INL SCHATKAMER VAN
DE NEDERLANDSE TAAL

Achtergronden bij de morfologische module van GiGaNT

Folgert B. Karsdorp

INL Working Papers - Taalbank Nederlands 2
Leiden, 2010

www.inl.nl

Achtergronden bij de morfologische module van GiGaNT

Folgert B. Karsdorp

Taalbank Nederlands, Instituut voor Nederlandse Lexicologie,
Matthias de Vrieshof 2-3, 2311 BZ Leiden

Abstract

This document describes the backgrounds of the morphological component of the GiGaNT lexicon. GiGaNT is a computational diachronic lexicon of Dutch which covers a period of over fourteen centuries. The lexicon is corpus-based which means that all entries in the lexicon are linked to attested language material. The benefits of a morphological component in GiGaNT are potentially numerous. For one, no diachronic lexicon for Dutch exists that contains extensive morphological information. GiGaNT thus paves the way for a whole new area of linguistic research. Moreover, the construction of a morphologically annotated diachronic lexicon is of vital importance to the development of linguistic tools that are able to analyze diachronic language material (like part of speech tagging and morphological segmentation) that has not been (extensively) investigated before.

Keywords: Computational lexicon, morphology, GiGaNT, diachronic

1. Inleiding

Binnen de afdeling Taalbank Nederlands van het Instituut voor Nederlandse Lexicologie (INL) wordt er gewerkt aan het 'Groot Geïntegreerd Lexicon van de Nederlandse Taal' (voortaan GiGaNT). GiGaNT is een computationeel lexicon van het Nederlands van de zesde eeuw tot het hedendaags Nederlands. Het lexicon heeft de vorm van een verzameling woorden, woordgroepen en woorddelen gekoppeld aan een modern Nederlands lemma. Alle geattesteerde woordvormen van een lemma (vorm- maar ook spellingvarianten) worden opgenomen in het lexicon. Elke ingang van het lexicon is gekoppeld aan een rijke bron van taalkundige informatie, zoals woordsoort (Part of Speech) en morfologische analyse.

Het lexicon is corpusgebaseerd. Dat betekent dat alle ingangen gekoppeld zijn aan daadwerkelijk taalmateriaal. De koppeling van het lexicon aan bewijsplaatsen verrijkt het lexicon in verschillende opzichten: (1) attestatie-informatie,

Email: folgert.karsdorp@inl.nl (Folgert B. Karsdorp)

(2) er kunnen frequentiegegevens bijgehouden worden en (3) de woorden kunnen voorzien worden van daterings- en lokaliseringsinformatie.

Het lexicon heeft een modulaire opbouw waarin de combinatie WOORDVORM, LEMMA en PART-OF-SPEECH de kern vormt. Aan deze kern kunnen complementair nieuwe modules worden toegevoegd met uitgebreidere informatie over de woorden.

In dit stuk beschrijf ik de toevoeging van een morfologische module aan GiGaNT. Het belang van een morfologische component in GiGaNT is groot. In de eerste plaats bestaat er voor het Nederlands nog geen morfologisch geannoteerd, diachroon lexicon, laat staan een corpus-gebaseerd diachroon morfologisch lexicon. Onderzoekers die onderzoek doen naar morfologische ontwikkelingen in het Nederlands (zie bijvoorbeeld Hüning en Van Santen 1994; Hüning 1999; Tálasi 2009), moeten het nu stellen met de schaarse morfologische informatie uit de historische woordenboeken (zoals het WNT, MNW en VMNW). De morfologische analyses in met name het WNT, zijn echter erg summier en bovendien ontbreken er vele morfologische procedés. Belangrijker is nog dat de woordenboeken niet bedoeld zijn voor morfologisch onderzoek en dat de morfoloog dus ook niet of slechts beperkt op morfologische elementen kan zoeken. Er bestaat bijvoorbeeld geen mogelijkheid om in het WNT alle adjectief+nomen-samenstellingen (van het type *groothandel, blauwdruk*) op te vragen. Een laatste bezwaar tegen de bestaande diachrone morfologische bronnen is dat de informatie uit de verschillende eeuwen verspreid is over diverse woordenboeken die bovendien elk hun eigen manier van coderen hebben. Hierdoor is het moeilijk zo niet onmogelijk om de ontwikkeling van morfologische processen door de eeuwen heen te kunnen volgen.

De vraag om een morfologisch-historische databank is dan ook groot onder morfologen van het Nederlands. De vraag is duidelijk verwoord door Van der Sijs en Van Santen (2006) die een concreet voorstel doen voor een historisch-morfologische database van het Nederlands, waarbij bovendien een beroep wordt gedaan op het INL:

> Voorzover ons bekend bestaat er voor geen enkele taal een historisch-morfologische database. Hier lijkt een mooie en oorspronkelijke taak weggelegd voor het INL, waar immers historische en etymologische woordenboeken zijn samengebracht (Van der Sijs en Van Santen 2006, 166).

De ontwikkeling van een morfologische component binnen GiGaNT is niet alleen belangrijk voor descriptieve en theoretische vragen over de ontwikkeling van de Nederlandse morfologie, maar kan ook ingezet worden bij de ontwikkeling van taalkundige gereedschappen. We moeten hierbij denken aan taken binnen de discipline van *Natural Language Processing* zoals Part of Speech-tagging, woordsegmentatie (het opsplitsen van een woord in de woorddelen), automatische mor-

fologische analyse en lemmatisering (zie bijvoorbeeld Kestemont et al. 2010). Dit gereedschap heeft met name het doel niet eerder of slechts summier bestudeerd taalmateriaal (verder) te ontsluiten.

Het ontwerp van het lexicon maakt GiGaNT bijzonder geschikt als bron voor informatie bij descriptieve en theoretische vraagstukken over de historische morfologie. Door de morfologische informatie uit verschillende eeuwen taalmateriaal te integreren in één centrale gegevensbank, ontstaat bijvoorbeeld de mogelijkheid de ontwikkeling van morfologische processen door de eeuwen heen te volgen. Met name voor studies naar taalverandering is dit van belang. Zonder een geïntegreerd morfologische lexicon is het moeilijk uitspraken te doen over taalkundige kwesties als grammaticalisatieprocessen of productiviteitsveranderingen van morfologische procedés. De morfologische module van GiGaNT probeert hier een oplossing voor te bieden.

De ontwikkeling van een morfologische component voor een lexicon dat meer dan 14 eeuwen taalmateriaal beslaat, is echter geen triviale opdracht. In dit stuk zal ik de achtergronden bespreken en de gemaakte keuzes verantwoorden bij de ontwikkeling van de morfologische module. Ik zal beginnen met de bespreking van de bestaande synchrone morfologische databanken van het Nederlands (§2). Vervolgens richt ik mij op de afbakening van het morfologische domein (§3). Hierbij komt de vraag aan bod wat het object van studie voor de morfologische component is. Vervolgens bespreek ik het ontwerp van de morfologische component (§4). In deze sectie komen ook de onderscheiden morfologische processen (§4.4.1) en het ontworpen codeerschema aan bod voor de hiërarchische segmentatie van de structuur van morfologisch complexe woorden (§4.4.3). In sectie 5 richt ik mij op een aantal kwesties die specifiek van belang zijn voor diachroon taalmateriaal. Tot slot zal ik in sectie 6 de technische kant bespreken van de morfologische module.

2. Overzicht bestaande morfologische databanken

Er bestaat geen historisch-morfologische databank van het Nederlands. Voor het hedendaags Nederlands zijn er drie databanken waarin morfologische informatie is opgenomen: de Lexicale databank CELEX (Baayen et al. 1995), het lexicon van het *Corpus Gesproken Nederlands* en *e-lex*. Zowel *e-lex* als het CGN-lexicon zijn sterk gebaseerd op de databank CELEX. Aangezien de documentatie van de morfologische componenten binnen *e-lex* en het CGN-lexicon erg summier is, beperk ik mij in deze paragraaf tot een beschrijving van de databank CELEX.

De lexicale databank CELEX is de belangrijkste bron van informatie voor de Nederlandse morfologie. De Nederlandse component van de databank bestaat uit 381.292 unieke woordvormen en 124.136 lemmata.[1] De databank is gebaseerd op

[1] De CELEX bevat naast een Nederlandse ook een Engelse en een Duitse component.

verschillende bronnen van het hedendaags Nederlands, waaronder materiaal uit woordenboeken en materiaal uit diverse corpora, zoals het *50 miljoen woorden corpus* van het INL.

De databank beschikt over gedetailleerde informatie over bijvoorbeeld spelling, fonologie en woordsoort. Voor wat betreft de morfologie levert de databank informatie over derivatie en samenstelling. Verder zijn alle gelede lemmata voorzien van een gedetailleerde hiërarchische morfologische analyse.

De verrijking van de structurele informatie van gelede lemmata is halfautomatisch tot stand gekomen. De door de computer gegenereerde morfologische analyses, zijn met de hand gecontroleerd. De handmatige controle heeft van de CELEX een (over het algemeen) betrouwbare databank gemaakt voor morfologische informatie van het hedendaags Nederlands.

