, zoals *parlevinker* of *parelmoer*. In mijn optiek is het echter moeilijk om een principiële grens te trekken tussen transparante en niet-transprante samen-

stellingen; vaak lijken beide toepassingen mogelijk. De aanwezigheid van de pseudo-samenstellingen leidde echter wel tot complicaties voor de MGBN, want doordat deze woorden meerdere (pseudo-)stammen bezitten, hebben ze ook meerdere ingangen in de MGBN. Dit bemoeilijkt de bewerking van deze lexemen (doordat elke verandering in ingang 1 ook bij ingang 2 moest worden doorgevoerd).

3) Door het werk aan de MGBN werd duidelijk dat een deel van de LGBN-woorden ten onrechte als samenstelling is geanalyseerd. Dit heeft verschillende oorzaken:

a) sommige morfeemcombinaties zijn per ongeluk als samenstelling geanalyseerd

b) scheidbaar samengestelde werkwoorden zijn consequent als samenstellingen behandeld, terwijl het eigenlijk om gelede basislexemen gaat

c) bij de markering van samenstellingen is geen consequent onderscheid gemaakt tussen compositionele samenstellingen en niet-compositionele (c.q. versteende) samenstellingen

d) De LGBN bevat enkele woordgroepen die als ongedeeld lexeem zijn geclassificeerd, zoals *anorexia nervosa*. Hier is duidelijk sprake van een fout in de structuurrepresentatie.

5.6.5 Demonstratie

De door mij gehanteerde analysemethode wordt gedemonstreerd aan de hand van het woordenlijstje in tabel 5-1. Dit lijstje bevat 48 woorden met het beginsegment *re* die alfabetisch zijn gesorteerd en met informatie over de syntactische woordklasse zijn verrijkt (N = nomen; A = adjectief; V = verbum; O = overige).[205] Wie probeert om deze woorden van stamgrenzen te voorzien, zal direct ervaren dat het beginsegment *re* soms wel en soms niet als prefix kan worden opgevat. Zonder inzicht in de betekenisstructuur van de woorden is dit probleem echter moeilijk op te lossen. De analyse van deze lijst vormt dan ook een probleem voor een automatische parser. Maar ook een handmatige analyse is niet eenvoudig, want er zijn geen standaardcriteria voor de identificatie van morfologische structuur.

cat	trefwoord	cat	trefwoord	cat	trefwoord	cat	trefwoord
A	reçu	N	reactivering	V	reageren	A	realistisch
N	reçu	N	reactiviteit	N	reagrarisatie	N	realiteit
N	rea	N	reactor	N	real (2)	O	realiter
N	reaal	N	reader	N	realgar	N	reallocatie
N	reach	N	reading	N	realia	V	realloceren
N	reactant	O	ready	N	realisatie	N	realo
N	reactantie	N	readymade	N	realisator	N	realpolitiek
N	reactie	N	reaffectatie	A	realiseerbaar	N	realpolitik
A	reactief	V	reaffecteren	V	realiseren	N	realpolitiker
A	reactionair	N	reaganomics	N	realisering	N	reanimatie
N	reactionair	N	reageerder	N	realisme	V	reanimeren
V	reactiveren	N	reagens	N	realist	N	reanimist

Tabel 5-1: Lijst van 48 alfabetisch gesorteerde trefwoorden met het beginsegment re*, inclusief aanduiding van woordklasse (cat): N = nomen, V = verbum, A = adjectief, O = overig.*

Zoals ik reeds heb aangegeven berust de grammaticale (deductieve) morfologiebenadering op het uitgangspunt dat morfemen uitsluitend gebruikt worden voor de vorming van nieuwe woorden. In deze visie is de analyse van bestaande c.q. lexicale woorden alleen interessant vanuit etymologische overwegingen. Toch zijn er diverse parsers gebouwd die als taak hebben

[205] De markering (2) achter *real* geeft aan dat dit woord twee verschillende uitspraken heeft; deze uitspraken corresponderen met verschillende betekenissen.

om de bestaande woordenschat te analyseren; het gaat dan om de markering van *productieve* morfemen, d.w.z. morfemen waarmee op regelgestuurde wijze nieuwe woorden kunnen worden gevormd. Toegepast op de woorden in tabel 5-1 zou men bijvoorbeeld op de representaties in tabel 5-2 kunnen uitkomen.

cat	trefwoord	cat	trefwoord	cat	trefwoord	cat	trefwoord
A	[reçu]	N	re[activ]ering	V	[reag]eren	A	[real]istisch
N	[reçu]	N	[react]iviteit	N	re[agrar]isatie	N	[real]iteit
N	[rea]	N	[react]or	N	[real] (2)	O	[realiter]
N	[reaal]	N	[reader]	N	[realgar]	N	re[alloc]atie
N	[reach]	N	[reading]	N	[real]ia	V	re[alloc]eren
N	[react]ant	O	[ready]	N	[realis]atie	N	[real]o
N	[react]antie	N	[readymade]	N	[realis]ator	N	[real][polit]iek
N	[react]ie	N	re[affect]atie	A	[realis]eerbaar	N	[real][politik]
A	[react]ief	V	re[affect]eren	V	[realis]eren	N	[real][politik]er
A	[react]ionair	N	[reaganomic]s	N	[realis]ering	N	re[anim]atie
N	[react]ionair	N	[reag]eerder	N	[real]isme	V	re[anim]eren
V	re[activ]eren	N	[reagens]	N	[real]ist	N	re[anim]ist

Tabel 5-2: Morfologische woordanalyse op basis van "productieve" morfemen. Bij een productieve derivatie staat de stam tussen vierkante haken.

Deze tabel is het resultaat van een werkwijze waarbij de woordstam is gedefinieerd als het kleinste woordinterne deel dat een zelfstandige (conceptuele) betekenis draagt en dat een doorzichtige relatie onderhoudt met de betekenis van het hele woord. Woorden met het karakter van een samenstelling hebben twee stammen gekregen, bijvoorbeeld *realpolitiek*. Bij de begrenzing van de stammen is het uitheemse segment *re* (dat potentieel met het prefix RE- correspondeert) doorgaans als deel van de stam opgevat; bij woorden als *reageren* lijkt de betekenis "antwoorden" namelijk gekoppeld te zijn aan de stam REAG, en niet rechtstreeks afleidbaar te zijn uit de betekenis van RE- en de wortel AG (die men ook aantreft in woorden als *ageren* en *agent*). Hetzelfde geldt voor de variant REACT, die voorkomt in *reactie*. Bij *reactiveren* daarentegen is duidelijk sprake van een afleiding op basis van het adjectief *actief*: hoewel dit woord in de verte weer gerelateerd is aan de wortel ACT, lijkt dit niet relevant voor de afleiding. Daarom is gekozen voor de structuur RE[ACTIV]EREN.

Deze redenering is ook gehanteerd bij het besluit om *reagrarisatie* en *reanimeren* af te leiden van de stammen AGRAR en ANIM. Er is slechts een gradueel verschil met de stam REAG/REACT. Dat geldt niet voor het besluit om *realiseren* terug te voeren op de stam REALIS, want er is geen enkele semantische of etymologische aanwijzing dat *realis* gebaseerd is op een wortel AL. Vergelijkbare vragen bestaan ten aanzien van de status van de suffixen -IS en -IV: soms zijn deze bij de stam getrokken en soms zijn ze deel van een suffixcluster. Bij het contrast tussen *reactiveren* en *reactiviteit* hangt dit bijvoorbeeld samen met de aanname dat het woord *actief* niet verder analyseerbaar is, in tegenstelling tot de stam REACTIV in *reactiviteit*, die uiteenvalt in een stam REACT en een suffix -IEF.

In de grammaticale benadering worden deze vragen meestal genegeerd: het enige wat telt, is of Nederlanders in staat zijn om woorden als *reagens* en *realia* zelf te construeren op basis van kleinere eenheden, of dat ze deze woorden integraal in hun lexicon opslaan. De discussie draait dus om de vraag waar de grens ligt tussen "productieve" en "niet-productieve" morfemen. In mijn visie is het onderscheid tussen productieve en niet-productieve morfemen echter nogal kunstmatig; ik denk namelijk dat van geen enkel niet-bestaand woord kan worden voorspeld of het "gemunt" zal worden, en of het dan een regelmatige betekenis zal dragen. Ook zeer weinig gebruikte affixen kunnen plotseling in nieuwe woorden opduiken,

zonder dat de spreker hoeft uit te leggen waarom hij juist deze vorm kiest. Dit is bijvoorbeeld het geval met woorden als *Reagonomics* (van *Reagan* en *economics*) en *euroforie* (van *euro* en *euforie*). Anderzijds hebben taalgebruikers zeer veel intuïties over de interne structuur van bestaande woorden. Wie de bekende natuurkundige wet *actie is min reactie* kent, zal bijvoorbeeld nooit meer vergeten dat *reactie* in feite de structuur RE+ACTIE heeft. En wie weet wat *import* betekent, kan waarschijnlijk ook wel bedenken wat met *export* wordt bedoeld. Hierbij doet het er niet toe doet of de onderscheiden morfemen productief zijn.

De morfologische transparantie van bestaande woorden kan wel beïnvloed worden door persoonsgebonden factoren als opvoeding, taalgevoel, opleiding en belezenheid. Bovendien kunnen mensen hun inzicht in de woordstructuur vergroten door erop te studeren. Het lijkt me daarom onjuist om het niet-productieve gedeelte van de woordvorming zomaar te negeren, bijvoorbeeld op grond van het argument dat deze kennis pas op school wordt opgedaan. Taalgebruikers communiceren nu eenmaal op verschillende niveaus van complexiteit, en dit moet te maken hebben met verschillen in de cognitieve representatie van het taalsysteem. Een compleet lexiconmodel zal dan ook de mogelijkheid moeten bieden om deze kennis systematisch te coderen. Dit is mogelijk indien men de functie van morfologische structuur niet beperkt tot de aanmaak van nieuwe woorden, maar een centrale rol laat spelen bij de lexicale opslag van deze woorden. In mijn visie kan elk door meerdere woorden gedeeld segment met een constant effect op betekenis en/of de grammaticale eigenschappen van de hiermee geconstrueerde woorden als een morfeem worden aangemerkt. Hierbij kan het ook om bouwstenen gaan die alleen herkend kunnen worden op basis van kennis van andere talen en/of oudere taalstadia. Deze structuur kan worden achterhaald door per vormeigenschap na te gaan of woorden analoog gedrag vertonen; in dat geval kan het betreffende segment als morfeem worden gecodeerd.

Bij de opbouw van de MGBN heb ik ernaar gestreefd om de kleinst mogelijke segmenten op te sporen die in enig stadium van het Nederlands (of een leentaal) als morfeem hebben gediend; de MGBN weerspiegelt dus niet alleen de hedendaagse morfeemstructuur, maar ook de historische (c.q. etymologische) morfeemstructuur. Of deze structuur ook op compositionele wijze bijdraagt aan bepaalde woordfuncties is een vraag die pas in een volgend stadium relevant wordt. Tabel 5-3 toont het resultaat van de door mij gehanteerde annotatiemethode.

cat	trefwoord	cat	trefwoord	cat	trefwoord	cat	trefwoord
A	re[çu]	N	re[act]iv;er;ing	V	re[ag]er;en	A	[real]ist;isch
N	re[çu]	N	re[act]iv;iteit	N	re[agrar]is;at;ie	N	[real]iteit
N	[rea]	N	re[act]or	N	[real] (2)	O	[real]iter
N	[reaal]	N	[read;er]	N	[realgar]	N	re;al[loc]at;ie
N	[reach]	N	[read]ing	N	[real]ia	V	re;al[loc]er;en
N	re[act]ant	O	[ready]	N	[real]is;at;ie	N	[real]o
N	re[act]ant;ie	N	[ready][made]	N	[real]is;at;or	N	[real][polit]iek
N	re[act]ie	N	re;af[fect]at;ie	A	[real]is;eer;baar	N	[real][polit]ik
A	re[act]ief	V	re;af[fect]er;en	V	[real]is;er;en	N	[real][polit]ik;er
A	re[act]ion;air	N	[reagan]om;ic;s	N	[real]is;er;ing	N	re[anim]at;ie
N	re[act]ion;air	N	re[ag]eer;der	N	[real]isme	V	re[anim]er;en
V	re[act]iv;er;en	N	re[ag]ens	N	[real]ist	N	re[anim]ist

Tabel 5-3: Morfologische woordanalyse op basis van etymologische morfemen. Hiertoe is de kleinste etymologische stam van vierkante haken ([,]) voorzien, terwijl opeenvolgende affixen door puntkomma's (;) worden gescheiden.

In deze lijst zijn alle woorden waarvan de etymologische structuur een uitheems prefix bevat, opgesplitst in een prefix en een wortel. Zo komt de stam REACT oorspronkelijk van het prefix RE- en de wortel ACT; daarom heb ik deze stam steeds opgesplitst, net als de verwante stam REAG in *reageren*. Het afsplitsen van suffixclusters berust op soortgelijke overwegingen; zo zijn *reactief*, *reactionair* en *reactor* etymologisch gezien duidelijk verwant; daarom moeten -IEF, -IONAIR en -OR als suffixen of suffixclusters worden opgevat. Suffixclusters als -IONAIR heb ik bovendien nog onderverdeeld in de morfemen -ION en -AIR. Bij dit soort beslissingen heb ik me meestal door vormanalogie laten leiden. Bij twijfelgevallen heb ik een naslagwerk geraadpleegd, bijvoorbeeld om na te gaan of *realiseren* ooit een prefix RE- heeft gekend. De hier gepresenteerde tabel geeft al met al een redelijk beeld van de wijze waarop de MGBN van structuur is voorzien. Deze onconventionele methode maakt het mogelijk om empirisch onderzoek te doen naar de combinatorische eigenschappen van potentiële morfemen.

5.7 De gerealiseerde gegevensbank

5.7.1 De veldstructuur van de MGBN-lemma's

Tabel 5.4 toont alle datavelden uit de MGBN, en specificeert de inhoud van elk veld aan de hand van twee morfologisch complexe lexemen, te weten het inheemse *gedachte* en het uitheemse *gradueel*. Tabel 5.5 geeft per veld een korte toelichting op de inhoud. Deze informatie vormt de basis voor de constructie van het MGBN-model. De voorbeeldlemma's zijn zo gekozen dat ze inzicht geven in de differentiatiemogelijkheden van de n1vorm, de n2vorm en de n3vorm. In voorbeeld 1 worden slechts twee van de drie opties benut, want de n1vorm valt hier samen met de n2vorm. In voorbeeld 2 zijn er op alle structuurniveaus verschillen (zowel voor de wortel als voor de affixen).

veldnaam	**voorbeeld 1**	**voorbeeld 2**
01. <lemma>	gedachte	gradueel
02. <semkey>	:1	:1
03. <syntcat>	N	A
04. <subtyp>	+Z,-,-	-,-,-
05. <n3stam>	denk.1	gred.1
06. <n2stam>	dacht	grAd
07. <n1stam>	dacht	grad
08. <n2stam0>	(dacht)	(grAd)
09. <n3vorm>	s: *ge(i)_[denk.1]_*e(i)	s: [gred.1]_*{aa,ee}l(u)
10. <n2vorm>	o: ge_[dacht]_e	o: [grAd]_eel
11. <n1vorm>	a: ge;[dacht];e	a: [grad];u:eel
12. <n2vorm0>	(*ge_[dacht]_*e)	([grAd]_*El)
13. <sylvorm>	ge@1dach=te	gra=du=eel
14. <fonvorm>	*[g @ - d A x - t @|010:]	*[g r A1 - d y - w0 e l|201:]
15. <finsuf>	e:N | *e(i)	eel:A | *{aa,ee}l(u)
16. <cmphist>	|-|	|-|
17. <wdinfo>	[+auto][+lp][-mp][+rp][+dep]	[+auto][-lp][-mp][-rp][-dep]
18. <wdfreq>	40;0;20	0;0;0
19. <wnnstat>	+nn	+nn
20. <inipiek>	-ini	-ini
21. <wnncat>	12	20
22. <taallab>	-	-
23. <tokfreq>	12465	15
24. <stamnum>	1	1
25. <synt1>	1	2
26. <synt0>	1	2
27. <avorm0>	ge[dacht]e	[grad]u:eel
28. <comm>	-	-

Tabel 5.4: De veldstructuur van de MGBN

veldnaam	omschrijving
01. <lemma>	spelvorm van lemma (c.q. basislexeem)
02. <semkey>	semantische key (c.q. betekenisindex); default = 1
03. <syntcat>	syntactische categorie
04. <subtyp>	subtype: [±M]=Mens of [±Z]=Zaak, [±N]=Naam, [±L]=Leenwoord
05. <n3stam>	niveau-3-stam: deze generaliseert over klankvarianten
06. <n2stam>	niveau-2-stam: deze generaliseert over spellingsvarianten
07. <n1stam>	niveau-1-stam: deze correspondeert met de spelvorm
08. <n2stam0>	oude niveau-2-stam
09. <n3vorm>	niveau-3-vorm van morfeemrepresentatie lemma (cf. veld 05)
10. <n2vorm>	niveau-2-vorm van morfeemrepresentatie lemma (cf. veld 06)
11. <n1vorm>	niveau-1-vorm van morfeemrepresentatie lemma (cf. veld 07)
12. <n2vorm0>	oude niveau-2-vorm (cf. veld 08)
13. <sylvorm>	syllabevorm c.q. afbreekrepresentatie lemma
14. <fonvorm>	fonologische vorm c.q. klankrepresentatie lemma
15. <finsuf>	n1,n2 en n3-vorm van het finale suffix (evt. '-')
16. <cmphist>	computationele historie: constructiegeschiedenis
17. <wdinfo>	woorddeel-informatie: lp = links, mp = midden, rp = rechts
18. <wdfreq>	woorddeel-frequentie van lexeem in positie lp, mp en rp
19. <wnnstat>	wel/niet opgenomen in nn-woordenboek (= wdb hedendaags ned.)
20. <inipiek>	wel/geen initiële stress-piek (> wel/niet scheidbaar prefix)
21. <wnncat>	categorie in nn-woordenboek (indien van toepassing)
22. <taallab>	taallabel (bij leenwoorden)
23. <tokfreq>	tokenfrequentie (in corpus)
24. <stamnum>	aantal stammen (c.q. sublexemen) binnen lemma
25. <synt1>	nummer van nieuwe syntactische categorie
26. <synt0>	nummer van oorspronkelijke syntactische categorie
27. <avorm0>	oude niveau-1-vorm van morfologische representatie
28. <comm>	commentaar

Tabel 5.5: Toelichting bij de datavelden in de MGBN

De inhoud van het sublexicon MB1

De lexeeminventarisatie van de MB1 omvat alle basislexemen uit de LGBN, d.w.z. alle lexeemconstitenten die zelfstandig bruikbaar zijn of die onderdeel zijn van de afbreekvorm van een samengesteld lexeem. Een deel van deze basislexemen kent twee gedaantes, namelijk met en zonder bindmorfeem. Er zitten ook ca. 6000 pseudo-samenstellingen tussen. Voor alle lexemen uit de MB1 is informatie over de morfeemstructuur beschikbaar; verder wordt informatie gegeven over kenmerken als de afbreekvorm, de uitspraak, de inflectiecategorie en de woorddeelpositie (indien mogelijk).

De inhoud van het sublexicon MB2

De MB2 bevat uitsluitend basislexemen die als zelfstandig woord kunnen worden gebruikt. Deze lexeeminventarisatie bestaat deels uit MB1-lexemen en deels uit nieuw geconstrueerde lexemen. De LGBN bevat namelijk tal van lexemen (ca. 15.000) die in mijn optiek ten onrechte als samenstelling zijn geanalyseerd. Hierbij gaat het in de eerste plaats om scheidbaar samengestelde werkwoorden (die hierdoor los zijn komen te staan van de nominale en adjectivale toepassingen van deze V-stammen), maar er zitten ook "samenstellingen" tussen waarvan het linkerdeel of rechterdeel duidelijk met een affix correspondeert, wat impliceert dat het niet om een samenstelling maar om een derivatie gaat. Deze geheranalyseerde lexemen zijn allemaal in de MB2 opgenomen. Bij de aanmaak van de MB2 heb ik de nieuwe basislexemen zoveel mogelijk langs automatische weg van morfologische structuur voorzien door gebruik te maken van reeds beschikbare structuurinformatie over de lexeemstam. Hierdoor kon een groot deel van de nieuwe lexemen tamelijk snel van morfologische structuurinformatie worden voorzien; het restant is vervolgens langs redactionele weg bewerkt.

6 Constructie, analyse en evaluatie van een L-KRING-model van de MGBN

6.1 Introductie

6.1.1 Doelstelling

Dit hoofdstuk bespreekt opzet, inhoud en kwaliteit van een reeks datarapporten die inzicht geven in de morfologische samenstelling van de Morfologische Gegevensbank van het Nederlands (MGBN). Meer specifiek geven deze rapporten kwalitatieve en kwantitatieve informatie over alle in de MGBN aangetroffen affixen en hun combinatorische eigenschappen (beperkt tot de syntagmatische dimensie). Ik heb deze rapporten vervaardigd door de geanalyseerde datadomeinen langs computationele weg in een op L-KRING-principes gebaseerd deellexicon c.q. MGBN-model om te zetten. Dit MGBN-model verschilt van de MGBN doordat het hiërarchisch is gestructureerd en een catgeoriale typering geeft van het inwaartse en uitwaartse selectiedomein van de affixen. De aan dit MGBN-model ontleende structuurinformatie leent zich goed voor een vergelijking met de reeds bestaande kennisbronnen, zoals het Morfologisch Handboek. Zo kan worden achterhaald of het MGBN-model betrouwbare informatie biedt over de morfologische representaties in het mentale lexicon.

Zoals ik in hoofdstuk 5 uiteen heb gezet, is het morfologische gegevensbestand dat ten grondslag ligt aan het MGBN-model het resultaat van een structureringsmethode waarbij een groot deel van de Nederlandse woordenschat langs inductieve weg van morfologische structuur is voorzien. Meer specifiek geldt dat deze gegevensbank indirect (namelijk via de samenstellende delen) alle woorden uit de Grote Van Dale (editie 1999) dekt en dat de hieraan toegekende structuurrepresentaties de mogelijkheid bieden om een nagenoeg complete inventarisatie op te bouwen van de (potentiële) orthografische morfemen van het Nederlands, hun combinatorische eigenschappen en enkele van de hiermee verbonden woordkenmerken (zoals de morfologische klasse en de inflectiecategorie). Hierbij moet echter wel de kanttekening worden geplaatst dat de huidige gegevensbank niet meer is dan een tussenstadium in een semi-automatisch ontwikkelingstraject dat uiteindelijk een integraal model van het Nederlandse lexicon moet opleveren. Inmiddels heeft dit proces een stadium bereikt waarin de datastructuur zo consistent is geworden dat de MGBN morfologisch onderzoek naar het Nederlands kan ondersteunen en kan bijdragen aan de systematisering van de woordkenmerken in VDL's lexicografische gegevensbank.

Bij de evaluatie van de MGBN heb ik me niet beperkt tot een vergelijking met het Morfologisch Handboek. In aanvulling op deze externe evaluatiemethode heb ik namelijk ook een interne evaluatiemethode beproefd. Hiertoe heb ik onderzoek gedaan naar de distributieverdeling in de data; deze maakt het mogelijk om na te gaan of de MGBN patroonklassen bevat die relatief onder- of oververtegenwoordigd zijn. Dit kan aanleiding zijn voor een nadere inspectie van deze patronen. Na aanpassing van deze patronen kunnen weer meer verfijnde structuurcriteria worden achterhaald, zodat een nieuwe evaluatieronde mogelijk wordt, en dit proces kan net zolang doorgaan totdat de gegevensbank met een welhaast volmaakte patrooninventarisatie correspondeert. De in dit hoofdstuk besproken datarapporten berusten dan ook op een expliciet hypothetisch lexiconmodel, en de hiermee opgebouwde inventarisatie van morfologische patronen kent dus eveneens een expliciet hypothetische status.

6.1.2 Analysevragen

Het MGBN-model leent zich zowel voor syntagmatische als voor paradigmatische structuuranalyses. In dit hoofdstuk zal ik me echter beperken tot de bespreking van datarapporten met syntagmatische analyses van de affixdimensie van het MGBN-model. Op deze manier wil ik een indruk geven van de analysemogelijkheden van de MGBN, en meer specifiek van de door mij gehanteerde analysemethode en de eigenschappen van de hieruit voortgekomen datarapporten. De in dit hoofdstuk besproken datarapporten richten zich op de volgende thema's:

a) de inventarisatie van wortels en prefixstammen
b) de inventarisatie en evaluatie van prefixen en hun combinatorische eigenschappen
c) de inventarisatie en evaluatie van suffixen en hun combinatorische eigenschappen
d) de inventarisatie en evaluatie van prefix-suffix-combinaties

Meer specifiek komen de volgende vragen aan de orde:

i. welke wortels zijn er? welke prefix-sequenties kunnen met deze wortels samengaan? hoeveel lexeemtoepassingen bezitten deze stammen?
ii. welke prefixen zijn er? op welke posities komen ze voor? welke prefix-combinaties kunnen ze aangaan?
iii. welke suffixen zijn er? op welke posities komen ze voor? welke suffix-combinaties kunnen ze aangaan? welke inwaartse en uitwaartse inflectiecategorieën selecteren ze?
iv. welke prefix-suffix-combinaties bestaan er? wat is het categoriale effect van de prefixen? hoe groot is hun stamdomein?
v. wat voor eigenschappen gelden voor sequenties met een begin- of eindvariabele?

In het kader van de evaluatie zal ik voor alle klassen van datarapporten nagaan hoe de hierin verzamelde structuurkenmerken zich tot de morfologische kennis in het Morfologisch Handboek van het Nederlands (MHB) verhouden. In aanvulling op deze externe evaluatiemethode, die alleen antwoord kan geven op de vraag in hoeverre de MGBN-patronen reeds in de vakliteratuur zijn beschreven, zal ik ook een interne evaluatiemethode proberen te ontwikkelen, d.w.z. een methode die inzicht geeft in de interne consistentie van de data. Om deze te beoordelen is kennis nodig over de onderliggende structuurcriteria (d.w.z. mijn onbewust gehanteerde opdelingscriteria). Deze kunnen worden achterhaald door op zoek te gaan naar algemene dataverbanden. Gegeven deze verbanden kan men nagaan in hoeverre het gedrag van afzonderlijke affixen hiermee spoort: indien sprake is van sterk afwijkend gedrag, is dit een aanwijzing dat het toepassingsdomein van het betreffende affix te ruim of te krap is en dus moet worden bijgesteld.

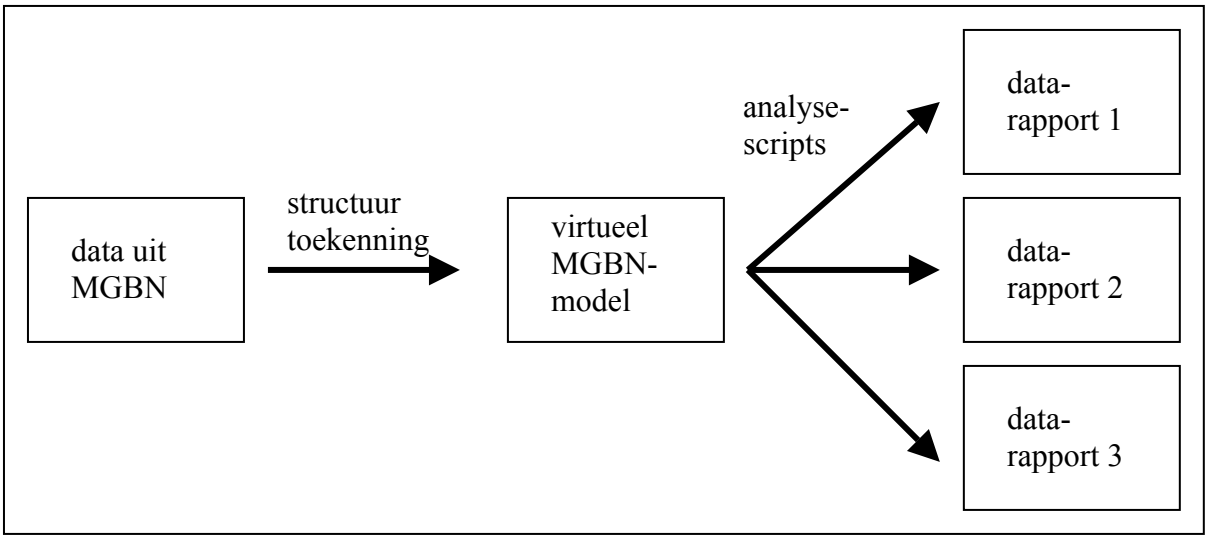
6.1.3 Indeling

Dit hoofdstuk kent de volgende indeling. In H6.2 wordt de onderzoeksmethode besproken. Hierbij zet ik in afzonderlijke secties uiteen welke beginselen ik heb gehanteerd bij de constructie, de analyse en de evaluatie van het MGBN-model. In H6.3 presenteer ik een op de L-KRING-theorie gebaseerde beschrijving van het analysedomein, waarbij zowel kwalitatieve als kwantitatieve basiskenmerken aan de orde komen. In de hierop volgende secties worden een aantal deelonderzoeken besproken die als doel hebben om antwoord te geven op de centrale analysevragen. Het gaat om de inventarisatie van wortels en kern-affix-combinaties (H6.4), de inventarisatie van prefixen en hun combinatoriek (H6.5), de inventarisatie van suffixen en hun combinatoriek (H6.6) en de inventarisatie van prefix-suffix-combinaties (H6.7). Het hoofdstuk wordt afgesloten met een conclusie (H6.8).

6.2 Methode

6.2.1 Introductie

De in dit hoofdstuk te bespreken analyserapporten berusten op een analysemethode waarbij de structuurinformatie in de Morfologische Gegevensbank niet rechtstreeks wordt geanalyseerd, maar op basis van een virtueel MGBN-model, namelijk een op L-KRING-principes gebaseerd (deel)model van het Nederlandse lexicon (zie het schema in figuur 6-1).



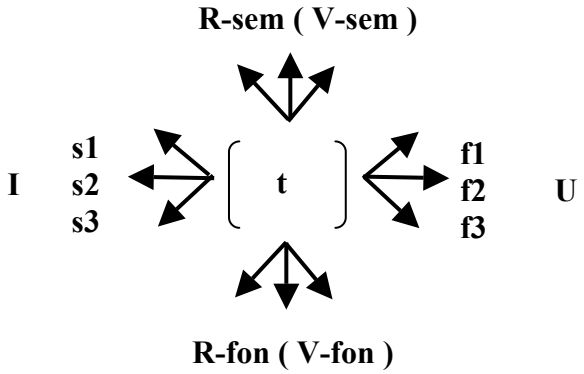
Figuur 6-1: De conceptuele basis van de in hoofdstuk 6 besproken analyserapporten.

Dit lexiconmodel is dus niet "fysiek" geïmplementeerd, maar bestaat alleen in de scripts waarmee de analyserapporten worden geproduceerd (en in de beschrijving van deze analyses). Bovendien construeren deze scripts slechts een klein deel van het virtuele MGBN-model, namelijk alleen die onderdelen die nodig zijn om de benodigde query te kunnen uitvoeren. Deze kunstgreep maakt het mogelijk om vooruit te lopen op een stadium waarin het virtuele MGBN-model daadwerkelijk in een kennissysteem is geïmplementeerd en hier onderzoek mee te doen naar de morfologische patronen die zo'n systeem kan ontsluiten. Dit heeft als bijkomend voordeel dat de langs deze weg tot stand gekomen analyserapporten duidelijk laten zien wat het door mij beoogde kennissysteem moet kunnen en waarom het nuttig is om een project op te zetten dat als doel heeft om zo'n systeem te realiseren. De meerwaarde van het beoogde kennissysteem schuilt in het feit dat het niet alleen statische datarapporten kan produceren (zoals de rapporten die centraal staan in dit hoofdstuk), maar dat het ook dynamisch te doorzoeken is. Zo stel ik me voor dat dit systeem de gebruiker in staat stelt om eenvoudig van de ene naar de andere representatie te switchen, bijvoorbeeld van woordvorm naar morfologische representatie en vervolgens van de hierin aangetroffen woordstam naar alle woorden waarin deze stam voorkomt; bovendien moet men voor elke eenheid (van morfeem tot lexeemcombinatie) aanvullende kenmerken kunnen opvragen (en desgewenst aanpassen of aanvullen), waaronder uitspraak, betekenis en frequentiegegevens.

6.2.2 De constructie van het MGBN-model

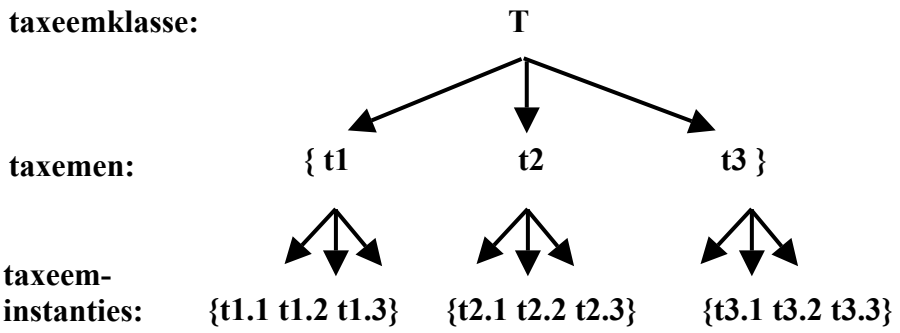
Bij de constructie van het MGBN-model heb ik me laten leiden door de principes van de L-KRING-theorie. Zoals in hoofdstuk 4 uiteen is gezet, stelt deze theorie dat het mentale lexicon met een netwerk van indexen correspondeert, waarbij elke index met een lexicale kenniseenheid correspondeert; in het morfologische domein kan het bijvoorbeeld om een morfeem gaan, maar ook om een morfeemklasse of een morfeemsequentie, zoals een affixcombinatie of een combinatie van een wortel, een prefix en een suffix. Indien sprake is van

een morfologische kenniseenheid (waarbij per definitie sprake is van een lexicale relatie tussen vorm en functie) spreek ik bij voorkeur van een taxeem, dit ter onderscheiding van eenheden die uitsluitend een fonologische of een semantische functie hebben. In het hier bedoelde netwerkmodel kan elk taxeem tal van relaties met andere taxemen onderhouden. Hierbij kan men onderscheid maken tussen inwaartse en uitwaartse compositierelaties (die de basis vormen voor de opbouw van mofotactische representaties), zijwaartse projectierelaties (die toegang geven tot het morfofonologische (R-fon) en het morfosemantische (R-sem) representatiedomein) en classificatierelaties (die de overerving van kenmerken beregelen).



Figuur 6-2: Domeinschema voor een taxeemindex t: het toont de compositionele dimensie (met domeinen I en U) en de projectiedimensie (met domeinen R-fon en R-sem).

Het partiële domeinschema in figuur 6-2 toont de relatie tussen een taxeem t en twee van de hier genoemde taxeemdimensies, te weten de compositiedimensie en de projectiedimensie. Hierbij correspondeert de compositiedimensie met de horizontale as, bestaande uit een inwaarts domein I (met de stammen s1, s2 en s3) en een uitwaarts domein U (met de functors f1, f2 en f3), en de projectiedimensie met de verticale as. De derde structuurdimensie, te weten de classificatiedimensie, staat loodrecht op de twee andere dimensies; hoewel deze niet is weergegeven in figuur 6-3 kan men zich hier een voorstelling van maken door het schema als een bovenaanzicht te interpreteren: het toont dus één van de dwarsvlakken van de suppositie-dimensie, te weten het t-vlak. Indien er meerdere t's bestaan (bijv. een stel suffixen die dezelfde stam kunnen selecteren), kunnen deze t's als subtypes van een klasse T worden opgevat. Elke klasse T bestaat namelijk uit taxemen die tot dezelfde functieklasse behoren (op basis van hun morfosyntactische en/of semantische kenmerken) en/of een vergelijkbaar U-domein bezitten; de t's onder T kunnen zelf weer als een klasse-index fungeren ten opzichte van de lexicaal opgeslagen instanties van dit taxeem (bijv. t2.1, t2.2 etc.), d.w.z. indexen die naar concreet waargenomen toepassingen verwijzen (inclusief gedetailleerde informatie over hun uitspraak en betekenis). Deze relaties zijn weergegeven in figuur 6-3.

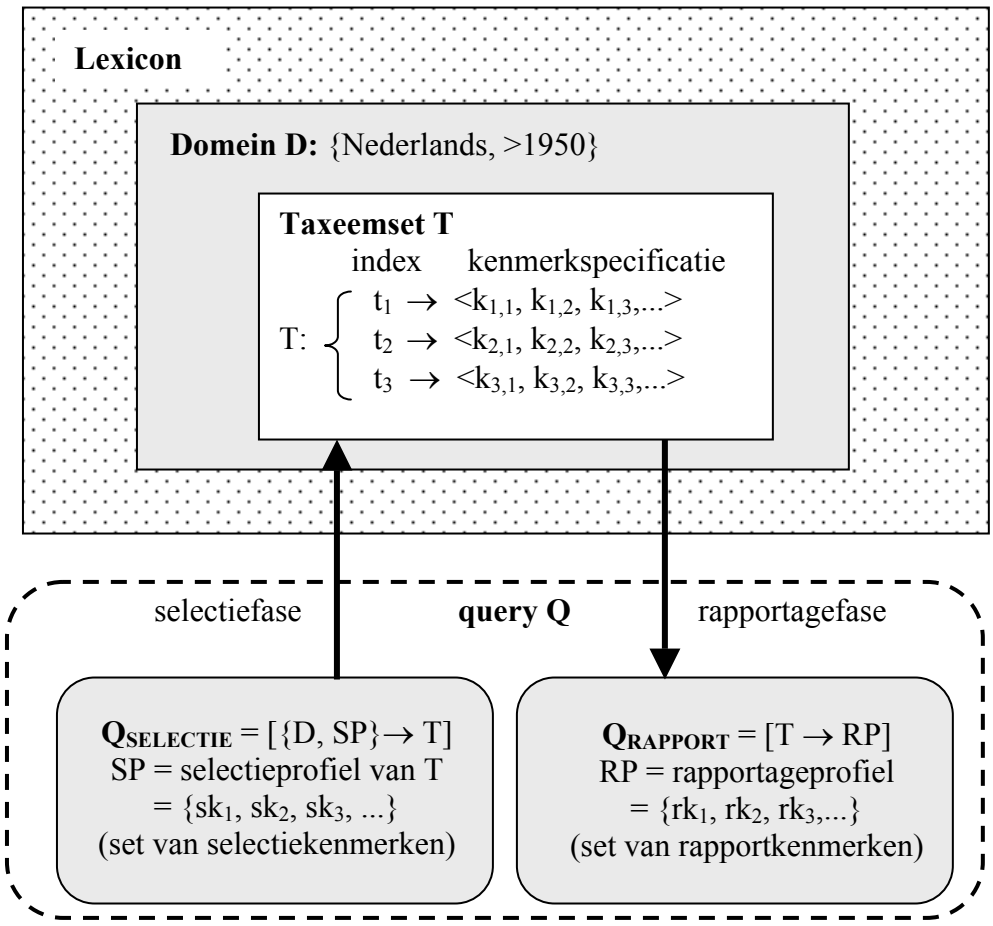


Figuur 6-3: Domeinschema voor de taxemen t1, t2 en t3; het toont zowel het opwaartse classificatiedomein (voor taxeemklassen) als het neerwaartse (voor taxeeminstanties).

6.2.3 De analyse van het MGBN-model

6.2.3.1 De structuur van een query

Alle in dit hoofdstuk beschreven MGBN-analyses berusten op een speciaal voor deze analysevraag geschreven computerscript (of reeks van scripts). Bij de beschrijving van de MGBN-analyses zal ik zoveel mogelijk van de feitelijke aanpak abstraheren door deze te beschrijven alsof mijn analyses met het zoeksysteem uit figuur 6-4 zijn uitgevoerd.[206]



Figuur 6-4: Het zoek- en rapportagesysteem van het L-KRING-lexicon.

In dit systeem is de informatie in het L-KRING-lexicon L (bijvoorbeeld het lexicon van mijn MGBN-model) uitsluitend toegankelijk via queries met een selectie-opdracht (Q-selectie) en een rapportage-opdracht (Q-rapportage). Hierbij bestaat elke selectie-opdracht uit twee componenten, namelijk een domeinspecificatie (D), bijvoorbeeld het Nederlands van na 1950, en een selectieprofiel (SP), dat uit een of meer selectiekenmerken (*sk*'s) dient te bestaan. In het geval van een morfotactische zoekopdracht zal één selectiekenmerk met de kenniseenheid "taxeem" moeten corresponderen. Indien L met een ideaal lexicon correspondeert zal de combinatie van D en SP altijd hetzelfde resultaat opleveren,[207] namelijk een nader te specificeren taxeemset T (waarbij elk taxeem met een aparte index correspondeert). Deze T vormt tevens de basis voor de definitie van een rapportageprofiel (RP) met één of meer rapportagekenmerken (*rk*'s). Dit profiel geeft aan hoe elk taxeem in taxeemset T in het datarapport moet worden opgeleverd (aangezien een taxeem niets anders is dan een index).

[206] Dit systeem werd in hoofdstuk 5 geïntroduceerd.
[207] Bij een mentaal lexicon is dit niet gegarandeerd; de performance is hier contextafhankelijk.

6.2.3.2 Demonstratie van een query

Ik zal een en ander illustreren aan de hand van een voorbeeld. Stel dat de gebruiker infomatie wil opvragen over alle hedendaagse stammen die in staat zijn om zowel het N-vormende suffix -AGE als het V-vormende suffix -EER te selecteren, met als aanvullende eis dat het om stammen op prefix-niveau dient te gaan (= [+P]-stammen), dus om stammen die naast de wortel ook één of meer prefixen kunnen omvatten. Dan zou de aan de Query verbonden selectie-opdracht (Q-selectie) als volgt kunnen worden gedefinieerd:

Q_1-Selectie = [$<D_1,SP_1> \rightarrow T_1$]
-D = Domeinkenmerken. Voorbeeld:
D_1 = {hedendaags Nederlands, basislexemen}
-SP_1 = Selectie Profiel. Voorbeeld:
{ X : X = taxeem, X = [+P]-stam, $[X + \text{AGE}]_N \in L|_{\text{LEXEEM}}$, $[X + \text{EER}]_V \in L|_{\text{LEXEEM}}$ }

Hierbij correspondeert de uitvoer met een taxeemset (T_1), d.w.z. een verzameling taxeem-ndexen (c.q. netwerklocaties) die aan de eisen van verzameling X voldoet. Er is een aparte rapportage-opdracht (Q-rapportage) nodig om aan te geven wat er over deze taxemen moet worden gerapporteerd. Stel dat de gebruiker de volgende rapportage-opdracht geeft:

Q_1-Rapportage = [$T_1 \rightarrow <RP_1, I_S, I_O>$]
-RP = Rapportage Profiel (nr. 1). Voorbeeld:
RP_1 = { nvorm(T-S), nvorm(T-W), spelvorm($[\text{T-S} + \text{AGE}]_N$), spelvorm($[\text{T-S} + \text{EER}]_N$) }
-I_S = Sorteer Instructies. Voorbeeld:
I_S = { Kenmerk 1: nvorm(T-W), S-alfa], [Kenmerk 2: INV(nvorm(T-S)), S-alfa]}
-I_O = Opmaak Instructies. Voorbeeld:
I_O = {veldscheiding met tabs: [specificatie tabposities, specificatie uitlijning] }

Dan dient een query-rapport te worden opgeleverd waarin voor elke stam uit T_1 wordt aangegeven wat zijn morfologische structuur is, wat zijn wortel is, en wat de spelvorm is van de lexemen met -AGE en -EER. Hierbij dienen de stammen eerst alfabetisch (= S-alfa) op de normaalvorm (c.q. citatievorm) van de taxeemwortel (nvorm T-W) te worden gesorteerd en dan alfabetisch (= S-alfa) op de inverse-morfeem-weergave van de normaalvorm van de stam (INV(nvorm T-S)). Tabel 6-1 laat zien wat deze query zou kunnen opleveren indien deze op het MGBN-model wordt losgelaten:

index T	**nvorm bij T-W**	**nvorm bij INV(T-S)**	**spelvorm bij $[\text{T-S} + \text{age}]_N$**	**spelvorm bij $[\text{T-S} + \text{eer}]_V$**
t3	cycl	re_[cycl]	recyclage	recycleren
t5	mas	[mas]	massage	masseren
t7	nom	re_[nom]	renommage	renommeren
t4	pas	[pas]	passage	passeren
t2	patr	[patr]_on	patronage	patroneren
t8	port	con_[port]	colportage	colporteren
t1	port	re_[port]	reportage	reporteren
t6	sabot	[sabot]	sabotage	saboteren

Tabel 6-1: Fictief rapport bij een mogelijke query op de MGBN.