Zonder af te willen doen aan de waarde van de CELEX voor taalkundig onderzoek, noem ik een aantal belangrijke tekortkomingen van de databank voor wat betreft de morfologie. Allereerst vinden we in de CELEX een behoorlijk aantal fouten en inconsequenties in de morfologische annotatie. Laureys et al. (2004) laten zien dat ongeveer 6.2% (7646 lemmata) verkeerd geanalyseerd is. Daarnaast zijn veel gelede woorden niet geanalyseerd. Enkele voorbeelden zijn *aardewerk*, *socialist* en *absurdisme*. Ook is er een reeks inconsistenties in de analyses. Een voorbeeld daarvan vormt de analyse van *blauwzijden, roodzijden* en *witzijden*. Deze woorden hebben dezelfde morfologische structuur, maar zijn allemaal verschillend geannoteerd:[2]

(1) (a) `((blauw)[A],(zijde)[N],(en)[A|AN.])[A]`
(b) `((rood)[A],((zijde)[N],(en)[A|N.])[A])[A]`
(c) `(witzijden)[A]`

Ook zijn er inconsistenties wat betreft het niveau van analyse bij tweedegraadsafleidingen. Een vorm als *analyseren* wordt onderverdeeld in *analyse* en *-eren*. De analyse van de complexere afleiding *analyseerbaar* erft deze onderverdeling en voegt er het affix *-baar* aan toe: `(((analyse)[N],(eer)[V|N.]),(baar)[A|V.])[A]`. De afleiding *manoeuvreerbaar* wordt echter geanalyseerd als `((manoeuvreer)[V],(baar)[A|V.])[A]`, waarbij *manoeuvreer* geen verdere analyse kent.

Voorbeelden van echte foutieve analyses zijn *papier*, *buitenspeelster* en *kwikkuur*:

(2) (a) `((paap)[N],(ier)[N|N.])[N]` (onterecht suffix)
(b) `((buiten)[N],(speel)[V],(ster)[N|NV.])[N]` (verkeerde woordsoort *buiten*)

[2] Voor een toelichting bij deze codering, zie §4.4.3

(c) `((kwik)[N],(uur)[N|N.])[N]` (verkeerde segmentatie)

Daarnaast zijn er ook een aantal analyses van vrij dubieuze aard, zoals die van *nazi*:

(3) `((((natie)[N],(ionaal)[A|N.])[A],(socialist)[N])[N])[N]`

Nazi is inderdaad een (Duitse) afkorting van *nationaal-socialist*, maar dat heeft niets met de morfologische structuur van *nazi* te maken.

Een tweede belangrijke tekortkoming van de databank CELEX is dat alle vormen met een frequentie van één (de zogenoemde *hapax legomena*) zijn weggelaten uit de databank. Dit is met name problematisch voor onderzoek dat gericht is op neologismen of op morfologische productiviteit. Ongeveer 5000 (4%) lemmata die wel in de INL-corpora voorkomen, ontbreken in de databank (Baayen 1991, 227).

Een foutloze databank is uiteraard een utopie. Toch streven we er binnen GiGaNT naar een databank te ontwikkelen met een lager foutpercentage dan de CELEX. Ook in GiGaNT zal er gewerkt worden met een automatische voorbewerking van de morfologische analyses. Daarnaast zullen alle automatisch gegenereerde analyses handmatig worden gecontroleerd. Om een zo hoog mogelijke annotatiekwaliteit te behalen, is er geëxperimenteerd met verschillende codeerschema's (zie Karsdorp 2010). De evaluatie van deze schema's geeft ons een redelijk zeker beeld dat ons streven om een kwalitatief betere morfologische databank te ontwikkelen, binnen het bereik ligt.

3. De afbakening van het morfologische domein

In deze sectie zal ik de afbakening van het morfologische domein binnen GiGaNT bespreken. De afbakening van het morfologische domein valt uiteen in drie stappen. De morfologie houdt zich bezig met woorden. Daarom moeten we eerst definiëren wat we onder een woord verstaan. Het gaat in de morfologie echter primair om een bepaald type woorden, te weten gelede woorden. Deze twee onderwerpen komen aan bod in paragraaf 3.1. Ook moeten we bepalen welke woordvormingsprocessen we tot de morfologie rekenen. Dat zal ik doen in paragraaf 3.2.

3.1. De afbakening van het woord

Het onderzoeksobject van de morfologie is de interne structuur van gelede woorden. Om tot een afbakening te komen van het morfologische domein, moeten we eerst definiëren wat we onder een woord en vervolgens wat we onder een geleed woord verstaan.

Onder een woord verstaan we taaleenheden met een vorm en een betekenis die zelfstandig functioneren. Met zelfstandig bedoelen we dat een woord zelfstandig kan voorkomen binnen een syntactische context en dat het geïsoleerd en verplaatst, (Van den Toorn 1975, 133) maar niet intern gescheiden kan worden (De Haas en Trommelen 1993, 3). Dit verbod op interne scheidbaarheid onderscheidt woorden van woordgroepen. Een problematische categorie vormt in dit verband de categorie samenkoppelingen of partikelwerkwoorden, zoals *opbellen* en *wegschrijven*. Deze werkwoorden kunnen namelijk wel intern gescheiden worden (*Hij belde haar op.*). Hoewel deze vormen een reeks eigenschappen delen met woordgroepen, vatten we ze op als woorden, omdat de delen weliswaar niet vormelijk maar wel semantische een eenheid vormen. Dit criterium biedt echter niet altijd steun bij de identificatie van samenkoppelingen. Zeker wat betreft diachroon taalmateriaal is het vaak moeilijk te bepalen of een combinatie van woorden als zelfstandige eenheid opereert (als een woord) of als woordgroep. De intuïties over de status van een woordgroep kunnen uiteenlopen en niet alle syntactische tests zijn (bijvoorbeeld plaatsing in de eindgroep en ontkenning van de samenkoppeling) op alle gevallen van toepassing.

Onder een GELEED WOORD verstaan we die woorden waarin een bepaalde structuur in de vorm correspondeert met een structuur in de syntactische functie en/ of betekenis. Dit in tegenstelling tot ONGELEDE woorden waarbij een dergelijke relatie niet aan te wijzen is. De vorm-functie-correspondenties binnen gelede woorden komen tot stand in de relatie tussen gelede woorden en de daarmee corresponderende minder gelede woorden, en in de relatie met andere gelede woorden (zie o.a. Schultink 1962; Booij en Van Santen 1998; Hüning 1999). Een woord als *groenig* bijvoorbeeld, gaat op dezelfde manier een relatie aan met *groen* als *lawaai* met *lawaaiig*. Door de vergelijking krijgt de vormovereenkomst (*-ig*) betekenis.

Deze paradigmatische benadering maakt ook dat we woorden die geen ongeleed correlaat (meer) hebben, kunnen opnemen in de categorie gelede woorden. Een woord als *vadsig* bijvoorbeeld, kent geen basiswoord **vads*. De betekenisovereenkomst met andere woorden op *-ig* echter, maakt dat we toch een vormelijke structuur met een daarmee corresponderend betekenismoment kunnen identificeren. Ook een woord als *bedriegen* waarvan geen ongeleed correlaat bestaat in het hedendaags Nederlands, kunnen we in deze benadering als geleed beschouwen. *Bedriegen* gedraagt zich namelijk syntactisch en morfologisch hetzelfde als alle andere werkwoorden met het prefix *be-*: het is transitief en het mist het prefix *ge-* in het deelwoord (*bedriegen – bedrogen*).

Het morfologische domein van GiGaNT kunnen we nu als volgt definiëren:

Definitie 1 (Morfologische domein). De morfologische module van GiGaNT beslaat alle woorden die een correspondentie vertonen tussen een bepaalde structuur in de vorm en een structuur in de syntactische functie en/ of betekenis,

hetzij in vergelijking met minder gelede hetzij in vergelijking met andere gelede woorden.

3.2. De afbakening van woordvormingsprocessen

De morfologische component van GiGaNT bestaat uit alle woordvormingsprocessen behalve flexie: derivatie, samenstelling, samenstellende afleiding etc. Flexie is geplaatst onder het domein van de morfosyntaxis waarin de paradigmata worden besproken. Het onderscheid tussen flexie en derivatie is echter niet altijd even duidelijk. Daarom zal ik hieronder het onderscheid vrij uitvoerig bespreken om tot een werkdefinitie te komen.

De Haas en Trommelen (1993) maken een strikt onderscheid tussen flexie en derivatie. Als logisch gevolg hiervan gebruiken ze voor flexie-affixen de term UITGANG en voor derivatie-affixen de term AFFIX. Het belangrijkste onderscheid tussen flexie en derivatie dat De Haas en Trommelen (1993, 8) noemen is distributioneel van aard: "[...] flexionele uitgangen zijn perifeer ten opzichte van derivationele affixen." Woorden als *dank-baar-e* en *voogd-es-en* zijn welgevormd in tegenstelling tot *dank-e-baar* en *voogd-en-es*. Andere in de literatuur genoemde verschillen tussen flexie en derivatie zijn (zie Booij en Van Santen 1998, 111):

1. derivatie is niet verplicht, flexie wel;
2. flexie is per definitie niet categorie-bepalend, terwijl derivatie dat wel kan zijn;
3. flexie heeft een hogere productiviteitsgraad dan derivatie;
4. afleidingen kenmerken zich doorgaans door een minder transparante betekenis dan flexie.

De genoemde verschillen tussen flexie en derivatie zijn echter niet absoluut van aard, maar gradueel. Het verschil wordt nog onduidelijker wanneer we een onderscheid aanbrengen tussen INHERENTE en CONTEXTUELE flexie. We spreken van contextuele flexie in die gevallen waarbij de syntactische context een bepaalde flexievorm domineert (denk aan naamvallen, verbuiging van het adjectief in attributieve positie en werkwoordsflexie). Alle overige typen zijn voorbeelden van inherente flexie. We moeten daarbij denken aan de keuze tussen een enkelvouds- of een meervoudsvorm van een nomen of aan de vorming van de vergelijkende of overtreffende trap van een adjectief: dergelijke keuzes worden niet bepaald door de syntactische context waarin ze gebruikt worden.

Met name inherente flexie en derivatie lijken veel op elkaar. Neem de vorming van de superlatief in het Nederlands, die in de literatuur als inherente flexie wordt beschouwd. Zoals gezegd is flexie nooit categorie-bepalend. Interessant is nu dat de superlatief in een enkel geval wel kan zorgen voor transpositie van de ene naar de andere woordsoort, zoals in het geval van *bovenste* en *buitenste*. *Boven* komt

los alleen voor als bijwoord of als prepositie, maar niet als adjectief (Booij en Van Santen 1998).

Een in de literatuur genoemd kenmerk van flexie is dat ieder woord van een bepaalde woordklasse een flexievorm heeft. Ook dat lijkt niet het geval bij de overtreffende trap. Zo is een overtreffende trap van niet-gradeerbare adjectieven zeer onwaarschijnlijk:

(4) *uitklapbaarst, *doodst, *Amerikaanste, *politiekste

Merk ook op dat de morfologische superlatief in een aantal gevallen alterneert met een omschrijvende superlatief met *meest*, zoals in *de meest fantastische gebeurtenissen...* (zie voor een beschrijving van deze alternatie Karsdorp en Beekhuizen 2010). Interessant is nu dat de omschrijvende superlatief vaak beter samengaat met niet-gradeerbare adjectieven: "De meest Amerikaanse stad van Nederland" (*de Volkskrant* 21-11-01). Een dergelijke interactie is typisch voor derivatie, maar niet voor flexie.

Ook met de genoemde verschillen blijft de afbakening van flexie en derivatie problematisch. Booij en Van Santen (1998, 83) geven de volgende definitie van flexie:

> [O]nder flexie verstaan we het geheel van morfologische relaties tussen woorden die als vormen van één lexeem worden beschouwd.

Bij derivatie hebben we te maken met verschillende lexemen, bijvoorbeeld *man* en *mannelijk* terwijl het bij flexie gaat om verschillende vormen van één lexeem. *Man* en *mannen*, *klein* en *kleine* behoren tot hetzelfde lexeem en dus is er sprake van flexie. Volgens deze regel vallen ook de verschillende verschijningsvormen van een werkwoord onder flexie. Bij *man* en *mannelijk*, *slons* en *verslons* daarentegen hebben we te maken met verschillende lexemen en daarom spreken we van derivatie (zie Booij en Van Santen 1998, 8).

Het onderscheid tussen flexie en derivatie kunnen we nu als volgt definiëren:

Definitie 2 (Flexie tegenover derivatie). Onder derivatie verstaan we het proces waarbij affigering leidt tot een nieuw lexeem. Onder flexie verstaan we die gevallen waarbij affigering niet leidt tot een nieuw lexeem, maar tot een nieuwe verschijningsvorm van hetzelfde lexeem. Een lexeem is gedefinieerd als een representatie waarbij geabstraheerd is van verschillen in vorm, betekenis en syntactische valentie.

Volgens deze opvatting moeten we de verbindingsklanken /ə/ of /s/ die kunnen optreden tussen twee leden van een samenstelling ook opvatten als flexie:

er bestaan alternerende vormen van een samenstellingslexeem met verschillende verbindingsklanken (vergelijk: *schaapskooi* tegenover *schapenkooi*).[3]

Voor de morfologische module binnen GiGaNT houden we de hierboven geformuleerde definitie aan. Zoals in paragraaf 4 aan bod komt, zal flexie ondergebracht worden bij de woordvormen en derivatie bij de lemmata.

4. Opbouw morfologische module

4.1. Basisonderdelen

De kern van GiGaNT wordt gevormd door de combinatie van WOORDVORM, LEMMA en PART-OF-SPEECH. Logische plaatsen om deze kern met morfologische informatie te verrijken zijn het lemma en de woordvorm. Als we de morfologische gegevens direct koppelen aan de woordvorm, moeten we echter vaak dezelfde informatie opnemen. Daarom is ervoor gekozen om de morfologische informatie direct te koppelen aan het lemma. Omdat het lemma en de woordvorm met elkaar verbonden zijn, is de morfologische informatie ook indirect verbonden met de woordvorm. Flexie wordt wel direct gekoppeld aan de woordvorm, aangezien flexie alleen op dat niveau een variabele is.

Hieronder zal ik de verschillende basisonderdelen bespreken van de morfologische module in GiGaNT. Vanzelfsprekend is deze structuur bedoeld voor zowel synchrone als diachrone morfologie. De diachrone morfologie verdient echter wat extra aandacht. Daarom zal ik in sectie 5 een aantal punten van de hier voorgestelde structuur nader beschrijven die met name van belang zijn voor historisch materiaal. Ik zal nu eerst een beschrijving geven van de derivationele en compositionele informatie (§4.1.1). Hierna zal ik in paragraaf 4.1.2 kort de flexie-informatie binnen GiGaNT bespreken. In paragraaf 4.2 komt de gebruikte tag-set aan bod. Paragraaf 4.3 behandelt de toegepaste methoden voor de hiërarchische segmentatie van de morfologische structuur. Ik sluit deze sectie af met een opsomming van een aantal extra verrijkingen van de morfologische module (§4.4).

[3] In de literatuur vinden we de opvatting waarin de verbindingsklanken worden geïdentificeerd met de meervoudsuitgangen *-s* en *-en* (zie bijvoorbeeld Neijt en Schreuder 2009). Een *-s* in *bloemetjesjurk* bijvoorbeeld, duidt in deze opvatting op de aanwezigheid van meerdere bloemen op een jurk. Volgens deze visie moeten we bindfonemen opvatten als voorbeelden van flexie.

Booij en Van Santen (1998, 158) stellen dat de bindfonemen niet met de meervoudsuitgangen geïdentificeerd kunnen worden. Als argument brengen ze naar voren dat de verbindingsklank in samenstellingen niet systematisch een meervoudsinterpretatie oplevert. *Dagjesmensen* gaan typisch slechts één dagje eropuit en het *torentjesoverleg* vindt plaats in slechts één torentje. Hoewel een meervoudsinterpretatie van het bindfoneem in een aantal samenstellingen niet onwaarschijnlijk lijkt, vatten we bindfonemen binnen GiGaNT omwille van consistentie niet op als meervoudsuitgangen.

4.1.1. Derivationele en compositionele elementen van het lemma

Een eenvoudige manier om informatie over derivationele en compositionele elementen van het lemma te verkrijgen is via de hiërarchische segmentatie van de morfologische structuur. Elke (vormelijk gelede) lemmavorm kunnen we voorzien van een morfologische structuur zoals onder (5):[4]

(5) (a) *huisdeur*: `(NOU (NOU huis) (NOU deur))`

(b) *versnelling*: `(NOU (VRB (AFF ver) (ADJ snel)) (AFF ing))`

(c) *blauwogig*: `(ADJ (ADJ blauw) (NOU oog) (AFF ig))`

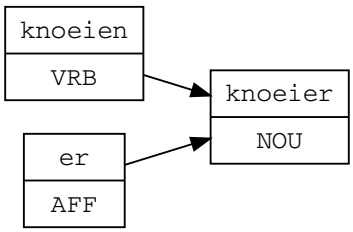
Voor een computationeel lexicon zijn morfologische structuren in dit formaat weinig functioneel. Voor een lexicon als GiGaNT is het functioneler om de morfologische structuur van het lemma te ontsluiten door de afzonderlijke delen van een geleed woord te koppelen aan hun corresponderende ingangen in het lexicon. Op deze manier wordt de mogelijkheid gecreëerd om op recursieve wijze informatie van de samenstellende delen te gebruiken (bijvoorbeeld bij zoekopdrachten naar morfologisch complexe woorden). Het voornaamste voordeel van deze structuur, is dat er zowel syntagmatische verbanden (binnen het woord) als paradigmatische verbanden (tussen woorden) beschikbaar zijn. Deze structuur verschilt wezenlijk van die van de databank CELEX waar alleen syntagmatische verbanden aanwezig zijn.