Hierbij correspondeert de eerste kolom (T) met de taxeem-indexen die het resultaat zijn van de selectie-opdracht, namelijk een verzameling van 8 stammen (waarvan er twee dezelfde wortel bezitten, te weten de kern PORT). Deze indexverzameling heeft echter een zeer abstract karakter, want de taxeemindexen hebben geen eigen vorm of betekenis, maar ze leggen uitsluitend een verband tussen elders gedefinieerde taxeemkenmerken. Daarom is de rapportage-stap onmisbaar. Hierbij wordt voor elke taxeemindex geïnventariseerd wat voor gegevens er met de rapportagekenmerken corresponderen, wat een rapportageprofiel c.q. RP-lijst

oplevert; vervolgens worden alle RP-lijsten gesorteerd volgens de ordeningseisen, waarna ze conform de opmaakeisen in het rapport worden opgenomen. Tot slot wordt het bestand aan de gebruiker opgeleverd, die het rapport kan raadplegen door het in een editor te openen. Zoals ik in hoofdstuk 1 en 4 heb betoogd, kan deze computationele procedure ook als model dienen voor de procedures die ten grondslag liggen aan de raadpleging van het mentale lexicon. Indien een taalgebruiker bijvoorbeeld naar een andere taalgebruiker luistert, fungeren de binnenkomende taaluitingen (klankreeksen) als selectie-opdracht voor het zoeken naar de lexicale eenheden die inzicht kunnen geven in de betekenis van de waargenomen woorden; omgekeerd kan een boodschap worden onderverdeeld in concepten op woordniveau, waarna het zoeksysteem de bijbehorende indexen kan zoeken om vervolgens hun klankvorm te activeren (waarbij de best passende vorm moet worden gekozen). Het hier gedefinieerde zoeksysteem is dan ook heel geschikt om de MGBN-queries zo te beschrijven dat ze ook als mentale queries kunnen worden geïnterpreteerd.

6.2.3.3 Overzicht van kennisdimensies

Hieronder volgt een overzicht van de belangrijkste kennisdimensies van de datarapporten die informatie geven over de samenstelling van het MGBN-model. Al deze kenmerken tezamen definiëren de potentiële queryruimte van het analysemodel.

1. Lexicale domeinen in termen van structuureenheden
   a) het domein van de [±samengestelde] woorden
   - wel/niet samengestelde woorden
   - wel/niet actuele woorden (actueel = vermelding in woordenboek)

   b) het domein van de basislexemen
   - [+auto] = lexemen met zelfstandige toepassing
   - [-auto] = lexemen zonder zelfstandige toepassing
   - [+dep] = lexemen met woorddeel-toepassing
   - [-dep] = lexemen zonder woorddeel-toepassing
   - [+lp] = lexemen die als linkerdeel fungeren
   - [+mp] = lexemen die als middendeel fungeren
   - [+rp] = lexemen die als rechterdeel fungeren

2. Lexicale domeinen in termen van bronparameters
   - [-] = geen restricties
   - [+nn] = lexemen uit Van Dale's Woordenboek Hedendaags Nederlands (WHN)
   - [+comp] = lexemen met computationeel toegevoegde kenmerken
   - [+mod] = [+nn]-lexemen uit speciaal informatieveld ([+comp]-status)
   - [+mhb] = affix dat in het Morfologisch Handboek (MHB) wordt vermeld

3) structuureenheden binnen de basislexemen uit de MGBN
   - sublexemen, morfemen
   - subclassificatie van morfemen: stammen, affixen en bindfonemen
   - subclassificatie van affixen: prefixen en suffixen, midden-affixen
   - affix-eenheden (lengte = 1) versus affixsequenties (lengte > 1)
   - ongelede stam (wortel) versus gelede stam (wortel + affixen)

4) niet-kwantitatieve kenmerken bij de structuureenheden
   - categoriale kenmerken: A, B, C, D, N, O, Q, R, T, V, X
   - etymologische kenmerken: [±leen], taallabel, inheems/uitheems (= i/u)
   - semantische kenmerken: [±naam], [±mens]
   - representatieniveau: n1, n2 of n3
   - positie van de structuureenheden in de basislexemen
     - links-rechts-telling of rechts-links-telling
     - telling vanaf lexeemgrens, stamgrens of wortelgrens

5) kwantitatieve kenmerken bij de structuureenheden
- tokenfrequentie = aantal voorkomens binnen een corpus
- typefrequentie = aantal voorkomens binnen een lexicon (binnen een domein D)
- lexeemfrequentie = aantal lexicale lexeemtoepassingen (bij stam of affix)
- stamfrequentie = omvang van het lexicale stamdomein (bij affix)
- u-frequentie = omvang van 1e uitwaartse taxeemdomein (bijv. affixen)
- i-frequentie = omvang van 1e inwaartse taxeemdomein (bijv. wortels)
- locale typefrequentie = omvang van 1e inwaartse of uitwaartse subdomein
- globale typefrequentie = omvang van hele inwaartse of uitwaartse subdomein

6.2.4 De evaluatie van het MGBN-model

In mijn visie op lexicale kennis dient een lexicografische gegevensbank een zo betrouwbaar mogelijk beeld te geven van de lexicale kennis van een *ideale* taalgebruiker. Met betrekking tot een morfologische gegevensbank geldt dus dat de hierin opgenomen morfologische kennis een zo goed mogelijke afspiegeling moet vormen van de morfologische structuurdimensie van het mentale lexicon van een ideale taalgebruiker. Dit is ook het beoogde eindstadium voor de Morfologische Gegevensbank. Maar zoals ik al uiteen heb gezet is dit doel niet eenvoudig te realiseren, aangezien er nog maar weinig bekend is over de mentale representatie van woordkennis, laat staan over de morfologische structuurdimensie van het mentale lexicon. Weliswaar is voor vele talen zeer gedetailleerd onderzoek gedaan naar de vraag wat de morfologische grammaticaregels zijn (zowel vanuit didactisch als vanuit taalpsychologisch perspectief), maar juist door de focus op grammaticaregels draagt dit type onderzoek weinig bij aan de vraag hoe woorden in het mentale lexicon zijn opgeslagen: in mijn optiek werkt het zelfs belemmerend voor dit inzicht.

Zoals in hoofdstuk 4 uiteen werd gezet, heb ik me bij de ontwikkeling van de Morfologische Gegevensbank niet door grammaticaregels, maar door intuïtieve structuuroordelen laten leiden. Hierdoor kan de MGBN een vrij direct inzicht geven in de mentaal relevante structuurkenmerken van de Nederlandse woordenschat. Maar tijdens het ontwikkelingstraject werd duidelijk dat het onduidelijk is waar men de grens moet trekken tussen psychologisch reële en niet-reële morfemen. Juist vanwege dit analyseprobleem heb ik ernaar gestreefd zoveel mogelijk formele (doorgaans etymologisch gemotiveerde) morfemen zichtbaar te maken, vanuit het idee dat deze opzet de beste uitgangspositie biedt voor een fundamenteel onderzoek naar de morfologische segmentatiecriteria van het Nederlands. In deze opzet dient de gegevensbank zijn eigen structuurcriteria te leveren door als onderzoeksdomein te dienen voor het identificeren van psychologisch gemotiveerde segmentatiecriteria en deze criteria vervolgens aan te wenden voor de evaluatie en verbetering van de aanwezige structuurrepresentaties. Men kan deze structuurcriteria opsporen door een grote verzameling structuurrepresentaties aan te leggen en op zoek te gaan naar gemeenschappelijke distributiekenmerken van de samenstellende morfemen. In de L-KRING-visie corresponderen morfemen namelijk met de kleinste bouwstenen die een systematische koppeling vertonen tussen vormkenmerken en (abstracte of concrete) distributiekenmerken en die (dus) potentieel een bijdrage kunnen leveren aan de compressie van het mentale lexicon.

Bij de evaluatie van de MGBN kan zowel een intern als een extern evaluatieperspectief worden gehanteerd. Er is sprake van een intern evaluatieperspectief als de gehanteerde structuurcriteria aan hetzelfde kennisdomein zijn ontleend als het domein dat men wil evalueren, dus als men uit is op maximalisering van de interne consistentie. Dit is mogelijk door het te beoordelen kennisdomein aan een onderzoek te onderwerpen dat als doel heeft om de bestaande structuurkenmerken te inventariseren en om via statistische methodes de onderliggende structuurprincipes te identificeren; hierbij geldt de vuistregel dat een structuur-

principe relevanter is naarmate hij een hogere gebruiksfrequentie kent. De resulterende structuurprincipes dienen vervolgens zo systematisch mogelijk te worden doorgevoerd, terwijl de structuurkenmerken die hier niet door ondersteund worden juist moeten worden verwijderd. In feite is de huidige versie van de MGBN reeds het product van zo'n werkwijze, want tijdens de ontwikkeling van de MGBN heb ik voortdurend consistentiechecks uitgevoerd. Dit resulteerde vaak in het versterken van frequente en het wegwerken van infrequente patronen.

Er kan ook een extern evaluatieperspectief worden gehanteerd: in dat geval dient de te beoordelen gegevensbank met een bestaande kennisbron te worden vergeleken, zodat men een indruk krijgt van de inhoudelijke overeenkomsten en verschillen. Maar zolang niet bekend is hoe betrouwbaar de externe kennisbron is, kan de uitkomst van deze vergelijking alleen inzicht opleveren in de onderlinge informatie-afstand (met betrekking tot het gemeenschappelijke domein). Hierbij kan men onderscheid maken tussen de informatie-afstand op patroonniveau (c.q. lexicaal typeniveau) en de informatie-afstand op representatieniveau (c.q. lexicaal tokenniveau). In het eerste geval volstaat een externe kennisbron met informatie over de algemeen geldige structuurregels, terwijl men in het tweede geval op een gegevensbank met concrete structuurrepresentaties is aangewezen.

Met betrekking tot de morfologische structuur van het Nederlands is het Morfologisch Handboek een duidelijk voorbeeld van een kennisbron met informatie op type-niveau, terwijl het CELEX-lexicon als kennisbron op token-niveau kan fungeren. Beide kennisbronnen kennen echter de fundamentele beperking dat ze uitgaan van grammaticale structuurregels. Hierdoor bieden deze kennisbronnen geen complete afspiegeling van de mentale kennis over morfologisch relevante structuurkenmerken. Deze beperking geldt voor alle kennisbronnen die lexiconbrede informatie geven over de Nedelrandse morfologie (zoals het CGN-lexicon).[208] De Morfologische Gegevensbank is namelijk de eerste kennisbron waarin geprobeerd is om deze mentale structuurkennis systematisch en lexiconbreed vast te leggen.

Men kan het externe evaluatieperspectief ook invullen door een relevante steekproef te nemen uit de te beoordelen dataverzameling en deze aan het oordeel van een aantal zorgvuldig geselecteerde proefpersonen te onderwerpen. Maar deze methode werkt alleen als de proefpersonen goed geïnstrueerd worden over de te hanteren uitgangspunten, waarmee deze evaluatiemethode een deel van zijn objectiviteit verliest. Een bijkomende moeilijkheid is dat het in mijn L-KRING-visie op lexicale kennis erg onwaarschijnlijk is dat twee taalgebruikers precies dezelfde structuuroordelen zullen hebben, aangezien de structuur van de lexicale representaties sterk afhangt van de samenstelling van dit lexicon en mogelijk ook van de cognitieve structuurcriteria (die per taalgebruiker kunnen verschillen).

De voorgaande uiteenzetting leert dat het niet mogelijk is om een externe kennisbron te vinden waarmee de MGBN compleet geëvalueerd kan worden. Om toch inzicht te krijgen in de externe datakwaliteit heb ik het MGBN-model op type-niveau aan het Morfologisch Handboek getoetst (en vice-versa). Hiertoe heb ik de affixkenmerken van het MGBN-model zo uitvoerig mogelijk met de affixkennis in het Morfologisch Handboek vergeleken, namelijk een vergelijking op alle kenmerken die door beide kennisdomeinen worden beschreven. Zo bleven de fonetische vorm en de betekenis uit het MHB noodgedwongen buiten beschouwing; hetzelfde geldt voor de staminformatie in de MGBN, evenals langere affix-sequenties (van meer dan 2 eenheden) en gedetailleerde informatie over de spelvormvarianten. Hieronder volgt een overzicht van de daadwerkelijk vergeleken kenmerken. Hierbij heb ik me op drie klassen van affixtypes gericht, namelijk prefixen, suffixen en prefix-suffix-combinaties. In de

[208] Er bstaan echter wel redactioneel ontlede deellexica; deze woden meestal ontwikkeld ten behoeve van het trainen en/of testen van morfologische parsers.

laatste analyse gaat het om het modificatie-effect van prefixen op de categorie van het bijbehorende lexeempatroon (prefix+suffix-combinatie).

- ✓ affixvorm: spelvorm van het affix (incl. vormvarianten)
- ✓ etymologische klasse: inheems, uitheems, of onbepaald
- ✓ ucat-type: uitwaartse morfeemcategorie (zoals V, N, A, B, T of P)
- ✓ icat-type: inwaartse morfeemcategorie + uitwaartse morfeemcategorie
- ✓ combinatorische eigenschappen op het niveau van de morfemen

Om inzicht te krijgen in de interne datakwaliteit, ben ik voor enkele klassen van structuureenheden nagegaan of hun lexicale distributie aan een herkenbaar patroon voldoet (onder meer door op zoek te gaan naar statistische verbanden). Deze kwantitatieve evaluatiemethode helpt niet veel bij de beoordeling van afzonderlijke patronen, maar wel om een indruk te krijgen van globale lexicale verbanden en de onderliggende structuurmechanismes. Hierbij gaat het me vooral om het opsporen van asymmetrische verdelingen, want deze zijn een indicatie dat de MGBN met een niet-triviale (want niet random toegekende) verzameling structuurrepresentaties correspondeert.

6.3 Basiskenmerken van het MGBN-model

6.3.1 Introductie

Deze sectie biedt informatie over de kwalitatieve en kwantitatieve basiskenmerken van het MGBN-model. Ik zal eerst de structuur van het MGBN-model behandelen (H6.3.2); hierbij zullen alle structuurkenmerken (en bijbehorende termen) worden behandeld die van belang zijn voor de beschrijving van de in dit hoofdstuk te bespreken datarapporten met deelanalyses van het MGBN-model. Vervolgens zal ik inzoomen op de fundamentele asymmetrie tussen prefixen en suffixen (zie H6.3.3), want deze heeft grote invloed op de wijze waarop ik de affixgerichte analyses heb opgezet. Tot slot volgt een sectie met kwantitatieve gegevens over de omvang van het MGBN-model in de vorm van kencijfers over diverse soorten eenheden in het woorddomein, het lexeemdomein en het morfeemdomein.

6.3.2 De structuur van het MGBN-model

De in dit hoofdstuk te presenteren analyses tonen de MGBN vanuit het perspectief van mijn indexgebaseerde lexiconmodel, namelijk het in H4 geïntroduceerde L-KRING-model. In dit model is morfologische structuur een bijverschijnsel van een opslagmechanisme dat als doel heeft om lexicale kennis zo gecomprimeerd mogelijk op te slaan zonder dat er informatieverlies optreedt. Het centrale uitgangspunt van dit model is dat het mentale lexicon met een netwerk van hiërarchisch gestructureerde indexrepresentaties correspondeert. In dit netwerk is voor elk bestaand lexeem een aparte index beschikbaar, terwijl de interne structuur van deze lexemen kan worden verantwoord door voor elke lexeeminterne bouwsteen een aparte index aan te maken en deze via lexicale compositierelaties (met de markering ⊕) aan zowel de uitwaartse (onderschikkende) als de inwaartse (ondergeschikte) indexen te koppelen. Ik zal dit toelichten aan de hand van een stapsgewijze analyse van het lexeem *constructivisme* (met lexeemindex L1 en lexeemklasse $N). Deze analyse wordt hieronder weergegeven:

indexstructuur	**morfeemstructuur**
L1	[con_struct_iv_isme]$_{\#N}$ ⊕ $N
(M2 ⊕$\|_F$ M1)	[[con_struct_iv]$_{\#A}$ ⊕ isme$\|_{\#N}$] ⊕ $N
(N2 ⊕$\|_F$ N1) ⊕$\|_F$ M1	[[[con_struct]$_{\#V}$ ⊕ iv$\|_{\#A}$] ⊕ isme$\|_{\#N}$] ⊕ $N
(O1 $_F\|$⊕ O2) ⊕$\|_F$ N1 ⊕$\|_F$ M1	[[[con$\|_{\#V}$ ⊕ [struct]] ⊕ iv$\|_{\#A}$] ⊕ isme$\|_{\#N}$] ⊕ $N

In de eerste analysestap wordt het lexeem *constructivisme* (met index L1) onderverdeeld in een morfeem [CON_STRUCT_IV] (met index M2 en morfeemklasse #A) en een morfeem -ISME

(met index M1 en morfeemklasse #N). Hierbij is M1 functor ten opzichte van M2 (blijkens de markering $\oplus|_F$), dus M1 bevindt zich in het inwaartse domein van M2 en omgekeerd bevindt M2 zich in het uitwaartse domein van M1. De stam M2 kan zelf weer in twee kleinere morfemen worden opgedeeld, te weten het morfeem [CON_STRUCT] (met index N2 en morfeemklasse #V) en een morfeem -IV (met index N1 en morfeemklasse #A); op lexeemfinale positie krijgt dit morfeem de vorm -IEF (hetgeen verantwoord kan worden door beide n1vormen aan dezelfde n2vorm te koppelen). Hierbij correspondeert het suffix (N1) met de functor van stam N2; deze stam bevindt zich dus in het I-domein van N1 (evenals de stammen DESTRUCT, RELAT en SENSIT), terwijl het suffix -ISME zich in het U-domein van N2 bevindt (evenals -IST, -ITEIT en [$A] (een A-vormend 0-suffix)). Het resultaat staat in rapport 1:

I-domein	**kern**	**U-domein**	(1)
[CONSTRUCT], [DESTRUCT]	-IV / -IEF	-ISME, -IST,	
[RELAT], [SENSIT]		-ITEIT, [$A]	

De stam [CON_STRUCT] ten slotte is onder te verdelen in een prefix CON- (met index O1 en prefixklasse #V) en een wortel [STRUCT] (met index O2); in dit geval staat de functor (O1) links van de stam (wat op indexniveau overigens niets uitmaakt).

Het hier gedemonstreerde structuurprincipe berust op het uitgangspunt dat morfologische representaties per definitie een asymmetrische opbouw hebben, waarbij de functor statistisch gezien bepalend is voor de eigenschappen van het geheel. Dit is in overeenstemming met de veel beschreven observatie dat het rechterhoofd van een willekeurig Nederlands woord doorgaans bepalend is voor de grammaticale eigenschappen van dit woord. Maar in het L-KRING-model is deze "wetmatigheid" een specifiek geval van het gegeven dat elke index zowel een inwaarts als een uitwaarts selectiedomein bezit, zodat structuuropbouw altijd samengaat met de combinatie van een stam en een functor. Hierbij vertoont het Nederlands (net als vele andere talen) de bijzonderheid dat suffixen vaak meer gedragsbepalend zijn dan prefixen, hetgeen samen lijkt te hangen met een verschil in semantische functie. In het bovenstaande voorbeeld is dit zichtbaar gemaakt door steeds de morfeemklasse te specificeren. Deze hiërarchische structuurvisie ligt ook ten grondslag aan de scripts waarmee ik de morfologische structuurrepresentaties in de MGBN heb geanalyseerd. Bij de beschrijving van de met deze scripts vervaardigde analyseraporten zal ik uiteraard intensief gebruik maken van het begrippenapparaat dat ik voor het L-KRING-model heb ontwikkeld. Hieronder wordt uitgelegd hoe dit begrippenapparaat zich tot de structuurrepresentaties in de MGBN verhoudt. Beschouw om te beginnen de resultaten voor het woord *gedachtestelsel*. Blijkens rapport 2a correspondeert dit samengestelde woord met de LGBN-index 1, de categorie N (nomen) en de lexeemstructuur #gedachte#+#stelsel#.

	woord	**index**	**categorie**	**lexeemstructuur**	(2a)
W0	|gedachtestelsel|	1	N	#gedachte#+#stelsel#	

Rapport 2b toont twee perspectieven op de samenstellende delen van dit woord (c.q. de woordinterne lexemen), namelijk een links-rechts-perspectief en een rechts-links-perspectief; de hier bedoelde lexemen worden zowel links als rechts door een #-symbool begrensd.

lr-analyse	**pos 0**	**pos 1**	**pos 2**	**pos 3**	(2b)
lexeem	-	#gedachte#	#stelsel#	-	
rl-analyse	**pos 0**	**pos -1**	**pos -2**	**pos -3**	
lexeem	-	#stelsel#	#gedachte#	-	

Uit rapport 2c blijkt dat de lexemen #gedachte# en #stelsel# ook zelfstandig voorkomen, d.w.z. als een woord met slechts 1 woorddeel c.q. lexeem; dit enkele woorddeel correspondeert met positie pos 0, zowel in de rl-analyse als in de lr-analyse):

	woord	**index**	**categorie**	**lexeemstructuur**	(2c)
W1	|gedachte|	1	N	#gedachte#	
W2	|stelsel|	1	N	#stelsel#	

Rapport 2d toont de morfeemstructuur van de hier besproken eenheden, te weten de lexemen *gedachte* (L1) en *stelsel* (L2); hierbij hanteer ik voor het gemak een links-rechts-nummering.

	structuur L1	**structuur L2**	(2d)
n0	#gedachte#	#stelsel#	= spelvorm zonder structuur
n1	ge;[dacht];e	[stel];sel	= structuur spelvormniveau
n2	ge_[dacht]_e	[stel]_sel	= structuur 1e abstractieniveau
n3	ge_[denk.1]_e(i)	[stel.1]_sel(i)	= structuur 2e abstractieniveau

n1	**n1-prefix-sequentie**	**n1-wortel**	**n1-suffix-sequentie**
L1	ge	dacht	e
L2	0	stel	sel

n2	**n2-prefix-sequentie**	**n2-wortel**	**n2-suffix-sequentie**
L1	ge	dacht	e
L2	0	stel	sel

n3	**n3-prefix-sequentie**	**n3-wortel**	**n3-suffix-sequentie**
L1	ge(i)	denk.1	e(i)
L2	0	stel.1	sel(i)

Hieronder zal ik dezelfde concepten toelichten voor een tweede voorbeeld, te weten het MGBN-woord *ingewikkeldheidsgraad*, dat net als het voorgaande voorbeeld met een samenstelling (c.q. lexeemcombinatie) correspondeert. Een verdere toelichting acht ik onnodig.

	woord	**index**	**categorie**	**lexeemstructuur**	(3a)
W0	|ingewikkeldheidsgraad|	1	N	#ingewikkeldheids#+#graad#	

lr-analyse	**pos 0**	**pos 1**	**pos 2**	**pos 3**	(3b)
lexeem	-	#ingewikkeldheids#	#graad#	-	
rl-analyse	**pos 0**	**pos -1**	**pos -2**	**pos -3**	
lexeem	-	#graad#	#ingewikkeldheids#	-	

	woord	**index**	**categorie**	**lexeemstructuur**	(3c)
W1	|ingewikkeldheid|	1	N	#ingewikkeldheid#	
W2	|graad|	1	N	#graad#	

De structuurrepresentaties in rapport 3d laten zien dat een lexeem als *ingewikkeldheids* zo goed als geen variatie vertoont tussen de n1vorm, de n2vorm en de n3vorm. Het lexeem *graad* daarentegen laat zien dat de lexeemeenheid ook kan samenvallen met de stameenheid, en dat deze op elk niveau een andere vorm kan aannemen (hier GRED.1, GRAD en GRAAD).

	structuur L1	**structuur L2**	(3d)
n0	ingewikkeldheids	#graad#	= spelvorm zonder structuur
n1	in_ge_[wikkel]_d_heid_s	[graad]	= structuur spelvormniveau
n2	in_ge_[wikkel]_d_heid_s	[grAd]	= structuur 1e abstractieniveau
n3	in_ge_[wikkel.1]_d_heid_s	[gred.1]	= structuur 2e abstractieniveau

n1	**n1-prefix-sequentie**	**n1-wortel**	**n1-suffix-sequentie**
L1	in_ge	[wikkel]	d_heid_s
L2	_	[graad]	_

n2	**n2-prefix-sequentie**	**n2-wortel**	**n2-suffix-sequentie**
L1	in_ge	[wikkel]	d_heid_s
L2	_	[grAd]	_
n3	**n3-prefix-sequentie**	**n3-wortel**	**n3-suffix-sequentie**
L1	$in(i)_ge(i)	[wikkel.1]	d(i)_heid(i)_s(i)
L2	_	[gred.1]	_

Rapport 3e illustreert de relatie tussen wortels en stammen (die behalve een wortel ook één of meer affixen kunnen omvatten). Men kan een (complete) prefixstam construeren door de wortel met alle eraan voorafgaande prefixen uit te breiden. Men kan een (complete) suffixstam construeren door de wortel met alle erop volgende suffixen uit te breiden. Deze mogelijkheden worden hieronder gedemonstreerd voor de wortel WIKKEL van het lexeem *ingewikkeldheids*. Uit de tabel blijkt dat elke constructiestap tot een ophoging van de stam-index leidt: zo correspondeert prefixstam-0 met de kale wortel, prefixstam-1 met de combinatie van een wortel en één prefix en prefixstam-2 met de combinatie van een wortel en twee prefixen.

		(3e)
prefixstam-0 = wortel	[wikkel]	
prefixstam-1 = wortel + 1 prefix	ge_[wikkel]	
prefixstam-2 = wortel + 2 prefixen	in_ge_[wikkel]	
suffixstam-0 = max. prefix-sequentie + wortel	{#_in_ge_[wikkel]}	
suffixstam-1 = wortel + 1 stamsuffix	{#_in_ge_[wikkel]}_d	
suffixstam-2 = wortel + 2 stamsuffixen	{#_in_ge_[wikkel]}_d_heid	
suffixstam-3 = wortel + 3 stamsuffixen	{#_in_ge_[wikkel]}_d_heid_s	
lexeem = max. prefix-seq + wortel + max. suffix-seq	{{#_in_ge_[wikkel]}_d_heid_s_#}	

Rapport 3f toont het resultaat van een frequentiebepaling op woorddeelniveau voor alle voorbeeldlexemen (waarbij ik uitga van de n2vorm en een rechts-links-perspectief).

rl-analyse	**niveau**	**max**	**pos 0**	**pos -1**	**pos -2**	**pos -3**	(3f)
#ge_[dacht]_e#	n2	123	1	50	62	10	
#[stel]_sel#	n2	305	1	298	6	-	
#in_ge_[wikkel]_d_heid#	n2	1	1	-	-	-	
#in_ge_[wikkel]_d_heid_s#	n2	1	-	-	1	-	
#[grAd]#	n2	117	1	105	10	1	

Rapport 3g toont het resultaat van een frequentiebepaling op wortelniveau voor alle voorbeeldlexemen (waarbij ik uitga van de n2vorm).

rl-analyse	**niveau**	**max**	**pos 0**	**pos -1**	**pos -2**	**pos -3**	**pos -4**	(3g)
denk.1 / dacht	n2	198	26	77	85	10	-	
stel .1 / stel	n2	2558	133	2146	262	17	-	
wikkel.1 / wikkel	n2	354	27	185	138	4	-	
gred.1 / grAd	n2	190	24	134	30	2	-	

De MGBN bevat ook lexemen met twee of meer wortels; hierbij gaat het vaak om lexemen die beter als samenstelling kunnen worden behandeld, zoals *gedachte+spinsel*. Maar er zijn ook "samenstellingen" waar dit minder evident is, zoals het lexeem *architect* en het hiermee verwante lexeem *architectuur*. Dit laatste lexeem heeft de volgende MGBN-structuur gekregen: [archi]+[tect];uur. Het verschil met normale samenstellingen is dat de twee sublexemen (*archi* en *tectuur*) niet bruikbaar zijn als zelfstandig woord. Om die reden is het beter om deze pseudo-samenstellingen (c.q. stam-samenstellingen) structureel te onderscheiden van normale samenstellingen (c.q. lexeem-samenstellingen). Dit is nog niet consequent doorgevoerd in de

MGBN, maar bij de modellering van de MGBN zijn de samenstellingen beneden lexeemniveau op een andere wijze gecodeerd dan de samenstellingen boven lexeemniveau.

6.3.3 Stamdomein versus lexeemdomein

In de L-KRING-visie op lexicale kennisrepresentatie bestaan er twee mogelijkheden om invulling te geven aan de lexicale typefrequentie van een bouwsteen in een gegeven structuurdomein (bijvoorbeeld de typefrequentie van een morfeem in het domein van de lexemen), namelijk via de omvang van het uitwaartse toepassingsdomein (normaliter binnen het gegeven structuurdomein) en via de omvang van het inwaartse toepassingsdomein (normaliter binnen het gegeven structuurdomein). Hierbij verdient de inwaartse benadering een sterke voorkeur, want de omvang van het inwaartse domein is bepalend voor de bijdrage die een bouwsteen kan leveren aan de compressie van lexicale informatie. De betekenis van de uitwaartse typefrequentie is veel minder duidelijk, al is deze frequentiemaat dominant in het psychologische onderzoek naar de mentale representatie van morfologische kennis.

Dit kan worden toegelicht aan de hand van twee concrete analysevoorbeelden met betrekking tot de typefrequentie van een prefix, namelijk GE-, en een suffix, namelijk EER_D. Beide maken gebruik van de morfologische representatie van het lexeem *gecomponeerd*:

Lexeem:	ge_com_[pon]_eer_d	(4)
Stam van GE-:	{com_[pon]}	
Stam van EER_D:	{ge_com_[pon] }	

Beschouw om te beginnen het prefix GE- (dat ik als een overte variant beschouw van de V-vormende functor [0/GE]). Volgens de hierboven gegeven definitie correspondeert de stam van dit prefix met de eenheid {COM_[PON]}, ook al is dit geen zelfstandig toepasbare morfeemconfiguratie. Indien men nu de stamfrequentie wil bepalen van GE-, dient men na te gaan hoe groot het stamdomein is, dus hoeveel verschillende substituties beschikbaar zijn voor de stam {COM_[PON]}, waarbij het zowel om ongelede stammen (c.q. wortels) als om gelede (prefix-initiële) stammen mag gaan. In het MGBN-model omvat het hier bedoelde stamdomein bijvoorbeeld de ongelede stam PON (in *gecomponeerd*) en de gelede stam PRO_[MIT] (in *gecompromitteerd*). In het geval van het suffix EER_D corespondeert de stam van het lexeem *gecomponeerd* met {GE_[COM]_PON}, en in *gecompromitteerd* met {GE_COM_PRO_[MIT]}. Men kan er natuurlijk over twisten of het prefix GE- wel deel moet uitmaken van de stam van lexemen met de V-toepassingen *compromitteren* en *componeren*. In mijn optiek is de voorgestelde stam echter correct, want zoals ik al aangaf is de vorm GE- een expliciete vorm van een functor die sowieso nodig is voor de toekenning van V-gerelateerde kenmerken. Zo is het lexeem *gecomponeerd* duidelijk verwant aan het V-lexeem *componeren*, zodat er ook een gemeenschappelijke functor moet worden aangenomen.

Hoewel mijn computationele definitie van stamfrequentie een groot bereik heeft, dekt hij niet alle situtaies, want de gegevensbank die ten grondslag ligt aan het MGBN-model (namelijk de MGBN) geeft alleen informatie over de lineaire morfeemstructuur, d.w.z. over de kleinste morfologische bouwstenen (c.q. etymologische bouwstenen) en hun onderlinge volgorde binnen een lexeem. Hierdoor schiet mijn definitie tekort bij lexemen waarvan de cognitieve stam met een ander lexeem correspondeert; in dat geval correspondeert de stam vaak met een MGBN-eenheid die onder meer een suffix bevat. Zo berust het lexeem *gecompartimenteerd* minimaal op de stam COM_[PARTI]_MENT (in *gecompartimenteerd*) en het lexeem *gecomplimenteerd* op de stam COM_[PLI]_MENT. Deze stammen vallen vooralsnog buiten het bereik van mijn computationele definitie van het stamdomein van een prefix.

Met inachtneming van deze beperking ga ik ervan uit dat de stamfrequentie op zich een goede indicator is van het lexicale belang van een affix of affixsequentie, d.w.z. van de mate waarin

het prefix bijdraagt aan de compressie van lexicale informatie en dus ook van de waarschijnlijkheid dat dit prefix wordt geactiveerd. Met het oog op deze doelstelling is het in elk geval een veel interessanter gegeven dan de uitwaartse typefrequentie (c.q. lexeemfrequentie), want bij die laatste maat wordt (als gevolg van de door mij veronderstelde functor-asymmetrie tussen prefixen en suffixen) niet alleen de stamvariatie geteld, maar ook de suffixvariatie. Hierdoor kan geen onderscheid worden gemaakt tussen prefixen die een groot aantal stammen kunnen selecteren, maar waarbij de stammen een laag gemiddelde kennen met betrekking tot de omvang van het suffixdomein, en prefixen waarbij dit precies omgekeerd is.

Dit probleem kan worden ondervangen door afzonderlijk informatie te geven over de gemiddelde stamproductiviteit en de gemiddelde lexeemproductiviteit. Het eerste gegeven kan worden bepaald door per prefix na te gaan wat de stamfrequentie (= omvang van het stamdomein) en wat de wortelfrequentie (= omvang van het worteldomein) is, en door vervolgens de stamfrequentie door de wortelfrequentie te delen. Het tweede gegeven kan worden bepaald door de lexeemfrequentie door de stamfrequentie te delen. Dit is niet nodig bij lexeemfinale suffixen, wanthier vallen stamfrequentie en lexeemfrequentie per definitie samen.

De door mij geconstrueerde datarapporten gaan altijd uit van de stamfrequentie, ook al is die veel moeilijker te achterhalen dan de lexeemfrequentie, want terwijl de lexeemfrequentie gelijk is aan het aantal lexemen dat aan een bepaald zoekpatroon voldoet, dient men voor de stamfrequentie inzicht te hebben in de interne structuur van de lexemen: men mag namelijk slechts 1 lexeem per stam meetellen, maar om dit filter te kunnen toepassen moet eerst bekend zijn welke lexemen dezelfde stam hebben. Indien deze informatie inderdaad voorhanden is, zijn drie stappen nodig om de stamfrequentie te bepalen: querystap 1 correspondeert met de identificatie van lexemen die aan het zoekpatroon voldoen, stap 2 met de extractie van de stam, en stap 3 met de telling van het aantal unieke stammen in de extractielijst.

6.3.4 Kencijfers bij het MGBN-model

Zie appendix B.1

6.4 Inventarisatie van wortels en prefixstammen

6.4.1 Introductie

Deze sectie biedt een eerste kennismaking met constructiewijze en samenstelling van een reeks datarapporten die zijn voortgekomen uit de doelstelling om een inventarisatie op te bouwen van alle wortels en hiermee opgebouwde prefix- en suffix-stammen die deel uitmaken van het MGBN-lexicon. Hierbij beperk ik me tot de bespreking van opzet (6.4.2) en resultaten (6.4.3) van enkele voorbeeldqueries, te weten een query naar de meest voorkomende wortels en een query naar de meest voorkomende prefixstammen. In het kader van de interne evaluatie (6.4.4) besteed ik ook enige aandacht aan de globale stamdistributie in de onderliggende datarapporten. De sectie eindigt met een conclusie (6.5).

6.4.2 Opzet

6.4.2.1 Introductie

In deze subsectie zal ik stilstaan bij de opzet en interpretatie van mijn datarapporten met betrekking tot de stamdimensie van het MGBN-model. Voor een goed begrip van de hierin verzamelde stamtellingen (en meer in het algemeen voor de in dit hoofdstuk besproken datarapporten) is het cruciaal om inzicht te hebben in de structuur van de door deze stammen ondersteunde lexeemparadigma's. Om dit te belichten zal ik enkele concrete voorbeelden bespreken. Voor dit doel heb ik twee omvangrijke lexeemparadigma's geselecteerd, te weten het lexeemparadigma van de inheemse n3wortel SCHIET.1, die onder meer in het werkwoord *schieten* wordt aangetroffen, en het lexeemparadigma van de uitheemse n3wortel DUC.1, die

bijvoorbeeld in het werkwoord *produceren* voorkomt. Deze paradigma's corresponderen met een op prefixstam geordende verzameling van lexemen met dezelfde etymologische wortel, waarbij voor elk lexeem is aangegeven wat de interne morfeemopbouw is (maar nog zonder hiërarchische structuurinformatie). Via deze voorbeeldparadigma's krijgt men beter inzicht in de morfologische eigenschappen van de inheemse en uitheemse structuurrepresentaties in de MGBN, wat nodig is om de in dit hoofdstuk besproken tabellen met affixkenmerken te interpreteren. Na de behandeling van deze voorbeeldparadigma's ga ik concreet in op de samenstelling en analysemogelijkheden van mijn databestanden met staminventarisaties.

6.4.2.2 Voorbeeldparadigma's

Voorbeeld van een inheems lexeemparadigma

Tabel 6-2 toont alle n1vormen uit het lexeemparadigma van de inheemse n3wortel (= wortel in de n3vorm) SCHIET.1, die een groot aantal vormvarianten kent, te weten *schiet*, *schot*, *schoot*, *scheut, schut* en enkele aan het Engels ontleende vormen, te weten *shot* en *shoot*. Het lexeemparadigma is onderverdeeld in drie subklassen van prefixstammen, te weten:

- subklasse A: lexemen waarvan de prefixstam de status [-prefix] heeft (= enkel een wortel)
- subklasse B: lexemen waarvan de prefixstam de status [+prefix] heeft (= wortel + prefix)
- subklasse C: lexemen waarvan de prefixstam uit twee wortels bestaat

A: [schiet], [schiet];en, [schiet];er, [schiet];er;s, [schiet];ing, [schiet];je, [schoot], [schoot];s, [schot], [schot];en, [schot];ig, [schot];je, [schot];loos, [schot];s, [schott];e:ling, [schott];er, [shoot], [shoot];er, [shot], [shott];en, [scheut], [scheut];el;ing, [scheut];en, [scheut];ig, [scheut];ist, [shott];er, [schut], [schut];s, [schut];ster, [schutt];er, [schutt];er;en, [schutt];er;ig, [schutt];er;ij, [schutt];er;lijk, [schutt];er;s

B: aan;[schiet];en, achter;[schot], achter;na;[schiet];en, achter;uit;[schiet];en, af;[schiet];en, af;[schot], be;[schiet];en, be;[schiet];er, be;[schiet];ing, be;[schot], be;[schot];en, bij;[schiet];en, door;[schiet];en, door;[schiet];er;s, ge;[schot], ge;[schot];en, ge;[schut], in;[schiet], in;[schiet];en, langs;[schot], mis;[schiet];en, mis;[schot], na;[scheut], na;[schiet];en, neer;[schiet];en, om;[schiet];en, onder;[schiet];en, ont;[schiet];en, ont;[schott];en, ont;[schott];ing, ont;[schutt];er;en, op;[schiet];en, op;[schiet];er, over;[schiet];en, over;[schoot], over;[schot], rond;[schiet];en, tegen;[schot], terug;[schiet];en, toe;[schiet];e:lijk, toe;[schiet];en, tussen;[schot], uit;[schiet];baar, uit;[schiet];en, uit;[schiet];er, uit;[schot], ver;[schiet], ver;[schiet];en, ver;[schiet];er, ver;[schiet];ing, ver;[schot], vol;[schiet];en, voor;[schiet];en, voor;[schoot], voor;[schot], voor;bij;[schiet];en, voor;uit;[schiet];en, voort;[schiet];en, weg;[schiet];en

C: [come]+[shot], [hot]+[shot], [mug]+[shot], [trouble]+[shoot];er

Tabel 6-2: Alle n1vormen uit de lexeemparadigma's van SCHIET.1.

Tabel 6-3 toont een selectie uit het hierboven gespecificeerde lexeemparadigma, namelijk een op prefixstam gesorteerde lexeeminventarisatie, waarbij alleen prefixstammen zijn weergegeven waarvan de n3vorm een typefrequentie van 4 of hoger bezit. Hierbij is elke n3wortel verder uitgesplitst naar n2vorm en n1vorm, terwijl voor elke [n3,n2,n1]-combinatie een lijst met de bijbehorende lexeemtoepassingen is gespecificeerd. Men kan deze stammenlijst nog uitbreiden met een n2variant van de n3stam SCHIET.1, namelijk met de n2stam SCHUT (in de betekenis van SCHIET.1); maar ik heb dit achterwege gelaten om duidelijk te maken dat de MGBN, ondanks de hoge detailleringsgraad, geen 100 procent betrouwbare stamclassificatie kent. De verbetering van deze informatie is echter extreem arbeidsintensief, terwijl het waarschijnlijk niet veel uitmaakt voor de resultaten van mijn onderzoek naar de syntagmatische en paradigmatische eigenschappen van affixen. Wel wenselijk is het coderen van stamkenmerken die een onderverdeling naar synchroon betekenisdomein mogelijk maken.

n3stam	n3freq	n2stam	n2freq	n1stam	n1freq	n1lexeem
[schiet.1]	30	[schiet]	6	[schiet]	6	---
[schiet.1]	30	[schiet]	6	[schiet]	6	[schiet]
[schiet.1]	30	[schiet]	6	[schiet]	6	[schiet];en
[schiet.1]	30	[schiet]	6	[schiet]	6	[schiet];er
[schiet.1]	30	[schiet]	6	[schiet]	6	[schiet];er;s
[schiet.1]	30	[schiet]	6	[schiet]	6	[schiet];ing
[schiet.1]	30	[schiet]	6	[schiet]	6	[schiet];je
[schiet.1]	30	[schOt]	4	[schoot]	2	---
[schiet.1]	30	[schOt]	4	[schoot]	2	[schoot]
[schiet.1]	30	[schOt]	4	[schoot]	2	[schoot];s
[schiet.1]	30	[schOt]	4	[schot]	2	---
[schiet.1]	30	[schOt]	4	[schot]	2	[schot];en
[schiet.1]	30	[schOt]	4	[schot]	2	[schot];ig
[schiet.1]	30	[schot]	6	[schot]	4	---
[schiet.1]	30	[schot]	6	[schot]	4	[schot]
[schiet.1]	30	[schot]	6	[schot]	4	[schot];je
[schiet.1]	30	[schot]	6	[schot]	4	[schot];loos
[schiet.1]	30	[schot]	6	[schot]	4	[schot];s
[schiet.1]	30	[schot]	6	[schott]	2	---
[schiet.1]	30	[schot]	6	[schott]	2	[schott];e:ling
[schiet.1]	30	[schot]	6	[schott]	2	[schott];er
[schiet.1]	30	[scheut]	5	[scheut]	5	---
[schiet.1]	30	[scheut]	5	[scheut]	5	[scheut]
[schiet.1]	30	[scheut]	5	[scheut]	5	[scheut];el;ing
[schiet.1]	30	[scheut]	5	[scheut]	5	[scheut];en
[schiet.1]	30	[scheut]	5	[scheut]	5	[scheut];ig
[schiet.1]	30	[scheut]	5	[scheut]	5	[scheut];ist
[schiet.1]	30	[shot]	6	[shot]	4	---
[schiet.1]	30	[shot]	6	[shot]	4	[come]+[shot]
[schiet.1]	30	[shot]	6	[shot]	4	[hot]+[shot]
[schiet.1]	30	[shot]	6	[shot]	4	[mug]+[shot]
[schiet.1]	30	[shot]	6	[shot]	4	[shot]
[schiet.1]	30	[shot]	6	[shott]	2	---
[schiet.1]	30	[shot]	6	[shott]	2	[shott];en
[schiet.1]	30	[shot]	6	[shott]	2	[shott];er
[schiet.1]	30	[shoot]	3	[shoot]	3	---
[schiet.1]	30	[shoot]	3	[shoot]	3	[shoot]
[schiet.1]	30	[shoot]	3	[shoot]	3	[shoot];er
[schiet.1]	30	[shoot]	3	[shoot]	3	trouble;[shoot];er
be(i)_[schiet.1]	5	be_[schiet]	3	be;[schiet]	3	be;[schiet];en
be(i)_[schiet.1]	5	be_[schiet]	3	be;[schiet]	3	be;[schiet];er
be(i)_[schiet.1]	5	be_[schiet]	3	be;[schiet]	3	be;[schiet];ing
be(i)_ [schiet.1]	5	be_[schOt]	1	be;[schot]	1	be;[schot];en
be(i)_ [schiet.1]	5	be_[schot]	1	be;[schot]	1	be;[schot]
ver(i)_[schiet.1]	5	ver_[schiet]	4	ver;[schiet]	4	ver;[schiet]
ver(i)_ [schiet.1]	5	ver_[schiet]	4	ver;[schiet]	4	ver;[schiet];en
ver(i)_[schiet.1]	5	ver_[schiet]	4	ver;[schiet]	4	ver;[schiet];er
ver(i)_[schiet.1]	5	ver_[schiet]	4	ver;[schiet]	4	ver;[schiet];ing
ver(i)_[schiet.1]	5	ver_[schot]	1	ver;[schot]	1	ver;[schot]
$uit(i)_[schiet.1]	4	uit_[schiet]	3	uit;[schiet]	3	uit;[schiet];baar
$uit(i)_[schiet.1]	4	uit_[schiet]	3	uit;[schiet]	3	uit;[schiet];en
$uit(i)_[schiet.1]	4	uit_[schiet]	3	uit;[schiet]	3	uit;[schiet];er
$uit(i)_[schiet.1]	4	uit_[schot]	1	uit;[schot]	1	uit;[schot]

Tabel 6-3: Een op prefixstam geordende selectie uit het lexeemparadigma van de inheemse n3wortel SCHIET.1, beperkt tot prefixstammen met een typefrequentie van 4 of hoger.

Voorbeeld van een uitheems lexeemparadigma

Tabel 6-4 toont alle n1vormen uit het lexeemparadigma van de uitheemse n3wortel DUC.1. Deze n3wortel kent net als de inheemse wortel uit het voorgaande voorbeeld meerdere vormvarianten, te weten de n1vormen *duc*, *duce* en *duct*. Het bijgaande lexeemparadigma bestaat uit dezelfde subklassen (A, B en C) als het inheemse lexeemparadigma in tabel 6-2.

lexeemparadigma van de n3wortel DUC.1:
A: [duce], [duct];iel, [duct];il;it;eit, [duct];us
B: ab;[duc];er;en, ab;[duct];ie, ab;[duct];or, ad;[duc];er;en, ad;[duct], ad;[duct];ie, ad;[duct];or, bij;pro;[duct], bio;pro;[duct];ie, con;[duct], con;[duct];eur, con;[duct];eur;s, con;[duct];ie, con;[duct];o:metr;ie, con;[duct];or, con;[duct];r:ice, de;[duc];er;en, de;[duct];ie, de;[duct];ief, e;[duct], her;intro;[duc];er;en, im;pro;[duct];ief, im;pro;[duct];iv;it;eit, in;[duc];er;en, in;[duct];ant;ie, in;[duct];ie, in;[duct];ief, in;[duct];or, in;pro;[duct], intro;[duc];é, intro;[duc];ent, intro;[duc];er;en, intro;[duct];ie, intro;[duct];ief, ir;re;[duct];ibel, na;pro;[duct], ob;[duc];ent, ob;[duc];er;en, ob;[duct];ie, on;pro;[duct];ief, onder;pro;[duct];ie, over;pro;[duct];ie, pro;[duc];en, pro;[duc];ent, pro;[duc];ent;en, pro;[duc];er;en, pro;[duc];er;end, pro;[duce];r, pro;[duct], pro;[duct];en, pro;[duct];ie, pro;[duct];ief, pro;[duct];iv;it;eit, pro;[duct];iv;it;eit;s, pro;[duct];schap, re;[ëduc];at;ie, re;[duc];eer, re;[duc];eer;baar, re;[duc];er;en, re;[duct];ie, re;[duct];ion;ism;e, re;[duct];ion;ist;isch, re;[duct];or, re;pro;[duc];eer;baar, re;pro;[duc];ent, re;pro;[duc];er;en, re;pro;[duct];ie, re;pro;[duct];ief, re;pro;[duct];iv;it;eit, se;[du];is;ant, se;[duct];ie, sub;[duct];ie, trans;[duc];ent, trans;[duc];er, trans;[duct];ie, tussen;pro;[duct], uit;pro;[duct], wan;pro;[duct]
C: [aqua]+[duct], [cervi]+[duct], [eco]+[duct], [edu]+[kin];es;i:o:log;ie, [edu]+[tain];ment, [man]+[duct];or

Tabel 6-4: Alle n1vormen uit het lexeemparadigma van de n3wortel DUC.1.

Tabel 6-5 toont weer een op prefixstam gesorteerde selectie van de n1vormen in het lexeemparadigma uit tabel 1a. Hierbij zijn alleen prefixstammen weergegeven waarvan de n3vorm een typefrequentie van 4 of hoger bezit. Verder is elke n3wortel weer verder uitgesplitst naar n2vorm en n1vorm, terwijl voor elke [n3,n2,n1]-combinatie een lijst met de bijbehorende lexeemtoepassingen is gespecificeerd.

n3stam	n3freq	n2stam	n2freq	n1stam	n1freq	n1lexeem
pre(u)_[duc.1]	14	pro_[duc]	4	pro;[duc]	4	pro;[duc];ent
pre(u)_[duc.1]	14	pro_[duc]	4	pro;[duc]	4	pro;[duc];er;en
pre(u)_[duc.1]	14	pro_[duct]	4	pro;[duc]	4	pro;[duc];ent;en
pre(u)_[duc.1]	14	pro_[duct]	4	pro;[duc]	4	pro;[duc];er;end
pre(u)_[duc.1]	14	pro_[duce]	2	pro;[duce]	2	pro;[duce];en
pre(u)_[duc.1]	14	pro_[duce]	2	pro;[duce]	2	pro;[duce];r
pre(u)_[duc.1]	14	pro_[duct]	8	pro;[duct]	8	[bio]+pro;[duct];ie
pre(u)_[duc.1]	14	pro_[duct]	8	pro;[duct]	8	pro;[duct]
pre(u)_[duc.1]	14	pro_[duct]	8	pro;[duct]	8	pro;[duct];ie
pre(u)_[duc.1]	14	pro_[duct]	8	pro;[duct]	8	pro;[duct];ief
pre(u)_[duc.1]	14	pro_[duct]	8	pro;[duct]	8	pro;[duct];iv;it;eit
pre(u)_[duc.1]	14	pro_[duct]	8	pro;[duct]	8	pro;[duct];schap
pre(u)_[duc.1]	14	pro_[duct]	8	pro;[duct]	8	pro;[duct];en
pre(u)_[duc.1]	14	pro_[duct]	8	pro;[duct]	8	pro;[duct];iv;it;eit;s
con(u)_[duc.1]	7	con_[duct]	7	con;[duct]	7	con;[duct]
con(u)_[duc.1]	7	con_[duct]	7	con;[duct]	7	con;[duct];eur
con(u)_[duc.1]	7	con_[duct]	7	con;[duct]	7	con;[duct];eur;s
con(u)_[duc.1]	7	con_[duct]	7	con;[duct]	7	con;[duct];ie
con(u)_[duc.1]	7	con_[duct]	7	con;[duct]	7	con;[duct];o:metr;ie
con(u)_[duc.1]	7	con_[duct]	7	con;[duct]	7	con;[duct];or

con(u)_[duc.1]	7	con_[duct]	7	con;[duct]	7	con;[duct];r:ice
re(u)_[duc.1]	7	re_[duc]	3	re;[duc]	3	re;[duc];eer
re(u)_[duc.1]	7	re_[duc]	3	re;[duc]	3	re;[duc];eer;baar
[re(u)_[duc.1]	7	re_[duc]	3	re;[duc]	3	re;[duc];er;en
re(u)_[duc.1]	7	re_[duct]	4	[duct]	4	re;[duct];ie
re(u)_[duc.1]	7	re_[duct]	4	[duct]	4	re;[duct];ion;ism;e
re(u)_[duc.1]	7	re_[duct]	4	[duct]	4	re;[duct];ion;ist;isch
[re(u)_[duc.1]	7	re_[duct]	4	[duct]	4	re;[duct];or
re(u)_pre(u)_[duc.1]	6	re_pro_[duc]	3	re;pro;[duc]	3	re;pro;[duc];eer;baar
re(u)_pre(u)_[duc.1]	6	re_pro_[duc]	3	re;pro;[duc]	3	re;pro;[duc];ent
re(u)_pre(u)_[duc.1]	6	re_pro_[duc]	3	re;pro;[duc]	3	re;pro;[duc];er;en
re(u)_pre(u)_[duc.1]	6	re_pro_[duct]	3	re;pro;[duct]	3	re;pro;[duct];ie
re(u)_pre(u)_[duc.1]	6	re_pro_[duct]	3	re;pro;[duct]	3	re;pro;[duct];ief
re(u)_pre(u)_[duc.1]	6	re_pro_[duct]	3	re;pro;[duct]	3	re;pro;[duct];iv;it;eit
im(u)_[duc.1]	5	in_[duc]	1	[in;duc]	1	in;[duc];er;en
im(u)_[duc.1]	5	in_[duct]	4	in;[duct]	4	in;[duct];ant;ie
im(u)_[duc.1]	5	in_[duct]	4	in;[duct]	4	in;[duct];ie
im(u)_[duc.1]	5	in_[duct]	4	in;[duct]	4	in;[duct];ief
im(u)_[duc.1]	5	in_[duct]	4	in;[duct]	4	in;[duct];or
ad(u)_[duc.1]	3	ad_[duct]	3	ad;[duct]	3	ad;[duct]
ad(u)_[duc.1]	3	ad_[duct]	3	ad;[duct]	3	ad;[duct];ie
ad(u)_[duc.1]	3	ad_[duct]	3	ad;[duct]	3	ad;[duct];or
ab(u)_[duc.1]	2	ab_[duct]	2	ab;[duct]	2	ab;[duct];ie
ab(u)_[duc.1]	2	ab_[duct]	2	ab;[duct]	2	ab;[duct];or

Tabel 6-5: Een op prefixstam geordende selectie uit het lexeemparadigma van de uitheemse n3wortel DUC.1, beperkt tot prefixstammen met een typefrequentie van 4 of hoger.

Discussie

Bij vergelijking van het inheemse en het uitheemse lexeemparadigma blijkt dat er tal van structurele overeenkomsten zijn. Zo geldt voor beide paradigma's dat de onderliggende wortels een groot aantal prefixen kunnen selecteren, waarvan sommige bovendien met een uitvoerig lexeemparadigma corresponderen. In dit lexeemparadigma is vrijwel altijd een werkwoordstoepassing te vinden en een hieraan gerelateerde nominalisatievorm. Verder is er een systematische koppeling tussen stamvormvarianten en suffixkeuze (of coverte typering van de lexeemfunctie). Deze parallellen zijn zo prominent dat ik geen reden zie om uitheemse lexeemparadigma's anders te behandelen dan inheemse lexeemparadigma's.

6.4.2.3 De samenstelling van de datarapporten

Ik zal nu concreet ingaan op opzet en samenstelling van mijn datarapporten met staminformatie. Hiertoe zal ik eerst uiteenzetten wat het globale idee is achter deze rapporten en welke analysedoelen hiermee gediend zijn, om vervolgens een overzicht te geven van de hierin opgenomen kenmerken en deze kort toe te lichten.

Mijn datarapporten met betrekking tot de stamdimensie van het MGBN-model hebben als doel om per stamklasse (zoals wortels, prefixstammen en suffixstammen) na te gaan welke stammen er bestaan, welke vormvarianten deze stammen kennen en wat hun typefrequentie is. Deze inventarisaties komen in de eerste plaats voort uit nieuwsgierigheid naar de vraag welke stammen het vaakst gebruikt worden (d.w.z. de meeste morfologische toepassingen kennen); tot nu toe is daar relatief weinig over bekend, omdat de bestaande gegevensbanken nauwelijks informatie geven over etymologische stamrelaties (zoals relaties tussen stammen met klinker-

variatie). Ten tweede vullen deze inventarisaties een belangrijke lacune in het MHB, want dit handboek behandelt de Nederlandse morfologie alleen vanuit een affixperspectief.

De staminventarisaties zijn ook van belang met het oog op de externe en interne evaluatie van het MGBN-model, en kunnen bovendien een zeer praktische functie vervullen bij de verdere ontwikkeling van de MGBN. Er zijn immers veel meer wortels dan affixen: zo geldt op het niveau van de n2vorm dat er ca. 20.000 wortels zijn tegenover niet meer dan 1.000 affixen, dus dat hun verhouding overeenkomt met 20:1. Hierdoor is de analyse van de stamdimensie veel ingewikkelder dan de analyse van de affixdimensie, al bestaat er de nodige interactie tussen deze analysetaken: alles wat van de wortel afgaat komt immers in het affixdomein terecht en vice versa. De inventarisatie van distributiegegevens voor diverse stamtypes kan een nuttige rol spelen bij de evaluatie en verbetering van de huidige segmentatiegrenzen.

Naast dit segmentatieprobleem (dat inmiddels flink is teruggedrongen) bestaat ook een reusachtig identificatieprobleem: het is namelijk een moeilijke opgave om alle wortels zodanig te clusteren dat wortels met dezelfde etymologische afkomst in hetzelfde cluster staan. Hiertoe dient men niet alleen overzicht te hebben over de reeds aangelegde wortelinventarisatie, maar ook over de onderliggende lexeemparadigma's. Deze analysetaak doet dus een enorm beroep op het geheugen, en kan daarom alleen worden volbracht door een cyclisch analysetraject te volgen. In dit traject is een cruciale rol weggelegd voor steeds vernieuwde lijsten met staminventarisaties. Mijn datarapporten met staminventarisaties bevatten relatief weinig informatie en zijn ook vrij eenvoudig te construeren. Het gaat om de volgende kenmerken:

i) stameenheden uit de gewenste stamklasse, zoals wortels, kale prefixstammen, prefixstammen met 1 suffix, etc.
ii) representatieniveaus: hoofdstam (n3vorm) en stamvarianten (n2vorm en n1vorm)
iii) absolute omvang van het lexeemdomein (voor elk representatieniveau)
iv) relatieve aandeel van de vormvarianten
v) optioneel: voorbeeldlexemen of integrale lexeemparadigma's

6.4.3 Resultaten

6.4.3.1 Inventarisatie van wortelstammen

De stammenlijst in tabel 6-6 correspondeert met een op de MGBN gebaseerde inventarisatie van de 60 hoogstfrequente wortelstammen (in n2vorm) op het niveau van de sublexemen. Uit deze lijst blijkt dat de 10 hoogstfrequente wortels in n2vorm een typefrequentie bezitten die net onder de 100 blijft, met waardes die oplopen van 81 tot 99. De hoogstfrequente wortel correspondeert met de uitheemse vorm PORT (n2freq = 99), die men kan aantreffen in lexemen als *rapport*, *importeren* en *supporter*. Plaats 2 correspondeert met de inheemse wortel TREK (n2freq = 95), die men aantreft in lexemen als *trekken*, *optrekje*, en *betrekking*. Plaats 3 is weer een uitheemse stam, namelijk ACT (n2freq = 88), die men aantreft in lexemen als *acteren*, *actief* en *acteur*. Als men ook de klankvarianten meetelt, zouden deze wortels zelfs hoger eindigen dan de wortel met n2vorm PORT, want hun n1freqs zijn (in de volgorde van hun rang) 99, 109 en 125. Zo kent de wortel met n2vorm TREK ook een klankvariant met de vorm TROK (bijv. in *betrokken*) en de wortel met de n2vorm ACT bezit een variant met de vorm AG (bijv. in *agent*). Omdat de hier gepresenteerde typefrequenties op etymologische structuurrepresentaties zijn gebaseerd, zijn ze niet maatgevend voor de synchrone productiviteit van deze wortels. Maar mogelijk is er een verband tussen het stamconcept en de etymologische typefrequentie: hoe hoger die typefrequentie, hoe groter het pragmatische belang van het onderliggende concept. Nader onderzoek moet uitwijzen of dit verband klopt.

n3vorm	n3freq	n2vorm	n2freq
[port.1]	99	[port]	99
[trek.1]	109	[trek]	95
[ag.1]	125	[act]	88
[son.1]	88	[sOn]	88
[leg.1]	183	[leg]	87
[haal.1]	97	[hAl]	85
[legi.1]	176	[lect]	83
[vert.1]	123	[vers]	83
[druk.1]	82	[druk]	82
[maak.1]	97	[mAk]	81
[dien.1]	78	[dien]	78
[geef.1]	113	[gEF]	76
[laat.1]	83	[lAt]	76
[neem.1]	104	[nEm]	75
[part.1]	84	[part]	75
[hand.1]	93	[hand]	74
[bouw.1]	74	[bouw]	74
[schrijf.1]	133	[schrijF]	73
[serv.1]	76	[serv]	73
[licht.1]	74	[licht]	73
[vaar.1]	164	[voer]	72
[steek.1]	109	[stEk]	72
[wijs.1]	81	[wijZ]	71
[voeg.1]	71	[voeg]	71
[trouw.1]	70	[trouw]	70
[fac.1]	173	[fect]	69
[snijd.1]	81	[snijd]	69
[een.1]	74	[En]	69
[log.1]	73	[lOg]	69
[staan.1]	174	[stand]	68

n3vorm	n3freq	n2vorm	n2freq
[leef.1]	90	[lEF]	68
[ord.1]	68	[ord]	68
[pon.1]	122	[pos]	67
[ken.1]	117	[ken]	67
[draag.1]	83	[drAg]	67
[par.1]	79	[par]	67
[zeg.1]	77	[zeg]	67
[staan.1]	174	[stAn]	65
[legi.1]	176	[leg]	64
[doen.1]	127	[doen]	64
[koop.1]	71	[kOp]	64
[speel.1]	68	[spEl]	64
[slaan.1]	122	[slAg]	63
[gaan.1]	115	[gAn]	63
[visi.1]	90	[vis]	62
[cip.1]	91	[cept]	61
[reken.1]	62	[reken]	61
[teken.1]	61	[teken]	61
[hang.1]	83	[hang]	60
[patr.1]	69	[patr]	60
[fer.1]	63	[fer]	60
[spreek.1]	102	[sprEk]	59
[laad.1]	99	[lAd]	59
[dek.1]	77	[dek]	59
[vang.1]	66	[vang]	59
[foon.1]	60	[fOn]	59
[werp.1]	75	[werp]	58
[pres.1]	70	[pres]	58
[breng.1]	61	[breng]	58
[zin.1]	58	[zin]	58

Tabel 6-6: De 60 hoogstfrequente wortels (met typefrequentie op lexeemniveau).

6.4.3.2 Inventarisatie van prefixstammen

Tabel 6-7 toont de 30 hoogstfrequente prefixstammen uit de MGBN (op basis van hun n2vorm) op het niveau van de sublexemen, beperkt tot stammen met een overt prefix.

n3stam	n3freq	n2stam	n2freq
per(i)_[son.1]	29	per_[sOn]	29
con(u)_[muun.1]	31	com_[mun]	27
con(u)_[legi.1]	28	col_[lect]	22
pre(u)_[fes.1]	18	pro_[fes]	18
ge(i)_[meen.1]	23	ge_[mEn]	18
se(u)_[creet.1]	18	se_[crEt]	17
im(u)_[form.1]	17	in_[form]	17
di(u)_[recht.1]	25	di_[rect]	16
con(u)_[serv.1]	17	con_[serv]	16
re(u)_[lati.1]	17	re_[lat]	15
re(u)_[spond.1]	17	re_[spons]	14
per(i)_[fac.1]	14	per_[fect]	14
im(u)_[stru.1]	14	in_[stru]	14
sub(u)_[sidi.1]	13	sub_[sidi]	13
im(u)_[tens.1]	17	in_[tens]	13
syn(u)_[chroon.1]	12	syn_[chrOn]	12
re(u)_[cip.1]	24	re_[cept]	12

im(u)_[pres.1]	15	im_[pres]	12
ge(i)_[lijk.1]	13	ge_[lijk]	12
con(u)_[trah.1]	13	con_[tract]	12
con(u)_[sul.1]	17	con_[sult]	12
$uit(i)_[vaar.1]	17	uit_[voer]	12
#uni(u)_[vert.1]	12	uni_[vers]	12
ver(i)_[een.1]	11	ver_[En]	11
ver(i)_[doem.1]	20	ver_[doem]	11
syn(u)_[bool.1]	11	sym_[bOl]	11
re(u)_[spic.1]	11	re_[spect]	11
re(u)_[ag.1]	22	re_[act]	11
ob(u)_[ject.1]	12	ob_[ject]	11
ge(i)_[nees.1]	13	ge_[nEZ]	11

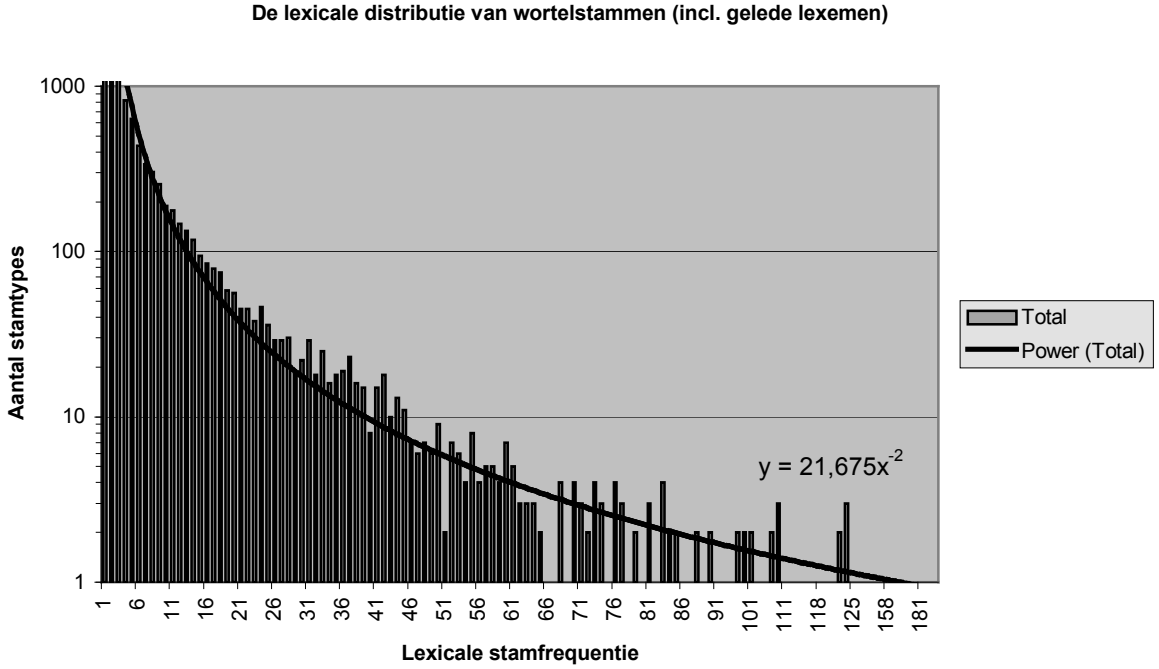
Tabel 6-7: De 30 hoogstfrequente prefixstammen (met prefix).

Hoewel de hier weergegeven stammen formeel een prefix bevatten, wijst een nadere inspectie uit dat het in alle gevallen om gelexicaliseerde prefix-wortel-combinaties gaat, dus dat er vanuit morfosemantisch oogpunt wel eens sprake zou kunnen zijn van ongelede stammen. Maar dit geldt mogelijk voor alle lexemen uit de LGBN. Dit neemt niet weg dat formeel gezien wel degelijk sprake is van een geleding, waarbij de aanwezigheid van het prefix blijkbaar een negatieve invloed heeft op de omvang van het beschikbare lexeemparadigma.

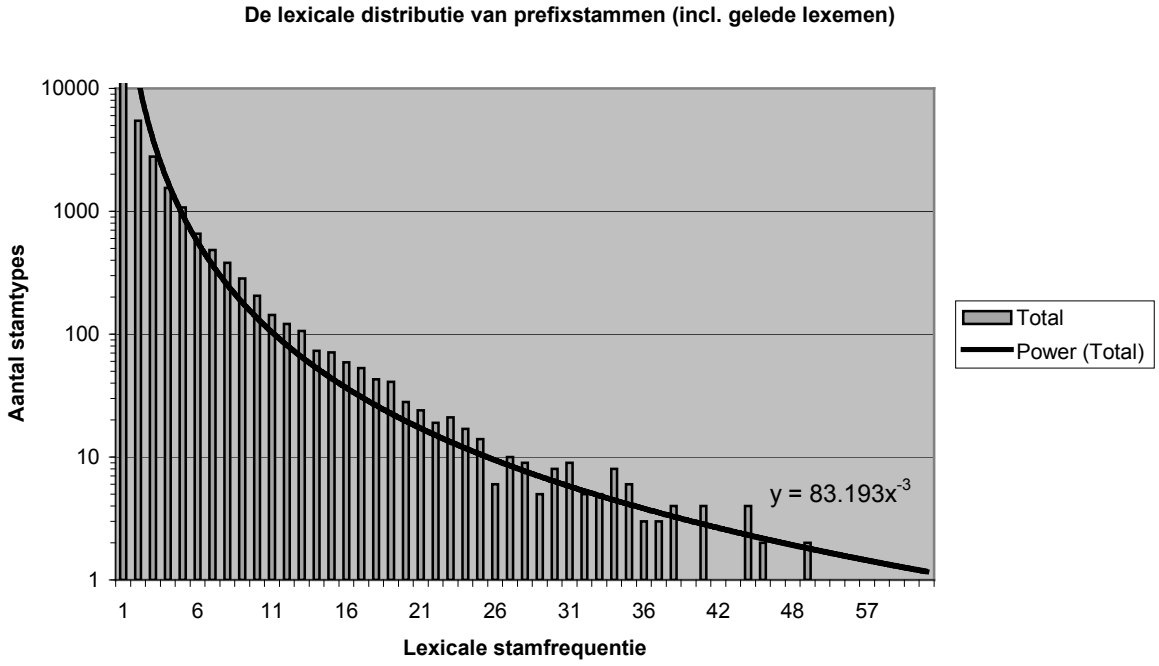
6.4.4 Interne evaluatie

6.4.4.1 Distributiepatronen

In deze paragraaf presenteer ik twee grafieken (in figuur 6-5 en 6-6) met informatie over de frequentieverdeling van de wortelstammen en de prefixstammen uit de hierboven behandelde datarapporten. Meer specifiek bieden deze grafieken inzicht in de relatie tussen de lexicale stamfrequentie (d.w.z. het aantal lexiconinterne lexeemtoepassingen van een stam) en het aantal stamtypes waar deze frequentie op van toepassing is (uitgezet op een logaritmische schaal). In de discussie zal ik nader ingaan op de interpretatie van deze grafieken.



Figuur 6-5: Grafiek met de lexicale distributie van wortelstammen



Figuur 6-6: Grafiek met de lexicale distributie van wortelstammen

6.4.4.2 Discussie

De hierboven weergegeven distributiegrafieken laten zien dat zowel de wortels als de wortelstammen geen willekeurige distributie vertonen, maar dat er een tamelijk betrouwbaar verband bestaat tussen de lexicale stamfrequentie en het aantal stammen waar deze frequentie op van toepassing is. In beide gevallen is sprake van een exponentiële functie, waarbij de functie voor de prefixstammen een wat steiler verval kent dan die voor de prefixstammen, namelijk c_1/x^3 versus c_2/x^2. Kwalitatief betekent dit dat er relatief veel prefixstammen zijn met een stamfrequentie van 1 (meer dan 10.000), maar heel weinig met een stamfrequentie van 40 of groter (namelijk 5 of minder stammen per frequentieklasse). De functie voor de wortels begint wat lager, maar loopt veel langer door; zo zijn er voor stamfrequentie 40 nog altijd minstens 10 stammen per frequentieklasse te vinden. Deze observaties zijn in overeenstemming met het te verwachten functiegedrag. De hier ontdekte functies vormen een aanwijzing dat de door mij aangebrachte structuur op een robuust parseringscriterium berust, al is het ook mogelijk dat de hier geconstateerde verdeling een gevolg is van de interne samenstelling van de Nederlandse woordenschat. Ongeacht het antwoord op deze vraag kunnen deze functies een praktische toepassing krijgen als intern evaluatiecriterium; via dit criterium kunnen stammen worden opgespoord die een te hoge of juist een te lage stamfrequentie vertonen; dit kan aanleiding zijn tot een herziening van het aan deze stam verbonden lexeemparadigma.

6.4.5 Conclusie

De mofologische structuurinformatie in de MGBN maakt het mogelijk om zeer gedetailleerd onderzoek te doen naar de distributie van Nederlandse wortels en de hierdoor ondersteunde stammen en lexeemparadigma's. In deze sectie heb ik dit gedemonstreerd door enkele voorbeeldtabellen te presenteren van lexeemparadigma's en van inventarisaties van wortels en prefixstammen. De bijbehorende datarapporten werden ook aan een interne evaluatie onderworpen. Deze evaluatie wees uit dat zowel de wortels als de prefixstammen een redelijk voorspelbare distributie vertonen, wat een aanwijzing is dat de staminformatie al een hoog kwaliteitsniveau heeft bereikt.

6.5 Inventarisatie van prefixen en hun combinatoriek

6.5.1 Introductie

Deze sectie biedt een eerste kennismaking met constructiewijze en samenstelling van een reeks databestanden die zijn voortgekomen uit de doelstelling om een inventarisatie op te bouwen van alle prefix-eenheden en prefix-sequenties die deel uitmaken van het MGBN-lexicon en onderzoek te doen naar hun combinatorische eigenschappen. In het kader van deze doelstelling heb ik behalve het complete lexeemdomein ook specifieke deeldomeinen geanalyseerd (zoals het domein van de als zelfstandig woord toepasbare lexemen), terwijl ik bij de analyse van de interne opbouw van de prefixsequenties zowel een links-rechts-perspectief als een rechts-links-perspectief heb toegepast. H6.5.2 biedt algemene informatie over de opzet en samenstelling van de resulterende datarapporten. H6.5.3 geeft een indruk van de aldus verkregen kenmerkinventarisaties aan de hand van enkele voorbeeldtabellen. In H6.5.4 wordt uiteengezet hoe de externe evaluatie is aangepakt en wat deze voor resultaten heeft opgeleverd. Tot slot volgt een conclusie.

6.5.2 Opzet

In deze subsectie zal ik stilstaan bij opzet en interne samenstelling van de datarapporten die de basis vormen voor mijn onderzoek naar de prefixdistributie in het MGBN-model. Hiertoe zal ik eerst uiteenzetten wat het globale idee is achter deze datarapporten, om vervolgens een overzicht te geven van de belangrijkste informatievelden, waarbij ik elk veld kort zal toelichten. Tot slot zal ik uiteenzetten welke keuzes mogelijk zijn met betrekking tot het te analyseren domein en de rapportagekenmerken (zoals sorteeropties en datafilters), waarbij ik ook zal aangeven welke analysemogelijkheden hieruit voortvloeien.

Mijn datarapporten met betrekking tot de prefixdimensie van het MGBN-model hebben als doel om deze dimensie zo gedetailleerd mogelijk te beschrijven en zo een basis te leggen voor externe evaluaties en voor statistisch onderzoek naar de onderliggende structuurcriteria (mede ten behoeve van interne evaluaties). Meer in het bijzonder was mijn analysemethode erop gericht om informatie te verzamelen over alle prefixeenheden en prefixsequenties die deel uitmaken van het MGBN-model en deze weer te geven door middel van morfologisch gestructureerde representaties van hun vormkenmerken (van spelvorm tot algemene vormsleutel), om per formele patroonklasse informatie te geven over de combinatorische eigenschappen (zowel vanuit een links-rechts-perspectief als vanuit een rechts-links-perspectief) en enkele morfosyntactische kenmerken (namelijk de morfologische klasse en de etymologische klasse), en om voor al deze kenmerken kwantitatieve gegevens te verstrekken, in het bijzonder de lexicale typefrequentie op lexeemniveau (zowel absoluut als relatief) en de omvang van het inwaartse en het uitwaartse toepassingsdomein.

Ter verduidelijking van de hier beschreven opzet bspreek ik een concreet analysevoorbeeld, namelijk de structuuranalyse van het A-lexeem *ongecompliceerd*. In de MGBN bezit dit lexeem de structuurrepresentatie ON_GE_COM_[PLIC]_EER_D. Het prefixdeel correspondeert dus met een 3-ledige prefixsequentie, namelijk ON_GE_COM. Deze sequentie kan zowel van links naar rechts als van rechts naar links worden geanalyseerd. Hierbij correspondeert elke analysestap met een aparte regel in het datarapport. In de tabellen 6-8 en 6-9 worden beide analyseperspectieven gedemonstreerd.

Verklaring afkortingen
maxseq = maximale prefixsequentie (vanaf het eerste prefix)
partseq = partiële prefixsequentie (ten opzichte van maxseq)
inv(maxseq) = inverse-weergave van maxseq-patroon (t.b.v. rl-positiebepaling)

lengte maxseq	lengte partseq	eerste prefix	lr-partseq	maxseq	inv(maxseq)
3	1	com	on	on_ge_com	com_ge_on
3	2	com	on_ge	on_ge_com	com_ge_on
3	3	com	on_ge_com	on_ge_com	com_ge_on

Tabel 6-7: demonstratie van de links-rechts-analyse van de prefixdimensie aan de hand van de n1-prefixsequentie ON_GE_COM in ongecompliceerd.

lengte maxseq	lengte partseq	eerste prefix	rl-partseq	maxseq	inv(maxseq)
3	1	com	com	on_ge_com	com_ge_on
3	2	com	ge_com	on_ge_com	com_ge_on
3	3	com	on_ge_com	on_ge_com	com_ge_on

Tabel 6-8: demonstratie van de rechts-links-analyse van de prefixdimensie aan de hand van de n1-prefixsequentie ON_GE_COM in ongecompliceerd.

Het is ook mogelijk om de eerste of laatste eenheid van een prefixsequentie door een variabele te vervangen. Deze variabele, die zowel hier als in mijn datarapporten met het teken @ correspondeert, komt altijd in de plaats van een bestaand prefix; bij een links-rechts-perspectief dient deze variabele aan de rechterkant te staan (bijv. a1+a2+@), bij een rechts-links-analyse aan de linkerkant (bijv. @+a2+a1). Op deze manier kunnen prefixtypes worden geconstrueerd waarvoor geldt dat de prefixen gegarandeerd een voorgaand of volgend affix selecteren (afhankelijk van de plaats van @), terwijl onbekend blijft hoeveel prefixen weer voor of achter die laatste positie kunnen worden aangehecht. Dergelijke prefixtypes zijn handig om te generaliseren over een reeks laagfrequente prefixcombinaties waarvan het eerste of laatste prefix wel hoogfrequent is, zoals de prefixsequentie @+GE (bijv. AAN_GE, VOOR_GE, UIT_GE, BE_GE etc.). Ik zal dit principe demonstreren voor het lexeem *onaangedaan* (met de structuur ON_AAN_GE_[DAAN]). Hierbij corresponderen de rechte haken (]) met de rechtergrens van de wortel en de hekjes (#) met de lexeemgrens. In de rapporten 5a en 5b wordt getoond op welke posities een variabele kan worden geplaatst en welk prefix (onder meer) als specificator kan dienen. Het patroon @+GE correspondeert dus met de optie prefix 3 van het lr-perspectief; ter verduidelijking van de positiebepaling onder de rl-analyse hebben de prefixpatronen een inverse weergave gekregen.

lr-analyse	**pos 0**	**pos 1**	**pos 2**	**pos 3**	**pos 4**	(5a)
prefix 1	#	on	aan	ge	[	
prefix 2	#	@	aan	ge	[	
prefix 3	#		@	ge	[	

rl-analyse	**pos -0**	**pos 1**	**pos 2**	**pos 3**	**pos 4**	(5b)
prefix 1	[	ge	aan	on	#	
prefix 2	[	@	aan	on	#	
prefix 3	[		@	on	#	

In de datarapporten is systematisch in kaart gebracht welke prefixen er voorkomen, welke vormvarianten elk prefix kent en welke prefixsequenties hiermee geconstrueerd kunnen worden. Bovendien wordt bij elke prefixtoepassing aangegeven wat de bijbehorende typefrequentie is (naast andere kwantitatieve gegevens), en welk aandeel deze patronen hebben in de typefrequentie van de centrale eenheid (d.w.z. de eenheid waar de analyse op is gericht). De inhoud van de datarapporten is afhankelijk van het gehanteerde queryprofiel. Dit profiel kent tal van vrij te kiezen parameters, waaronder het taxeemniveau (woorden, lexemen of sublexemen, wel/geen samenstellingen, zelfstandige versus niet-zelfstandige lexemen en wel/geen

beperking tot hedendaags Nederlands (c.q. WHN); maar het kan ook om frequentie-filters gaan. Hierdoor wordt het mogelijk om in te zoomen op specifieke deeldomeinen en om deze domeinen met elkaar te vergelijken. Binnen dit deeldomein correspondeert elke dataregel met een unieke patroonspecificatie (al kan het weergegeven patroon best meerdere keren voorkomen). Zo'n patroonspecificatie geeft aan welk perspectief is gehanteerd bij de selectie van de combinatorische kenmerken en bij de specificatie van de kwantitatieve eigenschappen. Deze patroonspecificaties bestaan (onder meer) uit de volgende kenmerken:

- vrije prefixlengte (= lengte 0) versus specifieke prefixlengte (lengte 1, 2, 3, etc.)
- aantal prefix-eenheden in de getoonde prefixsequentie
- positie van het eerste prefix (gegeven de analyserichting)
- status van laatste prefixpositie in prefixsequentie: variabele vs. specifiek prefix
- representatieniveau: n1 = spelvorm, n2 = 1e vormsleutel, n3 = 2e vormsleutel

Het op deze kenmerken gebaseerde classificatiesysteem maakt het mogelijk om per prefix zeer gedetailleerde informatie te verstrekken over de morfologische combinatiepatronen en de bijbehorende gebruiksfrequenties. In de volgende subsectie (6.5.3) zal ik dit demonstreren door enkele voorbeeldtabellen te presenteren met informatie uit de resulterende datarapporten.

6.5.3 Resultaten voor de prefixdimensie

Zie appendix B.2

6.5.4 Externe evaluatie

Om enig zicht te krijgen op de externe kwaliteit van de prefixdimensie van het MGBN-model heb ik een evaluatie-onderzoek uitgevoerd waarbij ik de informatie uit de in H6.6.3 gepresenteerde datarapporten langs computationele weg met de suffixgegevens in het Morfologisch Handboek heb vergeleken. Zoals ik reeds uiteen heb gezet, kennen deze informatiebronnen verschillende doelstellingen, waardoor het weinig zin heeft om ze integraal met elkaar te vergelijken. Om die reden heb ik me beperkt tot een onderzoek naar de wederzijdse dekking van prefixen op het niveau van de hoofdtypes (c.q. klankvorm). In het kader van deze vergelijking heb ik ook onderzoek gedaan naar de invloed van basisparameters als analyseperspectief, sequentielengte en typefrequentie.

Zie Appendix B.2.6 voor de evaluatieresultaten.

6.5.5 Conclusie

De morfologische structuurinformatie in de MGBN maakt het mogelijk om zeer gedetailleerd onderzoek te doen naar de distributie van Nederlandse prefixen en hun combinatorische eigenschappen. In deze sectie heb ik dit gedemonstreerd door enkele voorbeeldtabellen te presenteren met prefixkenmerken. De bijbehorende datarapporten zijn ook aan een externe evaluatie onderworpen. In dit kader heb ik onderzocht in hoeverre de prefixtypes uit de MGBN overeenkomen met die in het MHB op het punt van de orthografische vorminformatie en de productiviteitsgegevens; de categoriale structuurkenmerken zijn hier echter buiten beschouwing gebleven, want in mijn visie zijn deze alleen zinvol te definiëren in combinatie met de suffixcomponent (zie verder H6.8). De hier genoemde evaluatie wees uit dat de MGBN een veel groter bereik heeft dan het MHB en ook omvangrijker is in het meest relevante vergelijkingsdomein. Alle MHB-types zijn direct (63%), indirect (16%) of als laagfrequent prefix (21%) in de MGBN terug te vinden. Omgekeerd geldt dat de MGBN-inventarisatie slechts voor 50% resp. 60% door het MHB wordt gedekt (indien men de vergelijking beperkt tot losse prefix-eenheden met een frequentie van 5 resp. 10). Hiernaast heeft de MGBN een veel groter bereik, want behalve de prefix-eenheden en hun vormvarianten biedt de MGBN ook een complete inventarisatie van prefixcombinaties.

6.6 Inventarisatie van suffixen en hun combinatoriek

6.6.1 Introductie

Deze sectie (die op dezelfde wijze is opgezet als H6.5) biedt een eerste kennismaking met constructiewijze en samenstelling van een reeks databestanden die zijn voortgekomen uit de doelstelling om een inventarisatie op te bouwen van alle suffix-eenheden en suffix-sequenties die deel uitmaken van het MGBN-lexicon en onderzoek te doen naar hun combinatorische eigenschappen. In het kader van deze doelstelling heb ik behalve het complete lexeemdomein ook specifieke deeldomeinen geanalyseerd, en verschillende perspectieven toegepast. H6.6.2 biedt algemene informatie over de opzet en samenstelling van de resulterende datarapporten. H6.6.3 geeft een indruk van de aldus verkregen kenmerkinventarisaties aan de hand van enkele voorbeeldtabellen. In H6.6.4 wordt uiteengezet hoe de externe evaluatie is aangepakt en wat deze voor resultaten heeft opgeleverd. De sectie over de interne evaluatie (H6.6.5) richt zich op de vraag in hoeverre de prefixdistributie voorspelbare patronen vertoont. Tot slot volgt een conclusie (H6.6.6).

6.6.2 Opzet

In deze subsectie bespreek ik de opzet en samenstelling van de datarapporten die de basis vormen voor mijn onderzoek naar de suffixdistributie in het MGBN-model. Net als bij de behandeling van de prefixrapporten zal ik eerst uiteenzetten wat het globale idee is achter deze datarapporten, om vervolgens een overzicht te geven van de belangrijkste informatievelden, waarbij ik elk veld kort zal toelichten. Verder zal ik uiteenzetten welke keuzes mogelijk zijn met betrekking tot het te analyseren domein en de rapportagekenmerken (zoals sorteeropties en datafilters), waarbij ik ook angeef welke analysemogelijkheden hieruit voortvloeien.

Mijn datarapporten met betrekking tot de suffixdimensie van het MGBN-model hebben als doel om deze dimensie zo gedetailleerd mogelijk te beschrijven en zo een basis te leggen voor externe evaluaties en voor statistisch onderzoek naar de onderliggende structuurcriteria (mede ten behoeve van interne evaluaties). Meer in het bijzonder was mijn analysemethode erop gericht om informatie te verzamelen over alle suffixeenheden en suffixsequenties die deel uitmaken van het MGBN-model en deze weer te geven door middel van morfologisch gestructureerde representaties van hun vormkenmerken (van spelvorm tot algemene vormsleutel), om per formele patroonklasse informatie te geven over de combinatorische eigenschappen (zowel vanuit een links-rechts-perspectief als vanuit een rechts-links-perspectief) en enkele morfosyntactische kenmerken (namelijke de categoriale functie en de etymologische klasse), en om voor al deze kenmerken kwantitatieve gegevens te verstrekken, in het bijzonder de lexicale typefrequentie op lexeemniveau (zowel absoluut als relatief), de u-potentie van de stam en de omvang van het inwaartse en het uitwaartse toepassingsdomein.

Ter verduidelijking van de hier beschreven opzet zal ik nu een concreet analysevoorbeeld bespreken, namelijk de structuuranalyse van het N-lexeem *compositionaliteit*. In de MGBN bezit dit lexeem de structuurrepresentatie COM_[POS]_IT_ION_AL_IT_EIT. Het suffixdeel correspondeert dus met een 5-ledige suffixsequentie, namelijk IT_ION_AL_IT_EIT. Deze sequentie kan zowel van links naar rechts als van rechts naar links worden geanalyseerd. Hierbij correspondeert elke analysestap met een aparte regel in het datarapport. In de tabellen 6-10 en 6-11 worden beide analyseperspectieven gedemonstreerd.

Om het voorbeeld niet te ingewikkeld te maken, zal ik me beperken tot de analyse van een deelsequentie, namelijk de 4-ledige sequentie ION_AL_IT_EIT. Dit draagt tevens bij aan de uitleg van de analysemethode, want in de datarapporten wordt ook informatie gegeven over de distributie van deelsequenties. Om dit laatste mogelijk te maken heb ik een aparte parameter geïntroduceerd (namelijk [±var], als aanduiding van "variabele affixgrens"), waar-

mee kan worden aangegeven of de geanalyseerde affixsequentie met de hele affixcomponent correspondeert (= [-var]) of alleen met een deel van die affixcomponent (= [+var]). Indien sprake is van een [+var]-analyse blijft onbekend hoever de sequentie doorloopt, omdat het doel van deze analyse eruit bestaat om de frequentie te bepalen van alle toepassingen van de betreffende affixsequentie (in plaats van een analyse waarbij alleen de "complete" suffixtoepassingen worden geteld).

Bij een complete suffixsequentie hoeft de eerste suffixpositie van de subsequentie overigens niet overeen te komen met de eerste positie van de complete suffixsequentie (gegeven het gekozen analyseperspectief). Men kan de analyse namelijk ook vanaf een andere positie laten beginnen. Deze zelfde mogelijkheid bestaat ook bij analyses van niet-complete suffix-sequenties. In het hier gepresenteerde analysevoorbeeld begint de links-rechts-analyse bij-voorbeeld bij het tweede suffix, te weten ION (blijkens de specificatie [2] in de kolom "eerste positie"). Maar in de tabel met de rechts-links-analyse begint de analyse bij het eerste suffix, te weten -EIT. Toch bezitten beide voorbeeldanalyses het kenmerk [-tot]; dit geeft aan dat de opgegeven sequentielengte (namelijk "lengte maxseq") geen garantie biedt dat de eindpositie van de betreffende suffixsequentie tevens de eindpositie is van de suffixcomponent van de geanalyseerde lexemen. Dit is toevallig wel het geval bij de links-rechts-analyse (want de laatste lr-positie in de suffixsequentie ION_AL_IT_EIT valt samen met het laatste suffix in de suffixcomponent van de structuurrepresentatie van het lexeem *compositionaliteit*) maar niet bij de rechts-links-analyse, want in IT_ION_AL_IT_EIT wordt het suffix -ION nog gevolgd (c.q. voorafgegaan) door het suffix -IT.