Laat me deze structuur verduidelijken aan de hand van een voorbeeld. Een woord als *knoeier* heeft de volgende morfologische structuur: `(NOU (VRB knoei) (AFF er))`. De afzonderlijke delen (*knoei* en *-er*) zijn gekoppeld aan de ingangen *knoeien* en *-er* in het lexicon (zie Figuur 1). Deze delen zijn op hun beurt weer gekoppeld aan een PART-OF-SPEECH-tag en eventueel verder onderverdeelbare morfemen. Dit proces herhaalt zich tot het niveau waarop nog enkel atomaire morfemen te herkennen zijn.

Door nieuwe woorden aan het netwerk in Figuur 1 toe te voegen ontstaan er naast syntagmatische verbanden tussen de delen van een woord ook paradigmatische verbanden tussen dat woord en andere woorden. Uiteraard geldt hetzelfde voor de samenstellende delen binnen het woord. Figuur 2 geeft een voorbeeld van de paradigmatische verbanden die in de databasestructuur aanwezig zijn.

Met deze structuur ontstaan relaties tussen alle woorden die het element `(AFF er)` bevatten (*knoeier, zwammer, kletser*). Ook kunnen we een paradigmatische zoekopdracht formuleren waarmee we de familie van *knoeien* opvragen (*verknoeien, knoeierij, knoeiboel, knoeiwerk*). Naast relaties tussen lemmata, ontstaan er ook verbanden tussen PART-OF-SPEECH-tags en daarmee tussen woordstructuren. Zo kunnen we bijvoorbeeld zoeken naar alle zelfstandige naamwoor-

[4] Voor een verantwoording van de codering, zie paragraaf 4.4.3



Figuur 1: Voorbeeld morfologische structuur *knoeier*.

den met een werkwoord en een zelfstandig naamwoord als samenstellende delen, zoals *prutswerk, knoeiboel* en *knoeiwerk*.

De basis van de morfologische module binnen GiGaNT kan met deze structuur worden gelegd. De structuur biedt de mogelijkheid om talloze verbanden die impliciet in de structuur verborgen liggen te expliciteren. Hierop zal ik nader ingaan in paragraaf 4.4.

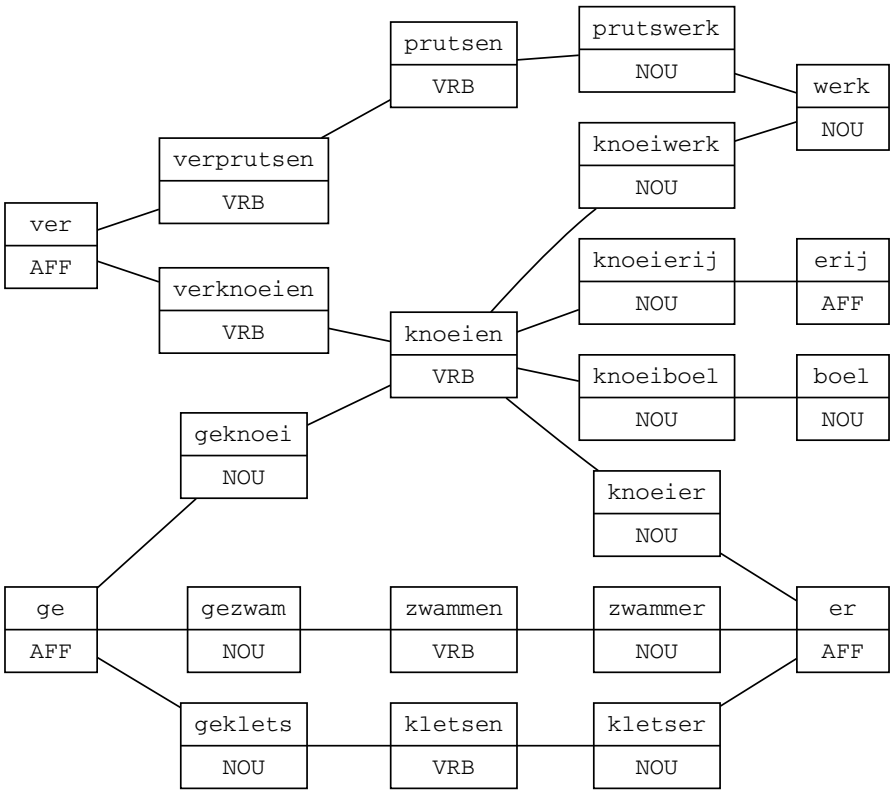
4.1.2. Flexie-elementen van de woordvorm

In het lexicon is de WOORDVORM gekoppeld aan attestatie-informatie, waaronder het citaat waarin deze woordvorm voorkomt. Aangezien flexie alleen in een syntactische context plaatsvindt, wordt de flexie-informatie van een woord gekoppeld aan de WOORDVORM. Voor elk woord wordt de flexievorm opgeslagen in een tabel met een bijbehorend label. Tabel 1 geeft een overzicht van de affixen die in de morfologische module van GiGaNT als flexie worden beschouwd.

Figuur 3 geeft een voorbeeld van de manier waarop flexie is gekoppeld aan de WOORDVORM.

4.2. De tag-set

De tag-set voor de morfologische module van GiGaNT is gebaseerd op de tag-set voor de morfosyntactische verrijking. In de morfosyntactische component van GiGaNT worden de volgende tags onderscheiden:



Figuur 2: Voorbeeld paradigmatische (en syntagmatische) verbanden binnen de databasestructuur.

	Inherente flexie	Contextuele flexie
N	*getal* (enkelvoud, meervoud) *verbindingsklank* (*-s-* of *-e(n)-*)	*naamval*
A	*trappen van vergelijking* (comparatief, superlatief)	*verbuiging in attributieve positie*
V	*tijd* (onvoltooid tegenwoordige tijd, onvoltooid verleden tijd, voltooid tegenwoordige/verleden tijd)	*getal* (enkelvoud, meervoud)
	modus (indicatief, conjunctief, imperatief)	*persoon* (1e, 2e, 3e)
	deelwoorden (voltooid, tegenwoordig)	
	infinitief	

Tabel 1: Overzicht onderscheiden flexiecategorieën per woordsoort binnen GiGaNT (naar Booij en Van Santen 1998, 84).

```
NOU = zelfstandig naamwoord  VRB = werkwoord  ADP = voorzetsel
AA  = Adjectief/bijwoord     ADV = bijwoord   NUM = telwoord
ADJ = bijvoeglijk naamwoord  CON = voegwoord  RES = residual
PRO = voornaamwoord          AFF = affix      DET = lidwoord
INT = tussenwerpsel
```

De morfologische module voegt daar nog de volgende tags aan toe:

```
* = deel van discontinu affix  BRM = gebonden basiswoord
```

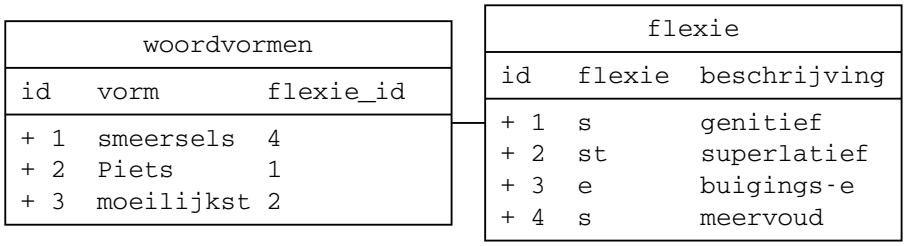
We spreken van een discontinu affix als twee affixen alleen in combinatie voorkomen (bijvoorbeeld *ge-* en *-te* in *gebergte*: zowel **bergte* als **geberg* zijn niet mogelijk).

In de lexicale databank Celex worden afleidingen die geen bestaand basiswoord (meer) hebben, voorzien van een zogenoemd hypothetisch basiswoord. Zo gaat de Celex consequent uit van een proces van affixsubstitutie bij affigering met bepaalde uitheemse suffixen, zoals *-atie*. Veel afleidingen met *-atie* hebben geen bestaand basiswoord en daarom wordt er in de analyse eerst een niet-gerealiseerd grondwoord aangenomen op basis waarvan de afleiding wordt gemaakt. Een voorbeeld is *acceptatie*, dat als volgt is gecodeerd:

(6) `(((accept)[N],(eer)[V|N.])[V],(atie)[N|V.])[N]`

Accepteren wordt afgeleid van het hypothetische nomen *accept* en vervolgens wordt het suffix *-eren* door affixsubstitutie vervangen door *-atie*.

De aanname van hypothetische basiswoorden is om meerdere redenen problematisch voor een lexicon als GiGaNT waarin meerdere eeuwen taalmateriaal



Figuur 3: Voorbeeld databasestructuur flexie-eigenschappen woordvormen.

bijeen zijn gebracht. In de eerste plaats is het voor oudere fasen van het Nederlands vaak moeilijk uit te maken, welk hypothetisch basiswoord aangenomen dient te worden. Daarnaast ligt ook het gevaar op de loer dat we een procesgerichte morfologische analyse geven, in plaats van een resultaatgerichte. Het probleem van een procesgerichte aanpak is dat het onduidelijk is tot waar (tot welk niveau van analyse of tot waar in de tijd) de analyse moet gaan. Het belangrijkste bezwaar is echter dat de opname van hypothetische basiswoorden het lexicon onvoorspelbaar maakt. We kunnen niet van de gebruiker verwachten dat h/zij weet bij welke woorden, welke hypothetische basiswoorden zijn aangenomen.