Verklaring veldnamen
grens-status: een [+var]-patroon heeft een variabele suffixgrens, een [-var]-patroon niet.
seqpos = absolute suffixpositie binnen een suffixsequentie
maxseq = maximale suffixsequentie (vanaf het eerste prefix)
partseq = partiële suffixsequentie (ten opzichte van maxseq)
inv(maxseq) = inverse-weergave van maxseq-patroon (t.b.v. rl-positiebepaling)

wel/niet totaal	eerste positie	lengte maxseq	lengte partseq	eerste prefix	rl-partseq	maxseq	inv(maxseq)
[+var]	2	4	1	ion	ion	ion_al_it_eit	eit_it_al_ion
[+var]	2	4	2	ion	ion_al	ion_al_it_eit	eit_it_al_ion
[+var]	2	4	3	ion	ion_al_it	ion_al_it_eit	eit_it_al_ion
[+var]	2	4	4	ion	ion_al_it_eit	ion_al_it_eit	eit_it_al_ion

Tabel 6-10: demonstratie van de links-rechts-analyse van de suffixdimensie aan de hand van de n1-suffixsequentie ION_AL_IT_EIT in het lexeem compositionaliteit.

grens-status	eerste seqpos	lengte maxseq	lengte partseq	eerste prefix	rl-partseq	maxseq	inv(maxseq)
[+var]	1	4	1	eit	eit	ion_al_it_eit	eit_it_al_ion
[+var]	1	4	2	eit	it_eit	ion_al_it_eit	eit_it_al_ion
[+var]	1	4	3	eit	al_it_eit	ion_al_it_eit	eit_it_al_ion
[+var]	1	4	4	eit	ion_al_it_eit	ion_al_it_eit	eit_it_al_ion

Tabel 6-11: demonstratie van de rechts-links-analyse van de suffixdimensie aan de hand van de n1-suffixsequentie ION_AL_IT_EIT in het lexeem compositionaliteit.

Het is ook interessant om de eerste of laatste eenheid van een suffixsequentie door een variabele te vervangen. Deze variabele, die zowel hier als in mijn datarapporten met het teken @ correspondeert, komt dus altijd in de plaats van een bestaand suffix; bij een links-rechts-perspectief dient deze variabele aan de rechterkant te staan (bijv. a1+a2+@), bij een rechts-

links-analyse aan de linkerkant (bijv. @+a2+a1). Op deze manier kunnen suffixtypes worden geconstrueerd waarvoor geldt dat de suffixen gegarandeerd een voorgaand of volgend suffix selecteren (afhankelijk van de plaats van @), terwijl onbekend blijft hoeveel suffixen weer voor of achter die laatste positie kunnen worden aangehecht. Dergelijke suffixtypes zijn handig om te generaliseren over een reeks laagfrequente suffixcombinaties waarvan het eerste of laatste suffix wel hoogfrequent is, zoals de suffixsequentie @+EEL (bijv. ION_EEL, FIC_(I)EEL, ANT_(I)EEL, MENT_EEL etc.). Ik zal dit principe demonstreren voor het lexeem *compositioneel* (met de structuur COM_[POS]_IT_ION_EEL). Hierbij corresponderen de rechte haken (]) met de rechtergrens van de wortel en de hekjes (#) met de lexeemgrens. In de onderstaande tabel wordt getoond op welke posities een variabele kan worden geplaatst en welk suffix (onder meer) als specificator kan dienen. Het patroon @+eel correspondeert dus met de optie suffix 3 van het rl-perspectief; ter verduidelijking van de positiebepaling onder de rl-analyse hebben deze suffixpatronen een inverse weergave gekregen.

lr-perspectief	**pos 0**	**pos 1**	**pos 2**	**pos 3**	**pos 4**	(6a)
suffix 1	]	it	ion	eel	#	
suffix 2	]	it	ion	@	#	
suffix 3	]	it	@		#	
suffix 4	]	@			#	

rl-perspectief	**pos 0**	**pos 1**	**pos 2**	**pos 3**	**pos 4**	(6b)
suffix 1	#	eel	ion	it	]	
suffix 2	#	eel	ion	@	]	
suffix 3	#	eel	@		]	
suffix 4	#	@			]	

Bij de analyse van de suffixsequenties heb ik ook informatie verzameld met betrekking tot de categoriale eigenschappen van de subsequenties (van eenheden tot complete sequenties). Zo ga ik er (op basis van mijn nieuwe visie op het rechterhoofdprincipe, zie hoofdstuk 3) vanuit dat de suffixsequentie IT_EIT in *compositionaliteit* met de lexeemcategorie N correspondeert, aangezien dit suffix op eindpositie staat, terwijl het bijbehorende lexeem de categorie N draagt. In mijn visie op morfologie correspondeert deze lexeemcategorie met een aparte functor, namelijk de begrenzer van het lexeemdomein. Hiërarchisch gezien staat deze domeinbegrenzer hoger dan alle lexeeminterne affixen. Voor het gemak vertaal ik dit in een representatie waarbij de suffxcomponent door een extra morfeem wordt gevolgd, namelijk de domeinbegrenzer; toegepast op het suffix IT_EIT leidt dit tot het patroon IT_EIT_#N, met de domeinbegrenzer #N. Er zijn ook suffixen die meerdere categorieën kunnen selecteren. In dat geval heb ik elke categorietoepassing apart geïnventariseerd.

In dit verband dient onderscheid te worden gemaakt tussen analyses op basis van affixsequenties met een "harde" (lexicale) categoriespecificatie en affixsequenties met een "zachte" (potentiële) categoriespecificatie. Dit is van belang met het oog op de categoriale analyse van subsequenties die niet op eindpositie staan, maar die wel deze mogelijkheid kennen. Zo treft men het suffix AAL vaak op eindpositie aan, blijkens lexemen als *tonaal* (met de structuur [TON]_ AAL) en *terminaal* (met de structuur [TERM]_IN_AAL), waarbij soms sprake is van de vormvariant EEL zoals in *rationeel* (met de structuur [RAT]_ION_EEL) en *compositioneel* (met de structuur COM_[POS]_IT_ION_EEL). In al deze gevallen is het mogelijk om een complexer lexeem te vormen door toevoeging van het suffix IT_EIT, bijvoorbeeld *tonaliteit* (met de structuur [TON]_AL_IT_EIT).

In de gangbare morfologiebenaderingen wordt aangenomen dat ingebedde suffixen dezelfde categorie markeren als op eindpositie. In mijn eigen visie is dit alleen correct met betrekking tot de morfeemcategorie, niet met betrekking tot de lexeemcategorie (die de basis vormt voor

inflectietoekenning), want deze kan alleen op lexeemeinde worden geactiveerd. Maar in de praktijk is het verschil niet zo groot, want ik ga ervan uit dat de morfeemcategorie en de lexeemcategorie in elkaars verlengde liggen. Gegeven deze aanname wordt het mogelijk om de categoriale eigenschappen van de affixen op middenpositie langs computationele weg af te leiden uit de categoriekenmerken van de morfemen op eindpositie. Maar omdat veel stammen en affixen met lexemen corresponderen die meerdere categorieën kunnen aannemen, heeft de op deze wijze verkregen categorie-informatie een zachte status, namelijk de status van potentiële categorie, in tegenstelling tot de categorie van eenheden op eindpositie, die een harde, lexicale status heeft.

In de datarapporten is systematisch in kaart gebracht welke suffixen er voorkomen, welke vormvarianten elk suffix kent en welke suffixsequenties hiermee geconstrueerd kunnen worden. Bovendien wordt bij elke suffixtoepassing aangegeven wat de bijbehorende typefrequentie is (naast andere kwantitatieve gegevens), en welk aandeel deze patronen hebben in de typefrequentie van de centrale eenheid (d.w.z. de eenheid waar de analyse op is gericht). De inhoud van de datarapporten is afhankelijk van het gehanteerde queryprofiel. Dit profiel kent tal van vrij te kiezen parameters, waaronder het taxeemniveau (woorden, lexemen of sublexemen, wel/geen samenstellingen, zelfstandige/niet-zelfstandige lexemen en wel/geen beperking tot hedendaags Nederlands (c.q. WHN); maar het kan ook om frequentie-filters gaan. Hierdoor wordt het mogelijk om in te zoomen op specifieke deeldomeinen en om deze domeinen met elkaar te vergelijken. Binnen dit deeldomein correspondeert elke dataregel met een unieke patroonspecificatie (al kan het weergegeven patroon best meerdere keren voorkomen). Zo'n patroonspecificatie geeft aan welk perspectief is gehanteerd bij de selectie van de combinatorische kenmerken en bij de specificatie van de kwantitatieve eigenschappen. Deze patroonspecificaties bestaan (onder meer) uit de volgende kenmerken:

- vrije suffixlengte (= lengte 0) versus specifieke suffixlengte (lengte 1, 2, 3, etc.)
- aantal suffix-eenheden in de getoonde suffixsequentie
- positie van het eerste suffix (gegeven de analyserichting)
- status van laatste suffixpositie in suffixsequentie: variabele vs. specifiek suffix
- suffixen met lexicale versus suffixen met potentiële lexeemcategorie
- representatieniveau: n1 = spelvorm, n2 = 1e vormsleutel, n3 = 2e vormsleutel

Het op deze kenmerken gebaseerde classificatiesysteem maakt het mogelijk om per suffix zeer gedetailleerde informatie te verstrekken over de morfologische combinatiepatronen en de bijbehorende gebruiksfrequenties. In de volgende subsectie (6.6.3) zal ik dit demonstreren door enkele voorbeeldtabellen te presenteren met informatie uit de resulterende datarapporten.

6.6.3 Resultaten

Zie appendix B.3

6.6.4 Externe evaluatie

Om enig zicht te krijgen op de externe kwaliteit van de suffixdimensie van het MGBN-model heb ik een evaluatie-onderzoek uitgevoerd waarbij ik de informatie uit de in H6.6.3 gepresenteerde datarapporten langs computationele weg met de suffixgegevens in het Morfologisch Handboek heb vergeleken. Zoals ik reeds uiteen heb gezet, kennen deze informatiebronnen verschillende doelstellingen, waardoor het weinig zin heeft om ze integraal met elkaar te vergelijken. Om die reden heb ik me beperkt tot de vergelijking van een aantal specifieke kenmerken, te weten:

i. de wederzijdse dekking van suffixen op het niveau van de hoofdtypes (c.q. klankvorm)
ii. de wederzijdse dekking van suffixen op het niveau van de ucat-types (c.q. klankvorm)
iii. de wederzijdse dekking van suffixen op het niveau van de icat-types (c.q. klankvorm)

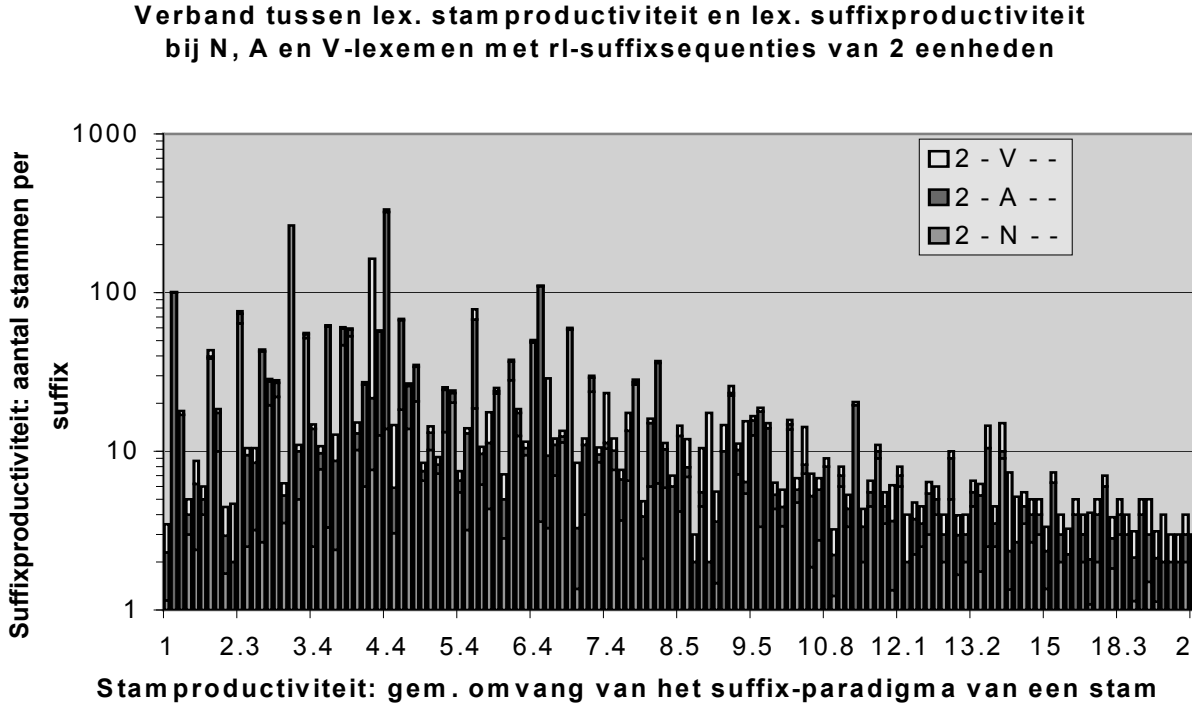
In het kader van deze vergelijking heb ik ook onderzoek gedaan naar de invloed van basis-parameters als analyseperspectief, sequentielengte en typefrequentie.

Zie Appendix B.3.6 voor de evaluatieresultaten.

6.6.5 Interne evaluatie

6.6.5.1 Distributiepatronen

De hieronder weergegeven grafieken zijn rechtstreeks afgeleid van de suffixkenmerken uit de in deze sectie besproken datarapporten. Deze grafieken spreken verder voor zich.

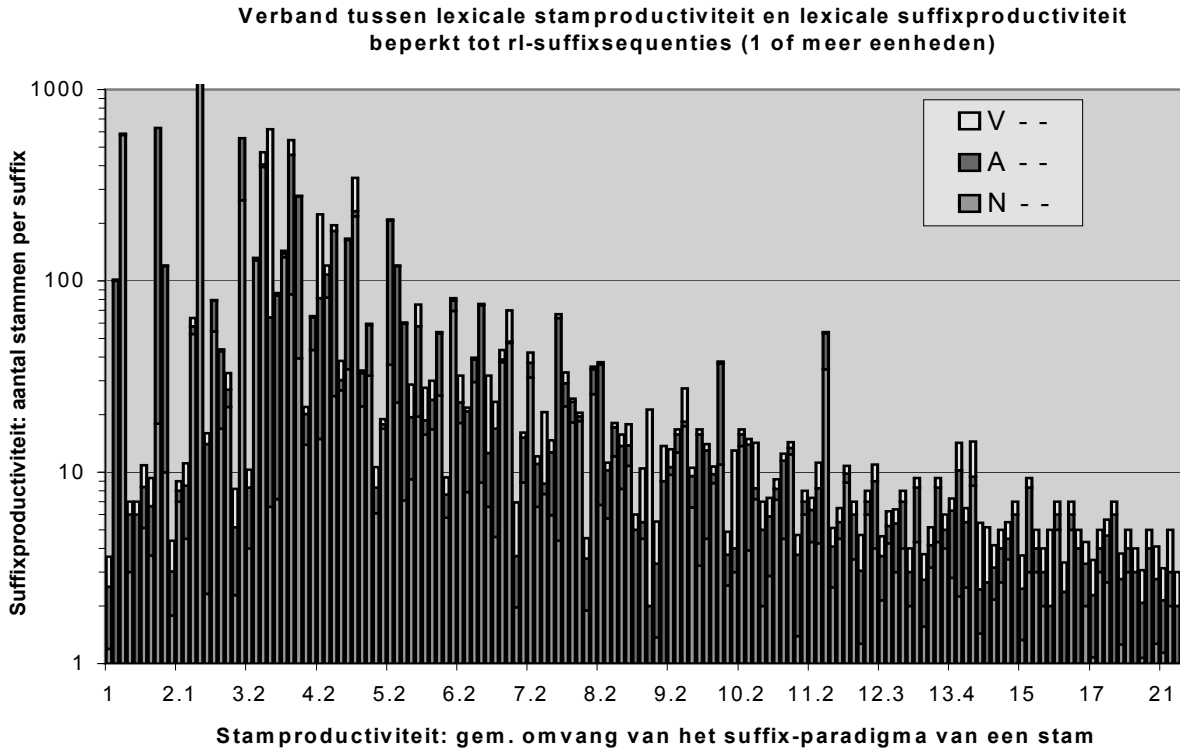


Figuur 6-9: Grafiek met de lexicale distributie van wortelstammen

6.6.5.2 Discussie

De hier gepresenteerde grafieken laten zien dat de suffixdistributie van de MGBN een vrij voorspelbaar patroon volgt, dat kan worden uitgedrukt in termen van een functie. Dit is een duidelijke aanwijzing dat de suffixkenmerken niet ad hoc zijn toegekend, maar op een cognitieve systematiek berusten. In dit verband zou men de volgende hypthese kunnen overwegen:

Hypothese Het lexicon van het MGBN-model kenmerkt zich door het feit dat er met betrekking tot suffixen een correlatie bestaat tussen de omvang van het stamdomein van het suffix en de gemiddelde omvang van de suffixparadigma's die aan de stammen van dit suffix zijn verbonden (c.q. de substitutiekans). Hoe kleiner het stamdomein, hoe hoger de substitiekans. Met andere woorden: hoe minder suffixen een stam kan selecteren, hoe groter de kans dat deze suffixen hoogfrequent zijn. Omgekeerd geldt voor stammen met een omvangrijk suffix-paradigma dat er vaak een of meer bijzondere (d.w.z. laagfrequente) suffixen tussen zullen zitten. Men zou ook kunnen zeggen dat de identiteit van een normaalfrequente stam bepaald wordt door de laagfrequente affixen. Bij de hoogfrequente stammen is waarschijnlijk een nadere uitsplitsing mogelijk naar betekenis.



Figuur 6-10: Grafiek met het verbamd tussen lexicale stamproductiviteit en lexicale suffixproductiviteit.

6.6.6 Conclusie

De morfologische structuurinformatie in de MGBN maakt het mogelijk om zeer gedetailleerd onderzoek te doen naar de distributie van Nederlandse suffixen en hun combinatorische eigenschappen. In deze sectie heb ik dit gedemonstreerd door enkele voorbeeldtabellen te presenteren met suffixkenmerken. De bijbehorende datarapporten zijn ook aan een externe evaluatie onderworpen. In dit kader heb ik onderzocht in hoeverre de suffixtypes uit de MGBN overeenkomen met die in het MHB op het punt van de orthografische vorminformatie en de categoriale specificaties. Deze evaluatie wees uit dat de MGBN een veel groter bereik heeft dan het MHB en ook omvangrijker is in het meest relevante vergelijkingsdomein (met suffixen van 1 of 2 eenheden en een minimum typefrequentie van 5). Alle MHB-types op het niveau van de hoofdvorm zijn namelijk direct (78%), indirect (12 %) of als laagfrequent type (8%) in de MGBN terug te vinden. Omgekeerd geldt dat de MGBN-inventarisatie slechts voor 26% door het MHB wordt gedekt (en 40% indien men het domein beperkt tot suffixen met een typefrequentie van 10 of meer). Op het niveau van de suffixtypes met categorie dekt de MGBN 85% van de MHB-types, terwijl 15% niet rechtstreeks is terug te vinden of een te lage frequentie heeft. Omgekeerd geldt dat het MHB slechts 25% (freq. 5) resp. 37% (freq. 10) van de MGBN-types met categorie dekt (en slechts 20% resp. 30% van de i-u-functies). Hiernaast heeft de MGBN een veel groter bereik, want behalve de 1- en 2-ledige suffixen en hun vormvarianten biedt de MGBN ook een complete inventarisatie van suffixsequenties.

6.7 Inventarisatie van prefix-suffix-combinaties

6.7.1 Introductie

Deze sectie biedt een eerste kennismaking met constructiewijze en samenstelling van de datarapporten die zijn voortgekomen uit de doelstelling om een inventarisatie op te bouwen van alle prefix-suffix-combinaties die deel uitmaken van het MGBN-lexicon en onderzoek te doen naar de categoriale effecten van de prefixcomponent. Bij de constructie van deze klasse van datarapporten heb ik een hoofdindeling gehanteerd die uitgaat van het aan de lexemen verbonden morfeempatroon, namelijk het totaal aantal morfemen en de interne verdeling van prefix-eenheden en suffix-eenheden. H6.7.3 geeft een indruk van het aldus tot stand gekomen datarapport aan de hand van enkele voorbeeldtabellen. Voor nadere informatie over de inhoud van deze tabellen en de door mij gehanteerde analysemethode kan men in H6.7.2 terecht. In het kader van de externe evaluatie (H6.7.4) heb ik onderzocht in hoeverre de via een speciale wegingsmethode geconstrueerde informatie over de categoriale invloed van de prefixen overeenkomt met de door het Morfologisch Handboek verstrekte informatie. De sectie eindigt met een korte conclusie (6.7.6).

6.7.2 Opzet

In deze subsectie zal ik stilstaan bij opzet en interne samenstelling van de datarapporten die de basis vormen voor mijn onderzoek naar prefix-suffix-combinaties in het MGBN-model. Hiertoe zal ik eerst uiteenzetten wat het globale idee is achter deze datarapporten en hoe dit doel werd gerealiseerd, om vervolgens een overzicht te geven van de belangrijkste informatievelden, waarbij ik elk veld kort zal toelichten. Verder zal ik enige aandacht besteden aan de analysemogelijkheden.

Mijn datarapporten met betrekking tot prefix-suffix-combinaties hebben als doel om een complete inventarisatie te bieden van de prefix-suffix-combinaties die deel uitmaken van het MGBN-model, om deze patronen te classificeren met betrekking tot hun interne morfeempatroon (door het aantal prefixen en suffixen te bepalen) en de lexeemcategorie, om voor elk patroon een lexeeemtoepassing te specificeren, om informatie te geven over de categoriale effecten van het prefix, en om voor elk patroon kwantitatieve gegevens te verstrekken, waaronder het totale aantal stamtoepassingen. Ik zal de hier beschreven opzet toelichten aan de hand van een concreet voorbeeld, namelijk het A-lexeem *ongecompliceerd* met de morfeemrepresentatie ON_GE_COM_[PLIC]_EER_D. Gegeven deze representatie kan men het prefix-suffix-patroon achterhalen door de wortel weg te laten. Dit levert een patroon op met de structuur ON_GE_COM_[-]_EER_D (A), waarbij de component [-] aangeeft dat de wortel is weggelaten en dus vrijelijk kan worden gespecificeerd.

Omdat het lexeem *ongecompliceerd* een antoniem is van het lexeen *gecompliceerd*, kan het hier onderzochte prefix-suffix-patroon als een gemodificeerde versie van het lexeempatroon GE;COM;[-];EER;D (A) worden geïnterpreteerd, namelijk modificatie door het negatie-prefix ON-. De categorie van het basislexeem is dus identiek aan die van het door ON- gemodificeerde lexeem. Hieruit volgt dat er minstens 1 instantie bestaat van het patroon {P + GE_COM_[-]_EER_D} waarvoor geldt dat het prefix P = ON- met de categoriale functie <A,A> zou kunnen corresponderen (wat ook de gangbare analyse is). Of dit ook een houdbare analyse is, hangt af van de vraag hoe systematisch dit verband is en hoe vaak het voorkomt. Indien het betreffende patroon slechts incidenteel een niet-gemodificeerde tegenhanger kent, is het geen goede kandidaat voor de definitie van een lexicaal patroon. Dit bezwaar geldt ook indien het betreffende patroon maar een klein toepassingsdomein kent (bijv. minder dan 10 wortels). Indien er echter sprake is van een constant prefix-effect op de categoriale functie is dit een aanwijzing dat de bijbehorende categoriefunctie een vast kenmerk is van het betref-

fende prefix. Dit kan worden geverifieerd door na te gaan voor hoeveel procent van de prefix-suffix-combinaties sprake is van een significant categorie-effect (met dezelfde categoriale functie). In mijn optiek biedt dit eenvoudig te berekenen gegeven een vrij solide basis voor een vergelijking met de MHB-informatie over prefix-effecten. Het MHB gaat ervan uit dat dergelijke prefix-effecten onafhankelijk zijn van de suffix-context, maar in mijn visie is dit geen relevant gegeven. Om een vergelijking mogelijk te maken (zie H6.7.4), heb ik ervoor gezorgd dat dit gegeven toch beschikbaar is.

6.7.3 Resultaten

Zie appendix B.4

6.7.4 Externe evaluatie

Zie appendix B.4.5

6.7.5 Conclusie

De morfologische structuurinformatie in de MGBN maakt het mogelijk om zeer gedetailleerd onderzoek te doen naar de morfeemstructuur van Nederlande lexemen en de distributie van Nederlandse prefix-suffix-combinaties (zowel op formeel klasseniveau als op type-niveau). In deze sectie heb ik dit gedemonstreerd door voorbeeldtabellen te presenteren met prefix-suffix-combinaties, namelijk voor lexemen met 4, 7 en 8 morfemen (wat het maximum is). Het door mij vervaardigde datarapport biedt ook informatie over het categoriale effect van het eerste prefix van lexemen met minimaal 1 prefix; dit wordt gecodeerd door middel van een functie van invoercategorie (te weten het lexeem zonder dit prefix) naar uitvoercategorie (te weten het lexeem met prefix). Door over deze suffixgevoelige functies te generaliseren kan een suffixonafhankelijk prefixeffect worden berekend; dit statistische concept biedt een aanknopingspunt voor externe toetsing aan de categoriale prefixinformatie uit het MHB. Ook op deze dimensie blijkt de MGBN aanzienlijk completer te zijn dan het MHB. Want de MGBN dekt 60% van de MHB-informatie over de prefixinvloed op de uitvoercategorie c.q. resulterende lexeemcategorie en 40% van de MHB-informatie over de categoriale relatie tussen invoer- en uitvoerdomein van het prefix. Omgekeerd dekt het MHB slechts 22% (of 47% voor suffixen met frequentie 10 of meer) van de MGBN-prefixen met een uitvoercategorie en slechts 19% (resp/. 50%) van de MGBN-prefixen met een categoriale i-u-functie.

6.8 Conclusie

In dit hoofdstuk heb ik laten zien hoe men de structuurinformatie in de MGBN kan aanwenden om een virtueel, op L-KRING-principes gebaseerd model van het mentale lexicon te construeren (in afwachting van een computationele implementatie van dit lexicon) en hoe men de inhoud van dit MGBN-model kan analyseren en evalueren. Omdat de resulterende datatabellen bijzonder omvangrijk zijn, bleef de bespreking ervan beperkt tot een toelichting op de doelen en opzet van de onderliggende datarapporten en de samenvatting van de resultaten in de vorm van kencijfers, topscorelijstjes, grafieken en opmerkelijke verbanden. Verder werd voor alle klassen van datarapporten een externe of interne evaluatie besproken. Deze evaluaties dienden inzicht te geven in de externe en/of interne kwaliteit van de op het MGBN-model gebaseerde datarapporten en daarmee van de MGBN.

In het kader van de externe evaluaties heb ik de affixgegevens uit het MGBN-model integraal vergeleken met de affixgegevens in het Morfologisch Handboek. Hierdoor ontstond een objectieve basis voor de evaluatie van zowel het MGBN-model als het MHB, waarbij ik de evaluatie beperkt heb tot het gemeenschappelijke deel van hun empirische domeinen, te weten de orthografische dimensie van de Nederlandse prefixen en suffixen en hun combinatie-

mogelijkheden (zowel op het niveau van concrete morfemen als op het niveau van de categoriale typering van de inflectieklasse). Deze evaluatiemethode heeft de volgende constateringen opgeleverd:

1) Het MGBN-model biedt een nagenoeg complete dekking van de affixtypes in het Morfologisch Handboek, maar het Morfologisch Handboek dekt slechts een deel van de affixtypes in het MGBN-model.

2) Er bestaat een duidelijk verband tussen de lexicale frequentie van de morfologische patronen in het MGBN-model (in termen van de omvang van het stamdomein) en de mate waarin deze patronen door het MHB worden gedekt: hoe hoger de patroonfrequentie, hoe hoger de MHB-gebaseerde dekkingsgraad.

Combinatie van deze twee gegevens leidt tot de conclusie dat het MGBN-model met een complete en betrouwbare inventarisatie van Nederlandse affixtypes correspondeert, en dat het op dit punt al in de richting zit van een ideaal lexiconmodel van het Nederlands.[209] Dit biedt echter geen garantie dat alle lexemen op correcte wijze van een morfologische structuurrepresentatie zijn voorzien, al blijkt uit de interne evaluaties dat de morfologische patronen een redelijk waarschijnlijke distributie kennen.

Als onderdeel van de evaluatiedoelstelling heb ik de vraag besproken in hoeverre de morfologische kenmerken van het MGBN-model met de kennis uit het Morfologisch Handboek van het Nederlands (MHB) overeenkomen. Hierbij heb ik zowel naar kwalitatieve als kwantitatieve kenmerken gekeken. Dit vergelijkende onderzoek leidt tot de conclusie dat de MGBN alle affixen (en affixcombinaties) in het MHB dekt, maar dat het MHB slechts 20-30% van de MGBN-affixen dekt; bij de hoogfrequente patronen stijgt dit percentage tot ca. 40%. Dit impliceert dat de MGBN-representaties niet toevallig tot stand zijn gekomen, maar een betrouwbaar beeld geven van de (potentiële) affixtypes in de Nederlandse woordenschat.

Dit neemt niet weg dat maar een klein deel van de affixkenmerken via het MHB geëvalueerd kan worden, dus dat er ook andere methodes moeten worden ingeschakeld. Om die reden heb ik voor de suffixgeoriënteerde datarapporten onderzocht of er statistische tendenzen zichtbaar zijn. Deze bieden immers een aanknopingspunt om afwijkend affixgedrag op het spoor te komen, wat een aanwijzing kan zijn dat de betreffende segmenten ten onrechte als affix zijn aangemerkt. In het kader van deze studie was het echter niet mogelijk om zulke criteria systematisch toe te passen.[210] In plaats daarvan heb ik me beperkt tot het formuleren van hypotheses over interessante dataverbanden en de bespreking van aan de MGBN ontleende voorbeelden die positieve of juist negatieve evidentie bieden voor deze hypotheses. Deze dataverbanden zijn niet alleen interessant met het oog op de evaluatie en uitbreiding van de MGBN, maar ook met het oog op de vraag wat de aard is van de kennis in het mentale lexicon en welke analysecriteria de basis vormen voor het cognitieve vermogen om woorden te onderscheiden en deze van morfologische structuur te voorzien.

[209] In een uitgebreide studie over synchrone en diachrone eigenschappen van Nederlandse partikelwerkwoorden laat Blom (2005) zien dat de MGBN inderdaad completer is dan andere databronnen (p. 246).

[210] Met het oog op deze vraagstelling kan het nuttig zijn om de MGBN met de diachrone gegevensbank van Nicoline van der Sijs (Van der Sijs, 2002) te verbinden, iets wat zijzelf ook al oppert (p.584). Haar studie laat overtuigend zien dat het morfologische gedrag van hedendaagse stammen vaak etymologisch te duiden is.

7 Conclusie

In deze studie is beargumenteerd dat het zowel conceptueel als empirisch mogelijk is om het mentale lexicon langs inductieve weg (d.w.z. zonder gebruik te maken van grammaticaregels) van morfologische structuur te voorzien. De hierbij ontwikkelde gegevensbank (de MGBN) biedt naar het zich laat aanzien interessante mogelijkheden voor empirisch onderzoek naar de morfologische eigenschappen van het Nederlands en kan ook bijdragen aan de systematisering van de woordkenmerken in Van Dale's lexicografische informatiesysteem. Dit afsluitende hoofdstuk biedt een overzicht van de belangrijkste resultaten met betrekking tot de linguïstische en lexicografische dimensie van het beschreven onderzoek.

7.1 Linguïstische resultaten

In deze studie is een algemeen raamwerk voor lexicologisch onderzoek geïntroduceerd, te weten een Integraal Dynamisch Lexiconsysteem (IDL-systeem). Dit systeem bouwt voort op de uitgangspunten van de Ideale Woordenboek-visie van Verkuyl & al. (1998) en legt een dynamisch verband tussen de individuele en de collectieve woordenschat. Uitgaande van het IDL-systeem heb ik een nieuwe visie op morfologische structuuranalyse ontwikkeld. Deze visie is formeel vastgelegd in een theorie die uitgaat van Lexicale KennisRepresentatie door Inductieve Naamgeving (de L-KRING-theorie). Deze theorie heeft als doel om een fundamentele verklaring te geven voor de verwerving en activatie van morfologische kennis, en kent onder meer de volgende eigenschappen:

- De L-KRING-theorie stelt dat het mogelijk is om de in een lexicon opgeslagen woorden te comprimeren zonder dat er informatieverlies optreedt. Hiertoe dienen de gemeenschappelijke bouwstenen door indexen te worden vervangen. Door deze compressietechniek ontstaat spontaan morfologische structuur.
- In de L-KRING-theorie correspondeert de morfologische "grammatica" niet met een verzameling abstracte regels, maar met een gedetailleerde inventarisatie van lexicaal vastgelegde morfeemcombinaties en hiervan afgeleide patronen. Door gebruik te maken van een speciaal algoritme voor patroongeneralisatie kunnnen ook morfologische productieregels worden geconstrueerd. Deze kenmerken zich door een open stamdomein.
- In de L-KRING-theorie bestaan ten minste drie niveaus van morfologische structuuropbouw, te weten het niveau van de morfeemsequenties, het niveau van de lexeemsequenties en het niveau van de woordsequenties. Elk niveau wordt afgebakend door een (soms expliciet gemarkeerde) domeinbegrenzer die voor de overgang naar het volgende domein zorgt. In deze benadering zijn woorden altijd morfologisch geleed, ook als dit niet fonologisch is gemarkeerd. Want elk woord bestaat minimaal uit een lexeem en een woordbegrenzer, terwijl elk lexeem minimaal uit een wortel en een lexeembegrenzer bestaat. In deze theorie is geen kunstmatig onderscheid nodig tussen inheemse morfologie (die een syntagmatische basis zou hebben) en uitheemse morfologie (die een paradigmatische basis zou hebben): alle derivaties zijn namelijk stamgebaseerd (hierbij geldt elke wortel als een stam, terwijl elke volgende stam-affix-combinatie wederom een stam oplevert).
- In de L-KRING-theorie worden morfemen niet langs syntactische weg, maar langs paradigmatische weg geclassificeerd. Ik heb dit morfologische classificatiesysteem gemotiveerd door een aantal fundamentele problemen te bespreken met het syntactische classificatiesysteem en aan te tonen dat het nieuwe systeem hier een eenvoudige oplossing voor biedt. Dit paradigmatische classificatiesysteem biedt een empirische basis voor de introductie van morfologische en syntactische structuurklassen, door een fundamenteel verband te leggen met de lexicaal opgeslagen affixparadigma's van concrete stammen.

7.2 Lexicografische resultaten

- Het in deze studie beschreven onderzoek heeft een morfologisch gestructureerd lexicon opgeleverd. Deze Morfologische Gegevensbank voor het Nederlands (MGBN) omvat alle 80.000 basislexemen die ten grondslag liggen aan de 250.000 woorden (inclusief samenstellingen) in VDL's WoordKenmerkenBank Nederlands. De door mij toegekende structuurrepresentaties geven informatie over de lexeeminterne morfeemgrenzen en de klasse van de morfemen (wortels, prefixen of suffixen). Elke structuurrepresentatie kent een spelvormniveau en twee niveaus voor morfeemkoppelingen, namelijk voor koppeling van voorspelbare en voor koppeling van onvoorspelbare vormvarianten. Door de lexeeminformatie automatisch door te genereren naar het niveau van de samenstellingen (wat niet triviaal is, wegens lexicale identificatieproblemen), kon de hele Grote Van Dale van morfologische structuur worden voorzien.
- Met het realiseren van de MGBN is aangetoond dat het mogelijk is om in een overzienbare tijd (ca. 2 jaar) een compleet woordenboek langs redactionele weg van morfeemstructuur te voorzien op basis van een datagestuurde, door de L-KRING-theorie gelegitimeerde analysemethode. De hierbij toegekende structuur berust niet op vooraf vastgelegde regels, maar op de oordelen van een redacteur. Hierbij is een paradigmatisch analysecriterium gehanteerd; volgens dit criterium kan een woordintern segment als een affix worden aangemerkt als het bij een significant aantal woorden door een ander segment kan worden gesubstitueerd en als het voor al die woorden dezelfde inflectiecategorie en/of selectie-eigenschappen bezit.
- Bij de ontwikkeling van de MGBN is met succes een semi-automatische structureringsmethode gehanteerd. Bij deze methode wordt afgewisseld tussen automatische bewerking en redactionele controle, zowel in de vorm van interactieve zoek- en vervangopdrachten als door afwisseling van scripts voor automatische gegevensbewerking en de redactionele controle van het resultaat (in combinatie met eenvoudige sorteer-, selectie- en opmaakinstructies). Deze aanpak vereist een cyclische bewerking en verbetering van de data.
- Als onderdeel van deze studie zijn zeer gedetailleerde analyserapporten vervaardigd over de combinatiemogelijkheden van alle in de MGBN opgenomen affixen; hierbij wordt onder meer gedifferentieerd naar vormniveau, morfologische structuurpositie en syntactische functie; verder kan voor elke lexicale eenheid informatie worden opgevraagd over de absolute en relatieve typefrequentie (die op de omvang van het stamdomein berust). Tot slot kan voor elke affixvariant worden nagegaan of deze een MHB-vermelding kent.
- De nu gerealiseerde gegevensbank bevat (abstraherend van categoriale klasse en voorspelbare variatie in de spelvorm) ca. 19.000 wortels en ca. 1000 affixen (250 prefixen en 750 suffixen). Voor de prefixen zijn ca. 950 verschillende sequenties aangetroffen, voor de suffixen waren dit er ca. 3750. Wat betreft prefix-suffix-combinaties (incl. patronen zonder prefix of suffix) zijn ca. 7500 verschillende patronen gevonden, waaronder 4550 met categorie N, 950 met categorie V, 1900 met categorie A en 150 met categorie B. In totaal waren er 68.000 lexemen met en 16000 zonder affixen.
- De MGBN is extern geëvalueerd door de hieraan verbonden affixkenmerken integraal met de informatie in het Morfologisch Handboek (MHB) te vergelijken. Deze vergelijking leert dat de affixinventarisatie van het MGBN veel uitvoeriger is dan die van het MHB (dat volledig wordt gedekt), zowel wat betreft de omvang van het affixlexicon als de informatie over allomorfen, syntactische functies en affixcombinaties. Van de suffixen wordt (afhankelijk van het detailniveau en gegeven een typefrequentie van 10 of meer) 30 tot 40 procent door het MHB gedekt, van de prefixen ca. 60 procent; omgekeerd zijn alle MHB-affixen direct of indirect terug te vinden in de MGBN. Bovendien biedt de MGBN een complete inventarisatie van stammen en de bijbehorende affixparadigma's.

Appendices

A De evaluatie van een betekenisdomein

A.1 Introductie

Deze appendix bespreekt opzet en uitkomsten van een verkennende studie naar de lexicografische kwaliteit van de GWNT (zie H1.4.3 voor de achtergrond van dit onderzoek). Met het oog op deze vraag heb ik een concreet betekenisdomein uit de Grote Van Dale (13e druk) geanalyseerd, namelijk het domein van de notensoorten. Hierbij is eerst uitgezocht welke woorden tot het te evalueren domein behoren (door de selectie van woorden die naar een notensoort verwijzen). Vervolgens is nagegaan in hoeverre de bijbehorende lemma's aan het consistentiecriterium voldoen, dus in hoeverre deze lemma's een identieke opbouw vertonen. Idealiter zouden alle nootnamen als een subsoort van de *noot* moeten worden gedefinieerd, en zou de bijbehorende definitie moeten bepalen wat de gemeenschappelijke kenmerken zijn, welke kenmerken nootspecifiek zijn en hoe deze kenmerken gespecificeerd moeten worden. In de praktijk blijken de onderzochte lemma's echter veel variatie te vertonen, zowel in de toekenning van het genus als in de vermelding van allerlei aanvullende kenmerken.

Deze appendix is als volgt opgezet. Sectie A.2 behandelt de voorbereidende werkzaamheden. In A.3 komen een aantal concrete analysevoorbeelden aan de orde. In A.4 wordt het notendomein als geheel geëvalueerd. In A.5 wordt uiteengezet wat deze verkennende studie aan inzichten oplevert m.b.t. de consistentie van het notendomein en de bruikbaarheid van de analysemethode. Verder wordt aangeven wat voor aanpassingen nodig zijn om het geanalyseerde voorbeelddomein lokaal consistent te maken.

A.2 De constructie van het domein

De Grote Van Dale (en de onderliggende gegevensbank, de WKB-Ned) kent geen expliciete domeinstructuur.[211] Hierdoor is het geen sinecure om dit woordenboek (c.q. metalexicon) aan een domeingericht evaluatieonderzoek te onderwerpen. Want voordat men de consistentie van een betekenisdomein kan toetsen, dient eerst bekend te zijn welke woorden tot het te evalueren domein behoren. Indien een lexicon reeds een expliciete domeinstructuur bezit, is deze vraag eenvoudiger te beantwoorden. Maar indien deze structuur ontbreekt, of indien men de kwaliteit van de bestaande domeinstructuur wil toetsen, zal men deze structuur zelf moeten aanbrengen door per domein een selectiecriterium te formuleren en het lexicon integraal te doorzoeken op woorden die (mogelijk) aan dit selectiecriterium voldoen. Dit was dan ook de eerste stap in mijn onderzoek naar de consistentie van het notendomein. Het door mij geformuleerde zoekcriterium luidt als volgt: selecteer alle woorden waarvan de spelvorm op *noot* eindigt of waarvan de betekenisdefinitie (of een ander veld) het woord *noot* bevat. Dit zoekcriterium vormde de basis voor het doorzoeken van de GWNT (op basis van een full-text-search). Hierbij werd al gauw duidelijk dat het criterium niet precies genoeg was, want naast de betekenis "boomvrucht" kan het woord *noot* ook de betekenis "aantekening" aannemen (met als afgeleide betekenis "muzikaal symbool"). En er zijn ook woorden waarin het segment *noot* onderdeel is van het woorddeel *genoot*. Dergelijke woorden moesten daarom achteraf worden uitgefilterd. Van de resterende woorden bleek een deel niet alleen naar een notensoort te kunnen verwijzen, maar ook naar een noten producerende boom of plant. In enkele gevallen was dit zelfs de enige betekenis, namelijk bij *bitternoot*, *kanarieboom*, *tovernoot* en *vleugelnoot*. Bij mijn evaluatieonderzoek heb ik deze laatste categorie buiten

[211] VDL beschikt inmiddels wel over een complete inventarisatie van metonymie-relaties.

beschouwing gelaten, terwijl ik me bij de andere woorden alleen op de "notensoort"-definitie heb gericht. Ik heb deze woorden met elkaar vergeleken door een systematische inventarisatie te maken van hun semantische en formele woordkenmerken. In de volgende paragraaf wordt deze inventarisatie toegelicht aan de hand van een aantal voorbeeldlemma's.

A.3 Enkele evaluatievoorbeelden

Indien een betekenisdomein consistent is gestructureerd, dient de definitie van het hoofd-concept een helder criterium te bieden voor de selectie van de subtypes. Omgekeerd dienen de langs deze weg verzamelde subtypes weer terug te verwijzen naar dit hoofdconcept. Het door mij geconstrueerde GWNT-domein blijkt echter niet aan deze eisen te voldaan. Beschouw om te beginnen de GWNT-beschrijving van het woord *noot* in de plantkundige zin:

> **noot** (de; noten) =
> 1. boomvrucht met harde schaal; – als verkorting van *okkernoot* of *aardnoot*
> 2. (plantk.) eenzadige vrucht met houtige of leerachtige, niet openspringende wand
> 3. (gew.) muskaatnoot

Uit deze definitie blijkt dat de plantkundige definitie drie sublemma's kent, waarbij sublemma 1 en 3 aangeven dat de woordvorm als verkorting kan worden aangemerkt van de nootnaam van een subtype, terwijl de sublemma's 1 en 2 allebei een betekenisdefinitie geven, waarbij onduidelijk is hoe deze definities zich tot elkaar verhouden. Geen van deze sublemma's maakt melding van het feit dat het morfeem *noot* ook naar een boom of plant kan verwijzen (maar alleen indien het als hoofd van een samenstelling wordt gebruikt). Verder wordt met geen woord gerept over het intuïtief belangrijke gegeven dat noten vaak eetbaar zijn. Deze informatie staat echter wel bij het woord *boomvrucht*, en kan dus worden overgeërfd:

> **boomvrucht** (de) = (eetbare) vrucht die aan bomen groeit

De haakjes rondom *eetbaar* maken de hier gegeven betekenisdefinitie enigszins onduidelijk: is de eetbaarheid van de boomvrucht nu een secundaire eigenschap of is de boomvrucht niet altijd eetbaar? Ook de definitie als geheel is nogal minimaal. Het woord *vrucht* op zijn beurt wordt omschreven als:

> **vrucht** (de; -en) =
> 1. (eig.,plantk.) het uit het vruchtbeginsel gegroeide orgaan van de zaadplanten dat, als regel, een tot vele zaden bevat
> 2. (oneig.) schijnvrucht of ander soortgelijk (eetbaar) deel van een (daarvoor geteeld) gewas

En de definitie van *schijnvrucht* luidt als volgt:

> **schijnvrucht** (de) = (plantkunde) vrucht die grotendeels bestaat uit een doorgegroeide bloembodem of bloemdek, waarop of waarin zich de eigenlijke vruchtjes bevinden

Deze laatste definitie is circulair, want bij *vrucht* stond juist als één van de betekenissen *schijnvrucht*. Hier schiet de gebruiker dus weinig mee op. Afgaande op de rest van de definitie lijkt het cruciale verschil met een echte vrucht te zijn dat de schijnvrucht kleine vruchtjes bevat, terwijl de echte vrucht één of meer zaden bevat. De conclusie luidt dat het begrip *noot* niet scherp is gedefinieerd. Hieronder volgt een overzicht van alle direct of indirect gespecificeerde kenmerken:

> Betekeniscomponenten van de *noot*:
> a) vrucht
> b) afkomstig van een boom of plant
> c) eenzadig

d) met houtige of leerachtige, niet openspringende wand
e) met harde schaal
f) soms eetbaar
g) soms opzettelijk geteeld

Het is niet op voorhand zeker of deze informatie voor alle notensoorten geldt. Hiervoor zullen de lemma's van de afzonderlijke noten moeten worden bekeken. In de beschrijving van *noot* worden al enige voorbeelden gegeven, te weten *okkernoot*, *aardnoot* en *muskaatnoot*. Deze woorden hebben de volgende betekenisdefinities:

okkernoot
I (de) = de bekende vruchtkern van de gewone notenboom, syn. walnoot
II (de (m.)) = okkernotenboom

aardnoot (de) =
1. vrucht van de aardnoot (2), syn. pinda
2. plant van het geslacht Arachis uit de vlinderbloemenfamilie die aardnoten (1) oplevert (A. hypogea), syn. pinda (2), grondnoot (2)

muskaatnoot (de) =
1. vrucht van de muskaatboom
2. (stofn.) nootmuskaat

Uit de definitie van *okkernoot* blijkt dat dit woord een synoniem is van *walnoot*. Omgekeerd wordt bij *walnoot* aangegeven dat het een synoniem is van *okkernoot*. Hier is dus sprake van een notensoort met meerdere namen. Dit komt overigens wel vaker voor; zo is de naam *aardnoot* equivalent aan *pinda* (en ook aan *apennootje*) (en als plantennaam is hij equivalent aan *grondnoot*); op dezelfde manier is *pistache(noot)* equivalent aan *pimpernoot* en (mogelijk) ook aan (groene) *amandel*; verder geldt de *bosnoot* als een subklasse van de *hazelnoot* (al is niet duidelijk of het hier om de hazelnoot als noot of als struik gaat). Soms is dergelijke informatie alleen indirect afleidbaar, bijvoorbeeld uit informatie over een gemeenschappelijk brongewas (d.w.z. de boom of plant van herkomst); dit geldt voor de relatie tussen *muskaatnoot* en *nootmuskaat*. Onderstaande tabel vat deze observaties samen:

naam	klasse	synoniem	brongewas	uiterlijk	consumptie
okkernoot	vruchtkern	walnoot	notenboom	--	--
walnoot	--	okkernoot	--	soms zwart	--
aardnoot	vrucht	pinda	plant (Arachis)	spinnenweb	--
grondnoot	(plant)	aardnoot	--	--	--
pinda	vrucht	apennootje	plant (Arachis)	--	--
apennootje	--	pinda	--	--	--
muskaatnoot	vrucht	--	muskaatboom	--	specerij
nootmuskaat	vrucht	--	muskaatboom	geurig	specerij
pistache	vrucht	amandel	pistacheboom	hazelnoot	snoepje
pimpernoot	--	pistache	plant	klappernoot	--
amandel	steenvrucht	--	amandelboom	plat, ovaal	eetbare pit
hazelnoot	vrucht	--	hazelaar	--	--
bosnoot	(noot / plant?)	hazelnoot	--	--	--

Uit deze tabel blijkt dat de hier geanalyseerde GWNT-lemma's nogal verschillen in de specificatie van de potentiële betekenisdimensies. Zo wordt slechts bij enkele noten melding gemaakt van een consumptietoepassing (wat de vraag oproept of de andere noten dan niet eetbaar zijn). Ook wordt lang niet altijd aandacht besteed aan het uiterlijk. Bij de meeste nootnamen wordt wel een klasse (of synoniem) en een brongewas genoemd (incl. Latijnse

naam). Wat betreft de gewasspecificatie valt op dat er meestal sprake is van een boom, maar ook wel eens van een plant. Dit wijst erop dat de hoofdsoort *noot* een onderverdeling kent in boomvruchten en plantenvruchten. Wat betreft de klasse valt op dat geen van de nootnamen als subklasse van de *noot* wordt getypeerd, maar alleen als subklasse van de *vrucht*, de *vruchtkern* of de *steenvrucht* (of als synoniem van een andere noot). Deze observaties laten zien dat het tot nu toe bekeken deel van het notendomein veel variatie vertoont met betrekking tot de opbouw van de lemma's en dus een geringe mate van consistentie bezit.

A.4 De integrale domeinevaluatie

Tabel A-1 toont de complete evaluatietabel voor het notendomein. Hierbij zijn namen die naar dezelfde notensoort verwijzen (blijkens de vermelding van synoniemen) bij elkaar geplaatst: dergelijke clusters vormen een naamfamilie. Binnen deze naamfamilies is de naam die het vaakst als synoniem fungeert (d.w.z. die het vaakst voorkomt in de betekenisdefinitie van de andere nootnamen) als familiehoofd aangemerkt, wat door een vetgedrukt lettertype wordt gemarkeerd. Deze familiehoofden zijn vervolgens alfabetisch geordend, waarna elk familiehoofd in de bijbehorende families is geplaatst (eveneens in alfabetische volgorde). De hier gehanteerde ordening maakt het mogelijk om te controleren of er binnen elke namenfamilie sprake is van consistentie in de toegekende kenmerken, en of er tenminste één naam is waarbij alle gespecificeerde kenmerken in de definitie zijn terug te vinden (wat meestal het hoofd van de familie zal zijn). De overige kolommen geven de volgende informatie:

-fam	familienummer (op dezelfde regel als de naam van het familiehoofd)
-naam	nootnaam waarvan het GWNT-lemma is geanalyseerd c.q. trefwoord
-soort	soortnaam (indien gespecificeerd): noot, vrucht(en), zaad etc.
-syn	+s = naam waarbij ook een synoniemvorm is gespecificeerd
	>s = naam waarvan de betekenis via een synoniem is gedefinieerd
-bron:	brongewas:
	[+p] = noot afkomstig van plant
	[+b] = noot afkomstig van boom
	L = vermelding van Latijnse plantnaam of boomnaam
-cons:	[+c] = vermelding van consumptiemogelijkheden
-func:	[+f] = vermelding van functie
-uk/ik:	[+u] = vermelding van uitwendig kenmerk (zoals kleur of omvang)
	[+i] = vermelding van inwendig kenmerk (zoals olierijk)
-etym	+e = vermelding van etymologische informatie
-rest	[+r] = vermelding van restkenmerk(en)

fam	naam	soort	syn	bron	cons	func	uk/ik	etym	rest
	aardaker	knolwortel	+s	+p,L	+c				
	aardeikel		+s						
1	**aardnoot**	vrucht	+s	+p,L					
	akkernoot		+s						
	ape(n)nootje		+s						
	grondnoot		>s						
	katjang tjina	peulvrucht	+s					+e	+r
	lombokker		>s					+e	
	olienoot		+s					+e	
	olienootje		+s						
	pinda	vrucht	+s	+p,L				+e	

	pindanootje		+s						
2	**amandel**	steenvrucht	+s	+b	+c			+e	
	areka	vruchten	+s	+b,L	+c		+u	+e	
3	**arekanoot**	vrucht	+s						
	betelnoot		+s					+e	
	pinang		+s	+b				+e	+r
	pinangnoot	vrucht		+b	+c		+u		
4	**behennoot**	vrucht	+s	+b			+u	+e	
	zalfnoot	vrucht	+s	+b,L		+f	+i		
	beukel		+s						+r
5	**beukennoot**	vrucht		+b					
	beukennootje		+s						
	bokkempje		+s					+e	
	bokkenpit		+s						+r
6	**boternoot**	vrucht		+b					
7	**braaknoot**	zaadkorrel		+b		+f			
	kraanoog		+s			+f			
	bombaynoot		+s					+e	
8	**cashewnoot**	schijnvrucht	+s	+b,L			+u	+e	+r
	cachounoot		+s					+e	
	olifantsluis		+s						+r
	galappel	uitwas	+s	+b			+u	+e	+r
9	**galnoot**	uitwas	+s	+b			+u		+r
	knikkergal	gal	+s	+b			+u		
	baardnoot		+s						
	bosnoot	hazelnoot		+b					
10	**hazelnoot**	noot		+p					+r
	haze(n)noot		+s						
	lambertsnoot		+s						
	lammernoot		+s						
	lammertjesnoot	vrucht	>s	+b,L			+u		
	lammetjesnoot		+s						
	sint-lambertsnoot	hazelnoot	+s	+p,L			+u		
	sint-lambertusnoot		+s						
	zinknoot	noot	+s				+i		+r
11	**ivoornoot**	vrucht		+b,L		+f	+u		
	steennoot	vrucht	+s	+b,L			+u		
	taguanoot		+s						
12	**kemirinoot**	zaad		+b,L	+c				
13	**kokelekonoot**	zaad	+s	+b,L			+u		
	paranoot	doosvrucht		+b,L			+i	+e	
	coco		+s					+e	
	klapper	vrucht	+s	+b				+e	+r
	klappernoot		+s						
14	**kokosnoot**	vrucht	+s	+b			+u	+e	+r
	kola		+s						
	kolanoot	zaad		+b	+c		+i		
	liplap	kokosnoot	+s		+c		+u		+r
15	**krapnoot**	vrucht		+b					
16	**macadamianoot**	vrucht	>s	+b				+e	

	kruidnoot		+s		+c				
	mannetjesnoot		>s				+u		
17	**muskaatnoot**	vrucht	+s	+b	+c			+e	
	nootmuskaat	vruchtkern		+b,L	+c			+e	+r
	Papoeanoot		>s						
	talkmuskaatnoot	zaad	+s	+b			+i		
	houtnoot		>s				+u		
18	**okkernoot**	vruchtkern		+b					
	paarde(n)noot		>s						+r
	palmnoot	vrucht		+b					
	telnoot		+s					+e	+r
	walnoot		+s	+b,L				+e	
19	**paradijsnoot**	noot		+b,L					
20	**pecannoot**	vrucht	>s	+b	+c			+e	
	pimpernoot		+s	+p				+e	+r
21	**pistache**	vrucht	>s	+b	+c		+i	+e	+r
	pistachenoot		+s						
	prikkelnoot	amandel	+s				+u		
22	**purgeernoot**	zaad		+b,L		+f			
	schijtnoot		+s						
	stekelnoot		+s	+b,L					+r
23	**waternoot**		+s	+b,L	+c		+u		+r
	waterkastanje		+s						
24	**zeepnoot**	vrucht		+b,L					

Tabel A-1: Betekenisanalyse van GWNT-lemma's uit het domein 'notensoorten'.

In totaal blijken er 85 verschillende nootnamen te zijn (inclusief vormvarianten) die tot 24 verschillende betekenisfamilies behoren. Dit zijn er aanzienlijk meer dan de 11 soorten die via de betekenisingang *noot* zijn te vinden. Van de 85 nootnamen eindigen er 65 op het woorddeel *noot*, wat aantoont dat het niet voldoende is om naar trefwoorden met de substring *-noot-* te zoeken. Onder de 24 hoofdsoorten bevinden zich 19 vruchten, 6 zaden en 1 noot zonder genusspecificatie (de kastanje). Het is mogelijk dat ook bepaalde pitten (zoals de pijnboompit en de amandelpit) tot de noten moeten worden gerekend. Bij geen enkele naam wordt het genus *noot* gegeven. Van alle notenfamilies blijken er 22 van een boom te komen, de overige twee van een plant (al zijn er enkele families met tegenstrijdige specificaties). Van 13 families wordt de nootvorm beschreven en bij 11 families wordt informatie gegeven over de consumptiemogelijkheden.

De evaluatietabel laat dus zien dat per notenfamilie grote variatie bestaat in het aantal gespecificeerde kenmerken. Alleen de kenmerken 'klasse' en 'herkomst' worden standaard vermeld. De kenmerken 'consumptie' en 'functie' worden echter in minder dan de helft van de gevallen gespecificeerd. Binnen de afzonderlijke families is nog meer variatie aanwezig, maar dit hangt voor een deel samen met het feit dat veel nootnamen als synoniem van de overkoepelende klassenaam worden gedefinieerd, waardoor de andere eigenschappen kunnen worden overgeërfd.

Voor alle nootnamen samen gelden de volgende cijfers: bij 43 noten wordt het vruchttype gespecificeerd (27 zijn vrucht, 6 zaad en 10 overig); bij 44 noten wordt de bijbehorende boom of plant gespecificeerd (waaronder 20 keer een Latijnse naam); bij 31 van de noten heeft de bronboom of bronplant een eigen lemma (waardoor overerving mogelijk wordt); bij 14 noten

wordt een consumptietoepassing genoemd, bij 5 noten een andere functie, bij 25 noten wordt de vorm beschreven (van 18 het uiterlijk, van 11 het innerlijk), van 19 noten wordt een ander kenmerk genoemd en van 27 nootnamen wordt de etymologie vermeld. Verder zijn er enkele nootnamen die niet met een vrucht corresponderen, maar met een bepaalde consumptietoepassing, namelijk *borrelnoot*, *sojanoot*, *tafelnoot* en *vanillenoot*. Uit de analyse blijkt ook dat het woorddeel *noot* meerdere betekenissen heeft. Zo is er een basiskeuze tussen *vrucht* (c.q. *zaad*) of *plant* (namelijk *boom* of *struik*); beide klassen bezitten weer de nodige subklassen.

A.5 Conclusie en consequenties

Dit verkennende onderzoek wijst uit dat het GWNT-domein van de notensoorten slechts een geringe mate van consistentie vertoont. Dit viel ook te verwachten, want de GWNT kent (nog) geen expliciete domeinstructuur, wat betekent dat de impliciet aanwezige domeinen waarschijnlijk niet systematisch op compleetheid en consistentie zijn gecontroleerd. Voor een woordenboek is dit ook minder noodzakelijk dan voor een metalexicon, want een woordenboek hoeft niet in staat te zijn om taaltechnologische applicaties te ondersteunen, maar heeft doorgaans alleen een adviserende functie; voor dit type toepassing kan worden volstaan met korte, informele betekenisaanduidingen (zoals een synoniem), al dan niet aangevuld met informatie over opvallende kenmerken of bijzondere gebruikscondities. Daar komt bij dat een gedrukt woordenboek ruimtetechnische beperkingen kent, zodat het niet wenselijk is om alle lemma's even uitgebreid te behandelen. Uit mijn voorbeeld blijkt echter dat dit streven naar beknoptheid ten koste gaat van de precisie, waardoor allerlei interpretatievragen ontstaan, zoals substitutievragen (is woord a equivalent aan concept b of aan concept c) en onderschikkingsvragen (is woord a een subsoort van concept b of van concept c?). Dit is natuurlijk een ongewenst bijverschijnsel. Voor een metalexicon is het sowieso belangrijk om naar een toename van de compleetheid en de consistentie te streven (conform de IW-visie).

De hier gesignaleerde problemen kunnen worden opgelost door het onderzochte metalexicon een domeingebaseerde structuur te geven. Voor dit doel dienen alle lemma's expliciet aan een betekenisdomein te worden gekoppeld, terwijl elk betekenisdomein een eigen sjabloon voor de lemmadefinities moet krijgen; dit sjabloon moet ervoor zorgen dat alle woorden die tot hetzelfde betekenisdomein behoren een vergelijkbare lemmastructuur krijgen, wat garandeert dat hun betekenisdefinities dezelfde opbouw en dezelfde mate van gedetailleerdheid zullen vertonen. Verder dient voor elk betekenisdomein een algemeen toepasbare definitie van het hoofdconcept te worden uitgewerkt, d.w.z. een definitie die gemeenschappelijke kenmerken vastlegt en die alle betekenisdimensies introduceert die relevant zijn voor de definitie van de subtypes die door dit concept verenigd worden. Men kan deze definitie langs inductieve weg construeren door de bestaande definities van hoofdconcept en subtypes aan een componentiële analyse te onderwerpen (zoals in de voorgaande secties is gebeurd) en de meest voorkomende eigenschappen in een nieuwe conceptdefinitie onder te brengen.[212]

Bij wijze van voorbeeld volgt hier een nieuwe lemma-opzet voor het woord *noot* (in plantkundige zin). Hierbij is per betekenis aangegeven wat het bijbehorende structuurniveau is (namelijk woord, woorddeel of afkorting):

NOOT [plantk.] =
1. [als woord of woorddeel] schijnvrucht, eenzadig, met harde schaal (d.w.z. een houtige of leerachtige, niet openspringende wand), boomvrucht of plantenvrucht, vaak eetbaar
> soorten: amandel, hazelnoot, okkernoot, muskaatnoot, pinda, pistache, etc.
2. [als woord] harde kern van de vrucht onder (1)

[212] Deze methode ligt (in een veel verder uitgewerkte vorm) ook ten grondslag aan het Algemeen Nederlands Woordenboek, dat momenteel op het INL wordt ontwikkeld (zie Moerdijk, 2002).

3. [als woorddeel] nootdragend gewas (boom of struik)
> soorten: bosnoot, hazelnoot, okkernoot, etc.
4. [afkorting] nootmuskaat, okkernoot

Volgens dit definitieschema kent het woorddeel *noot* als plantkundige term vier verschillende gebruiksmogelijkheden, namelijk als aanduiding van een bepaald type schijnvrucht, als aanduiding van de harde kern van deze vrucht, als aanduiding van een nootdragend gewas (boom of struik) en als afkorting van bepaalde nootnamen. Bij definitie 1 en 3 wordt bovendien een opsomming gegeven van alle noten of planten die tot de gedefinieerde klasse behoren. Elk van deze subsoorten dient in termen van de hoofdsoort te worden gedefinieerd, waarbij de betekenisdimensies in de hoofddefinitie zo mogelijk nader worden ingevuld. Zo dient voor elke notensoort te worden aangegeven wat zijn uiterlijk is, wat de naam is van het bijbehorende gewas en wat de consumptiemogelijkheden zijn. Elke notensoort kan overigens weer een eigen verzameling van subtypes en/of synoniemen introduceren. Indien alle MGBN-domeinen op deze wijze worden gestructureerd, zal dit een aanzienlijk hogere consistentiegraad opleveren. Toch zal ook de domeingebaseerde analysemethode op problemen stuiten, want de werkelijkheid laat zich in het algemeen niet tijdloos en objectief in onderling exclusieve, netjes hiërarchisch ingedeelde concepten en subconcepten indelen.

B Datatabellen met MGBN-analyses

B.1 Kencijfers bij het MGBN-model

B.1.1 Introductie

In deze subsectie geef ik algemene kencijfers over de omvang van de belangrijkste lexicale domeinen in het MGBN-model, te weten het woordniveau (met zelfstandig bruikbare lexemen en combinaties van lexemen), het lexeemniveau (met zelfstandige lexemen, rechterdelen, middendelen en linkerdelen) en het morfeemniveau (met wortels en affixen). Zoals ik al aangaf, berust het MGBN-model op computationeel bewerkte informatie uit de MGBN, d.w.z. de morfologisch verrijkte versie van de LGBN (die zelf weer op informatie uit de WKB-Ned berust). De in dit hoofdstuk gepresenteerde datarapporten hebben betrekking op de morfologische structuurkenmerken van de (basis)lexemen, d.w.z. de kleinste lexicale bouwstenen die alleen of in combinatie met andere lexemen een zelfstandig (al dan niet samengesteld) woord kunnen vormen. In de paragrafen over het lexeemniveau en het woordniveau zullen echter ook kencijfers worden verstrekt die betrekking hebben op de samenstellingsmogelijkheden van de lexemen.

B.1.2 Kencijfers bij het woordniveau

De onderstaande tabel verstrekt kencijfers over het hoogste structuurdomein van het MGBN-model, te weten het woorddomein. Dit domein omvat zowel enkelvoudige als samengestelde woorden. Die laatste groep kan weer worden onderverdeeld in lemma's met 2, 3 en 4 lexemen. De tabel specificeert twee soorten tellingen, namelijk tellingen voor lemma's exclusief syntactische categorie en voor lemma's inclusief syntactische categorie.

lemmaklasse	aantal lemma's (excl. categorie)	aantal lemma's (incl. categorie)
samengestelde lemma's	163.584	202.308
lemma's zonder sublexemen	82.062	85.462
Totaal	245.646	287.770

Tabel 1: Telling van het aantal MGBN-lemma's op woordniveau (wel/niet samengesteld)

er zijn 7211 basislexemen waarvoor nog geen MGBN-representatie bestaat.
er zijn 8323 restwoorden zonder categorie
er zijn 4366 restwoorden met categorie

B.1.3 Kencijfers bij het lexeemniveau

Het lexeemniveau van het MGBN-model correspondeert met de lexicale bouwstenen die de combinatorische basis vormen voor enkelvoudige en samengestelde woorden. In deze sectie zal ik voornamelijk kencijfers verstrekken voor het complete lexeemdomein (D_0), al kan men dit domein ook inperken tot [wnn]-lexemen, [+auto]-lexemen of [-auto]-lexemen (met de subopties lp, mp en rp); zie H6.2.3 voor een toelichting. In onderstaande tabel vindt men alleen kencijfers over de complete lexeeminventarisatie. Hierbij maak ik onderscheid tussen telniveau T1, dat betrekking heeft op unieke lexeemvormen, telniveau T2, voor lexemen met een unieke combinatie van vorm en betekenisindex (die alleen bedoeld zijn voor het scheiden van etymologisch verschillende hoofdbetekenissen), telniveau T3, waar ook wordt gedifferentieerd naar categorie en telniveau (T4), dat correspondeert met een specificatie van het aantal dataregels in de MGBN, met subtellingen voor het aantal [+auto]-lexemen en het aantal [+dep]-lexemen (met aparte informatie voor de opties 'altijd' en 'soms'). Onder de eerste tabel wordt voor enkele telcategorieën een nadere uitsplitsing gegeven.

Algemene lexeemtelling

lexeemklasse (domein D0)	aantal
T1. unieke vormen	82.062
T2. vormen met betekenisindex	82.231
T3. vormen met index en categorie	84.786
T4. dataregels (met subkenmerken)	85.462

specificatie bij T4	100% kenmerk	≤ 100% kenmerk
[+auto]-lexemen	62.185	80.679
[+dep]-lexemen	4.783	23.277

specificatie bij T1	aantal unieke lemmavormen
zonder extra condities	82062
met wnn-status	42033
pseudosamenstelling	4666
met naamstatus	1709
met leenstatus	7398

Tabel 2: Telling van het aantal MGBN-lemma's op lexeemniveau en uitsplitsing van enkele subkenmerken

hoofdcategorieën	V	N	A	overig	**totaal**
aantal lexemen	17350	49210	13643	3168	**83371**

overige categorieën	B	P	C	D	T	R	**totaal**	[-]	**totaal**
aantal lexemen	1071	159	83	13	87	139	**1552**	1616	**3168**

Tabel 3: Telling van het aantal MGBN-lemma's per syntactische categorie

B.1.4 Kencijfers bij de morfologische structuurrepresentaties

Deze paragraaf biedt kencijfers over het aantal morfologische structuurrepresentaties per structuurniveau en over de lexicale frequentie van diverse soorten morfemen in diverse soorten contexten, d.w.z. cijfers over het aantal verschillende (basis)lexemen waarin een morfeem opduikt. Hieronder staat een korte toelichting op de classificatiekenmerken.

Morfologische representatieniveaus:
-n0vorm = niet-morfologische lexeemrepresentatie op het niveau van de spelvorm
-n1vorm = morfologische lexeemrepresentatie op het niveau van de spelvorm
-n2vorm = generalisatie over n1vormen: bundeling van regelmatige spelvormvarianten
-n3vorm = generalisatie over n2vormen: bundeling van (etymologische) klankvarianten

Lexeemklassen:
- [±index] geeft aan of de lexemen op het niveau van de lexeemvorm worden geteld of op het niveau van de hieraan ondergeschikte semantische index (= [+index])
- [±synt] geeft aan of de lexemen op het niveau van de lexeemvorm met semantische index worden geteld of op het niveau van de syntactische categorie (= [+synt])

Onderscheid tussen woorden, lexemen en sublexemen:
- *woorden*: syntactische basiseenheden (c.q. inflectie-dragers) die uit één of meer lexemen bestaan

- *lexemen*: lexicale eenheden die als bouwsteen dienen voor de samenstellingen in de LGBN
- *sublexemen*: lexeem-interne deeleenheden die volgens de MGBN een eigen wortel bezitten

B.1.5 Kencijfers over de morfologische lexeemrepresentaties in domein D0

De onderstaande tabellen bieden een kwantitatieve specificatie van het aantal morfologische lexeemrepresentaties in het D0-domein (het niveau van de niet-samengestelde lexemen) van de MGBN. Deze informatie wordt opnieuw uitgesplitst naar lexeemtype (via de parameter [±synt], die aangeeft of er rekening wordt gehouden met de syntactische klasse) en naar morfologisch representatieniveau.

lexeemtype T2: [-synt]	lexemen (voor opdeling)	sublexemen (na opdeling)
aantal n3-lexemen	77.694	75.690
aantal n2-lexemen	81.822	80.351
aantal n1-lexemen	82.830	81.543
aantal n0-lexemen	82.231	-

lexeemtype T3: [+synt]	lexemen (voor opdeling)	sublexemen (na opdeling)
aantal n3-lexemen	80.974	80.137
aantal n2-lexemen	84.398	84.136
aantal n1-lexemen	85.256	85.169
aantal n0-lexemen	84.786	-

Tabel 4: Telling van het aantal morfeemrepresentaties per structuurniveau ([±synt]*)*

B.1.6 Kencijfers per morfeemtype in domein D0

Deze kencijfers geven inzicht in het totaal aantal vormeenheden per morfologische klasse (te weten, basislexemen, stammen, prefix/suffix-sequenties en prefix/suffix-eenheden, met een uitsplitsing naar morfologisch representatieniveau (n3, n2 of n1). Deze inventarisatie heeft betrekking op de sublexemen van het D0-niveau van de MGBN.

taxeemtype	**niveau n1**	**niveau n2**	**niveau n3**
basislexemen	81.476	80.277	75.583
lexeemwortels	24018	20287	15560
prefixen	408	365	229
suffixen	1211	722	433
affixen	1619	1087	662
prefixsequenties	976	944	732
suffixsequenties	4535	3742	3351
affixsequenties	5511	4686	4083

Tabel 5: aantal MGBN-items per taxeemtype

B.2 Resultaten van de prefix-analyses

B.2.1 Introductie

Mijn computationele onderzoek naar de eigenschappen van het MGBN-model met betrekking tot de prefixdistributie heeft een hele reeks digitale datarapporten opgeleverd. In deze sectie zal ik me beperken tot de presentatie van enkele voorbeeldlijsten, namelijk een lijst met de hoogstfrequente prefixen (6.5.3.1), een lijst met de laagstfrequente prefixen (6.5.3.2) en fragmenten van een rechts-links gesorteerde prefixlijst (6.5.3.3) en een links-rechts gesorteerde prefixlijst (6.5.3.4). Alle lijsten richten zich primair op de eigenschappen bij prefixen in de n2vorm. Hieronder volgt een korte toelichting op de bijbehorende veldstructuur. Samen met deze toelichting spreken de lijsten verder voor zich.

Veldstructuur van de op frequentie gesorteerde prefixlijsten

1 aantal prefix-eenheden
2 n2vorm van mgbn-prefix
3 inheems (i) of uitheems (u) prefix
4 mhb-vorm van prefix (indien beschikbaar, anders '??')
5 mhb-status: productief (+p) of improductief (-p)
6 stamfrequentie bij de n3vorm
7 stamfrequentie bij de n2vorm
8 u-ratio: gemiddeld aantal uitwaartse lexeemspecificaties per prefix (negatief als f < 6)
9 i-ratio: gemiddeld aantal inwaartse stamspecificaties per prefix (negatief als f < 6)
10 voorbeeld van lexeemtoepassing (in n0vorm)

Veldstructuur van de rechts-links en links-rechts gesorteerde prefixlijsten:

1 wel/geen vermelding in handboek (= [+/-hb])
2 maximale sequentielengte (0 = vrij)
3 aantal prefix-eenheden
4 n2vorm van mgbn-prefix
5 inheems (i) of uitheems (u) prefix
6 mhb-vorm van prefix (indien beschikbaar, anders '??')
7 mhb-status: productief (+p) of improductief (-p)
8 stamfrequentie bij de n3vorm
9 stamfrequentie bij de n2vorm
10 u-ratio: gemiddeld aantal uitwaartse lexeemspecificaties per prefix (negatief als f < 6)
11 i-ratio: gemiddeld aantal inwaartse stamspecificaties per prefix (negatief als f < 6)
12 voorbeeld van lexeemtoepassing (in n0vorm)

B.2.2 De hoogstfrequente prefixen

1	2	3	4	5	6	7	8	9	10
1	ge	i	ge	[+p]	1258	1872	1.4	1.5	geploeter
1	ver	i	ver	[+p]	1108	1342	2.3	1.2	verslechtering
1	be	i	be	[+p]	739	1100	2.2	1.5	bekennen
1	af	i	af	[+p]	981	1007	1.5	1.0	afscheuring
1	uit	i	uit	[+p]	783	861	1.5	1.2	uitscheppen
1	on	i	on	[+p]	831	836	1.5	1.4	onwillige
1	over	i	over	[+p]	752	792	1.4	1.1	overschrijver
1	op	i	op	[+p]	736	764	1.6	1.0	oppositie
1	voor	i	voor	[-p]	616	632	1.4	1.2	voorziend
1	in	i	in	[+p]	605	620	1.6	1.0	inschrijden
1	aan	i	aan	[+p]	557	594	1.6	1.1	aanlappen
2	on_@	i	??		548	547	1.4	1.3	ongebleekt

1	2	3	4	5	6	7	8	9	10
1	na	i	na	[+p]	416	457	1.3	1.1	nascheut
1	onder	i	onder	[+p]	435	443	1.5	1.0	ondertrouwde
1	om	i	om	[+p]	404	413	1.4	1.0	omhokken
1	ont	i	ont	[+p]	370	398	1.7	1.0	ontmoedigings
1	door	i	door	[+p]	373	379	1.4	1.0	doorzakken
1	achter	i	achter	[-p]	347	350	1.2	1.2	achterlopen
1	re	u	re	[+p]	284	311	2.6	1.2	rescontreren
1	toe	i	toe	[+p]	291	300	1.4	1.0	toewijzing
1	de	u	de	[+p]	258	284	2.1	1.1	defeceren
1	bij	i	bij	[+p]	250	262	1.3	1.1	bijvoeging
1	weg	i	weg	[+p]	260	261	1.1	1.0	wegloodsen
1	tegen	i	tegen	[-p]	251	254	1.2	1.0	tegenkoning
1	rond	i	rond	[+p]	220	222	1.1	1.0	rondvragen
1	con	u	??		252	220	2.8	1.4	consortium
2	on_ge	i	??		217	217	1.3	1.0	ongekroond
1	in	u	in	[+p]	312	213	2.4	1.1	infinitivus
1	her	i	her	[+p]	184	193	1.6	1.2	herroepbaar
2	voor_@	i	??		171	171	1.2	1.4	voorgedeelte
1	tussen	i	??		166	166	1.1	1.0	tussenwaar
2	on_be	i	??		160	160	1.5	1.0	onbezorgd
1	a	u	a	[-p]	175	153	1.6	1.1	Azoïcum
1	terug	i	terug	[-p]	151	153	1.3	1.0	terugspoelen
1	voort	i	voort	[-p]	148	148	1.1	1.0	voortzeuren
1	mee	i	mee	[+p]	134	136	1.1	1.0	meekrap
2	ge_@	i	??		134	134	1.1	1.2	geïntimeerde
1	mis	i	mis	[+p]	131	132	1.4	1.0	misjaar
1	pro	u	pro	[+p]	156	130	2.8	1.5	professoraal
1	in	u	in	[+p]	199	128	1.7	1.0	indispositie
2	achter_@	i	??		127	127	1.0	1.5	achteropkomen
1	neer	i	neer	[-p]	124	124	1.1	1.0	neerstoten
1	wel	i	??		114	117	1.2	1.1	welbekend
1	vol	i	vol	[-p]	108	116	1.4	1.1	volautomatisch
1	ex	u	ex	[+p]	177	114	2.7	1.1	exploitatie
1	samen	i	samen	[+p]	105	106	1.6	1.0	samenschrapen
2	on_ver	i	??		96	95	1.4	1.0	onverstoorbaar
1	pre	u	pre	[-p]	156	92	2.0	1.1	prelinguaal
1	com	u	??		252	88	3.3	1.8	commensalisme
1	inter	u	inter	[-p]	76	86	1.7	1.1	interventie

B.2.3 Selectie uit de laagstfrequente prefixen

1	2	3	4	5	6	7	8	9	10
2	ver_aan	i	??		1	1	-2.0	-1.0	veraangenamen
3	ver_aan_@	i	??		1	1	-2.0	-1.0	veraangenaming
3	ver_aan_ge	i	??		1	1	-2.0	-1.0	veraangenaming
2	ver_af	i	??		1	1	-2.0	-1.0	verafgoding
2	ver_bij	i	??		1	1	-2.0	-1.0	verbijzonderen
2	ver_tegen	i	??		1	1	-2.0	-1.0	vertegenwoordigers
2	ver_vol	i	??		1	1	-2.0	-1.0	vervolledigen
2	ver_ab	i	??		1	1	-2.0	-1.0	verabsolutering
2	ver_be	i	??		1	1	-1.0	-1.0	verbestendigen
2	ver_com	i	??		1	1	-2.0	-1.0	vercommercialiseren
2	ver_di	i	??		1	1	-1.0	-1.0	verdiverteren
2	ver_ex	i	??		1	1	-2.0	-1.0	verexcuseren
2	ver_in	i	??		1	1	-1.0	-1.0	verinlandsen
2	ver_per	i	??		1	1	-2.0	-1.0	verpersoonlijking

2	ver_pro	i	??	1	1	-1.0	-1.0	verprocederen
2	ver_al	i	??	1	1	-4.0	-1.0	veralgemenisering
3	ver_al_@	i	??	1	1	-4.0	-1.0	veralgemening
3	ver_al_ge	i	??	1	1	-4.0	-1.0	veralgemenisering
2	a_@	u	??	1	1	-2.0	-1.0	apropos
2	a_pro	u	??	1	1	-2.0	-1.0	apropos
1	c'	u	??	1	1	-1.0	-1.0	c'est
1	ei	i	??	1	1	-1.0	-1.0	eilieve
1	el	u	??	1	1	-1.0	-1.0	eldorado
1	r	i	??	1	1	-1.0	-1.0	rommendom
1	ev	u	??	1	1	-1.0	-1.0	evviva
1	ib	u	??	1	1	-1.0	-1.0	ibidem
1	ie	i	??	1	1	-1.0	-1.0	ievallig
1	jo	u	??	1	1	-1.0	-1.0	johoe
2	l'_@	u	??	1	1	-1.0	-1.0	l'improviste
2	l'_im	u	??	1	1	-1.0	-1.0	l'improviste
1	mon	u	??	1	1	-1.0	-1.0	monseigneur
1	no	u	??	1	1	-1.0	-1.0	nobody
2	o_@	u	??	1	1	-1.0	-1.0	odontotherapie
2	o_dont	u	??	1	1	-1.0	-1.0	odontotherapie
3	o_dont_@	u	??	1	1	-1.0	-1.0	odontotherapie
3	o_dont_o	u	??	1	1	-1.0	-1.0	odontotherapie
1	off	i	??	1	1	-2.0	-1.0	offsetter
1	ol	u	??	1	1	-1.0	-1.0	olfactief
1	o'	u	??	1	1	-1.0	-1.0	o'clock
1	que	u	??	1	1	-1.0	-1.0	quebracho

B.2.4 Prefixlijst met rechts-links-sortering

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
-mb	0	1	as	u	??		214	23	3.1	1.6	geassureerd
-mb	0	2	@_as	u	??		52	9	1.6	1.8	rassurant
-mb	0	2	des_as	u	??		4	1	-1.0	-1.0	desassimilatie
-mb	0	2	co_as	u	??		3	1	-1.0	-1.0	coassuradeur
-mb	0	2	ge_as	u	??		15	4	-1.5	-1.0	geassumeerde
-mb	0	2	her_as	u	??		3	1	-1.0	-1.0	herassurantie
-mb	0	2	r_as	u	??		9	1	-2.0	-1.0	rassureren
-mb	0	2	re_as	u	??		9	1	-3.0	-1.0	reassureren
-mb	0	2	ver_as	u	??		5	1	-1.0	-1.0	verassureren
-mb	0	1	at	u	??		214	10	3.0	1.2	attraperen
-mb	0	2	@_at	u	??		52	2	-1.0	-2.0	inattent
-mb	0	2	in_at	u	??		5	1	-1.0	-1.0	inattent
-mb	0	2	on_at	u	??		2	1	-1.0	-1.0	onattent
-mb	0	1	an	u	??		40	2	-2.0	-1.0	anodisatie
+mb	0	1	ana	u	ana	[-p]	40	38	2.1	1.1	anaptyxis
-mb	0	2	@_ana	u	??		4	4	-2.0	-2.0	cryptoanalyse
-mb	0	2	crypt_ana	u	??		1	1	-2.0	-1.0	cryptanalytisch
-mb	0	2	crypto_ana	u	??		1	1	-2.0	-1.0	cryptoanalytisch
-mb	0	2	met_ana	u	??		1	1	-2.0	-1.0	metanalyse
-mb	0	2	micro_ana	u	??		1	1	-1.0	-1.0	microanalyse
-mb	0	2	ep_ana	u	??		1	1	-1.0	-1.0	epanalepsis
+mb	0	1	a	u	a	[-p]	30	2	-2.5	-1.0	avulsie
+mb	0	1	apo	u	apo	[-p]	30	28	2.0	1.0	apocalyptisch
-mb	0	2	@_apo	u	??		1	1	-1.0	-1.0	achterapostel
-mb	0	2	achter_apo	u	??		1	1	-1.0	-1.0	achterapostel
+mb	0	1	be	i	be	[+p]	1093	1093	2.2	1.5	bespraakt
-mb	0	2	@_be	i	??		365	365	1.4	1.8	overbejagen
-mb	0	2	mis_be	i	??		1	1	-1.0	-1.0	misbedeeld

-mb	0	2	on_be	i	??	160	160	1.5	1.0	onberoerd
-mb	0	2	un_be	i	??	160	2	-1.0	-1.0	unberufen
-mb	0	2	wan_be	i	??	9	9	1.1	1.0	wanbesluit
-mb	0	2	anti_be	i	??	1	1	-1.0	-1.0	antibeweging
-mb	0	2	aan_be	i	??	9	9	1.4	1.1	aanbetalen
-mb	0	3	@_aan_be	i	??	1	1	-1.0	-1.0	onderaanbesteding
-mb	0	3	onder_aan_be	i	??	1	1	-1.0	-1.0	onderaanbesteding
-mb	0	2	achter_be	i	??	1	1	-1.0	-1.0	achterbeslag
-mb	0	2	af_be	i	??	4	4	-1.5	-1.3	afbericht
-mb	0	3	@_af_be	i	??	1	1	-1.0	-1.0	voorafbetaling
-mb	0	3	voor_af_be	i	??	1	1	-1.0	-1.0	voorafbetaling
-mb	0	2	bij_be	i	??	8	8	1.3	1.0	bijbetrekking
-mb	0	2	door_be	i	??	2	2	-2.0	-1.0	doorberekening
-mb	0	2	in_be	i	??	5	5	1.2	1.0	inbegrip
-mb	0	2	mee_be	i	??	5	5	1.0	1.0	meebeleven
-mb	0	2	na_be	i	??	18	18	1.4	1.0	naberouw
-mb	0	2	onder_be	i	??	13	13	1.6	1.0	onderbewuste
-mb	0	2	over_be	i	??	23	23	1.4	1.0	overbekend
-mb	0	2	rond_be	i	??	1	1	-1.0	-1.0	rondbezorgen
-mb	0	2	tegen_be	i	??	13	13	1.0	1.0	tegenbewijs
-mb	0	2	terug_be	i	??	3	3	-1.6	-1.0	terugbetalen
-mb	0	2	toe_be	i	??	7	7	1.7	1.0	toebedenken
-mb	0	2	tussen_be	i	??	4	4	-1.0	-1.0	tussenbedrijf
-mb	0	2	uit_be	i	??	4	4	-2.7	-1.3	uitbesteden
-mb	0	3	@_uit_be	i	??	1	1	-2.0	-1.0	vooruitbetalen
-mb	0	3	voor_uit_be	i	??	1	1	-2.0	-1.0	vooruitbetalen
-mb	0	2	voor_be	i	??	25	25	1.7	1.0	voorbereidend
-mb	0	3	@_voor_be	i	??	1	1	-1.0	-1.0	onvoorbereid
-mb	0	3	on_voor_be	i	??	1	1	-1.0	-1.0	onvoorbereid
-mb	0	2	voort_be	i	??	2	2	-1.5	-1.0	voortbewegen
-mb	0	2	weg_be	i	??	1	1	-1.0	-1.0	wegbezuinigen
-mb	0	2	wel_be	i	??	23	23	1.1	1.0	welbevolkt
-mb	0	2	her_be	i	??	22	22	1.4	1.0	herbestemming
-mb	0	2	ver_be	i	??	1	1	-1.0	-1.0	verbestendigen

B.2.5 Prefixlijst met links-rechts-sortering

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
+mb	0	1	ana	u	ana	[-p]	36	39	2.0	1.1	metanalytisch
+mb	0	1	a	u	a	[-p]	29	2	-2.5	-1.0	aforist
+mb	0	1	apo	u	apo	[-p]	29	29	2.0	1.0	apograaf
+mb	0	1	be	i	be	[+p]	739	1100	2.2	1.5	bekanen
-mb	0	2	be_@	i	??		16	16	2.3	1.6	beïnvloeden
-mb	0	2	be_oor	i	??		2	2	-2.5	-2.0	herbeoordelen
-mb	0	2	be_dis	i	??		1	1	-2.0	-1.0	bediscussiëren
-mb	0	2	be_na	i	??		1	1	-1.0	-1.0	benadeligen
-mb	0	2	be_toe	i	??		1	1	-3.0	-1.0	betoelager
-mb	0	2	be_voor	i	??		3	3	-1.6	-1.5	bevoorradings
-mb	0	3	be_voor_@	i	??		1	1	-1.0	-1.0	onbevooroordeeld
-mb	0	3	be_voor_oor	i	??		1	1	-1.0	-1.0	onbevooroordeeld
-mb	0	2	be_com	i	??		2	1	-2.0	-1.0	becommentariëring
-mb	0	2	be_con	i	??		2	1	-2.0	-1.0	beconcurrering
-mb	0	2	be_ge	i	??		3	3	-3.6	-1.5	begeleid
-mb	0	2	be_in	i	??		1	1	-2.0	-1.0	beïnvloeding
-mb	0	2	be_ant	i	??		2	2	-2.0	-2.0	beantwoordings
+mb	0	1	bi	u	bi	[-p]	39	39	1.4	1.0	bitonaal
-mb	0	2	bi_@	u	??		1	1	-1.0	-1.0	bicommunautair
-mb	0	2	bi_com	u	??		1	1	-1.0	-1.0	bicommunautair

-mb	0	1	bis	u	??		39	1	-1.0	-1.0	bissectrice
+mb	0	1	co	u	co	[-p]	252	34	1.7	1.1	coherent
-mb	0	2	co_@	u	??		15	8	1.3	1.0	coëfficiënt
-mb	0	2	co_a	u	??		3	1	-1.0	-1.0	coacervaat
-mb	0	2	co_ad	u	??		3	1	-1.0	-1.0	coadjutor
-mb	0	2	co_as	u	??		3	1	-1.0	-1.0	coassuradeur
-mb	0	2	co_di	u	??		1	1	-1.0	-1.0	codimeer
-mb	0	2	co_ef	u	??		1	1	-1.0	-1.0	coëfficiënt
-mb	0	2	co_in	u	??		1	1	-2.0	-1.0	coïncidentie
-mb	0	2	co_pre	u	??		4	1	-1.0	-1.0	coprecipitatie
-mb	0	2	co_pro	u	??		4	1	-3.0	-1.0	coprofilie
-mb	0	1	cog	u	??		252	1	-1.0	-1.0	cognaat
-mb	0	1	col	u	??		252	22	3.0	1.2	rondcolporteren
-mb	0	1	com	u	??		252	88	3.3	1.8	commensalisme
-mb	0	2	com_@	u	??		15	1	-6.0	-1.0	compromissoir
-mb	0	2	com_pro	u	??		4	1	-6.0	-1.0	compromitterend
-mb	0	1	con	u	??		252	220	2.8	1.4	consortium
-mb	0	2	con_@	u	??		15	4	-1.0	-1.0	concomitant
-mb	0	2	con_co	u	??		1	1	-1.0	-1.0	concomitant
-mb	0	2	con_de	u	??		1	1	-1.0	-1.0	condescendentie
-mb	0	2	con_pro	u	??		4	1	-1.0	-1.0	conproportionering
-mb	0	2	con_sub	u	??		1	1	-1.0	-1.0	consubstantiatie
-mb	0	1	cor	u	??		252	13	3.0	1.4	correlair
-mb	0	2	cor_@	u	??		15	2	-5.5	-1.0	correspondentschap
-mb	0	2	cor_re	u	??		2	2	-5.5	-1.0	correspondentschap
-mb	0	1	kom	u	??		252	4	-1.7	-1.0	kompres

B.2.6 Resultaten van de externe evaluatie

De volgende subsecties bieden een samenvatting van twee exteren evaluatie-onderzoeken, namelijk een onderzoek aan het prefixdomein dat het resultaat is van een rechts-links-analyse van maximale prefisequenties, dus een vergelijking op het integrale prefixdomein (= domein 1), en een vergelijking op een domein dat optimaal aan het bereik van het MHB is aangepast (= domein 2), te weten het domein van (ongelede) prefixen met typefrequentie 5 of hoger.