Vanwege deze problemen hanteren we binnen GiGaNT een 'WYSIWYG'-methode (What You See Is What You Get). Dit houdt in dat alleen vormelijke elementen die aanwezig zijn in het woordbeeld worden opgenomen in de morfologische analyse van een geleed woord. Basiswoorden die niet zelfstandig voorkomen, worden gemarkeerd met de tag `BRM` wat staat voor 'gebonden basiswoord' ('Bounded Root Morpheme'). *Acceptatie* wordt dan geanalyseerd als `(NOU (BRM accept) (AFF atie))`. Een voordeel van deze methode is dat de opvraag van woorden en woorddelen uit de morfologische analyse wordt vereenvoudigd.[5]

4.3. Hiërarchische segmentatie van de morfologische structuur

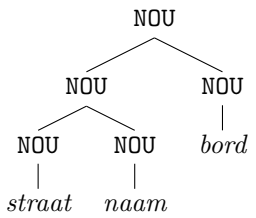
Voordat het lexicon gevuld kan worden met de samenstellende delen van gelede woorden, moeten de woorden eerst hiërarchisch worden gesegmenteerd. Bij

[5]Voor een uitgebreidere beschrijving van deze tag en de toepassing ervan op diachroon taalmateriaal, zie paragraaf 5.1.

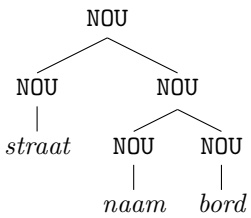
de meerderheid van morfologisch gelede woorden is segmentatie onproblematisch. Een nominale samenstelling als *keukenmes* kunnen we opdelen in de delen *keuken* en *mes*. Beide onderdelen zijn zelfstandige naamwoorden, waarbij de linker constituent een specificeerder is van de rechter. We kunnen dit weergeven in een hiërarchische structuur als `(NOU (NOU keuken) (NOU mes))`.

De segmentatie van woorden met een complexere geleding is echter niet altijd even eenduidig. Neem een woord als *straatnaambord*. Is een *straatnaambord* een 'bord met daarop een straatnaam' of 'een naambord van een straat'. De eerste betekenis heeft de segmentatie onder analyse (7a); de tweede als onder (7b):

(7) (a)



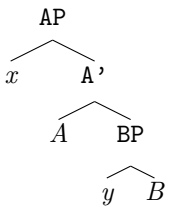
(b)



Gezien de omvang van GiGaNT is het van belang te onderzoeken of er gebruik gemaakt kan worden van een algemeen geldend onderverdeelprincipe. Een dergelijk principe zou bijvoorbeeld automatisch door een computer kunnen worden toegepast, waarmee de arbeidsintensiviteit aanzienlijk wordt verminderd. Daarnaast kan het ingezet worden om de consistentie van de handmatige morfologische analyses te bevorderen.

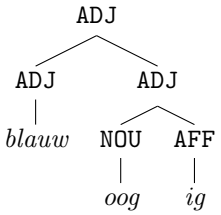
Voor de segmentatie van complexe(re) woorden zoals hierboven kan echter maar beperkt gebruik gemaakt worden van een generaliserend onderverdeelprincipe, zoals bijvoorbeeld onder (8) waarbij het principe van 'right branching' wordt toegepast:

(8)

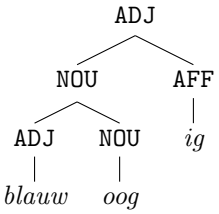


Een dergelijke onderverdeelstrategie is om verschillende redenen niet altijd toepasbaar op morfologische structuren. Zo zijn er verschillende voorbeelden van gelede woorden die niet binair kunnen worden onderverdeeld omdat dit resulteert in niet-bestaande basiswoorden, zoals onder (9a–b):

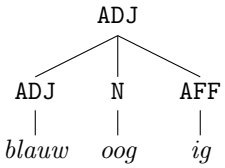
(9) (a)



(b)



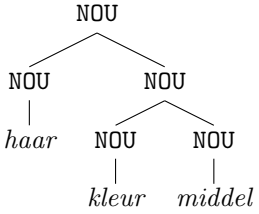
(c)



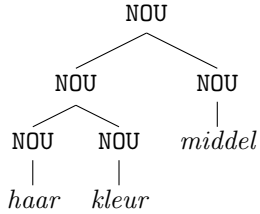
Zowel **ogig* als **blauwoog* zijn geen bestaande woorden in het Nederlands, die echter wel aangenomen dienen te worden als we een puur vormelijk segmentatieprincipe volgen. Omdat beide onderverdeelstrategieën een niet-bestaand basiswoord opleveren, is het beter om een drieledige analyse aan te nemen zoals in analyse (9c).

Ook de semantiek van een geleed woord kan een onderverdelingsschema tegenspreken. Zo is een *haarkleurmiddel* een 'kleurmiddel voor haar' (analyse 10a) en geen 'middel voor haarkleur' (om bijvoorbeeld de kleur bij te houden) zoals in analyse (10b).

(10) (a)

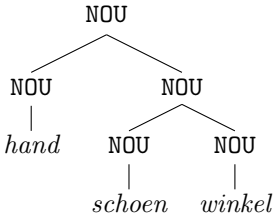


(b)

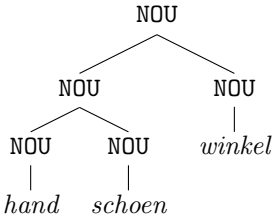


Maar niet alle woorden zijn rechtsvertakkend. Een *handschoenwinkel*, bijvoorbeeld, is geen 'schoenwinkel voor handen' (analyse 11a) maar een 'winkel waarin handschoenen worden verkocht' (analyse 11b).

(11) (a)



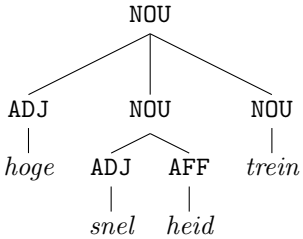
(b)



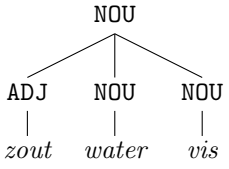
Wederom is de semantiek bepalend voor de vertakkingsstrategie.

Samenstellingen waarin een woordgroep voorkomt (zoals *hogesnelheidstrein* en *zoutwatervis*) worden vaak drieledig geanalyseerd, zoals onder (12):

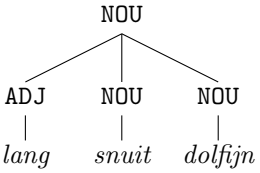
(12) (a)



(b)

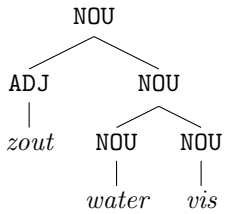


(c)

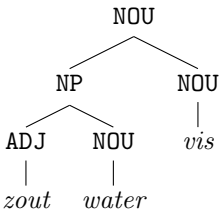


De reden hiervoor is dat in bijvoorbeeld *hogesnelheidstrein* noch *snelheidstrein* noch *hogesnelheid* bestaande woorden zijn. Toch is er wel wat voor te zeggen om een binaire structuur aan te brengen. Een *hogesnelheidstrein* is weliswaar geen 'snelheidstrein van hoog formaat' maar wel een 'trein die een hoge snelheid heeft' waarbij *hoge* en *snelheid* duidelijk een semantische eenheid vormen. Hetzelfde geldt voor *zoutwatervis* waar niet analyse (13a) maar analyse (13b) in overeenstemming is met de betekenis:

(13) (a)

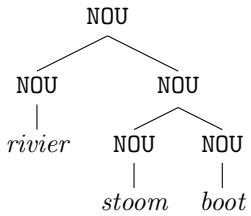


(b)

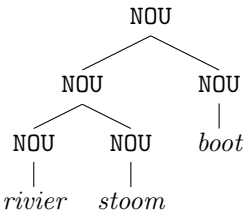


Een andere vertakkingsstrategie is segmentatie op basis van welke combinatie van delen van een samenstelling een bestaand basiswoord oplevert. Een samenstelling als *openluchtzwembad* kan opgesplitst worden in *openlucht* + *zwembad* en in *open* + *luchtzwembad*. Aangezien *luchtzwembad* geen bestaand basiswoord is en *openlucht* wel, wordt de strategie van 'left branching' toegepast, dus *openlucht* + *zwembad*. Volgens hetzelfde principe moet *rivierstoomboot* geanalyseerd worden als (14a) en niet als (14b):

(14) (a)



(b)



Als geen van de combinaties van de opeenvolgende samenstellende delen van een drieledige samenstelling een bestaand woord oplevert, zou gekozen kunnen worden voor een ternaire structuur (type *blauwogig*, *breedneuzig*).

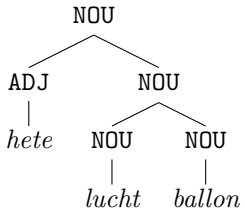
Deze methode is echter om verschillende redenen onwenselijk. In de eerste plaats zijn er voorbeelden van drieledige gelede woorden waarbij zowel de combinatie van de twee linker delen als de combinatie van de rechter delen een bestaand woord oplevert, zoals in tabel 2.

Een ander bezwaar op een dergelijke segmentatiemethode is dat het analyses oplevert die in strijd zijn met de betekenis van het woord. Neem een woord als *heteluchtballon*. *Hetelucht* komt niet los voor als samenstelling (alleen als woordgroep), *luchtballon* wel. Dit betekent dat we *heteluchtballon* moeten analyseren als in (15a):

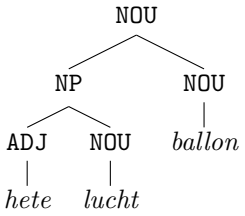
autobus+halte ⇔ auto+bushalte	avondmaal+tijd ⇔ avond+maaltijd
boekdruk+kunst ⇔ boek+drukkunst	brandweer+man ⇔ brand+weerman
diepvries+kast ⇔ diep+vrieskast	diepvries+kist ⇔ diep+vrieskist
eindexamen+cijfer ⇔ eind+examencijfer	eindexamen+klas ⇔ eind+examenklas
feestmaal+tijd ⇔ feest+maaltijd	grasmaai+machine ⇔ gras+maaimachine
grondwater+peil ⇔ grond+waterpeil	halfedel+steen ⇔ half+edelsteen
kantoorboek+handel ⇔ kantoor+boekhandel	kindbed+tijd ⇔ kind+bedtijd
zweefvlieg+tuig ⇔ zweef+vliegtuig	kunstijs+baan ⇔ kunst+ijsbaan
kunstmest+stof ⇔ kunst+meststof	landbouw+bedrijf ⇔ land+bouwbedrijf
landbouw+beleid ⇔ land+bouwbeleid	landbouw+grond ⇔ land+bouwgrond
landbouw+kunde ⇔ land+bouwkunde	leeftijd+genoot ⇔ leef+tijdgenoot
luchtpost+papier ⇔ lucht+postpapier	noordpool+cirkel ⇔ noord+poolcirkel
noordwesten+wind ⇔ noord+westenwind	overbuur+man ⇔ over+buurman
overbuur+vrouw ⇔ over+buurvrouw	schuldbewust+zijn ⇔ schuld+bewustzijn
slaapkamer+muur ⇔ slaap+kamermuur	sleutelbeen+breuk ⇔ sleutel+beenbreuk
steenkool+mijn ⇔ steen+koolmijn	straatnaam+bord ⇔ straat+naambord
strafwet+boek ⇔ straf+wetboek	tuinbouw+bedrijf ⇔ tuin+bouwbedrijf
voetbal+schoen ⇔ voet+balschoen	voetbal+spel ⇔ voet+balspel
waarborg+som ⇔ waar+borgsom	wijdwater+vat ⇔ wijd+watervat
zeeschip+vaart ⇔ zee+schipvaart	zuidwesten+wind ⇔ zuid+westenwind

Tabel 2: Voorbeelden van drieledige samenstellingen waarbij een linker en een rechter vertakking een bestaand woord oplevert.

(15) (a)



(b)



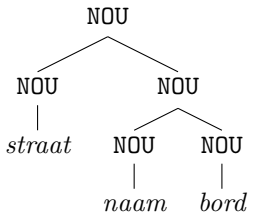
Maar hoewel het best voorstelbaar is dat de temperatuur in een heteluchtballon aan de hoge kant is, gaat het er bij een heteluchtballon om dat de ballon op hete lucht vaart. Bij deze interpretatie sluit een analyse als onder (15b) beter aan.

Het lijkt dus moeilijk om een algemeen geldend vertakkingsprincipe op te stellen waarmee alle woorden voorzien kunnen worden van een plausibele vormelijke en daarmee corresponderende semantische segmentatie. Daarom is er voor gekozen om bij de hiërarchische segmentatie van gelede woorden in GiGaNT de semantiek van de woorden als uitgangspunt te nemen.[6] Als voorbeeld neem ik de

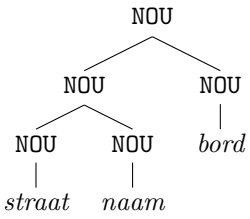
[6]Dit heeft echter geen gevolgen voor de manier waarop verbindingsklanken worden geanaly-

samenstelling waarmee ik de paragraaf ben begonnen: *straatnaambord*. *Straatnaambord* kunnen we op drie manieren onderverdelen:

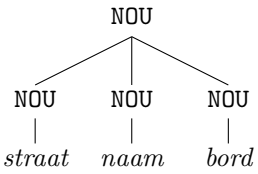
(16) (a)



(b)



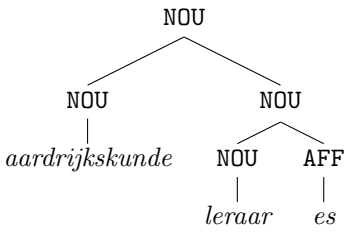
(c)



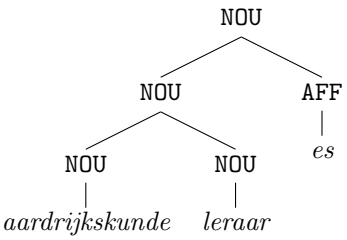
Omdat een *straatnaambord* niet zozeer 'een naambord is van een straat', maar eerder 'een bordje waarop de naam van een straat is aangebracht', kiezen we voor analyse (16b).

Helaas biedt deze methode vaak, maar niet altijd uitkomst. De voorbeelden die ik hier heb gegeven zijn allemaal samenstellingen waarbij een bepaalde vertakkingsstrategie gekozen dient te worden. Bij deze voorbeelden is het bovendien mogelijk om een keuze te maken. Anders ligt dat bij vormen die zowel een samenstelling als een afleiding kunnen zijn. Neem een woord als *hooglerares*. *Hooglerares* is evident een afleiding (op basis van *hoogleraar*), omdat het niet een bepaald type lerares is. In het geval van *aardrijkskundelerares* echter, is het minder eenduidig welke afleidingsrelatie gekozen dient te worden. Voor de betekenis maakt het niets uit of we *aardrijkskundelerares* onderverdelen als onder (17a) of als onder (17b):

(17) (a)



(b)



In dergelijke gevallen waarbij de semantiek geen uitsluitsel kan bieden wat betreft de hiërarchische segmentatie van de morfologische structuur, worden beide mogelijkheden opgenomen.

4.4. Extra morfologische verrijking

Naast de basisonderdelen is de morfologische module verrijkt met extra informatie, die ik hieronder zal bespreken.

seerd. Met uitzondering van de meervoudsvormen in samenstellingen met een woordgroep als linkerlid (denk aan *luiewijvensoep*), worden alle tussenklanken enkel opgevat als verbindingsklank en niet als meervoudsuitgang.

4.4.1. Morfologische processen

De morfologische processen waarmee woorden gevormd zijn, worden apart opgenomen in de morfologische module. Daarbij moeten we denken aan processen als woordsoortverandering door affigering en affixsubstitutie. Binnen GiGaNT worden de volgende morfologische processen onderscheiden:

Affigering: Bij het proces van AFFIGERING gaat het om de toevoeging van een affix aan een woord. Daarbij verandert soms maar niet altijd de woordsoort van het geheel (zie expliciete transpositie beneden). Enkele voorbeelden zijn gegeven onder (18):

(18) (a) `(AFF on)` + `(ADJ diep)` → `(ADJ ondiep)`;
(b) `(AFF be)` + `(NOU bos)` → `(VRB bebossen)`.

Affixsubstitutie: Vervanging van een affix door een ander affix noemen we AFFIXSUBSTITUTIE. Bij affixsubstitutie zijn de afleiding en het basiswoord in dezelfde mate geleed. Enkele voorbeelden zijn gegeven onder (19):

(19) (a) *veroveren* → *heroveren*;
(b) *ontstoppen* → *verstoppen*.

Impliciete transpositie: Gelijkvormige woorden die zich onderscheiden wat betreft betekenis en syntactische functie, zijn gevallen van IMPLICIETE TRANSPOSITIE. Impliciete transpositie is het proces waarbij de ene woordsoort overgaat in de andere zonder dat daarbij sprake is van een vormverschil. Dit proces wordt ook wel CONVERSIE of NULAFLEIDING genoemd. Enkele voorbeelden zijn gegeven onder (20):

(20) (a) `(NOU douche)` → `(VRB douche)`;
(b) `(ADJ gek)` → `(NOU gek)`.

Stamverandering: STAMVERANDERING is het proces waarbij woorden een semantische relatie met elkaar aangaan en deze relatie vormelijk ondersteund wordt door klinker- en/of medeklinkerwisseling. Dit proces is niet productief in het hedendaags Nederlands. Enkele voorbeelden zijn gegeven onder (21):

(21) (a) *spreken* → *spraken*;
(b) *spreken* → *spraak*;
(c) *klinken* → *klank*.