Domein 1: prefixen, rechts-links-analyse zonder beperkingen

Domeinkenmerken

- vrije sequentielengte
- geen minimum-frequentie

Algemene kencijfers

lexeemfrequentie	aantal types	aantal tokens
f = 50+	65	19974
f = 10+	188	22798
f = [0,10]	949	2000
f = 0+	1137	24798

Tabel 6: aantal prefixen per frequentieklasse (zowel op type-niveau als op token-niveau)

productiviteits-klasse (i of u)	aantal items in u-klasse	aantal items in i-klasse
2.0 of meer	43	7
1.5 of meer	92	20
1.2 of meer	146	53
0 of meer	188	188

Tabel 7: aantal prefixen per inwaartse (i) en uitwaartse (u) productiviteitsklasse voor prefixen met een lexeemfrequentie van 10 of hoger

Evaluatie van de MGBN in termen van het aantal MHB-treffers

frequentieklasse	aantal types	aantal hb-treffers
0+	1137	106 (9%)
10+	188	76 (40%)

Tabel 8: aantal hb-treffers per frequentieklasse (absoluut en relatief)

Evaluatie van het MHB in termen van het aantal MGBN-treffers

prefixklasse	aantal hb-types	aantal mgbn-treffers	aantal mgbn-missers
freq = 0+	126	106 (84 %)	20 (16 %)

Tabel 9: mgbn-dekking van hb-prefix-eenheden (absoluut en relatief)

lijst van onvindbare (want anders gecodeerde) MHB-prefixen:
aaneen, aarts, b, bijeen, binnen, boven, buiten, hecto, loco, oer, omhoog, omlaag, opper, oud, semi, terecht, thuis, turbo, uiteen

Domein 2: prefixen, rechts-links-analyse met beperkingen

Domeinkenmerken

- maximale sequentielengte = 1
- minimum suffixfrequentie = 5

Algemene kencijfers

lexeemfrequentie	aantal types	aantal tokens
f = 50+	54	17965
f = 10+	120	19585
f = [5,10]	44	287
f = 5+	164	19872

Tabel 10: aantal prefixen per frequentieklasse (op type-niveau en op token-niveau)

productiviteits-klasse (i of u)	aantal items in u-klasse	aantal items in i-klasse
minstens 2.0	42	1
minstens 1.5	80	7
minstens 1.2	109	28
- (geen eis)	120	120

Tabel 11: aantal prefixen per inwaartse (i) en uitwaartse (u) productiviteitsklasse voor prefixen met een lexeemfrequentie van 10 of hoger

Evaluatie van de MGBN in termen van het aantal MHB-treffers

frequentieklasse	aantal types	aantal hb-treffers
5+	164	80 (48 %)
10+	120	71 (59 %)

Tabel 12: aantal hb-treffers per frequentieklasse (absoluut en relatief)

Evaluatie van het MHB in termen van het aantal MGBN-treffers

prefixklasse	aantal hb-types	aantal mgbn-treffers	aantal mgbn-elders	aantal mgbn-missers
freq = 5+	126	80 (63 %)	20 (16%)	26 (21 %)

Tabel 13: mgbn-dekking van hb-prefix-eenheden (absoluut en relatief)

lijst van weggefilterde (want laag-frequente) MHB-prefixen:
aaneen, achteraan, achteraf, achterna, achterom, achterop, achteruit, ambi, amfi, circum, crypto, d, etno, intra, non, omver, onderuit, pluri, pseudo, retro, vooraan, vooraf, voorbij, voorin, voorop, voorover, vooruit

B.3 Resultaten van de suffix-analyses

B.3.1 Introductie

Mijn computationele onderzoek naar de eigenschappen van het MGBN-model met betrekking tot de suffixdistributie heeft een hele reeks digitale datarapporten opgeleverd. Appendix F biedt gedetailleerde informatie over de beschikbare bestanden en hun vindplaats. In deze sectie zal ik me beperken tot de presentatie van enkele voorbeeldlijsten, namelijk een lijst met de hoogstfrequente suffixen (6.6.3.1), een lijst met de laagstfrequente suffixen (6.6.3.2), een lijst met de hoogstfrequente suffixklassen (6.6.3.3) en voorbeelden van een rechts-links gesorteerde suffixlijst (6.6.3.4) en een links-rechts gesorteerde suffixlijst (6.6.3.5). Alle lijsten richten zich primair op de eigenschappen bij suffixen in de n2vorm. Hieronder volgt een korte toelichting op de bijbehorende veldstructuur. Samen met deze toelichting spreken de lijsten verder voor zich.

Veldstructuur van de op frequentie gesorteerde suffixlijsten in B3.2 en B3.3
1: aantal suffix-eenheden
2: n2vorm van mgbn-suffix
3: inheems (i) of uitheems (u) suffix
4: mhb-vorm van suffix (indien beschikbaar, anders '??')
5: u-frequentie van de wortel (= aantal suffixen dat direct na de wortel kan staan)
6: i-categorie = invoer-categorie = (potentiële) lexeemcategorie van i-stam
7: relatie-teken (X>Y betekent functie van X naar Y)
8: u-categorie = uitvoer-categorie = (potentiële) lexeemcategorie van u-stam
9: mhb-status: productief (+p) of improductief (-p)
10: stamfrequentie bij de n3vorm
11: stamfrequentie bij de n2vorm
12: het aandeel van de n2vorm met categorie in de stamfrequentie van de n3vorm
13: u-ratio: gemiddeld aantal uitwaartse lexeemspecificaties per suffix (negatief als f < 6)
14: i-ratio: gemiddeld aantal inwaartse stamspecificaties per suffix (negatief als f < 6)
15: voorbeeld van lexeemtoepassing (in n0vorm)

Veldstructuur van de op vorm gesorteerde suffixlijsten (-icat) in B3.4
1: aantal suffix-eenheden
2: n2vorm van mgbn-suffix
3: inheems (i) of uitheems (u) suffix
4: mhb-vorm van suffix (indien beschikbaar, anders '??')
5: u-frequentie van de wortel (= aantal suffixen dat direct na de wortel kan staan)
6: u-categorie = uitvoer-categorie = (potentiële) lexeemcategorie
7: mhb-status: productief (+p) of improductief (-p)

8: stamfrequentie bij de n3vorm
9: stamfrequentie bij de n2vorm
10: stamfrequentie bij de n3vorm met u-categorie
11: stamfrequentie bij de n2vorm met u-categorie
12: het aandeel van de n3vorm met categorie in de stamfrequentie van de n3vorm
13: het aandeel van de n2vorm met categorie in de stamfrequentie van de n3vorm
14: u-ratio: gemiddeld aantal uitwaartse lexeemspecificaties per suffix (negatief als f < 6)
15: i-ratio: gemiddeld aantal inwaartse stamspecificaties per suffix (negatief als f < 6)
16: voorbeeld van lexeemtoepassing (in n0vorm)

Veldstructuur van de op vorm gesorteerde suffixlijsten (+icat) in B3.5
6: i-categorie = uitvoer-categorie = (potentiële) lexeemcategorie
7: relatie-teken (X>Y betekent functie van X naar Y)
8: u-categorie = uitvoer-categorie = (potentiële) lexeemcategorie
9: mhb-status: productief (+p) of improductief (-p)
10: stamfrequentie bij de n3vorm
11: stamfrequentie bij de n2vorm
12: stamfrequentie bij de n3vorm met u-categorie
13: stamfrequentie bij de n2vorm met u-categorie
14: het aandeel van de n3vorm met categorie in de stamfrequentie van de n3vorm
15: het aandeel van de n2vorm met categorie in de stamfrequentie van de n3vorm
16: u-ratio: gemiddeld aantal uitwaartse lexeemspecificaties per suffix (negatief als f < 6)
17: i-ratio: gemiddeld aantal inwaartse stamspecificaties per suffix (negatief als f < 6)
18: voorbeeld van lexeemtoepassing (in n0vorm)

B.3.2 De hoogstfrequente suffixen

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	er	i	(d)er	4.7	-	>	N	[+p]	2319	2319	100%	1.1	1.5	vaster
1	s	i	s	3.4	-	>	N		1938	1938	100%	1.0	1.4	genots
1	ig	i	ig	3.8	-	>	A	[+p]	1803	1803	100%	1.0	1.4	spikkelig
1	er	i	(d)er	5.3	V	>	N	[+p]	2319	1790	77%	1.1	1.6	plamodder
1	isch	u	isch	3.1	-	>	A	[+p]	1742	1742	100%	1.0	1.6	encyclopedisch
1	s	i	s	3.6	N	>	N		1938	1733	89%	1.0	1.4	beleefdheids
1	er	i	(d)er	6.4	N	>	N	[+p]	2319	1243	53%	1.1	1.2	vinker
1	heid	i	heid	3.7	-	>	N	[+p]	1235	1235	100%	1.0	1.7	minzaamheid
1	d	i	d	1.7	-	>	A	[+p]	1202	1202	100%	1.0	1.6	geconsolideerd
1	je	i	X:je	6.1	-	>	N	[+p]	1146	1146	100%	1.0	1.1	sprietje
1	heid	i	heid	3.8	A	>	N	[+p]	1235	1117	90%	1.0	1.6	louterheid
2	at_ie	u	atie	3.1	-	>	N	[?p]	1030	1030	100%	1.0	1.6	rehabilitatie
1	ing	i	ing	6.3	N	>	N		3969	1002	25%	1.0	1.3	hearing
1	en	i	en	4.3	-	>	N	[-p]	992	992	100%	1.0	1.1	tunicaten
1	je	i	X:je	6.7	N	>	N	[+p]	1146	988	86%	1.0	1.0	koopje
1	e	i	e	5.6	N	>	N	[+p]	2441	960	39%	1.0	1.1	bate
1	ig	i	ig	6.3	N	>	A	[+p]	1803	869	48%	1.1	1.1	volhoevig
1	ie	u	ie	5.1	N	>	N	[+p]	3522	768	21%	1.1	1.4	francofilie
1	eer'en	i	eer	6.1	N	>	V	[+p]	2521	766	30%	1.0	1.2	gronderen
1	isch	u	isch	4.1	N	>	A	[+p]	1742	757	43%	1.0	1.3	naturalistisch
1	ig	i	ig	6.6	V	>	A	[-p]	1803	728	40%	1.1	1.1	seuterig
1	s	i	s	3.9	-	>	A	[+p]	703	703	100%	1.1	1.1	franciscaans
2	ism_e	i	isme	4.0	-	>	N	[-p]	699	699	100%	1.0	1.2	objectivisme

B.3.3 De laagstfrequente suffixen

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	aard	i	aard	1.0	-	>	B		1	1	100%	-1.0	-1.0	uiteraard
1	ast	u	ast	1.0	-	>	A		1	1	100%	-1.0	-1.0	enthousiast
1	d	i	d	1.0	-	>	?		1	1	100%	-1.0	-1.0	gemengdslachtig
2	aar_d	i	aard	1.0	-	>	A		1	1	100%	-1.0	-1.0	ongeëvenaard
2	eer_d	i	eerd	1.0	-	>	B		1	1	100%	-1.0	-1.0	gemoedereerd
2	eer_d	i	eerd	1.0	-	>	N		1	1	100%	-1.0	-1.0	gedistilleerd
2	d_e	i	de	1.0	-	>	O		1	1	100%	-1.0	-1.0	vernonde
1	heid	i	heid	1.0	-	>	O		1	1	100%	-1.0	-1.0	inzonderheid
1	lijk	i	(e)lijk	1.0	-	>	T		1	1	100%	-1.0	-1.0	gedrieënlijk
1	et	u	et	1.0	-	>	?		1	1	100%	-2.0	-1.0	tetterettet
2	in_o	u	ino	1.0	-	>	A		1	1	100%	-1.0	-1.0	solferino
2	ij_e	i ?	ije	1.0	-	>	N	[-p]	1	1	100%	-1.0	-1.0	commanderije
1	beet	u	??	1.0	-	>	A		1	1	100%	-1.0	-1.0	analfabeet
1	fant	u	??	1.0	-	>	N		1	1	100%	-1.0	-1.0	sycofant
2	ec_o:fiel	u	??	1.0	-	>	A		1	1	100%	-1.0	-1.0	myrmecofiel
2	ec_o:fiel	u	??	1.0	-	>	N		1	1	100%	-1.0	-1.0	myrmecofiel
1	geen	u	??	1.0	-	>	O		1	1	100%	-1.0	-1.0	estrogeen
1	meer	u	??	1.0	-	>	V		1	1	100%	-1.0	-1.0	eximeer
1	wijl	i	??	1.0	-	>	B		1	1	100%	-1.0	-1.0	middelerwijl
2	lijk_aard	i	??	1.0	-	>	N		1	1	100%	-1.0	-1.0	lelijkaard
1	wijs	i	??	1.0	-	>	C		1	1	100%	-1.0	-1.0	gelijkerwijs
1	\>ere	u	??	1.0	-	>	A		1	1	100%	-1.0	-1.0	bajad\>ere
1	\>ere	u	??	1.0	-	>	O		1	1	100%	-1.0	-1.0	arri\>ere
2	ent_a	u	??	1.0	-	>	A		1	1	100%	-1.0	-1.0	irredenta
2	ar_a	u	??	1.0	-	>	A		1	1	100%	-1.0	-1.0	alcantara
2	@_a	u	??	1.0	-	>	B@		1	1	100%	-1.0	-1.0	allottava
2	av_a	u	??	1.0	-	>	B		1	1	100%	-1.0	-1.0	allottava
2	in_ade	u	??	1.0	-	>	N		1	1	100%	-1.0	-1.0	harlekinade
2	@_aar	i	??	1.0	-	>	O@		1	1	100%	-1.0	-1.0	alveolaar
2	ol_aar	i	??	1.0	-	>	O		1	1	100%	-1.0	-1.0	alveolaar
2	in_aat	u	??	1.0	-	>	A		1	1	100%	-1.0	-1.0	subordinaat
2	ur_aat	u	??	1.0	-	>	N		1	1	100%	-1.0	-1.0	barbituraat
2	il_aat	u	??	1.0	-	>	N		1	1	100%	-1.0	-1.0	antranilaat
2	isc_aat	u	??	1.0	-	>	N		1	1	100%	-1.0	-1.0	lemniscaat
2	end_iaat	u	??	1.0	-	>	N		1	1	100%	-1.0	-1.0	stipendiaat

B.3.4 Voorbeeldlijst met links-rechts-perspectief (excl. cat-markering)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	ief	u	ief	3.4	-	[+p]	395	**366**	395	**366**	100%	**100%**	1.2	1.7	combattief
1	ief	u	ief	7.0	**?**		395	**366**	1	**1**	0%	**0%**	-2.0	-1.0	intensief
1	ief	u	ief	3.2	**A**	[+p]	395	**366**	322	**321**	81%	**87%**	1.2	1.6	participatief
1	ief	u	ief	5.0	**B**		395	**366**	3	**3**	0%	**0%**	-1.0	-1.0	formatief
1	ief	u	ief	3.8	**N**	[-p]	395	**366**	118	**115**	29%	**31%**	1.7	1.3	immutatief
1	ief	u	ief	1.0	**V**		395	**366**	1	**1**	0%	**0%**	-1.0	-1.0	aperitief
1	iev	u	??	3.4	-		395	**18**	395	**18**	100%	**100%**	1.1	1.1	actieveling
1	iev	u	??	3.2	**A**		395	**18**	322	**17**	81%	**94%**	1.1	1.0	actieven
1	iev	u	??	5.0	**B**		395	**18**	3	**1**	0%	**5%**	-2.0	-1.0	respectievelijk
1	iev	u	??	3.8	**N**		395	**18**	118	**7**	29%	**38%**	1.2	1.0	primitieven
1	iev'en	u	??	3.4	-		395	**1**	395	**1**	100%	**100%**	-1.0	-1.0	aperitieven
2	iv_at	u	??	13.0	-		2	**2**	2	**2**	100%	**100%**	-1.0	-1.0	cultivator
3	iv_at_or	u	??	4.0	-		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	cultivator
3	iv_at_or	u	??	4.0	**N**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	cultivator

3	iv_at_ie	u	??	22.0	-		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	motivatie
3	iv_at_ie	u	??	22.0	**N**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	motivatie
3	iv_at_@	u	??	13.0	**@**		2	**2**	2	**2**	100%	**100%**	-1.0	-1.0	motivatie
2	iev_e	u	??	8.5	-		4	**4**	4	**4**	100%	**100%**	-1.0	-1.0	exclusieve
2	iev_e	u	??	8.5	**N**		4	**4**	4	**4**	100%	**100%**	-1.0	-1.0	executieve
2	iv_e	u	??	4.0	-		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	cultivé
2	iv_e	u	??	4.0	**?**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	cultivé
2	if_eer	u	??	5.4	-		25	**1**	25	**1**	100%	**100%**	-1.0	-1.0	cruciferen
2	iv_eer	u	??	5.4	-		25	**18**	25	**18**	100%	**100%**	1.2	2.0	overgecultiveerd
2	iv_eer	u	??	6.1	**V**		25	**18**	19	**13**	76%	**72%**	1.3	1.4	activerings
2	iv_eer'en	u	??	5.4	-		25	**19**	25	**19**	100%	**100%**	1.0	1.7	activeren
2	iv_eer'en	u	??	6.1	**V**		25	**19**	19	**19**	76%	**100%**	1.0	1.7	desactiveren
3	iv_eer_baar	u	??	4.0	-		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	objectiveerbaar
3	iv_eer_baar	u	??	4.0	**A**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	objectiveerbaar
3	iv_eer_d	u	??	1.4	-		5	**5**	5	**5**	100%	**100%**	1.0	2.5	ongemotiveerd
3	iv_eer_d	u	??	1.4	**A**		5	**5**	5	**5**	100%	**100%**	1.0	2.5	overgecultiveerd
3	if_eer_en	u	??	13.0	-		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	cruciferen
3	if_eer_en	u	??	13.0	**N**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	cruciferen
3	iv_eer_der	u	??	2.0	-		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	deactiveerder
3	iv_eer_der	u	??	2.0	**N**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	deactiveerder
3	iv_eer_ing	u	??	8.0	-		13	**13**	13	**13**	100%	**100%**	1.2	1.4	activerings
3	iv_eer_ing	u	??	8.0	**N**		13	**13**	13	**13**	100%	**100%**	1.2	1.4	archivering
4	iv_eer_ing_s	u	??	11.0	-		3	**3**	3	**3**	100%	**100%**	-1.0	-1.5	reactiverings
4	iv_eer_ing_s	u	??	11.0	**N**		3	**3**	3	**3**	100%	**100%**	-1.0	-1.5	relativerings
4	iv_eer_ing_@	u	??	11.0	**@**		3	**3**	3	**3**	100%	**100%**	-1.0	-1.5	relativerings
3	if_eer_@	u	??	6.5	**@**		19	**1**	19	**1**	100%	**100%**	-1.0	-1.0	cruciferen
3	iv_eer_@	u	??	6.5	**@**		19	**18**	19	**18**	100%	**100%**	1.2	2.0	overgecultiveerd
2	iev_en	u	??	6.0	-		5	**5**	5	**5**	100%	**100%**	1.0	1.2	actieven
2	iev_en	u	??	6.0	**N**		5	**5**	5	**5**	100%	**100%**	1.0	1.2	inactieven
2	ief_heid	u	??	10.0	-		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	massiefheid
2	ief_heid	u	??	10.0	**N**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0	massiefheid
2	if_iek	u	??	7.0	-		2	**2**	2	**2**	100%	**100%**	-1.0	-1.0	pacifiek
2	if_iek	u	??	7.0	**A**		2	**2**	2	**2**	100%	**100%**	-1.0	-1.0	pacifiek
2	if_isch	u	??	5.3	-		8	**1**	8	**1**	100%	**100%**	-1.0	-1.0	signifisch
2	if_isch	u	??	5.0	**A**		8	**1**	6	**1**	75%	**100%**	-1.0	-1.0	signifisch
2	iv_is	u	??	5.3	-		8	**2**	8	**2**	100%	**100%**	-3.0	-2.0	collectivisatie
2	iv_isch	u	??	5.3	-		8	**5**	8	**5**	100%	**100%**	1.0	1.0	adjectivisch
2	iv_isch	u	??	5.0	**A**		8	**5**	6	**5**	75%	**100%**	1.0	1.0	recitativisch
3	iv_is_at	u	??	6.5	-		2	**2**	2	**2**	100%	**100%**	-1.0	-2.0	collectivisatie
4	iv_is_at_ie	u	??	6.5	-		2	**2**	2	**2**	100%	**100%**	-1.0	-2.0	decollectivisatie
4	iv_is_at_ie	u	??	6.5	**N**		2	**2**	2	**2**	100%	**100%**	-1.0	-2.0	collectivisatie

B.3.5 Voorbeeldlijst met rechts-links-perspectief (incl. cat-markering)

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1		ief	u	ief	3.6	-	>	**A**	[+p]	322	**321**	322	**321**	100%	**100%**	1.0	1.6	communicatief
1		ief	u	ief	6.5	**N**	>	**A**	[+p]	322	**321**	70	**70**	21%	**21%**	1.0	1.4	attributief
1		ief	u	ief	6.8	**A**	>	**A**		322	**321**	26	**26**	8%	**8%**	1.0	1.3	subjectief
1		ief	u	ief	12.0	**?**	>	**A**		322	**321**	10	**10**	3%	**3%**	1.1	1.0	actief
1		ief	u	ief	9.7	**V**	>	**A**		322	**321**	9	**9**	2%	**2%**	1.0	1.1	sportief
1		ief	u	ief	8.7	**O**	>	**A**		322	**321**	4	**4**	1%	**1%**	-1.0	-1.0	foutief
1		ief	u	ief	23.0	**P**	>	**A**		322	**321**	1	**1**	0%	**0%**	-2.0	-1.0	actief
1		ief	u	ief	6.0	**B**	>	**A**		322	**321**	1	**1**	0%	**0%**	-1.0	-1.0	decisief
1		if	u	??	3.6	-	>	**A**		322	**1**	322	**1**	100%	**100%**	-1.0	-1.0	captif
1		ive	u	??	6.8	**A**	>	**A**		322	**1**	26	**1**	8%	**100%**	-1.0	-1.0	intensive
1		ive	u	??	3.6	-	>	**A**		322	**1**	322	**1**	100%	**100%**	-1.0	-1.0	intensive
2	@_ief	u	??	4.6	-	>	**A@**		158	**158**	158	**158**	100%	**100%**	1.0	1.2	interpretatief

2	@_ief	u	??	10.2	**N**	>	**A@**		158	**158**	34	**34**	21%	**21%**	1.0	1.0 limitatief
2	@_ief	u	??	13.9	**?**	>	**A@**		158	**158**	14	**14**	8%	**8%**	1.0	1.0 charitatief
2	@_ief	u	??	9.9	**A**	>	**A@**		158	**158**	13	**13**	8%	**8%**	1.0	1.0 ultimatief
2	@_ief	u	??	11.9	**V**	>	**A@**		158	**158**	12	**12**	7%	**7%**	1.0	1.0 duratief
2	@_ief	u	??	13.8	**O**	>	**A@**		158	**158**	5	**5**	3%	**3%**	1.2	1.0 potestatief
2	@_ief	u	??	13.0	**P**	>	**A@**		158	**158**	3	**3**	1%	**1%**	-1.3	-1.0 possessief
2	o:cept_ief	u	??	9.0	**-**	>	**A**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0 proprioceptief
2	o:cept_ief	u	??	9.0	**A**	>	**A**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0 proprioceptief
2	it_ief	u	itief	6.6	**-**	>	**A**	[?p]	21	**21**	21	**21**	100%	**100%**	1.0	1.1 acquisitief
2	it_ief	u	itief	12.8	**N**	>	**A**		21	**21**	8	**8**	38%	**38%**	1.0	1.0 partitief
2	it_ief	u	itief	20.0	**?**	>	**A**	[?p]	21	**21**	3	**3**	14%	**14%**	-1.0	-1.0 factitief
2	it_ief	u	itief	24.0	**A**	>	**A**		21	**21**	2	**2**	9%	**9%**	-1.0	-1.0 primitief
2	it_ief	u	itief	21.5	**V**	>	**A**		21	**21**	2	**2**	9%	**9%**	-1.0	-1.0 partitief
2	it_ief	u	itief	28.0	**O**	>	**A**		21	**21**	1	**1**	4%	**4%**	-1.0	-1.0 factitief
2	at_ief	u	atief	4.3	**-**	>	**A**	[?p]	129	**128**	129	**128**	100%	**100%**	1.0	1.2 declaratief
2	at_ief	u	atief	8.8	**N**	>	**A**		129	**128**	25	**25**	19%	**19%**	1.0	1.0 curatief
2	at_ief	u	atief	11.2	**?**	>	**A**	[?p]	129	**128**	10	**10**	7%	**7%**	1.0	1.0 summatief
2	at_ief	u	atief	10.0	**V**	>	**A**		129	**128**	10	**10**	7%	**7%**	1.0	1.0 registratief
2	at_ief	u	atief	7.2	**A**	>	**A**		129	**128**	10	**10**	7%	**7%**	1.0	1.0 imperatief
2	at_ief	u	atief	10.2	**O**	>	**A**		129	**128**	4	**4**	3%	**3%**	-1.0	-1.0 potestatief
2	at_ief	u	atief	8.0	**P**	>	**A**		129	**128**	2	**2**	1%	**1%**	-1.0	-1.0 conatief
2	uat_ief	u	??	4.3	**-**	>	**A**		129	**1**	129	**1**	100%	**100%**	-1.0	-1.0 evaluatief
2	ent_ief	u	??	23.0	**-**	>	**A**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0 agentief
2	ess_ief	u	??	13.0	**-**	>	**A**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0 possessief
2	est_ief	u	??	1.0	**-**	>	**A**		1	**1**	1	**1**	100%	**100%**	-1.0	-1.0 intempestief
2	ut_ief	u	??	2.6	**-**	>	**A**		5	**5**	5	**5**	100%	**100%**	1.0	1.6 distributief

B.3.6 Resultaten van de externe evaluatie

De volgende subsecties bieden een samenvatting van twee externe evaluatie-onderzoeken naar de sufifxdimensie, namelijk een onderzoek aan het suffixdomein dat het resultaat is van een rechts-links-analyse van suffixsequenties op eindpositie (dus met een expliciete lexeem-categorie), dus een vergelijking op het integrale suffixdomein (= domein 1), en een vergelijking op een domein dat optimaal aan het bereik van het MHB is aangepast (= domein 2), te weten het domein dat beperkt is tot suffixsequenties met 1 of 2 eenheden en tot hoofdtypes met typefrequentie 5 of hoger.

Domein 1: suffixen, rechts-links-analyse zonder beperkingen

Domeinkenmerken

- suffixsequenties op eindpositie
- rechts-links-perspectief
- vrije sequentielengte
- geen minimum-frequentie

Algemene kencijfers

lexeemfrequentie	aantal types	aantal tokens
f = 50+	154	54237
f = 10+	488	62342
f = [0,10]	4063	8802
f = 0+	4352	71144

Tabel 14: aantal suffixen per frequentieklasse (zowel op type-niveau als op token-niveau)

productiviteits-klasse (i of u)	aantal items in u-klasse	aantal items in i-klasse
minstens 2.0	1	10
minstens 1.5	4	43
minstens 1.2	13	158
- (geen eis)	488	488

Tabel 15: aantal suffixen per inwaartse (i) en uitwaartse (u) productiviteitsklasse voor suffixen met een lexeemfrequentie van 10 of hoger

lexeemfrequentie	aantal ucat-types	aantal ucat-tokens
f = 50+	176	?
f = 10+	570	?
f = [0,10]	4561	?
f = 0+	5131	?

Tabel 16: aantal ucat-suffixen per frequentieklasse (op type-niveau en op token-niveau)

productiviteits-klasse (i of u)	aantal items in u-klasse	aantal items in i-klasse
minstens 2.0	1	10
minstens 1.5	4	43
minstens 1.2	13	165
- (geen eis)	570	570

Tabel 17: aantal ucat-suffixen per inwaartse (i) en uitwaartse (u) productiviteitsklasse voor suffixen met een lexeemfrequentie van 10 of hoger

categorie	aantal	categorie	aantal
N	3180	P	16
V	333	T	24
A	999	O	204
B	184	@	1058

Tabel 18: aantal ucat-suffixen per u- categorie

Evaluatie van de MGBN in termen van het aantal MHB-treffers

frequentieklasse	aantal suffixtypes	aantal hb-treffers
0+	4352	215 (4%)
10+	488	163 (33%)

Tabel 19: aantal hb-treffers per frequentieklasse

suffixklasse	aantal suffixtypes	aantal hb-treffers
hoofdtype	4352	233 (5 %)
ucat-type	5131	268 (5 %)
icat-type	10687	388 (4 %)

Tabel 20: aantal hb-treffers per suffixklasse, zonder frequentieconditie

suffixklasse	aantal suffixtypes	aantal hb-treffers
hoofdtype	488	176 (36 %)
ucat-type	570	191 (33 %)
icat-type	812	236 (29 %)

Tabel 21: aantal hb-treffers per suffixklasse, beperkt tot suffixen met frequentie 10+

Evaluatie van het MHB in termen van het aantal MGBN-treffers

suffixklasse	aantal hb-types	aantal mgbn-treffers	aantal mgbn-missers
hoofdtype	245	215 (87 %)	30 (12 %)
ucat-type	246	232 (94 %)	14 (5 %)
icat-type	377	350 (92 %)	27 (7 %)

Tabel 22: MGBN-dekking van hb-suffixen (per suffixklasse)

lijst van onvindbare (want anders gecodeerde) MHB-suffixen (7):
se, taria, gogie, lude, gewijs, isering ,t

lijst van niet-terugvindbare MHB-suffixen onder gegeven cat-specificatie (6 items):
elijk (P>A), erwijs (A), erwijs (A>B), et (T>N), eut (N>N), ied (N>N)

Domein 2: suffixen, rechts-links-analyse met beperkingen

Domeinkenmerken

- suffixsequenties op eindpositie
- rechts-links-perspectief
- maximale sequentielengte = 2
- minimum suffixfrequentie = 5

Algemene kencijfers

lexeemfrequentie	aantal types	aantal tokens
f = 50+	145	53401
f = 10+	422	60420
f = [5,10]	350	2394
f = 5+	717	62814

Tabel 23: aantal suffixen per frequentieklasse (op type-niveau en op token-niveau)

productiviteits-klasse (i of u)	aantal items in u-klasse	aantal items in i-klasse
minstens 2.0	1	10
minstens 1.5	4	40
minstens 1.2	13	131
- (geen eis)	422	422

Tabel 24: aantal suffixen per inwaartse (i) en uitwaartse (u) productiviteitsklasse voor suffixen met een lexeemfrequentie van 10 of hoger

lexeemfrequentie	aantal ucat-types	aantal ucat-tokens
f = 50+	167	?
f = 10+	504	?
f = [5,10]	373	?
f = 5+	877	?

Tabel 25: aantal ucat-suffixen per frequentieklasse (op type-niveau en op token-niveau)

productiviteits-klasse (i of u)	aantal items in u-klasse	aantal items in i-klasse
minstens 2.0	1	10
minstens 1.5	4	40
minstens 1.2	13	138
- (geen eis)	504	504

Tabel 26: aantal ucat-suffixen per inwaartse (i) en uitwaartse (u) productiviteitsklasse voor suffixen met een lexeemfrequentie van 10 of hoger

categorie	aantal	categorie	aantal
N	534	P	5
V	54	O	33
A	182	@	153
B	32	?	30
T	5		

Tabel 27: aantal ucat-suffixen per u-categorie

Evaluatie van de MGBN in termen van het aantal MHB-treffers

suffixklasse	aantal suffixtypes	aantal hb-treffers
hoofdtype	717	202 (28 %)
ucat-type	877	224 (25 %)
icat-type	1405	289 (20 %)

Tabel 28: aantal hb-treffers per suffixklasse, zonder aanvullende frequentieconditie

suffixklasse	aantal suffixtypes	aantal hb-treffers
hoofdtype	422	176 (41 %)
ucat-type	504	191 (37 %)
icat-type	768	236 (30 %)

Tabel 29: aantal hb-treffers per suffixklasse, beperkt tot suffixen met frequentie 10+

Evaluatie van de MHB in termen van het aantal MGBN-treffers

suffixklasse	aantal hb-types	aantal mgbn-treffers	aantal mgbn-elders	aantal mgbn-missers
hoofdtype	245	193 (78 %)	30 (12 %)	22 (8 %)
ucat-type	246	210 (85 %)	14 (5 %)	22 (8 %)
icat-type	377	275 (72 %)	15 (3 %)	87 (23 %)

Tabel 30: mgbn-dekking van hb-suffixen (per suffixklasse)

lijst van weggefilterde (want laag-frequente) MHB-suffixen (21 items): *anda, ande, droom, egge, ele, enda, etie, ied, ieur, ikoos, ineer, ioen, itsa, izie, (e)lijks, ooi, rama, sofie, staat, uleer, waarts*

B.4 Resultaten van de analyse op prefix-suffix-combinaties

B.4.1 Introductie

De voorbeeldlijsten met informatie uit het datarapport met prefix-suffix-combinaties kennen de volgende veldstructuur:

1. lengteklasse: morfeemlengte van de geanalyseerde lexeemrepresentaties
2. formele structuurklasse: P + W + S (= aantal prefixen + 1 wortel + aantal suffixen)
3. totaal aantal lexemen in de lengteklasse uit veld 1
4. aantal lexemen in de formele structuurklasse uit veld 2
5. aandeel van de structuurklasse in de lengteklasse
6. eerste prefix in morfeemsequentie (indien bestaand)
7. inheems / uitheems
8. lexeemrepresentatie zonder wortel (= prefix-suffix-patroon)
9. i-categorie (= categorie van complement-lexeem bij eerste prefix)
10. relatie-teken (X>Y betekent functie van X naar Y)
11. u-categorie (= resulterende lexeemcategorie)
12. wel/niet productief (volgens informatie in MHB)
13. voorbeeld van een lexeem uit de geanalyseerde lexeemklasse

In B.4.4 bevat de tabel twee extra velden:

12. aantal stamtoepassingen van de centrale morfeemcombinatie
13. aandeel van specifieke morfeemcombinatie in structuurklasse
14. wel/niet productief (volgens informatie in MHB)
15. voorbeeld van een lexeem uit de geanalyseerde lexeemklasse

B.4.2 Voorbeeldlijst: lexemen met 8 morfemen

1	2	3	4	5	6	7	8	9	10	11	12	13
8	2 + 1 + 5	7	3	42%	de	u	de;con;[-];ion;al;is;eer;ing	-	>	N	[-p]	de;con;[fess];ion;al;is;eer;ing
8	2 + 1 + 5	7	3	42%	de	u	de;pro;[-];ion;al;is;eer;en	-	>	V	[-p]	de;pro;[fess];ion;al;is;eer;en
8	2 + 1 + 5	7	3	42%	de	u	de;pro;[-];ion;al;is;eer;en	V	>	V	[-p]	pro;[fess];ion;al;is;eer;en
8	2 + 1 + 5	7	3	42%	de	u	de;con;[-];ion;al;is;eer;en	-	>	V	[-p]	de;con;[fess];ion;al;is;eer;en
8	1 + 1 + 6	7	4	57%	im	u	in;[-];ut;ion;al;is;eer;ing	-	>	N	[-p]	in;[stit];ut;ion;al;is;eer;ing
8	1 + 1 + 6	7	4	57%	im	u	in;[-];ut;ion;al;is;eer;en	-	>	V	[-p]	in;[stit];ut;ion;al;is;eer;en
8	1 + 1 + 6	7	4	57%	per	i	per;[-];ic;ul;ar;is;eer;en	-	>	V	[-p]	per;[pend];ic;ul;ar;is;eer;en
8	1 + 1 + 6	7	4	57%	im	u	in;[-];ut;ion;al;is;at;ie	-	>	N	[-p]	in;[stit];ut;ion;al;is;at;ie

B.4.3 Voorbeeldlijst: lexemen met 7 morfemen

1	2	3	4	5	6	7	8	9	10	11	12	13
7	3 + 1 + 3	58	7	12%	!on	i	on;ge;dis;[-];in;eer;d	-	>	A	[-p]	on;ge;dis;[cipl];in;eer;d
7	3 + 1 + 3	58	7	12%	!on	i	on;ge;dis;[-];in;eer;d	A	>	A	[-p]	ge;dis;[cipl];in;eer;d
7	3 + 1 + 3	58	7	12%	ge	i	ge;des;il;[-];ion;eer;d	-	>	A	[-p]	ge;des;il;[lus];ion;eer;d
7	3 + 1 + 3	58	7	12%	!on	i	on;ge;con;[-];ion;eer;d	-	>	A	[-p]	on;ge;con;[dit];ion;eer;d
7	3 + 1 + 3	58	7	12%	!on	i	on;ge;con;[-];ion;eer;d	A	>	A	[-p]	ge;con;[dit];ion;eer;d
7	3 + 1 + 3	58	7	12%	!on	i	on;ge;pre;[-];ic;i:eer;d	-	>	A	[-p]	on;ge;pre;[jud];ic;i:eer;d

7	3 + 1 + 3	58	7	12%	!on	i	on;ge;com;[-];eer;d;heid		-	>	**N**	[-p]	on;ge;com;[plic];eer;d;heid
7	3 + 1 + 3	58	7	12%	!on	i	on;ge;com;[-];eer;d;heid		**N**	>	**N**	[-p]	ge;com;[plic];eer;d;heid
7	3 + 1 + 3	58	7	12%	ver	i	ver;al;ge;[-];is;eer;ing		-	>	**N**	[-p]	ver;al;ge;[men];is;eer;ing
7	3 + 1 + 3	58	7	12%	ver	i	ver;al;ge;[-];is;eer;en		-	>	**V**	[-p]	ver;al;ge;[men];is;eer;en
7	2 + 1 + 4	58	20	34%	!on	i	on;ge;[-];ic;ul;eer;d		-	>	**A**	[-p]	on;ge;[art];ic;ul;eer;d
7	2 + 1 + 4	58	20	34%	!on	i	on;ge;[-];ic;ul;eer;d		**A**	>	**A**	[-p]	ge;[art];ic;ul;eer;d
7	2 + 1 + 4	58	20	34%	!on	i	on;ge;[-];il;is;eer;d		-	>	**A**	[-p]	on;ge;[civ];il;is;eer;d
7	2 + 1 + 4	58	20	34%	!on	i	on;ge;[-];il;is;eer;d		**A**	>	**A**	[-p]	ge;[civ];il;is;eer;d

B.4.4 Voorbeeldlijst: lexemen met 1 prefix en 2 suffixen

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
4	1 + 1 + 2	8983	4859	54%	her	i	her;[-];er;ing	-	>	**N**	1	100%	[-p]	her;[inn];er;ing
4	1 + 1 + 2	8983	4859	54%	ver	i	ver;[-];d:er;ing	-	>	**N**	2	100%	[-p]	ver;[min];d:er;ing
4	1 + 1 + 2	8983	4859	54%	ver	i	ver;[-];d:er;ing	**N**	>	**N**	2	100%	[-p]	[min];d:er;ing
4	1 + 1 + 2	8983	4859	54%	ver	i	ver;[-];er;ing	-	>	**N**	11	100%	[+p]	ver;[wild];er;ing
4	1 + 1 + 2	8983	4859	54%	ver	i	ver;[-];er;ing	**N**	>	**N**	1	9%	[-p]	[snipp];er;ing
4	1 + 1 + 2	8983	4859	54%	$uit	i	uit;[-];er;ing	-	>	**N**	1	100%	[-p]	uit;[waai];er;ing
4	1 + 1 + 2	8983	4859	54%	ont	i	ont;[-];er;ing	-	>	**N**	2	100%	[-p]	ont;[mask];er;ing
4	1 + 1 + 2	8983	4859	54%	be	i	be;[-];d;ing	-	>	**N**	2	100%	[-p]	be;[wei];d;ing
4	1 + 1 + 2	8983	4859	54%	ver	i	ver;[-];d;ing	-	>	**N**	2	100%	[-p]	ver;[blij];d;ing
4	1 + 1 + 2	8983	4859	54%	$na	i	na;[-];ig;ing	-	>	**N**	1	100%	[-p]	na;[rein];ig;ing
4	1 + 1 + 2	8983	4859	54%	$na	i	na;[-];ig;ing	**N**	>	**N**	1	100%	[-p]	[rein];ig;ing
4	1 + 1 + 2	8983	4859	54%	be	i	be;[-];ig;ing	-	>	**N**	20	100%	[+p]	be;[zuin];ig;ing
4	1 + 1 + 2	8983	4859	54%	be	i	be;[-];ig;ing	**N**	>	**N**	3	15%	[-p]	[macht];ig;ing
4	1 + 1 + 2	8983	4859	54%	$af	i	af;[-];ig;ing	-	>	**N**	2	100%	[-p]	af;[vaard];ig;ing
4	1 + 1 + 2	8983	4859	54%	$vol	i	vol;[-];ig;ing	-	>	**N**	1	100%	[-p]	vol;[eind];ig;ing
4	1 + 1 + 2	8983	4859	54%	$vol	i	vol;[-];ig;ing	**N**	>	**N**	1	100%	[-p]	[eind];ig;ing
4	1 + 1 + 2	8983	4859	54%	$aan	i	aan;[-];ig;ing	-	>	**N**	3	100%	[-p]	aan;[moed];ig;ing
4	1 + 1 + 2	8983	4859	54%	$aan	i	aan;[-];ig;ing	**N**	>	**N**	2	66%	[-p]	[mat];ig;ing
4	1 + 1 + 2	8983	4859	54%	$in	i	in;[-];ig;ing	-	>	**N**	2	100%	[-p]	in;[will];ig;ing
4	1 + 1 + 2	8983	4859	54%	$in	i	in;[-];ig;ing	**N**	>	**N**	1	50%	[-p]	[huld];ig;ing
4	1 + 1 + 2	8983	4859	54%	her	i	her;[-];ig;ing	-	>	**N**	1	100%	[-p]	her;[en];ig;ing

Resultaten van de externe evaluatie

Domeinkenmerken

- prefix-suffix-combinaties (zonder wortel)
- lexemen met maximaal 10 morfemen
- minimumfrequentie = 10

Algemene kencijfers

in totaal zijn er 83414 lexemen die aan de filtercriteria voldeden
hieronder zijn 67653 lexemen met minstens 1 prefix of suffix

lexeemfrequentie	aantal types	aantal tokens
f = 50+	34	14557
f = 10+	63	19314
f = 0+	227	21542

Tabel 31: aantal hb-prefixen per frequentieklasse (op type-niveau en op token-niveau)

lexeemfrequentie	aantal ucat-types
f = 50+	83
f = 10+	374
f = 0+	7994

Tabel 32: aantal MGBN-patronen (= prefix-suffix-combinaties) per frequentieklasse

categorie	aantal	categorie	aantal
N	4558	T	9
V	922	P	35
A	1899	O	227
B	156	?	136

Tabel 33: aantal ucat-patronen per u-categorie

aantal morfemen	aantal lexemen	aandeel in totaal
8	7	0 %
7	58	0 %
6	401	0 %
5	2046	3 %
4	8983	13 %
3	28791	42 %
2	27367	40 %
totaal	**67653**	**77 %**
1	15761	23 %
totaal	**83414**	**100 %**

Tabel 34: aantal lexemen per lexeemklasse (op basis van aantal morfemen)

m	m-sub-pat	m-freq	m-perc	m-sub-patfreq	m-sub-patperc
8	? + 1 + ?	7	0 %	7	100 %
8	4 + 1 + 3	7	0 %	0	0 %
8	3 + 1 + 4	7	0 %	0	0 %
8	2 + 1 + 5	7	0 %	3	42 %
8	1 + 1 + 6	7	0 %	4	57 %
7	? + 1 + ?	58	0 %	58	100 %
7	5 + 1 + 1	58	0 %	0	0 %
7	4 + 1 + 2	58	0 %	0	0 %
7	3 + 1 + 3	58	0 %	7	12 %
7	2 + 1 + 4	58	0 %	20	34 %
7	1 + 1 + 5	58	0 %	25	43 %
7	0 + 1 + 6	58	0 %	6	10 %
6	? + 1 + ?	401	0 %	401	100 %
6	4 + 1 + 1	401	0 %	3	0 %
6	3 + 1 + 2	401	0 %	15	3 %
6	2 + 1 + 3	401	0 %	118	29 %
6	1 + 1 + 4	401	0 %	194	48 %
6	0 + 1 + 5	401	0 %	71	17 %
5	? + 1 + ?	2046	3 %	2046	100 %
5	4 + 1 + 0	2046	3 %	0	0 %
5	3 + 1 + 1	2046	3 %	36	1 %
5	2 + 1 + 2	2046	3 %	572	27 %
5	1 + 1 + 3	2046	3 %	940	45 %
5	0 + 1 + 4	2046	3 %	498	24 %
4	? + 1 + ?	8983	13 %	8983	100 %
4	3 + 1 + 0	8983	13 %	9	0 %
4	2 + 1 + 1	8983	13 %	1696	18 %
4	1 + 1 + 2	8983	13 %	4859	54 %
4	0 + 1 + 3	8983	13 %	2419	26 %
3	? + 1 + ?	28791	42 %	28791	100 %
3	2 + 1 + 0	28791	42 %	435	1 %
3	1 + 1 + 1	28791	42 %	18347	63 %
3	0 + 1 + 2	28791	42 %	10009	34 %
2	? + 1 + ?	27367	40 %	27367	100 %
2	1 + 1 + 0	27367	40 %	4661	17 %
2	0 + 1 + 1	27367	40 %	22706	82 %
1	? + 1 + ?	15761	23 %	15761	100 %
1	0 + 1 + 0	15761	23 %	15761	100 %
#	? + 1 + ?	67653	100 %	67653	100%

Tabel 35: telling van het aantal basislexemen per mofologisch patroon (totaal aantal morfemen en prefix-suffix-verhouding)

Evaluatie van de MGBN in termen van het aantal MHB-treffers

frequentieklasse	aantal mgbn-types	aantal hb-treffers
0+	227	86 (37 %)
10+	63	57 (90 %)

Tabel 36: aantal hb-treffers per frequentieklasse (absoluut en relatief) (alle patronen)

prefixklasse	aantal mgbn-types	aantal hb-treffers
hoofdtype	227	87 (38 %)
ucat-prefix	7994	1774 (22 %)
icat-prefix	5424	1068 (19 %)

Tabel 37: aantal MGBN-types per prefixklasse en hun MHB-dekking (alle patronen)

prefixklasse	aantal mgbn-types	aantal hb-treffers
hoofdtype	63	57 (90 %)
ucat-prefix	374	179 (47 %)
icat=prefix	114	58 (50 %)

Tabel 38: aantal MGBN-types per prefixklasse en MHB-dekking (patronen met freq 10+)

Evaluatie van het MHB in termen van het aantal MGBN-treffers

prefixklasse	aantal hb-types	aantal mgbn-treffers	aantal mgbn-missers
kaal prefix	106	84 (79 %)	22 (21 %)
ucat-prefix	128	79 (61 %)	49 (39 %)
icat-prefix	184	78 (42 %)	106 (58 %)

Tabel 39: MGBN-dekking van hb-prefix-eenheden (absoluut en relatief)

lijst van onvindbare MGBN-suffixen (maar wel verklaarbaar):
aaneen, aarts, b, bijeen, binnen, boven, buiten, hecto, loco, oer, omhoog, omlaag, opper, oud, semi, terecht, thuis, turbo, uiteen

weggefilterde MHB-prefixen:
aaneen, achteraan, achteraf, achterna, achterom, achterop, achteruit, ambi, amfi, circum, crypto, d, etno, intra, non, omver, onderuit, pluri, pseudo, retro, vooraan, vooraf, voorbij, voorin, voorop, voorover, vooruit

Notatieconventies

De onderstaande tabel geeft informatie over de notatiewijze van twee veel gebruikte klassen van structuureenheden, te weten morfemen (incl. lexeemstammen) en woorden:

lexicale structuureenheid	morfeem-notatie	woord-notatie
morfotactische indexen (taxemen)	BE-, √KROON, -ING	*bekroning*
morfofonologische indexen (f-indexen)	*be-, kroon, -ing*	*bekroning*
-uitspraak van f-indexen	/be-/, /kroon/, /-ing/	/bekroning/
-spelling van f-indexen	|be-|, |kroon|, |-ing|	|bekroning|
morfosemantische indexen (s-indexen)	be, kroon, ing	bekroning

Het in deze tabel gespecificeerde notatiesysteem is primair bedoeld voor de eenheden die de basis vormen van mijn op L-KRING-principes gebaseerde lexiconsysteem (zie hoofdstuk 4), namelijk de lexicale *indexen*. Maar deze notatiewijze wordt ook toegepast op vergelijkbare structuureenheden bij de bespreking van andere taalmodellen. Om die reden heb ik deze indexen met algemene, modelonafhankelijke termen proberen aan te duiden. Lexicale indexen (in feite namen) kunnen met verschillende structuurniveaus corresponderen, waaronder morfemen (namelijk wortels en affixen), (al dan niet zelfstandige) lexemen en (al dan niet samengestelde) woorden. Zo correspondeert de lexeemindex *bekroning* met de in (1) getoonde compositiestructuur van de morfeemindexen BE-, KROON en -ING. Bij deze morfemen kan (conform de gangbare conventie) onderscheid worden gemaakt tussen drie subklassen, te weten de prefixklasse (X-), de wortelklasse (√X) en de suffixklasse (-X), waarbij de wortel door het teken √ wordt gemarkeerd; dit teken zal overigens vaak achterwege blijven.