Compositie: Bij het proces van COMPOSITIE of samenstelling gaat het om het combineren van twee of meer vrije morfemen tot één geheel van vorm en betekenis. Enkele voorbeelden zijn gegeven onder (22):

(22) (a) `(NOU koek)` + `(VRB hap)` → `(VRB koekhappen)`;
(b) `(ADJ rood)` + `(NOU huid)` → `(NOU roodhuid)`;
(c) `(VRB slaap)` + `(VRB wandel)` → `(VRB slaapwandelen)`.

Samenkoppeling: Het woordvormingsprocedé SAMENKOPPELING lijkt sterk op het proces van compositie. Het verschil is dat de leden van een compositum niet en die van een samenkoppeling wel van elkaar gescheiden kunnen worden door morfologische of syntactische processen (zie De Haas en Trommelen 1993, 31). Enkele voorbeelden zijn gegeven onder (23):

(23) *pianospelen, koffiezetten, liefhebben, opbellen*

Samenstellende afleiding: We spreken van een SAMENSTELLENDE AFLEIDING in die gevallen waarbij het proces van affigering en compositie simultaan plaatsvindt. Enkele voorbeelden zijn:

(24) (a) *breedneuzig* (**breedneus*; **neuzig*);
(b) *tweetakker* (**tweetak*; **takker*);

Samenstellende samenstelling: SAMENSTELLENDE SAMENSTELLINGEN zijn samenstellingen waarbij de combinatie van het eerste en het tweede lid noch de combinatie van het tweede en het derde lid een bestaande samenstelling oplevert. Enkele voorbeelden zijn:

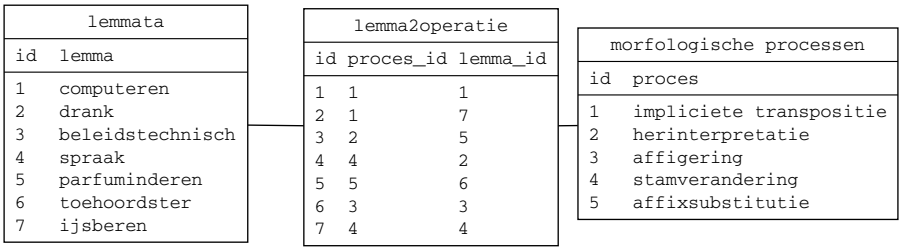
(25) *langsnuitdolfijn, tweepersoonsbed, dubbelloopsgeweer, driesterrenhotel, brandblusapparaat*

Splintercompositum SPLINTERCOMPOSITUM is het proces waarbij twee of meer vormen gecombineerd worden tot een geheel van vorm en betekenis en waarbij minimaal één vorm een gebonden inkorting is van een woord en/of de verschillende vormen elkaar overlappen (Meesters 2004, 106–107). Enkele voorbeelden zijn gegeven onder (26):

(26) (a) *relipop* uit *religieus* en *popmuziek*;
(b) *infotainment* uit *informatie* en *entertainment*;
(c) *omacipatie* uit *oma* en *emancipatie*;
(d) *consuminderen* uit *consumeren* en *minderen*.

Back-formation: Incidenteel komt het voor dat een meer geleed woord het basiswoord vormt voor een minder geleed woord. Dit proces noemen we BACK-FORMATION. Enkele voorbeelden zijn gegeven onder (27):

(27) (a) *stofzuiger* → *stofzuigen*;
(b) *onbesuisdheid* → *besuisdheid*;



Figuur 4: Databasestructuur morfologische operaties

(c) *onnozelheid* → *nozelheid*.[7]

De morfologische processen waardoor een geleed woord is ontstaan worden gekoppeld aan het corresponderende lemma van dat woord. In figuur 4 is deze koppeling schematisch weergegeven.

4.4.2. Morfologische familie

In verschillende morfologische studies is gewezen op het belang van dwarsverbanden tussen woorden (zie bijvoorbeeld De Jong et al. 2000; Krott et al. 2001). Men spreekt in dit verband ook wel over de familie van een woord. De familie van een woord of woorddeel bestaat uit alle woorden in het lexicon waarin dat woord voorkomt. *Rat* heeft bijvoorbeeld de volgende familie:

(28) *rattegezicht, rattehol, rattekop, rattengif, rattenkoning, rattenkruit, rattenplaag, rattenprobleem, rattenvanger, rattenverdelger, ratteval, beverrat, buidelrat, hotelrat, kerkrat, muskusrat, waterrat, woelrat, woestijnrat*

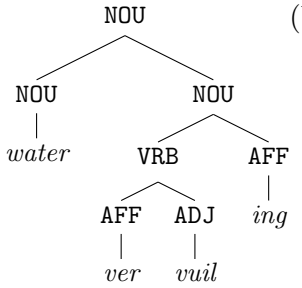
Deze familierelaties liggen in de in paragraaf 4.1 voorgestelde structuur impliciet besloten. Ze moeten alleen nog expliciet gemaakt worden en toegankelijk voor de gebruiker. Uit studies naar de invloed van woordfamilies op morfologische processen is vaak gebleken dat de familiegrootte de belangrijkste factor is.

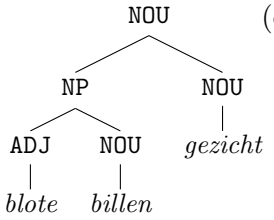
[7] *Nozelheid* is echter alleen vanuit synchroon perspectief en voorbeeld van back-formation. In het Middelnederlands komen *nozelheid* en *onnozelheid* naast elkaar voor en is het niet eenduidig vast te stellen welke vorm de basis heeft gevormd voor de ander.

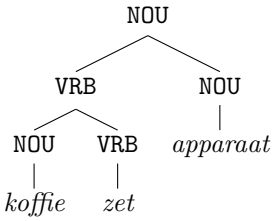
Daarom is naast de familierelaties van een woord ook de familiegrootte opgenomen.

4.4.3. Hiërarchische representatie van de morfologische structuur

Met de gelegde verbanden tussen de samenstellende delen van een geleed woord in de voorgestelde structuur is impliciet ook een meer traditionele hiërarchische representatie van de morfologische structuur van gelede woorden aanwezig, zoals in (29).

(29) (a) 

(b) 

(c) 

Voor een computationeel lexicon is een dergelijke representatie van de structuur in platte tekst niet erg functioneel, maar voor de eindgebruiker is een dergelijke representatie wel degelijk nuttig.

De morfologische structuur van een woord kan op verschillende manieren worden weergegeven. Het codeerschema van de databank CELEX is een (officieuze) standaard om morfologische structuren te representeren. Alle bestaande lexicale databanken van het Nederlands hanteren het codeerschema van de CELEX. Voor GiGaNT hebben we echter voor een ander codeerschema gekozen, dat meer (zoals hieronder duidelijk zal worden) in overeenstemming is met de principes van de morfologische module in GiGaNT. Voordat ik het codeerschema van GiGaNT presenteer, zal ik –om een context te scheppen– eerst het schema van de CELEX beschrijven.

Coderingschema in CELEX. In de CELEX is de representatie van de morfologische analyse telkens gebaseerd op de lemmavorm. Er zijn dan ook geen flexie-elementen van de woordvormen opgenomen, maar alleen derivationele en compositionele. Onder (30) zijn een aantal voorbeelden gegeven van de morfologische codering in de CELEX.

(30) (a) `((huis)[N],(deur)[N])[N]`
(b) `((meisje)[N],(s)[N|N.N],(boek)[N])[N]`
(c) `((on)[A|.A],((houd)[V],(baar)[A|V.])[A])[A]`

De hiërarchische segmentatie wordt weergegeven met behulp van ronde haakjes. Morfemen zijn gescheiden door komma's. Elk niveau van segmentatie (van

ondeelbare morfemen tot de gehele woordvorm) is voorzien van een woordsoort-aanduiding tussen rechte haakjes. De CELEX onderscheidt de volgende woordsoorten:

```
N = substantief    A = adjectief      Q = telwoord
V = werkwoord      D = lidwoord       O = voornaamwoord
B = bijwoord       P = voorzetsel     C = voegwoord
I = tussenwerpsel  X = restcategorie  . = affix
x = deel van discontinu affix
```

Gebonden morfemen (morfemen die alleen in combinatie met andere morfemen kunnen voorkomen) worden aangegeven door punten of de letter 'x' in het geval van een discontinu affix waarbij een punt wordt gebruikt voor het tweede lid (van het type *gebergte*: `((ge)[N|.Nx], (berg)[N],(te)[N|xN.])[N]`). De affixvalentie is weergegeven door een verticale streep. De woordsoorten achter de verticale streep staan voor de 'input' van het morfologische proces, de woordsoort vóór de streep voor de output-categorie. Een vorm als *onvindbaar* wordt als volgt geanalyseerd:

(31) `((on)[A|.A],((vind)[V],(baar)[A|V.])[A])[A]`

waarbij de achtervoeging van *-baar* leidt tot de transpositie van werkwoord naar adjectief (`[A|V.]`) en de aanhechting van *on-* geen verdere nieuwe woordsoort oplevert (`[A|.A]`).