(1) $[\ [\ \text{BE-} \oplus [\ \text{KROON}\ ]_{M0}\ ]_{M1} \oplus \text{-ING}\ ]_{M2} + \$_L\ ]_L \rightarrow [\ \textit{bekroning}\ ]_L$

De onder (1) weergegeven compositiestructuur berust op herhaalde toepassing van het basispatroon $[S \oplus F]_f \rightarrow S_f'$. Hierbij correspondeert $\oplus$ met een compositie-operator; deze zorgt ervoor dat stam S met functor F wordt gecombineerd onder vorming ($\rightarrow$) van een compositieproduct S'; dit compositieproduct correspondeert met dezelfde structuurklasse als de functor F (namelijk f). Omdat de combinatie van twee morfeemindexen altijd tot een morfeem leidt, is een aparte operator nodig om een hogere eenheid te construeren. Hiervoor is een lexeemoperator nodig (gemarkeerd door $\$_L$). Voor verdere uitleg dient men hoofdstuk 4 te raadplegen. Het gaat hier alleen om de notationele conventies.

Uit de tabel blijkt dat morfeemindexen een andere notatiewijze kennen dan de indexen voor lexemen en woorden. Verder blijkt dat de aan deze indexen verbonden notatievorm gevoelig is voor de modaliteit van deze eenheden. Voor deze studie is de *morfotactische* modaliteit het belangrijkst, d.w.z. de modaliteit waar vorm en betekenis met elkaar verbonden worden. In de gangbare grammaticamodellen valt de morfotactische representatiedimensie uiteen in morfologische representaties (die uit morfemen bestaan) en syntactische representaties (die uit lexemen of woorden bestaan). Om makkelijk over deze structuurniveaus te kunnen generaliseren, heb ik ervoor gekozen om de overkoepelende representatie met de term *morfotactisch* aan te duiden. De bijbehorende kenniseenheden (waaronder morfemen, lexemen en woorden) noem ik *taxemen*. Naast de morfotactische modaliteit onderscheid ik ook een *morfofonologische* modaliteit (met fonologische f-eenheden) en een *morfosemantische* modaliteit (met semantische s-eenheden).[213] De morfofonologische modaliteit integreert informatie uit twee submodaliteiten met specifiekere representaties, te weten de orthografische representatie (c.q. spelvorm) en de fonetische representatie (c.q. uitspraak).

[213] Omwille van de leesbaarheid gebruik ik vaak de termen *morfologisch*, *semantisch* en *fonologisch*.

Abbreviatorium

IW-model **Ideaal Woordenboek-model**: het IW-model is een door Verkuyl & al. (1997) ontwikkelde leidraad voor de ontwikkeling van woordenboeken die een goede afspiegeling vormen van de kennis in het mentale lexicon; het IW-model kent een structuur die vergelijkbaar is met het L-model van Verkuyl (1978).

IL-model **Ideaal Lexicon-model**: een op het IW-model vortbordurend metamodel voor het opzetten en beoordelen van lexicografische kennisbanken. Het IL-model kent een beter uitgewerkte structuur en beschrijft extra functies.

IDL-systeem **Integraal Dynamisch Lexiconsysteem**: aanduiding voor een lexicaal kennissysteem dat in beginsel alle functies van het mentale lexicon kan verantwoorden en dat daarom een goed vertrekpunt vormt voor de opzet van een lexicografische kennisbank die aan de eisen van een Ideaal Woordenboek voldoet.

L-KRING **Lexicale KennisRepresentatie door Inductieve NaamGeving**, het in deze studie gepresenteerde model voor lexicale kennisrepresentatie. Dit model biedt een formele uitwerking van de algemene richtlijnen uit het IDL-model.

L-model **Lexicon-model**: het semantische lexiconmodel van Verkuyl (1978)

LGBN **Lexicale Gegevensbank van het Nederlands**:
i) speciaal voor de MGBN ontwikkelde gegevensbank met een omvangrijke, deels bewerkte selectie van woorden en woorddelen uit de WKB-Nederlands, en met informatie over hun interne structuur, categorie en vormkenmerken;
ii) in bredere zin is de LGBN een aanduiding voor het informatiesysteem waarmee deze kennisbank (en de hierin op te nemen informatie uit de MGBN) toegankelijk wordt gemaakt;

LGBN-L **LGBN op Lexeem-niveau**: gegevensbestand met informatie over de woordkenmerken van alle samenstellende delen (c.q. basislexemen) uit de LGBN

LGBN-W **LGBN op Woord-niveau**: gegevensbestand met informatie over de woordkenmerken en lexeemstructuur van alle (±complexe) woorden uit de LGBN

MGBN **Morfologische Gegevensbank van het Nederlands**
een op de LGBN gebaseerd gegevensbestand met morfologische structuurinformatie over de basislexemen uit de LGBN; deze structuurinformatie is langs computationele weg naar het woordniveau uit te breiden;

MGBN-L MGBN op **Lexeem-niveau**: een op de LGBN-L gebaseerde gegevensbank met morfologische structuurrepresentaties over alle basislexemen

MGBN-W MGBN op **Woord-niveau**: een op de LGBN-W gebaseerde gegevensbank met morfologische structuurrepresentaties over alle (±complexe) woorden

VDL **Van Dale lexicografie**, uitgever van woordenboeken (o.a. de Grote Van Dale).

WKB-Ned **WoordKenmerkenBank Nederlands**, ook wel aan te duiden als VDL's Woordkenmerkenbank Nederlands: geïntegreerd vormkenmerkenbestand dat uitgaat van de informatie in VDL's beheersysteem voor Nederlandstalige Woordenboeken. De naam WKB-Ned komt alleen in deze studie voor.

WHN **Van Dale's Groot Woordenboek Hedendaags Nederlands** (1 band)

GWNT **Van Dale's Groot Woordenboek der Nederlandse Taal** (13e druk)

-GWNTb **boekeditie** van de GWNT, beter bekend als de Grote Van Dale (3 banden)

-GWNTe **elektronische** editie van de GWNT, die op een CD-ROM is uitgegeven

MHB **Morfologisch Handboek van het Nederlands** (De Haas & Trommelen, 1993)

WNT **Woordenboek der Nederlandse Taal** (bestaande uit 40 boekbanden); zeer omvangrijke inventarisatie van het Nederlandse taalgebruik tussen 1500 en 1976; tevens belangrijke wetenschappelijke bron voor woordenboekuitgevers.

Bibliografie

Reeksen, artikelenbundels en collectieve standaardwerken

CCR: *Concepts. Core Readings*, 1999. E. Margolis & S. Laurence (eds.). Bloemlezing, met uitgebreide introductie. MIT Press, Cambridge.
ILB: *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, 2000. Marantz, Alec, Yasushi Miyashita and Wayne O'Neil (eds.). MIT Press, Cambridge.
LIN: *Linguistics in the Netherlands*. Yearbook of AVT. John Benjamins, A'dam/Philadelphia.
LOED: *Lexicography and the OED, Pioneers in the Untrodden Forest*, 2000. Lynda Mugglestone (ed.), Oxford University Press, Oxford.
LDSLP: *Lexicon Development for Speech and Language Processing, 2000.* Frank van Eynde & Dafydd Gibbon (eds.). Kluwer Academic Publishers. Dordrecht, Boston, London.
MALP: *Morphological Aspects of Language Processing*, 1995. L.B. Feldman (ed.). Lawrence Erlbaum Inc., New Jersey.
MIH: *Morphologie / Morphology. Ein internationales Handbuch zur Flexion und Wortbildung / An International Handbook on Inflection and Word formation*, 2000. G.E. Booij, Ch. Lehman & J. Mugdan (eds.), i.s.m. W. Kesselheim en S. Skopeteas. Vol 1. Berlin: Walter de Gruyter. 996 p.
MSLP: *Morphological Structure in Language Processing*, 2003. R.H. Baayen (ed.). Mouton de Gruyter, Berlin.
PCA: *Performance & Competence in second language acquisition*, 1996. Gillian Brown, K. Malmkjaer & J. Williams (eds.). Cambridge Univ. Press, Cambridge.
RLR: *The reality of linguistic rules*. 1994. S.D. Lima, R.L. Corrigan & G.K. Iverson (eds.). Studies in Language Companion Series 26. Amsterdam/Philadelphia: Benjamins Publ.
YoM: *Yearbook of Morphology*, Geert Booij & Jaap Van Marle (eds). Kluwer, Dordrecht.

Individuele publicaties

Abney, Steven (1987), *The English Noun Phrase in its Sentential Aspect*. Dissertatie. MIT, Cambridge.
Ackema, Peter (1995), *Syntax below zero*. OTS Dissertation Series, Utrecht.
Ackerman, Farrell & Gert Webelhuth (1998), *A theory of predicates*. CSLI Lecture Notes No. 67. Stanford, California.
Aksu, Ayhan & Dan Slobin (1984), "The acquisition of Turkish morphology". In: D. Slobin (ed.). *The cross-linguistic study of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum.
Al, B.P.F. & Booij, G.E. (1981), "De productiviteit van woordvormingsregels. Enige kwantitatieve verkenningen op het gebied van nomina actionis", *Forum der Letteren* 22, p. 26-38.
Albright, Adam & Bruce Hayes (1999), *An automated learner for phonology and morphology*. Zie: www.humnet.ucla.edu/humnet/linguistics/people/hayes
Albright, Adam & Bruce Hayes (2003), "Rules versus Analogy in English Past Tenses: A Computational/Experimental Study", *Cognition* 90, p. 119-161. Zie ook:
Anderson, S.R. (1982), "Where's Morphology?", *Linguistic Inquiry* 13, p. 571-612, MIT Press, Cambridge.
Andrews, Sally (1986), "Morphological influences on lexical access: Lexical or nonlexcial effects?", *Journal of Memory and Language* 25, p. 726-740.
Andrews, Sally & Colin Davis (1999), "Interactive Activation Accounts of Morphological Decomposition: Finding the Trap in Mousetrap?", *Brain and Language* 68, p. 355-361.
Anshen, Frank & Mark Aronoff (1988), "Producing morphological complex words", *Linguistics* 26, p. 641-655.

Anshen, Frank & Mark Aronoff (1999), "Using dictionaries to study the mental lexicon", *Brain and Language* 68, p. 16-26.
Aronoff, Mark (1976), *Word Formation in Generative Grammar*. Cambridge, MIT Press, Cambridge.
Aronoff, Mark (1994), *Morphology By Itself*. MIT Press, Cambridge.
Audring, Jenny & Geert Booij (2005, ms.), *The interdependency of syntax and morphology in constructions*. Vrije Universiteit, Amsterdam.
Baayen, R. Harald (1989), *A corpus-based approach to morphological productivity*. Dissertatie. Vrije Universiteit, Amsterdam.
Baayen, R. Harald (1990), "Corpusgebaseerd onderzoek naar morfologische productiviteit", *Sprektator* 19-3, p. 213-233.
Baayen, R. Harald (1991), "De CELEX Lexicale Databank", *Forum der Letteren 33*, p. 220-231.
Baayen, R. Harald (1991a), "Quantitative aspects of morphological productivity". In: *YoM*, p. 109-149.
Baayen, R. Harald & Rochelle Lieber (1991), "Productivity and English derivation: a corpus-based study", *Linguistics* 29, p. 801-843.
Baayen, R. Harald (1992), "On frequency, transparency and productivity". In: *YoM*, 181-208.
Baayen, R. Harald (1992a), "Taalsystematiek, taalgebruik, semantiek en productiviteit." *Forum der Letteren* 33, p. 214-224.
Baayen, R. Harald, T. Dijkstra & R. Schreuder (1997), "Singulars and plurals in Dutch: evidence for a parallel dual-route model", *Journal of Memory and Language*, 27, p. 94-117.
Baayen, R. Harald & Robert Schreuder (1999), "War and Peace: Morphemes and full forms in a noninteractive activation parallel dual-route model", *Brain and Language* 68, p. 27-32.
Baayen, R. Harald & Robert Schreuder (2000), "Towards a psycholinguistic computational model for morphological parsing". In: *Philosophical Transactions of the Royal Society of London (A: Mathematical, Physical and Engineering Sciences)*, vol. 358, p. 1281-1293.
Baayen, R. Harald, R. Schreuder, N. de Jong & A. Krott (2002), "Dutch Inflection: The rules that prove the exception". In: S. Nooteboom (eds.), *Storage and Computation in the Language Faculty*, p. 61-92. Kluwer Academic Publishers, Dordrecht.
Backhuys, Kees-Jan (1986), *De morfologie van romaanse woordvorming in het Nederlands*. Doctoraalscriptie Utrecht. Uitgeverij Alexandrië, Utrecht.
Baroni, Marco (2000). *Distributional cues in morpheme discovery: A computational model and empirical evidence*. Diss. UCLA, California.
Baroni, Marco (2003), "Distribution-driven morpheme discovery: A computational/ experimental study" In: *YoM*, p. 213-248. Zie ook: http://sslmit.unibo.it/~baroni
Beard, Robert (1991), *Lexeme-Morpheme Base Morphology*. Albany: SUNY Press.
Beelen, Hans (2004), "Van leenwoord tot inheemse nieuwvorming. De herkomst van neoklassieke composita op -cratie". In: web-tijdschrift *Neerlandistiek.nl*. Zie: http://www.neerlandistiek.nl/publish/articles/000078/article.html
Bergen, Benjamin K. (2004), "The psychological reality of phonaesthemes", *Language* 80 (2).
Bergman, M.W., P.T.W. Hudson & P.A.T. Eling (1988). "How simple complex words can be: Morphological processing and word representation". *Quarterly Journal of Experimental Psychology*, 40A, p. 41-72.
Berko, Jean (1958), "The child's learning of English morphology", *Word* 14, p. 150-177.
Berman, Ruth Aronson (1981), Regularity vs anomaly: the acquistion of Hebrew inflectional morphology, *Journal of Child Language* 8, p. 265-282.
Bertram, Raymond, R.H. Baayen & R. Schreuder (2000), "Effects of family size for complex words", *Journal of Memory and Language* 42, p. 390-405.

Bertram, Raymond, R. Schreuder & R.H. Baayen (2000), "The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity", *Journal of Experimental Psychology: Memory, Learning, and Cognition* 26, p. 419–511.

Bierwisch, Manfred (1996), "Lexical Information from a Minimalist Point of View". In: C. Wilder, H-N Gärtner and M. Bierwisch (eds.), *The Role of Economy Principles in Linguistic Theory*. Studia Grammatica 40, Akademie Verlag, Berlin.

Blom, Corrien (2005), *Complex Predicates in Dutch. Synchrony and Diachrony*. PhD. Diss. (VUA). LOT Dissertation Series 111, Utrecht.

Bloomfield, Leonard (1933), *Language*. London: Allen & Unwin.

Bochner, Harry (1993), *Simplicity in Generative Morphology*. Publications in Language Sciences: 37. Mouton de Gruyter, Berlin/New York.

Bod, Rens (1995), *Enriching Linguistics with statistics: Performance models of natural language*. Diss. ILLC, Amsterdam.

Bolinger, Dwight L. (1948), "On defining the morpheme". In: Bolinger, D.L. (ed.), *Forms of English. Accent, Morpheme, Order*. Cambridge, Mass. Harvard Univ. Press, p.183-189.

Bolinger, Dwight L. (1975). *Aspects of Language*, 2nd edition. New York: Harcourt Brace Jovanovich.

Booij, Geert E. (1977), *Dutch Morphology. A study of Word Formation in Generative Grammar*. Peter De Ridder Press Publication on Dutch 2, Lisse.

Booij, Geert E. (1994), "Against Split Morphology", In: *YoM* 1993, p. 27-49.

Booij, Geert E. (1997), "Allomorphy and the Autonomy of Morphology", *Folia Linguistica* 31, p. 25-56.

Booij, Geert E. & Ariane van Santen (1998), *Morfologie. De woordstructuur van het Nederlands*. 2e geheel herziene druk. A'dam Univ. Press, Amsterdam.

Booij, Geert E. (2002), *The Morphology of Dutch*. Oxford University Press.

Booij, Geert E. (2002a), "Constructional idioms and the lexicon", *Journal of Germanic Linguistics* 14:4.

Booij, Geert (2005a, ms.), "Construction morphology". Vrije Universiteit, Amsterdam.

Booij, Geert (2005b, ms.), "Construction-dependent morphology". Te verschijnen in *Lingue e Linguaggio*.

Borer, Hagit (2000), *The impoverished lexicon*. Lecture Notes of UiL OTS Course.

Van den Bosch, Antal & Walter Daelemans (1999). Memory-based morphological analysis. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, ACL'99, University of Maryland, USA, 20-26 July 1999, p. 285-292. Zie ook: http://ilk.uvt.nl/~antalb

Van den Bosch, Antal (1999), "Careful Abstraction from Instance Families in Memory-Based Language Learning". *Journal of Experimental and Theoretical Artificial Intelligence*, 11:3, p. 339-368. Zie ook: http://ilk.uvt.nl/~antalb

Bouma, Gosse & Ineke Schuurman (1998), *De positie van het Nederlands in taal- en spraaktechnologie*. Een rapport in opdracht van de Nederlandse Taalunie. Beschikbaar via: http://odur.let.rug.nl/~gosse/taalunie/webrapport/rapport.html

Bouma, Gosse, Frank Van Eynde & Dan Flickinger (2000), "Constraint-Based Lexica." In: *LDSLP* (Van Eynde & Gibbon, 2000).

Brandt Corstius, Hugo (1978), *Computer-taalkunde*. Randgebieden No.3. Coutinho, Bussum.

Braine, Martin D.S. (1976), *Children's first word combinations*. Monographs of the Society for Research in Child Development, 41 (1, Serial No. 164).

Bresnan, Joan (1982), *The Mental Representation of Grammatical Relations*. Cambridge, Mass. MIT Press

Brown, Roger (1973), *A first language: the early stages*. Cambridge, Mass.: Harvard Press.

Burani, C. & Caramazza, A. (1987), "Representation and processing of inflected words", *Language and Cognitive Processes* 2, 217-227.
Burani, C. & A. Laudanna (1993). Units of representation for derived words in the mental lexicon. In: R. Frost & L. Katz (eds.), *Orthography, phonology, morphology, and meaning*, Amsterdam: Elsevier.
Butterworth, B. (1983), "Lexical Representation." In: B. Butterworth (ed.), *Language Production,* Vol. II: *Development, writing and language processes,* p. 257-294. London: Academic Press.
Bybee, Joan L. & Dan I. Slobin (1982), "Rules and Schemes in the Development and Use of the English Past Tense", *Language* 58, p. 265-289.
Bybee, Joan L. (1985), *Morphology: A study of the relation between meaning and form*. Typological Studies in Language 9. John Benjamins, A'dam/Philadelphia.
Bybee, Joan L. (1988), "Morphology as Lexical Organization". In: M. Hammond and M. Noonan (eds.), *Theoretical Morphology*, p. 119-141. San Diego, CA: Academic Press.
Bybee, Joan L. (1995), "Regular Morphology and the lexicon". *Language and Cognitive Processes* 10, p. 425-455.
Bybee, Joan L. (2001), *Phonology and Language Use*. Cambridge Studies in Linguistics 94. Cambridge Press, Cambridge.
Caramazza, A., A. Laudanna & C. Romani (1988), "Lexical access and inflectional morphology", *Cognition*, 28, 297-332.
Cassirer, Ernst (1972), "Structuralism in Modern Linguistics". In: *Readings in Modern Linguistics, An Anthology by Bertil Malmberg*. Stockholm.
Chomsky, Noam (1956), *Syntactic Structures*. Mouton, Den Haag.
Chomsky, Noam & Morris Halle (1968), *The sound pattern of English*. Harper & Row, New York.
Chomsky, Noam (1970), "Remarks on nominalization". In: Jacobs & Rosenbaum (eds.), *Readings in English Transformational Grammar*, Waltham, MA: Blaisdell.
Chomsky, Noam (1981), *Lectures on Government and Binding*. Dordrecht: Foris.
Chomsky, Noam (1982), *Some concepts and consequences of the Theory of Government and Binding*. Cambridge, Mass.: MIT Press.
Chomsky, Noam (1995), *The Minimalist Program*. MIT Press, Cambridge, Massachusetts.
Clahsen, Harald (1999), "Lexical Entries and Rules of Language: A Multidisciplinary Study of German Inflection", *Brain and Behavioral Sciences* 22, p. 991-1013.
Coppen, Peter-Arno & Crit Cremers (2002), "Parseren in de Polder. Nederlandse taal-technologie in perspectief." In: *Nederlandse Taalkunde* 7, p. 305-311.
Cornelis, Louise H. (1997), *Passive and Perspective*. Studies in Language and Communication, 10. Amsterdam/Atlanta. Rodopi, Utrecht
Cremers, Crit (2002), "('n) Betekenis berekend". In: *Nederlandse Taalkunde* 7, p. 375-395.
Daelemans, W., A. van den Bosch & J. Zavrel (1999), "Forgetting exceptions is harmful in language learning", *machine Learning* 34, p. 11-43.
Daelemans, Walter & Helmer Strik (2002), *Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen*. Een rapport in opdracht van de Nederlandse Taalunie. Beschikbaar via: http://taalunieversum.org/taal/technologie/docs/daelemans-strik.pdf
Daniels, Wim (2001), *Komkom, tuuttuut, hoho. Herhalingswoorden in het Nederlands en andere talen*. Uitgeverij Veen, Utrecht.
Daugherty, Kim G. & Mark S. Seidenberg (1994), "Beyond rules and exceptions: a connectionist approach to inflectional morphology". In *RLR*, p. 353-388.
Derwing, Bruce L. (1973), *Transformational Grammar as a theory of language acquisition: a study in the empirical, conceptual and methodological foundations of contemporary linguistics*. Cambridge Univ. Press, Cambridge.

Derwing, Bruce L. (1974), "Review of Fred W. Householder, *Linguistic speculations*", *Language Sciences* 30 (April), p. 25-32.
Derwing, Bruce L. & Royal Skousen (1989), "Morphology in the lexicon: a new look at analogy". In: *YOM*: 55-71.
Derwing, Bruce L. & Royal Skousen (1994), "Productivity and the English Past Tense: Testing Skousen's Analogy Model". In: *The reality of linguistic rules* (RLR), p. 193-218.
Deutsch, Avital, R. Frost, A. Pollatsek & K. Rayner (2000), "Early morphological effects in word recognition in Hebrew: Evidence from parafoveal preview benefit", *Language and Cognitive Processe* 15, p. 487-506. Zie ook: http://icnc.huji.ac.il/Files/word.pdf
Dijkstra, Ton, J. Grainger & W.J.B. van Heuven (1999), "Recognition of Cognates and Interlingual Homographs: The Neglected Role of Phonology", *Journal of Memory and Language* 41, p. 496–518. Web-link: http://www.andrew.cmu.edu/user/natashat/bilingualism/dijkstra.pdf
Domenig, Marc & Pius ten Hacken (1992), *Word Manager: A system for Morphological Dictionaries*. Georg Olms Verlag. Hildesheim, Zürich, New York.
Don, Jan (1993), *Morphological Conversion*. OTS Dissertations, Utrecht.
Don, Jan & al. (1994), *Inleiding in de generatieve morfologie*. Coutinho, Bussum.
Don, Jan (2003), "A note on conversion in Dutch and German". In: *LIN*, p. 33-44.
Dowty, David (1979), *Word Meaning and Montague Grammar. The semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Reidel: Dordrecht.
Dowty, David (2000), "The dual analysis of adjuncts/complements in categorial grammar." In: *ZAS-papers in Linguistics 17*, ed. C. Fabriciuis-Hansen, E. Lang and C. Maienborn, Zentrum für Allgemeine Sprachwissenschift, Typologie, Universalienforschung, Berlin. Ook beschikbaar via: http:/ling.osu.edu/~dowty
Dowty, David (2001), "The semantic asymmetry of 'arguments alternations' and why it matters". In: G. van der Meer & A.G.B. ter Meulen (eds.), *Groninger Arbeiten zur germanistischen Linguistik*, nr. 44, Centre for Language and Cognition, Groningen. Web-link: http://ling.osu.edu/~dowty
Drijkoningen, Frank (1995), "On the antisymmetry of words: circumfixation." In: *OTS Yearbook* 1995. Jan Don, Bert Schouten, Wim Zonneveld (eds.), Universiteit Utrecht.
Evans, Roger & Gerald Gazdar (1996), "DATR: A language for lexical knowledge representation", *Computational Linguistics* 22.2, p. 167-216
Evans, Roger & al. (2003), "A large-scale inheritance-based morphological lexicon for Russian." In: *Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages*. Web-link: ftp://ftp.itri.brighton.ac.uk/reports/ITRI-03-02.pdf
Everaert, Martin (1993), "Morfologische vaste verbindingen: bestaande woorden". In: *Tabu*, 23, 1-2, p. 29-40.
Everaert, Martin (2003), *Wijzen van zeggen*. Tekstuitgave van een rede. Univ. Nijmegen.
Fabb, Nigel (1988), "English suffixation is constrained only by selectional restrictions", *Natural Language and Linguistic Theory* 6, p. 527-539.
Fikkert, Paula (2003), "Papa, mag het donker aan? Kindertaal verzameld en geordend". In: *Onze Taal* 4, p. 80-83.
Fillmore, Charles G. (1978), "On the organization of semantic information the lexicon". In: Donka Farkas (ed.), *Papers from the parasession on the lexicon*. Chicago Linguistic Society
Fillmore, Charles J. (1988), "The mechanisms of 'Construction Grammar'". In: *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, p. 35-55. Berkeley.
Fillmore, Charles G. and Paul Kay (1996, ms.), *Construction Grammar*. University of California, Berkeley. Verkrijgbaar via: www.icsi.berkeley.edu/~kay/bcg/ConGram.html
Ford, A. & R. Singh (1991), "Propedeutique morphologique", *Folia Linguistica* 25: 3-4, p. 549-575

Ford, A., R. Singh & G. Martohardjono (1997), *Pace Panini*. Peter Lang, New York
Frauenfelder, Uli H. & Robert Schreuder (1991), "Constraining psycholinguistic models of morphological processing and representation: The role of productivity". In: *YoM*, 165-183.
Frege, Gottlob (1892), "On Sense and Meaning". In: J. van Heyenoort (ed.), *From Frege to Gödel: A Sourcebook in Mathematical Logic 1879-1931*. Cambridge, Mass.: Harvard Univ. Press, 1967. (Originele titel: "Über Sinn und Bedeutung")
Freyd, P. & J. Baron (1982), "Individual differences in acquisition of derivational morphology", *Journal of Verbal Learning and Verbal Behavior* 21, p. 282-295.
Frijn, Jacqueline & Ger De Haan (1990), *Het taallerend kind*. ICG Publications, Dordrecht.
Frost, Ram & Jonathan Grainger (2000), "Cross-linguistic perspectives on morphological processing: An introduction", *Language and Cognitive Processes* 15 (4/5), 321–328. Zie: http://www.up.univ-mrs.fr/wlpc/pagesperso/grainger/pubpdf/p321frost.pdf
Gamut, L.T.F. (1991), *Intensional Logic and Logical Grammar*. The University of Chicago Press, Chicago/Londen.
Geeraerts, Dirk, Stefan Grondelaers & Peter Bakema (1994), *The structure of lexical variation. Meaning, naming, and context*. Berlin: Mouton de Gruyter.
Gentilhomme, Yves (1964), *Manuel de Russe. A l'usage des scientifiques*. Dunod, Paris.
Giraudo, Helene & Jonathan Grainger (2001), "Priming complex words: Evidence for supralexical representation of morphology", *Psychonomic Bulletin & Review* 8**,** p. 127-131.
Giraudo, Helene & Jonathan Grainger (2003), "A supralexical model for French derivational morphology". In: D., Sandra, & A. Assink (eds.) *Reading complex words*. Kluwer, A'dam.
Goeman, A. & J. Taeldeman (1996), "Fonologie en morfologie van de Nederlandse dialecten. Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten". In: *Taal en Tongval* 48: 38-59. Zie ook: www.meertens.knaw.nl/projecten/mand/MANDpublicaties.html
Gold, E. Mark (1967), "Language identification in the limit", *Information and Control* 10, p. 447-474.
Goldsmith, John (2000), "Linguistica: An Automatic Morphological Analyzer". In: *The Proceedings from the Main Session of the Chicago Linguistic Society's Thirty-sixth Meeting*, 36-1. Arika Okrent and John Boyle (eds.).
Goldsmith, John (2001), "Unsupervised Learning of the Morphology of a natural language". In: *Computational Linguistics*, vol. 27-2, p. 153-198. Zie ook: http://humanities.uchicago.edu/faculty/goldsmith/
Gonnerman, L.M., M.S. Seidenberg & E.S. Andersen (2004, ms.). "Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology" (ingediend). Web-link naar voorpublicatie: http://lcnl.wisc.edu/people/marks/pubs/GonnermanSeidenbergAndersen.submitted.pdf
Graft, Kenneth Alan (1990), *Paradigmatic configurations and the synchronic lexicon: Theory and application (Volumes I and II)*. Dissertation. UMI, Ann Arbor.
Gruber, J.S. (1976), *Lexical Structures in Syntax and Semantics*. North-Holland: Amsterdam.
De Haas, Wim & Mieke Trommelen (1993), *Morfologisch Handboek van het Nederlands. Een overzicht van de woordvorming*. SDU, Den Haag.
Ten Hacken, Pius (1994), *Defining Morphology. A Principled Approach to Determining the Boundaries of Compounding, Derivation, and Inflection*. Georg Olms AG, Hildesheim.
Haegeman, Liliaene (1991), *Introduction to Government & Binding Theory*. Blackwell, Oxford.
Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn (1997), *Algemene Nederlandse Spraakkunst*. Tweede, geheel herziene druk, 1997. Groningen/Deurne, Martinus Nijhoff uitgevers/Wolters Plantyn. 2 banden + register.
Halle, Morris (1973), "Prolegomena to a theory of word formation", *Linguistic Inquiry* 4, p. 3-16.

Halle, Morris & Alec Marantz (1993), "Distributed morphology and the pieces of inflection." In: K. Hale & S.J. Keyser (eds.), *The view from building 20: Essays in honor of Sylvain Bromberger*. Cambridge, MA: MIT Press.
Harley, T. (2002), *The psychology of language*, 2e editie. Hove: Erlbaum.
Harris, Zellig S. (1955), "From phoneme to morpheme", *Language* 31, p. 190-222.
Harris, Zellig S. (1967), "Morpheme boundaries within words: report on a computer test." In: *Transformations and Discourse Analysis Papers*, Vol. 73.
Hay, Jennifer & Harald Baayen (2002), "Parsing and productivity". In: *YoM*, p. 203-235.
Heemskerk, Josée (1993), "A probabilistic context-free grammar for disambiguation in morphological parsing". In: *Proceedings of the sixth conference of the EACL*, p. 183-192.
Heemskerk, Josée & Vincent van Heuven (1993), "MORPA: A morpheme lexicon based morphological parser". In: V. van Heuven & L. Pols (eds.), *Analysis and synthesis of speech. Strategic research towards high-quality text-to-speech generation*, p. 68-85.
Van Heuven, V.J., A.H. Neijt and M. Hijzelendoorn (1994), "Automatische indeling van Nederlandse woorden op basis van etymologische filters". *Spektator* 23:4, p. 279-291.
Heynderickx, Priscilla & Jaap van Marle (1994), "Over het hybride karakter van -isch: Op de grens van inheems en uitheems", *Spectrum* 23: p. 229-239.
Heyvaert, E., A. Moerdijk & al. (eds.) (1998), *Het grootste woordenboek ter wereld. Een kijkje achter de kolommen van het Woordenboek der Nederlandse Taal (WNT)*. SDU, Den Haag en Standaard Uitgeverij, Antwerpen.
Hockett, Charles (1958), *A course in Modern Linguistics*. New York: Academic Press.
Hoeksema, Jacob (1984), *Categorial Morphology*. Dissertation. Groningen.
Hoeksema, Jacob (1988), "Head-types in morpho-syntax", In: *YoM*, p. 123-38.
Hoeksema, Jacob (2000), "Compositionality of meaning". In: *MIH*, sectie 82.
Van der Hulst, Harry & Michael Moortgat (1980), *ALEX*. INL Working Paper 2, INL, Leiden.
Iacobini, Claudia (2000), "Base and direction of derivation". In: *MIH*, sectie 84.
Jackendoff, Ray (1975), "Morphological and Semantic Regularities in the Lexicon", *Language* 51: 639-671
Jackendoff, Ray (1990), *Semantic Structures*. Current Studies In Linguistics 18. MIT Press.
Jackendoff, Ray (1997), *The architecture of the Language Faculty*. Linguistic Inquiry Monograph 28. MIT Press, Cambridge.
Jackendoff, Ray (2002), *Foundations of Language. Brain, Meaning, Grammar, Evolution*. New York: Oxford University Press, Oxford.
Janssen, Maarten (2002). *SIMuLLDA. A Multilingual Lexical Database Application using a Structured Interlingua*. PhD Thesis CKI, Universiteit Utrecht.
Jescheniak, Jörg D. & W.J.M. Levelt (1994), "Word frequency effects in speech production: Retrieval of syntactic information and of phonological form", *Journal of Experimental Psychology*, *Learning, Memory and Cognition* 20 (4), p. 824-843.
Jespersen, Otto (1928), *An international language*. Web-link: http://www.geocities.com/Athens/Forum/5037/AIL.html
Jespersen, Otto (1949-1958), *A Modern English Grammar on Historical Principles*. London, George Allen & Unwin. Vol. II, 1.15.
De Jong, Nivja H., R. Schreuder & R.H. Baayen (2000), "The morphological family size effect and morphology", *Language and Cognitive Processes*, 15 (4/5), 329-365.
De Jong, Nivja H., *Morphological families in the mental lexicon*. Dissertatie. MPI Series in Psycholinguistics. Max Planck Institute for Psycholinguistics, Nijmegen.
Kager, René (2001), "Stem Stress and Peak Correspondence in Dutch". In: *Optimality Theory*, p. 121-150
Kamp, Hans & Uwe Reyle (1993), *From Discourse To Logic*. Dordrecht: Reidel.

Kay, Paul (1997), *An Informal Sketch of a Formal Architecture for Construction Grammar*. Beschikbaar via: http://www.icsi.berkeley.edu/~kay/bcg/ConGram.html
Kelly, M.H. (1992). "Using sound to solve syntactic problems: The role of phonology in grammatical category assignments". *Psychological Review, 99*, 349-364.
Kerstens, Johan, E. Ruys, M. Trommelen & F. Weerman (1997), *Plato's probleem. Een inleiding in de generatieve taalkunde*. Coutinho, Bussum.
Kiparsky, Paul (1982), "Lexical Morphology and Phonology". In: I.-S. Yang (ed.), *Linguistics in the Morning Calm*, p. 3-91. Hanshin: Seoul.
Kiparsky, Paul (1982a), "From cyclic to lexical phonology". In: H. van der Hulst & N. Smith (eds.), *The structure of phonological representations*. Part 1, p. 131-176. Dordrecht: Foris.
Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum, and Ivan A. Sag (1985), *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge, MA.
Koornwinder, Oele (1997, ms.), *Gelaagde Kwantificatie*. Doctoraalscriptie, Univ. Utrecht.
Koornwinder, Oele & Henk Verkuyl (2000), "Morphological effects of lexical aspect". In: *LIN* 2000, p. 143-158.
Kostić, Aleksandar (1995), "Information Load Constraints on Processing Inflected Morphology." In: *MALP* (Feldman, 1995), p. 317-344.
Kostić, Aleksandar, T. Marković & A. Baucal (2003), "Inflectional Morphology and Word Meaning: Orthogonal or Co-Implicative Cognitive Domains?" In: *MSLP* (Baayen, 2003).
Kripke, Saul (1972), "Naming and Necessity". In: D. Davidson & G. Harman (eds.), *Semantics of Natural Language*, p. 253-355. Reidel, Dordrecht.
Krott, Andrea (2001), *Analogy in Morphology. The selection of Linking Elements in Dutch Compounds*. Dissertation, Radboud Universiteit, Nijmegen.
Krott, Andrea, R.H. Baayen & R. Schreuder (2001), "Analogy in morphology: modeling the choice of linking morphemes in Dutch", *Linguistics* 39(1), p. 51-93.
Landauer, T.K. & S.T. Dumais (1997), "Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge". *Psychological Review* 104 (2), p. 211-240. Zie ook: http://lsa.colorado.edu/
Landauer, T.K. (2002). "On the computational basis of learning and cognition: Arguments from LSA". In: N. Ross (ed.), *ThePsychology of Learning and Motivation* 41, p. 43-84. Zie ook: http://lsa.colorado.edu/
Laudanna, A. & C. Burani (1985), "Address Mechanisms to Decomposed Lexical Entries", *Linguistics* 23, p. 775-792.
Laudanna, A. & C. Burani (1995), "Distributional properties of derivational affixes: Implications for processing". In: *MALP*, p. 345-364.
Laureys, T., G. de Pauw, H. van Hamme, Walter Daelemans & D. van Compernolle (2004), "Evaluation and Adaptation of the Celex Dutch Morphological Database". In: M.T. Lino e.a. (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation*, p. 1247-1250. Zie ook: http://cnts.uia.ac.be/cnts/ps/20040615.7610.text.pdf
Levelt, Willem J.M. (1989), *Speaking: From Intention to Articulation*. MIT Press.
Lieber, Rochelle (1980), *On the organization of the Lexicon*. Dissertation. Bloomington: IULC.
Lieber, Rochelle & R. Harald Baayen (1993), "Verbal prefixes in Dutch: a study in lexical conceptual structure". In: *YoM*, p. 51-78.
Lieber, Rochelle & R. Harald Baayen (1997), "A semantic principle of auxiliary selection in Dutch". In: *Natural Language and Linguistic Theory* 15, p. 789-845.
Lieber, Rochelle & R. Harald Baayen (1998), "Nominalizations in a calculus of lexical semantic representations". In: *YoM*, p. 175-197.

Lieske, C. (1994), *Object- and Database-oriented Integration of the CELEX Lexical Data in a System for Natural Language Grammar Engineering*. Doctoraalscriptie. Universität Koblenz-Landau, Duitsland.

Loonen, Nard (2003), *Stante pede gaande van dichtbij langs AF bestemming @.* Proefschrift, Universiteit Utrecht. CD-ROM-publicatie, in eigen beheer uitgegeven: ljml@tiscali.nl. Beschikbaar via http://www.library.uu.nl/digiarchief/dip/diss/2003-0709-125214/AF.HTM

Lowie, Wander (1998), *The acquisition of interlanguage morphology: a study into the role of morphology in the L2 learner's mental lexicon*. Diss. Groningen. Zie ook: http://www.ub.rug.nl/eldoc/dis/arts/w.m.lowie/

Lukatela, G., B. Gligorijević, A. Kostić & M.T. Turvey (1980). "Representation of Inflected Nouns in the Internal Lexicon." In: *Memory and Cognition*, 8, 415-423.

Lyons, J. (1977). *Semantics*. Cambridge Un. Press.

MacWhinney, Brian (1978), *The acquisition of morphophonology*. Monographs of the Society for Research in Child Development, 43 (1-2, Serial No. 174).

MacWhinney, Brian & Jared Leinbach (1978), "Implementations are not conceptualizations: Revising the verb learning model", *Cognition* 40, p. 121-157.

Marantz, Alec (1997), "No Escape from Syntax: Don't Try Morphological Analysis in the Privacy of Your Own Lexicon". In: A. Dimitriadis, L. Siegel & al. (eds.), *University of Pennsylvania Working Papers in Linguistics*, vol. 4.2. Proceedings of the 21st Annual Penn Linguistics Colloquium, p. 201-225.

Marantz, Alec (2001), *Words*. Lecture notes bij PhD-cursus in de LOT-zomerschool.

Marantz, Alec (2003), *Brain Waves and Button Presses: The Role for Experiments in Theoretical Linguistics*. Lezing t.g.v. jubileumviering UiL-OTS. Tekst en slides beschikbaar via: http://web.mit.edu/marantz/Public/Utrecht/

Marchand, Hans (1969), *The Categories and Types of Present-day English Word-formation*. München: C.H. Beck.

De Marcken, Carl (1995), *Unsupervised Language Acquisition*. Diss. MIT, Cambridge.

Marcus, Gary, U. Brinkman, H. Clahsen, R. Wiese & S. Pinker (1995), "German inflection: The exception that proves the rule", *Cognitive Psychology* 29, p. 189-256.

Marcus, Gary (1999), *The algebraic mind*. Cambridge, MA: MIT Press.

Margolis, E. & S. Laurence (eds.) (1999), *Concepts. Core Readings*. Bloemlezing, met uitgebreide introductie. MIT Press, Cambridge.

Van Marle, Jaap & G.A.T. Koefoed (1980), "Over Humboldtiaanse taalveranderingen, morfologie en de creativiteit van taal". In: *Spektator* 10, p. 111-147. Zie ook: http://www.dbnl.org/tekst/marl002humb01/index.htm

Van Marle, Jaap (1985), *On the paradigmatic dimension of morphological creativity*. Diss. Utrecht. Dordrecht: Foris.

Van Marle, Jaap (1986), "The Domain Hypothesis: The Study of Rival Morphological Processes", *Linguistics*, 24: 601-627.

Marslen-Wilson, W.D., L.K. Tyler, R. Waksler, & L. Older (1994), "Morphology and meaning in the English mental lexicon", *Psychological Review* 101, p. 3-33.

Mattens, Willy (1970), *De indifferentialis: Een onderzoek naar het numerieke gebruik van het substantief in het Algemeen Bruikbaar Nederlands*. Assen: Van Gorcum, Prakke & Prakke.

Matthews, P.H. (1972), *Inflectional Morphology: A theoretical study based on aspects of Latin Verb Conjugation*. Cambridge Univ. Press, Cambridge .

Matthews, P.H. (1974), *Morphology*. Cambridge Univ. Press, Cambridge .

McCarthy, J.J. & A. Prince (1993, ms.), *Prosodic morphology I. Constraint interaction and satisfaction*. University of Amherst and Rutgers University.

McClelland, J.L. & D.E. Rumelhart (1981), "An interactive activation model of context effects in letter perception: Part 1. An account of basic findings", *Psychological Review* 88, p. 375-405. Zie ook: http://www.itee.uq.edu.au/~cogs2010/cmc/chapters/IAC/#Intro

McKinnon, Richard, Mark Allen and Lee Osterhout (2003), "Morphological decomposition involving non-productive morphemes: ERP evidence", *Cognitive Neuroscience and Neuropsychology*, Vol. 14 No 6, p. 883-886. Ook beschikbaar via: http://faculty.washington.edu/losterho/fulltext.pdf

Meesters, Gert (2002, ms.), *Marginale morfologie in het Nederlands. Paradigmatische samenstellingen, neoklassieke composita en splintercomposita*. Dissertation, Leuven.

Meijs, W.J. (1985), "Morphological meaning and the structure of the mental lexicon." In: T. Weyters (ed.), *Meaning and the lexicon*. Dordrecht: Foris Publications.

Moerdijk, Fons (2002), *Het woord als doelwit*. Oratiereeks. Vossiuspers UvA, Amsterdam.

Montague, Richard (1974), "The Proper Treatment of Quantification in Ordinary English". In: R.H. Thomason (ed.), *Formal Philosophy. Selected papers of Richard Montague*. Yale.

Moortgat, Michael (1981), "Subcategorization and the notion 'lexical head'", *LIN*, p. 45-54.

Moortgat, Michael (1985), *Kasimir, A Categorial Grammar Parser*. INL Working Paper.

Moortgat, Michael (1987), "Compositionality and the Syntax of Words". In: J. Groenendijk, D. de Jongh, M. Stokhof (eds.), *Foundations of Pragmatics & Lexical Semantics*, p. 41-62. Foris, Dordrecht.

Moortgat, Michael (1999), "Constants of grammatical reasoning". In: Bouma, Hinrichs, Kruijff & Oehrle (eds.), *Constraints and Resources in Natural Language Syntax and Semantics*, p. 195-219. CSLI, Stanford.

Moortgat, Michael & Harry van der Hulst (1981), "Geïnterpreteerde Morfologie", *Glot* 4-2/3, p. 179-214.

Moscoso del Prado Martín, Fermin (2003), *Paradigmatic Structures in Morphological Processing: Computational and Cross-Linguistic Experimental Studies*. Dissertatie. MPI Series in Psycholinguistics. Max Planck Institute for Psycholinguistics, Nijmegen. Zie ook: www.mrc-cbu.cam.ac.uk/~fermin.moscoso-del-prado-martin/

Moscoso del Prado Martin, F., A. Kostić & R.H. Baayen (2004), "Putting the Bits Together: an Informational Perspective on Morphological Processing", *Cognition* 94 (1), p. 1-18.

Muysken, Pieter (1999), *Talen. De toren van Babel*. Amsterdam University Press, A'dam.

Napps, S.E. (1985). Morphological, Semantic, and Formal Relationships in the Organization of the "Mental Lexicon". PhD dissertation. Dartmouth College, Massachusetts.

Napps, S.E. & C.A. Fowler (1987). Formal relationships among words and the organization of the mental lexicon. *Journal of Psycholinguistic Research* 16, p. 257-272.

Napps, S.E. (1989). Morphemic relationships in the lexicon: Are they distinct from semantic and formal relationships? *Memory and Cognition* 17, p. 729-739.

Neef, Martin (1999), "A declarative approach to conversion into verbs in German." In: *YoM*, p. 199-224.

Neeleman, Ad & Joleen Schipper (1992), "Verbal prefixation in Dutch: thematic evidence for conversion". In: *YoM*, p. 57-92.

Neijt, A.H. & J.J. Zuidema (1994), *Spellingdossier. Deel I. Spellingrapport*. SDU, Den Haag.

Neijt, Anneke, R. Schreuder & R.H. Baayen (2003), "Verpleegsters, ambassadrices and masseuses: Stratum differences in the comprehension of Dutch words with feminine agent suffixes". In: *LIN* 2003, p. 117-128.

Neuvel, Sylvain (2001), "Whole Word Morphologizer: Expanding the Word-Based Lexicon: A non-stochastic computational approach", *Brain and Language* 81, p. 454-463.

Neuvel, Sylvain & Sean A. Fulop (2002), "Unsupervised Learning of Morphology Without Morphemes". In: *Proceedings of the ACL Workshop on Morphological and Phonological Learning 2002.* ACL Publications. Of: www.neuvel.net

Neuvel, Sylvain & R. Singh (2002), "Vive la difference! What morphology is about", *Folia Linguistica* 35: 3-4, p. 313-320. Of: www.neuvel.net
Newman, S. (1948), "English Suffixation: A descriptive approach", *Word* 4, p. 24-36.
Nida, Eugene (1949), *Morphology. The descriptive analysis of words*. University of Michigan Press, Ann Arbor, MI.
Nunn, Anneke (1998), *Dutch Orthography; A Systematic Investigation of the Spelling of Dutch Words*. Dissertation, Radboud Universiteit, Nijmegen.
Nunn, Anneke (2000), "Automatic hyphenation of Dutch words based on linguistic rules." In: *Proceedings of CLIN* 1999.
Oehrle, Richard T. (2000, ms.), *Logics for intercalation*. Preprint, Universiteit Utrecht.
Van Oostendorp, Marc (1998), "Dutch Orthography". Review van Nunn (1998). In: *Nederlandse Taalkunde* 4.3. Zie ook: www.vanoostendorp.nl
Ordelman, Roeland (2003), *Dutch speech recognition in multimedia information retrieval*. Dissertatie. CTIT, Enschede. Taaluitgeverij Neslia Paniculata.
Van Parreren, C.F. (1971), *Psychologie van het leren I*. Van Loghum Slaterus, Deventer.
Peters, Ann M. (1976), "Language learning strategies: Does the whole equal the sum of the parts?", *Language* 53, p. 560-573.
Peters, Ann M. (1983), *The units of language acquisition*. Cambridge Monographs and Texts in Applied Psycholinguistics. Cambridge Univ. Press, Cambridge.
Petruck, Miriam R. L. (1996): Frame Semantics. In: Jef Verschueren, Jan-Ola Östman, Jan Blommaert, and Chris Bulcaen (eds.), H*andbook of Pragmatics*. Philadelphia: John Benjamins. Beschikbaar via: http://www.icsi.berkeley.edu/~framenet/
Pianesi, Fabio & Achille C. Varzi (1996), "Events, Topology and Temporal Relations". In: *The Monist*. Vol. 79, no. 1, p. 89-116.
Pinker, Steven & Alan Prince (1988), "On language and connectionism: Analysis of a Parallel Distributed Processing model of language acquisition", *Cognition* 28, p. 73-193.
Pinker, Steven & Alan Prince (1994), "Regular and irregular morphology and the psychological status of rules of grammar". In: *The reality of linguistic rules* (RLR), p. 321-352.
Pinker, Steven (1998), "Words and Rules", *Lingua 106*, 219-242.
Plag, Ingo (1996), "Selectional restrictions in English suffixation revisited: a reply to Fabb (1988)", *Linguistics* 34, p. 769-798.
Plag, Ingo (1998), "The polysemy of *-ize* derivatives: on the role of semantics in word formation". In: *YoM*, p. 219-242.
Plag, Ingo (1999), *Morphological Productivity. Structural Constraints in English Derivation*. Mouton de Gruyter. Berlin, New York.
Plag, Ingo (2002, ms.), "The role of selectional restrictions, phonotactics and parsing in constraining suffix ordering in English". Max Planck Instituut, Nijmegen.
Plag, Ingo, C. Dalton-Puffer & R.H. Baayen (1999), "Morphological productivity across speech and writing", *English Language and Linguistics* 3.2, p. 209-228.
Plaut, David C. & Laura M. Gonnerman (2000), "Are non-semantic morphological effects incompatible with a distributed connectionist approach to language processing?", *Language and Cognitive Processes* 15, p. 445-485. Web-link: http://www.cnbc.cmu.edu/~plaut/papers/pdf/PlautGonnerman00LCP.morph.pdf
Plunkett, K. & V. Marchman (1991), "U-shaped learning and frequency effects in a multi-layered perceptron", *Cognition* 38, p. 43-102.
Pollard, Carl and Ivan A. Sag (1987), *Information-Based Syntax and Semantics*. CSLI, Stanford, California.
Pollard, Carl and Ivan A. Sag (1994), *Head-Driven Phrase Structure Grammar*. CSLI, Stanford, California.
Popma, Jildou (1992), "Suffixparen in het Nederlands". In: TABU 1992, Groningen.

Posthumus, Jan (1997), "Een overzicht van de veranderingen in inhoud en inrichting van Koenens Verklarend Handwoordenboek". In: *Honderd Jaar Koenen*, met bijdragen van Jan Posthumus, Siemon Reker en Arie de Ru. Van Dale Lexicografie, Utrecht-Antwerpen.

Prasada, Sandeep & Steven Pinker (1993), "Generalizations of Regular and Irregular Morphological Patterns", *Language and Cognitive Processes* 8, p. 1-56.

Prince, Alan & Paul Smolensky (1993), *Optimality Theory: Constraint Interaction in Generative Grammar*. Interne publicatie, Rutgers University Cognitive Science Center, New Brunswick, NJ. MIT Press.

Pustejovsky, James (1991), "The Generative Lexicon", *Computational Linguistics* 17, p. 409-441.

Rastle, Kathleen, M.H. Davis & B. New (2004), "The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition", *Psychonomic Bulletin & Review* 11. p. 1090-1098. Web-link: http://www.borisnew.org/ressources/Morpho-orthographic%20segmentation-2004.pdf

Rohde, D.L.T & D.C. Plaut (2003), "Connectionist models of language processing", *Cognitive Studies, Japan* 10(1), p. 10-28. Web-link: http://tedlab.mit.edu/~dr/Papers/RohdePlaut03.pdf

Richter, Frank (2000), *A mathematical formalism for linguistic theories with an application in head-driven phrase structure grammar*. Dissertatie, Universiteit Tübingen.

Riehemann, Suzanne Z. (1998), "Type-based Derivational Morphology", *Journal of Comparative Germanic Linguistics* 2, p. 49-77.

Riehemann, Suzanne Z. (2001), *A constructional approach to idioms and word formation*. Dissertatie, Stanford University.

Rosch, Eleanor (1978), "Principles of Categorization", *CCR* (1999). Oorspronk.: E. Rosch & M. Munitz (eds.), *Languages, Belief and Metaphysics,* vol. I, 1970. New York Press.

Rumelhart, David & James McClelland (1986), "On learning the past tenses of English Verbs. Implicit Rules or Parallel Distributed Processing?" In: J. McClelland, D. Rumelhart and the PDP Research Group, *Parallel Distributed Processing: Explorations of the Microstructure of Cognition*. Cambridge, MA: MIT Press.

Sadler, Louise and Andrew Spencer (2001), "Syntax as an exponent of morphological features". In: YoM 2000, p. 71-96.

Sandra, Dominiek (1990), "On the representation and processing of compound words. Automatic access to constituent morphemes does not occur", *Quarterly Journal of Experimental Psychology* 42A, p. 529-567.

Sandra, Dominiek (1994), "The morphology of the mental lexicon: internal word structure viewed from a psycholinguistic perspective", *Language and Cognitive Processes* 9 (3), p. 227-269.

Sandra, Dominiek, S. Frisson & Fr. Daems (1999). "Why simple verb forms can be so difficult to spell: The influence of homophone frequency and distance in Dutch", *Brain and language* 68, 277-283

Van Santen, Ariane (1992), *Productiviteit in taal en taalgebruik. Een studie op het gebied van de Nederlandse woordvorming*. Diss. Leiden.

Van Santen, Ariane (1995), "Beschrijving en theorie in het Morfologisch Handboek", *Leuvense Bijdragen: Tijdschrift voor Germaanse Filologie* 84, p. 543-560.

De Saussure, Ferdinand (1916), *Cours de Linguistique Générale*. Paris, Payot.

Scha, Remko (1990), "Taaltheorie en taaltechnologie; competence en performance". In: R. de Kort en G.L.J. Leerdam (ed.), *Computertoepassingen in de Neerlandistiek*. Almere: LVVN, 1990, pp. 7-22. Zie ook: http://iaaa.nl/rs/Leerdam.html

Schaerlakens, A.M. & S. Gillis (1987), *De Taalverwerving van het kind: Een hernieuwde oriëntatie in het Nederlandstalig onderzoek*. Groningen: Wolters-Noordhoff.

Schone, Patrick & Daniel Jurafsky (2001). "Knowledge-free induction of inflectional morphologies". In: *2nd Meeting of the North American Chapter of the ACL*, p. 183–191. Association for Computational Linguistics, Morgan Kaufman.
Schreuder, Robert (1990), "Lexical Processing of verbs with separable particles". In: *YoM*, p. 65-79.
Schreuder, Robert & R. Harald Baayen (1994), "Prefix Stripping Re-Revisited", *Journal of Memory and Language* 33, 357-375.
Schreuder, Robert & R. Harald Baayen (1995), "Modeling morphological procesing." In: *MALP* (Feldman, 1995), p. 131-154.
Schreuder, Robert & R. Harald Baayen (1997), "How complex simplex words can be", *Journal of Memory and Language* 37, p. 118-139.
Schreuder, Robert, C. Burani & R.H. Baayen (2003). "Parsing and semantic opacity." In E. Assink & D. Sandra, *Reading complex words*. Cross-language studies (pp. 159-189). Dordrecht: Kluwer.
Schultink, Hans (1961), "Productiviteit als morfologisch fenomeen." *Folia der Letteren* 2, p. 110-125.
Schultink, Hans (1962), *De morfologische valentie van het ongelede adjectief in modern Nederlands*. Diss. Den Haag: Van Goor. Herdruk, 1980, HES Publishers, Utrecht.
Schultink, Hans (1978), "Ambassadrice contra masseuse. Afgeleide, [+vrouwelijke], Nederlandse nomina en hun beschrijving", *De nieuwe Taalgids 71*, p. 594-601.
Schultink, Hans (1994), "Een eeuw Nederlandse morfologie; de ontwikkelingsgang van een discipline", *Spectator* 23-1, p. 45-77.
De Schutter, G. & P. van Hauwermeiren (1983), *De structuur van het Nederlands. Taalbeschouwelijke grammatica.* De Sikkel, Malle.
De Schutter, G. & S. Gillis (1990), "Structurele aspecten van het Nederlandse lexicon", *Antwerp Papers in Linguistics*, vol. 64.
De Schutter, G. (1994), "Recensie: W. de Haas en M. Trommelen: Morfologisch handboek van het Nederlands", *Taal en tongval* 46, p. 89-97.
Seidenberg, Marc (1987), "Sublexical structures in visual word recognition: Access units or orthographic redundancy?". In: M. Colthaert (ed.), *Attention and performance* XII. Hove: Lawrence Erlbaum Associates Ltd.
Seidenberg, Marc & Laura Gonnerman (2000), "Explaining derivational morphology as the convergence of codes", *Trends in Cognitive Sciences* 4(9), 353-361.
Siegel, Doris (1974), *Topics in English Morphology*. Dissertatie. MIT, Cambridge, Mass.
Van der Sijs, Nicoline (2001), *Chronologisch woordenboek, De ouderdom en herkomst van onze woorden en betekenissen*. Dissertatie. Veen, Amsterdam/Antwerpen.
Skousen, Royal (1979), "Empirical interpretations of psychological reality". In: E. Fischer-Jorgensen, J. Rischel and N. Thorsen (eds.), *Proceedings of the Ninth Internat. Congress of Phonetic Sciences*, vol. 2, p. 121-128. Institute of Phonetics, Univ. of Copenhagen.
Skousen, Royal (1989), *Analogical Modeling of Language*. Kluwer, Dordrecht.
Smedts, Willy (1979), *Lexicale morfologie: de beheersing van de woordvorming door Vlaamse 'brugklassers'*. Dissertatie, KU Leuven.
Spencer, Andrew (1991), *Morphological Theory*. Uitgave van 1993. Blackwell Publishers.
Sproat, Richard (1992), *Morphology and Computation*. MIT Press Series in Natural-Language Processing.
Stanners, R.F., J.J. Neiser, W.P. Hernon & R. Hall (1979a). "Memory representation for related words", *Journal of Verbal Learning and Verbal Behavior*, 18, 399-412
Stanners, R.F., J.J. Neiser, & S. Painton (1979b), "Memory representation for prefixed words", *Journal of Verbal Learning and Verbal Behavior*, 18, 733-743.

Stemberger, Joseph P. & Brian MacWhinney (1988), "Are inflected forms stored in the lexicon?" In: M. Hammond & M. Noonan (eds.), *Theoretical Morphology: Approaches in modern linguistics*, p. 101-116. London: Academic Press.
Stemberger, Joseph P. (1994), "Rule-Less Morphology at the Phonology-Lexicon Interface". In: *The Reality of Linguistic Rules* (RLR), p. 147-170.
Taft, Marcus & K.I. Forster (1975), "Lexical Storage and retrieval of prefixed words", *Journal of Verbal Learning and Verbal Behavior* 14, p. 271-294.
Taft, Marcus (1979), "Recognition of affixed words and the word frequency effect", *Memory & Cognition* 7, 263-272.
Taft, Marcus (1988), "A Morphological Decomposition Model of Lexical Representation", *Linguistics* 26, 657-667.
Taft, Marcus (1994), "Interactive-activation as a Framework for Understanding Morphological Processing", *Language and Cognitive Processes* 9 (3), 271-294
Taft, Marcus (1994a), "Prefix Stripping Revisited", *Journal of Verbal Learning and Verbal Behavior*, 20, 289-297.
Trommelen, Mieke & Wim Zonneveld (1986), "Dutch Morphology: Evidence for the Righthand Head Rule", *Linguistic Inquiry* 17, p. 147-169.
Uhlenbeck, E.M. (1953), "The study of Word-Classes in Javanese", *Lingua* 2, p. 322-354. [Herdrukt in Uhlenbeck (1978), *Studies in Javanese Morphology*, p. 40-68.]
Uhlenbeck, E.M. (1977), "The concepts of productivity and potentiality in morphological descriptions and their psycholinguistic reality", *Salzburger Beiträge zur Linguistik* 4, p. 379-391.
Uhlenbeck, E.M. (1979), "Hoe een linguïst omgaat met ambassadrices en masseuses." In T. Hoekstra & H. van der Hulst (eds.), *Morfologie in Nederland* (Glot-special), p. 7-20.
Vennemann, Theo (1974), "Words and Syllables in Natural Generative Phonology". In: *Papers from the Parasession on Natural Phonology*. Chicago Linguistic Society.
Verhey, A.J.C. (2000), *Bits, Bytes, and Binyanim. A quantitative study of verbal lexeme formations in the hebrew bible.* Orientalia Lovaniensia Analecta 93. Peeters, Leuven.
Verkuyl, Henk J. (1978), "Lexicon en werkelijkheid", *Forum der letteren* 19, 1, p. 20-39.
Verkuyl, Henk J. (1993), "Hoe goed of hoe fout is Van Dale?, *De Nieuwe Taalgids* 86, I: 212-237, II: 303-327. Ook beschikbaar via http://www.let.uu.nl/~Henk.Verkuyl/personal.
Verkuyl, Henk J. (1993a), *A theory of aspectuality. The interaction between temporal and atemporal structure*. Cambridge Studies in Linguistics 64. Cambridge Univ. Press.
Verkuyl, Henk J. (1996), "Komt er een fusie tussen Van Dale en Winkler Prins?". In: *Trefwoord* 13, *Jaarboek Lexicografie 1998-1999,* SDU Uitgevers: Den Haag, p. 135-151. Of: www.let.uu.nl/~Henk.Verkuyl/personal (list of publications, year 1996)
Verkuyl, Henk J. (1996b), *De schouders waarop wij staan. Taalfilosofische grondslagen voor taalkundig onderzoek.* Openingscollege. Faculteit der Letteren, Universiteit Utrecht.
Verkuyl, Henk J. & al. (1998), *The OTS Dictionary Project*. Working Paper van de Werkgroep Lexicon. UiL OTS, Utrecht.
Verkuyl, Henk J. (1999), Stereotyping, Prototyping, and Figurative Use: Towards a Proper Semantic Analysis. In: T.F. Shannon & J.P. Snapper (eds.), *The Berkeley Conference on Dutch Linguistics 1997. The Dutch Language at the Millennium.* Univ. Press of America: Lanham, New York and Oxford, 2000, p. 21-43.
Verkuyl, Henk J. (2000), *Semantiek. Het verband tussen taal en werkelijkheid.* Amsterdam Univ. Press.
Verkuyl, Henk J. (2003), *Woorden, woorden, woorden.* Afscheidsrede. Interne publicatie van de Faculteit der Letteren, Universiteit Utrecht.

Voga, Madeleine & Jonathan Grainger (2004), "Masked Morphological Priming with Varying Levels of Form Overlap: Evidence from Greek Verbs". In: *Current Psychology Letters* 13, Vol. 2. Zie: http://cpl.revues.org/document422.html
De Vries, J.W. (1975), *Lexicale morfologie van het werkwoord in modern Nederlands*. Leiden: Univ. Pers.
Wijk, Judith van (2002), "The Dutch plural landscape", in: *LIN 19*, p. 211-221.
Williams, Edwin (1981), "On the notions 'Lexically Related' and 'Head of a Word' ", *Linguistic Inquiry*, Vol. 12(2), p. 245-274.
Wittgenstein, Ludwig (1953), *Philosophical Investigations*. 3e, vertaalde editie. Sectie 65-78, in *CCR* (1999).
Wong Fillmore, Lily (1976), *The second time around: cognitive and social strategies in second language acquisition*. Dissertatie,. Stanford University.
Van der Wouden, Ton (1988), "Automatic Morphology for Lexical Databases", *GRAMMA, tijdschrift voor taalkunde* 12 (1988), 1, p. 27-40.
Zipf, G.K. (1935). *Psycho-Biology of Languages*. Houghton-Mifflin.
Zonneveld, Wim (1980), "Autonome spelling", *De Nieuwe Taalgids* 73, 518-537.
Zuidema, Johan, Anneke Neijt & Jeroen Weber (1998), "Hiërarchieën op de knieën", *Spektator* 23, p. 137-163.
Zuidema, Johan (1988). *Efficiënt spellingonderwijs: een leer- en expertmodel voor het spellen*. Diss. Utrecht. ACCO: Leuven/Amersfoort.

Woordenboeken, grammatica's en elektronische datapublicaties

ANS (1997): *Algemene Nederlandse Spraakkunst*. Tweede, geheel herziene druk. Onder redactie van W. Haeseryn, K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn, Groningen/Deurne, Martinus Nijhoff uitgevers/Wolters Plantyn. 2 banden + register. Raadpleegbaar via de E-ANS (zie aldaar).
Augst, Gerhard (1998), *Wortfamilienwörterbuch der deutschen Gegenwartssprache*. In zusammenarbeit mit K. Müller, H. Langner, A. Reichmann. Max Niemeyer Verlag.
Baayen, R.H., R. Piepenbrock & L. Gulikers (1995), *The CELEX Lexical Database.* CD-ROM. Linguistic Data Consortium, Univ. of Pennsylvania, Philadelphia, PA.
Battus (2002), *Opperlans! Taal- & Letterkunde*. Uitgeverij Querido, Amsterdam.
Brouwers, L. (1989), *Het juiste woord. Standaard betekeniswoordenboek der Nederlandse taal*, 7de druk, bewerkt door F. Claes. Antwerpen: Standaard Uitgeverij.
Canoo Dictionary of German Morphology (2000-2005). Canoo Engineering AG: Basel, Switzerland. Permanent toegankelijk via het Canoo-Net: http://www.canoo.net/index.html
Corpus Gesproken Nederlands (2004). Versie 1.0. Ontwikkeld door de Nederlandse Taalunie. Beschikbaar via de TST-centrale: http://www.tst.inl.nl/
Cranshoff, Betty & Johan Zuidema (2002), *De Lijsterbij 3*. Uitgeverij Zwijsen, Maarssen.
Dr. Verschuyl (2003), *Grote puzzelencyclopedie*. Uitgever, Kosmos Z&K
E-ANS (2004): Elektronische versie van de ANS, versie 1.1. Zie: http://oase.uci.kun.nl/~ans/
Heemskerk, Josée & Wim Zonneveld (2000), *Uitspraakwoordenboek*. Ontwikkeld voor de Nederlandse Taalunie. Uitgeverij Het Spectrum, Utrecht.
Huizinga, A. (1998), *Huizinga's Complete lijst van namen. Vraagbaak voor de afkomst van de Nederlandse en Vlaamse familienamen*. Tirion, Baarn.
Kostić, Đ. (1999), *Frequency Dictionary of Contemporary Serbian Language*, vol. I-VII. Belgrado. Zie ook: http://www.serbian-corpus.edu.yu/indexns.htm
MAND (2005): *Morfologische Atlas van Nederlandse Dialecten*. G. De Schutter & al. (eds.), Meertens Instituut, Amsterdam. Amsterdam University Press.
MHB (1993): *Morfologisch Handboek van het Nederlands. Een overzicht van de woordvorming*. Wim de Haas & Mieke Trommelen. SDU, Den Haag.

Nieuwborg, E.R. (1978), *Retrograde woordenboek van de Nederlandse Taal*. 2e druk. Deventer/Antwerpen. Kluwer Technische boeken.

Schludermann, B. & al. (2004): *The Hague Miscellany: Koninklijke Bibliotheek MS 128 E 2. Facsimile and Transcription, Concordance and Finding Lists.* Bewerkt door Brigitte Schludermann, John Dawson & Heinz Bück. Turnhout, Belgium: Brepols Publishers NV, due 2004). Boeken + CD-ROM. Web-link: www.hull.ac.uk/denhaagKB/index.html.

De Schutter, G. & al. (2005), *Morfologische atlas van de Nederlandse dialecten [MAND]. Deel 1: meervoudsvorming bij zelfstandig naamwoorden, vorming van verkleinwoorden, geslacht bij zelfstandig naamwoord, bijvoeglijk naamwoord en bezittelijk voornaamwoord.* Onder redactie van: Georges De Schutter, Boudewijn van den Berg, Ton Goeman en Thera de Jong. Meertens Instituut, Amsterdam. Amsterdam University Press. Boek en CD-ROM. Zie ook: www.meertens.nl/projecten/mand/MAND

Uit den Boogaert, P.C. (1975), *Woordfrequenties - In geschreven en gesproken Nederlands*. Werkgroep Frequentie-Onderzoek van het Nederlands (WFON): Oosthoek, Scheltema & Holkema.

Van Dale (1984a), *Groot Woordenboek der Nederlandse Taal*, 11e herz. druk. Onder redactie van Guido Geerts en H. Heestermans m.m.v. C. Kruyskamp. VDL, Utrecht-Antwerpen.

Van Dale (1984b), *Groot Woordenboek van Hedendaags Nederlands*, 1e druk. Onder redactie van P.G.J. van Sterkenburg & W.J.J. Pijnenburg. VDL, Utrecht-Antwerpen.

Van Dale (1988), *Lexitron*. Elektronisch woordenboek. VDL, Utrecht-Antwerpen.

Van Dale (1991a), *Groot woordenboek van Hedendaags Nederlands*, 2de druk. Onder redactie van P.G.J.van Sterkenburg, in samenwerking met G.E. Booij en P.R.F. Verhoeven. VDL, Utrecht/Antwerpen.

Van Dale (1991b), *Groot woordenboek van Synoniemen en andere betekenisverwante woorden.* Onder redactie van P.G.J.van Sterkenburg, in samenwerking met M. van Dalen, M.J.M. Hooyman en M.E. Verburg. VDL, Utrecht-Antwerpen.

Van Dale (1992), *Groot Woordenboek der Nederlandse Taal*, 12e druk. Onder redactie van Guido Geerts en Ton den Boon. VDL, Utrecht-Antwerpen.

Van Dale (1997a), *Groot Woordenboek der Nederlandse Taal*, 12e druk, nieuwe spelling. Onder redactie van Guido Geerts en H. Heestermans. VDL, Utrecht-Antwerpen.

Van Dale (1997b), *Etymologisch woordenboek. De herkomst van onze woorden*. Onder redactie van P.A.F. Veen en Nicoline van der Sijs. VDL, Utrecht-Antwerpen.

Van Dale (1999), *Groot Woordenboek der Nederlandse Taal*, 13e, herz. druk. Onder redactie van Guido Geerts en Ton den Boon. VDL, Utrecht-Antwerpen.

Van Dale (2000), *Groot Woordenboek der Nederlandse Taal op CD-ROM*. Versie 1.0. Gebaseerd op de 13e druk van de Grote Van Dale. VDL, Utrecht-Antwerpen.

Van Dale (2005), *Groot Woordenboek der Nederlandse Taal*. 14e druk. VDL, Utrecht-Antwerpen.

WNT (1864-1998), *Woordenboek der Nederlandsche taal*. Den Haag. M. Nijhoff, A.W. Sijthoff e.a., afl. 1-686 (1864-1998), Deel I-XXIX (1882-1998), 40 banden.

Curriculum Vitae

Oele Koornwinder werd in 1972 geboren als zoon van een mathematicus en een mediaeviste. Hij bezocht van 1978 tot 1984 de Koningin-Emmaschool te Bussum en was van 1984 tot 1989 leerling aan het Gemeentelijk Gymnasium te Hilversum, waar hij met goed gevolg examen deed in een natuurwetenschappelijk georiënteerd vakkenpakket. Hiernaast ontwikkelde hij zich tot een enthousiast amateurpianist (op de 'pianowerkplaats Gert') en kreeg hij grote belangstelling voor literatuur en kunst, filosofie en maatschappelijke vraagstukken.

In 1989 begon Oele aan de studie natuur- en wiskunde aan de Universiteit Utrecht, maar kwam in de loop van het propedeusejaar tot de conclusie dat hij zich meer interesseerde voor de taalkundige fundamenten van wetenschappelijke kennis en stapte daarom over naar de studie Nederlandse Taal- en Letterkunde. Hier raakte hij al snel in de greep van Verkuyl's logische benadering van het taalsysteem. Hij specialiseerde zich in de Syntaxis en Semantiek en studeerde cum laude af op een uitgebreide doctoraalscriptie naar collectieve kwantificatie (ter verklaring van de betekenisverschillen tussen universele kwantoren als 'elke', 'alle' en 'al de'). In deze periode werd hij ook maatschappelijk actief, eerst in facultaire inspraakorganen en later in de (jongeren)politiek.

Met ingang van 1 januari 1998 kreeg Oele de kans om een promotieproject uit te voeren bij het UiL OTS; dit lexicologische project had als doel om een morfologische gegevensbank te ontwikkelen voor het Nederlands op basis van de lexicale kennis bij Van Dale Lexicografie. Het bood hem de kans om zijn theoretische en empirische blikveld te verruimen en een brug te slaan tussen taalkundige en lexicografische onderzoekstradities.

Oele is nu werkzaam bij GridLine BV, een ICT-bedrijf dat zich onder meer toelegt op het opbouwen en ontsluiten van thesaurussystemen. Zijn lexicale expertise wordt hier aangewend voor de ontwikkeling van tools op het terrein van de automatische extractie van terminologie.

Summary in English

This dissertation reports about a theoretical and empirical study to the morphological structure of the Dutch vocabulary. The purpose of this study was to develop a better insight in Dutch morphology by developing a Morphological Databank of Dutch (to be referred to as the MGBN), using the lexicographic resources of a well-known publisher of Dutch dictionaries: Van Dale Lexicografie. As an additional requirement, the databank needed a design which could contribute to the systematic treatment of the word features in the original data resources. This project constitutes the central theme of my dissertation. With this study I intend to integrate two disciplines which have been living apart for a long time: the linguistic (cognitive-grammatical) approach and the lexicographic approach to language knowledge. In general, the cognitive-grammatical studies are focused on explaining the hidden patterns behind the wealth of language data that speakers are processing everyday. As a consequence, this discpline tends to neglect the systematic inventarization of data, which is the core business of lexicographic researchers, who, although having an academic background, often work for a company, either in the field of dictionary making or in the related field of automatic speech analysis and synthesis. Due to this difference in purpose, a long time has passed in which both groups were not very interested in each other's results. In the recent past, however, this situation started to change, at least at the Utrecht University. Here a number of linguists realized that the computer era has brought new challenges to the world of dictionaries, as there is a growing interest in tools for automatic language processing. This kind of applications require a very rich and well-organized lexicon.

The ideal dictionary

In this context, a working group at the UiL OTS invented the concept of an ideal dictionary ("ideaal woordenboek") and wrote a manifest about it (Verkuyl & al, 1998). This name is a metaphor for the kind of system that is required to support the needs of artificial intelligent systems: these meta-minds are only satisfied if they have unlimited access to a huge memory with very systematic descriptions of the linguistic data a human being knows by heart. While human beings would soon end up bored by such a non-fancy type of dictionary, a machine cannot have enough of it. Even a stereotypical lexicographer, which is known to be very obsessed in searching for almost trivial language data, would prefer to read a romantic novel if confronted with such a dictionary. For the linguist, however, it is a real challenge to develop and fill a data representation system that can meet the requirements of such an ideal dictionary. According to the Utrecht Lexicon group, this concept requires a database which is complete, consistent and is corpus-based, which is equivalent to having a statistical basis: only if an automatic language processor has access to information about the plausible and the implausible data patterns of the language task it has to deal with, one may hope this processor to be successful in its task. As the cognitive-grammarian scholar is trained in making generalizations over fragments of language behaviour, he can help the lexicographer with encoding the available knowledge about the field that has to be described. He even might use experimental techniques to find out more about the underlying representation system, like psycholinguists are used to do. This, at least, is the perspective I take in this study. In my view, the first criterion of a scientific theory about linguistic knowledge is whether the theory can be empirically tested against a given body of data, which I call the spectrum of the theory. Only if the theory defines an explicit relation between the rule system and the required data structure, this criterion can be satisfied. This implies that a theory can become better if one confronts it with a (preferably large) set of unseen data.

A Morphological Database of Dutch

The joint effort of the UiL OTS and Van Dale to develop a complete morphological database of Dutch can be seen as a consequence of the recent insight that linguists can profit from the data and the analysis techniques in the lexicographic field, while the dictionary makers can profit from the mathematical approach of the linguist. This also implies that both parties can profit from a joined effort to develop databases with systematic encoded information about language data, like the morphological dimension. Therefore, Van Dale's invitation to develop a language broad data base of Dutch word formation patterns by the semi-automatic enrichment of their existing data resources, offered a very interesting chance to work out this new approach to cognitive-linguistic research.

The Morphological Data Base of Dutch contains all 80.000 base lexemes which underly the 250.000 words (including compound words) that belong to Van Dale's Large Dictionary of Dutch (c.q. Grote Van Dale), the largest public dictionary of present day Dutch, with a time span of more than 100 years. The assigned representations provide information about the lexeme internal morpheme boundaries and their distributional class (i.e. prefix, root or suffix). Each structure representation consists of three layers: a spell form layer and two derived layers which classify the basic morpheme segments by assigning class indices (i.e. unique meta-forms) to segments with the same morphological function. These representations only have been assigned to basic lexemes, but as Van Dale's knowledge base contains information about the lexeme structure of all Dutch compounds, it was possible to obtain their morphological representations by means of a compositional construction method.

To realize a database like this, it is important to think of the question how one can legitimate the structure one assigns to it. If one just starts out with an existing framework of morphological rules, like the Dutch Handbook of Morhology (De Haas en Trommelen, 1993), it probably will turn out to be unworkable, because all grammatical systems are based on the assumption that the rules are already known (c.q. native). But if one has to analyse new data, one continuously has to make decisions about the question which segments are relevant and which segments are not. Therefore the best way to proceed seems to be to invert the question: first start to analyse the data by assigning them an intuitive structure and then try to find out what mechanisms are responsible for these subconscious decisions.

This exactly is the way my project was organized. It harmonized with my general ideas about the fundamental nature of the language system. Nevertheless, the process of morphological structure assigment was a very adventurous job, as I really had no idea what I could expect from the resulting lexicon. Meanwhile, I learned to recognize a huge amount of frequent and remarkable morphological patterns (i.e. typical sequences and clusters of stems and affixes) and I got a clear insight in all the factors that influenced my choice between the different options. By this process of structure assignment, the lexical theory which is presented in this study more or less came into existence by itself: it reflects my experience with the mental organization of my own lexicon. Therefore the main result of this project neither is the morphological database of Dutch neither the linguistic theory that has been derived of it, but the insight that this inductive approach to theory formation can be a very fruitful way to make progress in the language sciences.

The MGBN has been developed by using a paradigmatic structure criterion, which is motivated by the L-KRING theory. According to this criterion, a lexeme internal segment can be identified as a morpheme if it can be substituted by other morphemes without changing the function of the complement. To make this idea more concrete, take a word like *development*: this word can be assigned the structure [DEVELOP]+MENT, which consists of a root DEVELOP and a suffix -MENT. This analysis can be defended by the observation that the suffix -MENT

determines the word category noun and by the fact that it can be substituted by other morphemes, like -ER (which constructs the N *developer*), -ING (which constructs the N *developing*) and a zero "suffix" -0 (which "constructs" the V *develop*). From this paradigm one can conclude that the complement of -MENT, DEVELOP, is a morpheme too, being a root that allows for a lot of morphological contexts c.q. semantic functions. For a native speaker of the invested language, it is easy to judge about the relevance of such formal relations, as he has access to an incredible rich network of word paradigms. Therefore, it is possible to analyse large amounts of words without inspecting all the words individually: a native speaker simply can predict a lot of morphological properties from just a word internal substring. In my project, this type of knowledge has been applied on an industrial scale, resulting in a completely analysed lexicon.

Below I provide an overview of the most important results with respect to the linguistic dimension and the lexicographic dimension of this study.

Linguistic results

This study introduces a general framework for modeling the mental lexicon. It is called an Integral Dynamic Lexicon System (IDL-system). This system can be seen as a more specific implementation of the Ideal Dictionary view of Verkuyl & al. (1998). Its most important feature is its ability to create a dynamic relation between the individual and the collective vocabulary. Departing from the IDL-system, I developed a new perspective on morphological structure. This perspective is formally embedded in a theory which is based on the principle of Lexical Knowledge Representation by Inductive Name Giving (L-KRING). The purpose of this L-KRING theory is to provide a fundamental explanation of the acquisition and activation of morphological knowledge. This formal representation theory is based on the following assumptions:

- the lexicon is able to compress word knowledge without loosing information. To make this possible, the shared units of a set of words have to be substituted by indices (c.q. name based reference items). As a consequence, each morphological complex word can be replaced by a (hierarchical) sequence of indices which refer to the lexical locations where the corresponding morphemes are defined. This compression technique leads to the spontanuous emergence of morphological structure (in a way that is similar to the proposal of Bybee (1985; 1988)). Technically this proposal is realized by defining a lexical inheritance system along the lines of DATR (Evans & Gazdar, 1996).
- there is no morphological grammar in the sense of a fixed system of rules. Instead, the mental lexicon contains a detailed inventarization of stored morpheme sequences and the generalized combination schemes (c.q. redundancy rules) that can be derived from them by stem abstraction. In addition, there is a special algorithm for the extraction of productive combination schemes, which can be used for the construction and analyis of new words. Here I call a scheme productive if it can be applied to an open set of stems (which is defined in an *intensional* way), as opposed to lexical generalizations, which only apply on the lexical domain these schemes have been derived from (i.e. their *extensional* definition). This study only introduces the basic ideas behind these algorithms. The definition of concrete algorithms requires further study.
- there are at least three domains of morphological structure building, i.e. the domains of morpheme sequences, lexeme sequences and word sequences. Each morphological domain is closed by a boundary operator, which may be explicitly indicated, e.g. by inflection features. In this approach, words always are morphologically complex, even if this is not indicated in a phonological way, as each word minimally consists of a lexeme

and a word boundary. In the same way, each lexeme minimally consists of a morphological root and a lexeme boundary, which reflects the idea that all affix morphology is root based (paradigmatic) instead of lexeme based (syntagmatic). In this respect, native and non-native lexemes thus are assumed to behave similarly. In the present version of the L-KRING-theory, each set of domain boundaries is expressed in terms of domain specific variants of the traditional main categories (N, V and A); e.g. #v denotes a V-root, $v a V-lexeme (still available for compound usage) and V a "syntactic" word unit with inflectional category V. This hierarchical classification system clearly is in need of a more differentiated set of sematically and phonologically motivated distribution categories. For a first exploration of the consequences, this system nevertheless suffices.

- In the L-KRING theory, morphemes are not identified in a syntagmatic, but in a paradigmatic way. I motivate this choice by discussing a number of fundamental problems with the syntagmatic approach, which only takes into account the linear selection properties, as opposed to the paradigmatic approach, which relates these selection properties to the whole paradigm of derivation options (similar to the paradigms of arabic word stems).
- This whole approach is made possible by the central assumption that the mental lexicon may store all words that a language user ever encounters, and that each stored word reflects its internal structure by means of index based reference relations to its lexical constituents. This fundamental property of the lexicon opens a fascinating world of unsupervised analysis options.

Lexicographic results

- The realization of the MGBN demonstrates that it is possible to enrich a complete lexicon with morphological structure representations by applying a semi-automatic, intuition based annotation method, without the use of a predefined rule system. This labor-intensive method only required two man years of annotation work.
- As part of this study, a number of data reports is discussed with very detailed information about the morphological properties of the MGBN lexicon. These statistical reports analyse the word internal morpheme patterns by differentiating these patterns with respect to unit type (e.g. roots, prefixes and suffixes), structural position, sequence length, morphological class features and combinatorial class features. In addition, one can inspect statistical information about a number of frequency measures for each analysed unit, among which a stem based type frequency and a paradigm based type frequency.
- To facilitate the validation of the database, the data reports also provide information about the linguistic status of the analysed units; for this purpose, all units are compared with the morpheme information in the Morphological Handbook of Dutch (MHB) of De Haas & Trommelen (1993), again differentiating the forms for the properties listed above. This detailed comparison information is directly available, so that one can easily analyse the differences between the MGBN and the MHB.
- This study claims that the MGBN covers all form information in the MHB, and that it extends the set of known morphemes with a significant amount of new units (both at the form level and at the combinatorial level). Therefore, one can conclude that the MGBN clearly improves the MHB with respect to the the coverage of the Dutch morphology, both with respect to the number of described affixes as with respect to the information about their frequency, their combinatorial patterns and their lexem extension (i.e. the set of lexemes in which the affixes and stems are used).

Below the present state of the MGBN is characterized by means of statistical facts about its size and the coverage of the Morphological Handbook of Dutch (MHB):

- The MGBN contains 80.000 non-compound lexemes, which consist of 19.000 etymological roots and 1000 uniqe affixes, among which 250 prefixes and 750 suffixes. For the prefixes one can find 950 different sequences, for the suffixes even 3750 different sequences. As for their combination, a total of 7500 different prefix-suffix-patterns can be found, among which 4550 of category noun, 950 of category verb, 1900 of category adjective and 150 of category adverb. Of all 80.000 basic lexemes, 16000 only consisted of a root.
- The MHB based evaluation learns that all affixes and affix sequences in the MHB are covered by the MGBN (although a few are retrieved only indirectly). The reverse evaluation learns that only 30% to 40% of the MGBN suffixes is covered by the MHB (given a type frequency of 10 lexeme applications or more), and 60% of the prefixes. Further the MHB provides only mentions a few dozens of affix sequences while the MGBN covers several thousands of affix sequences, as specified above.
- As part of the evaluation I also looked at the statistical distribution of the affixes, both within the MGBN and with respect to the MHB. These internal evaluations provided evidence that the MGBN has a non-random distribution, and that the more frequent patterns have a better chance to be covered by the MHB. Integrating all these facts, one can conclude that the MGBN is a large and reliable data source about Dutch morphology.

The realization of the MGBN proves the cognitive plausability of the lexicological framework presented in this study. According to this framework, the MGBN itself may be a valuable tool for the construction of a completer L-KRING-model of the Dutch lexicon, while the development of this model may in turn lead to a better quality of the data in the MGBN: their development thus can proceed in parallel. In addition, the MGBN might become a valuable data source for psycholinguistic and neurological experiments, and for the devlopment of better tools for the automatic analysis and synthesis of the Dutch language.