Coderingsschema morfologische structuur GiGaNT. Voor GiGaNT is gekozen voor een resultaat-gebaseerde structuurrepresentatie van gelede woorden. Dat wil zeggen dat er in de representatie geen morfologische processen worden opgenomen die tot het eindproduct hebben geleid (zoals bijvoorbeeld de valentieregels in de databank CELEX). Dit is in overeenstemming met de in paragraaf 4.2 voorgestelde WYSIWYG-methode.

Op een aantal punten wijkt de structuurrepresentatie van gelede woorden binnen GiGaNT af van die van de CELEX. Ten eerste is gekozen om het scheiden van morfemen (in de CELEX door komma's) te laten samenvallen met het aanbrengen van hiërarchie (dus door middel van haakjes). Daarnaast is gekozen voor een simpelere haakjesstructuur die de representaties van de structuren visueel aantrekkelijker maar vooral inzichtelijker moet maken. Deze haakjesstructuur is gebaseerd op de syntaxis van de programmeertaal *Common Lisp*. In deze syntaxis wordt consequent gebruik gemaakt van een linker- en een rechterkant, waarbij de rechterkant een argument is bij de functie van de linkerkant. In het geval van de morfologische codering is de functie van de linkerkant het specificeren van de morfologische/ syntactische categorie van de rechterkant. Dit proces wordt herhaald tot op atomair morfeemniveau. Hieronder zijn enkele voorbeelden gegeven:

(32) (a)
```
(ADJ
    (AFF on)
    (ADJ
      (VRB vind) (AFF baar)))
```
(b)
```
(NOU
    (VRB
      (AFF ver) (ADJ snel))
    (AFF ing))
```
(c)
```
(NOU
    (NOU
      (NOU bloem) (NOU bol))
    (AFF en)
    (NOU
      (VRB kweek) (AFF er)))
```
(d)
```
(VRB
    (NOU
      (NOU pols) (NOU stok))
    (VRB
      (ADJ hoog) (VRB spring)))
```
(e)
```
(NOU (BRM ep) (AFF iek))
```

Het voordeel van dit codeerschema is niet alleen dat de representaties sterk zijn vereenvoudigd (wat zowel de gebruiker als de codeerder ten goede komt), maar ook dat de structuren eenvoudiger gebruikt kunnen worden bij computationele toepassingen.

5. Diachrone morfologie

In de vorige sectie heb ik de structuur besproken van de morfologische module in GiGaNT. Ik heb dat gedaan met behulp van enkel hedendaags Nederlandse voorbeelden. Maar zoals gezegd in de inleiding (§1), is GiGaNT een lexicon waarin het Nederlands van de zesde eeuw tot nu is opgenomen.

Het is een gemeenplaats dat de synchronie en de diachronie nauw met elkaar verweven zijn. Ook in de besproken databasestructuur zien we dat terug, bijvoorbeeld in de keuze voor een gebonden basiswoord (zie §4.2). In deze paragraaf zal ik een aantal keuzes verantwoorden die specifiek betrekking hebben op historisch morfologisch materiaal.

5.1. Gebonden basiswoord

Zoals in paragraaf 4.2 is besproken, onderscheiden we binnen GiGaNT een zogenoemd gebonden basiswoord. Gebonden basiswoorden zijn grondwoorden die

niet zónder verdere afleiding voorkomen (type *vadsig*: `(ADJ (BRM vads) (AFF ig))`). Het komt echter regelmatig voor dat een afleiding waarvan het basiswoord in het hedendaags Nederlands niet langer bestaat, in oudere fasen wel een bestaand basiswoord had. Een voorbeeld is het werkwoord *vermommen*. In het hedendaags Nederlands vinden we het basiswoord *mom* alleen nog terug in de versteende uitdrukking 'onder het mom van…' en in de samenstelling *mombakkes*, maar niet meer zelfstandig en in de betekenis 'masker'. In de zeventiende eeuw functioneert *mom* echter nog wel als zelfstandig element, zoals in:

> "De godsdienst is het mom, waar zich zijn list meê tooit, […]" (Wiselius, *Meng.* 5, 21).

Omdat *mom* in de zeventiende eeuw nog zelfstandig gebruikt wordt, analyseren we *vermommen* als `(VRB (AFF ver) (NOU mom))`, dus zonder gebonden basiswoord. Een aanwijzing dat *mom* in het zeventiende-eeuws nog los voorkomt is ook te vinden in de grotere hoeveelheid combinatorische mogelijkheden. Zo vinden we in het WNT onder het artikel MOM II, afleidingen en samenstellingen als:

(33) *mommen, ontmommen, mommig, mommendans, momaanzicht, mombakkes, mommekans, mommetuig, momsgewijs, mommegrijn, mommenhoofd, mommekleederen, momlijst, mommepak, mommeschijn, mommetronie, momverdek, mommenvolk, mombanket, momdienst.*

Het criterium om een bepaald basiswoord als gebonden of zelfstandig te beschouwen is dat het basiswoord van een afleiding als zelfstandige vorm geattesteerd moet zijn binnen GiGaNT. Het is vervolgens aan de gebruiker om te bepalen of een woord nog slechts vormelijk of ook nog semantisch geleed is.

5.2. Verbindingsklanken

De klanken /ə/ en /s/ kunnen in het hedendaags Nederlands optreden als verbindingsklank tussen twee leden van een samenstelling. Historisch gezien zijn deze klanken echter geen verbindingsklanken maar veelal naamvalsuffixen. Een nominale samenstelling als *koningskroon* is ontstaan uit de woordgroep *die coninx crone* waarbij de specificeerder in de genitief staat. Andere voorbeelden zijn (overgenomen uit Van Loey 1964, 180):

(34) (a) *die lants here* → *die lantshere* → *de landsheer*;
(b) *die siele rust* → *die siele rust* → *de zielerust*;
(c) *dat heren huis* → *dat herenhuis* → *het herenhuis.*

Heeft een woordgroep zich eenmaal ontwikkeld tot samenstelling, dan kunnen we echter niet zonder meer spreken van naamvalsuffixen. Delen van een woord

kunnen normaliter namelijk geen naamvalsuitgang krijgen. Binnen GiGaNT vatten we de klanken /ə/ en /s/ tussen twee leden van een samenstelling daarom niet op als naamvalsuitgangen (ook niet in oudere fasen van het Nederlands) maar alleen als verbindingsklank. De kwestie behoort daarmee eerder tot de morfosyntaxis waarbij uitgemaakt moet worden, welke woordgroepen zich gedragen als één woord, dat wil zeggen, een zelfstandig, onscheidbaar talig element (zie §3.1).

5.3. Leenwoorden en leenaffixen

Leenwoorden en leenaffixen vormen een problematische categorie in een diachroon lexicon als GiGaNT. Leenwoorden kunnen bijvoorbeeld affixen dragen die op een bepaald moment in de taalgeschiedenis nog niet als zodanig ervaren werden maar op andere momenten wel. Het is daardoor vaak onduidelijk waar we de grens moeten leggen: is een uitheems affix ook in het Nederlands een affix, of (nog) niet?

Voor een enkel uitheems talig element vinden we in de literatuur een nauwgezette beschrijving over wanneer het onderdeel is geworden van het Nederlands. Voorbeelden daarvan zijn het suffix *-erij* dat ontleend is aan het Frans (Hüning 1999) en de recente opkomst van het suffix *-gate*, ontleend aan het Engels (Hüning 2000). Van andere oorspronkelijk uitheemse affixen is wel bekend dat ze ontleend zijn, maar vaak ontbreekt een gedetailleerde beschrijving over wanneer de ontlening zich heeft voltrokken. Zo vinden we in het *Leenwoordenboek* (Van der Sijs 1996) wel de opmerking dat het suffix *-es* (zoals in *prinses* en *dienares*) ter vorming van vrouwelijke persoonsnamen ontleend is aan het Frans en in de 16de/17de eeuw algemeen in gebruik is genomen. We vinden echter niet terug wanneer dit talige element de status suffix heeft gekregen in het Nederlands. Bij andere affixen, bijvoorbeeld *-matig* (zoals in *beleidsmatig*), vinden we alleen de mededeling dát ze ontleend zijn (hier aan het Duits).

Zonder een uitvoerige detailstudie per geval, is het onderscheid tussen directe ontlening en afleiding ook erg lastig (nog los van de vraag of we het onderscheid überhaupt kunnen volhouden). Meestal worden in eerste instantie complete woorden ontleend aan een taal. Later kunnen deze woorden eventueel ontleed worden in samenstellende delen. Maar het kan ook zijn dat taalgebruikers een affix gebruiken om woorden te vormen binnen het uitheemse deel van de woordenschat. Zoals Van Bree (1996) stelt:

> Het is mogelijk dat ook een woord als bijv. *massage* los van het gelijkluidende franse woord in het Nederlands is ontstaan als afleiding van *masseren*.

Om met de hierboven geschetste problematiek om te gaan, kunnen we de pragmatische keuze maken om wat betreft de morfologische analyse geen onderscheid te maken tussen leenwoorden en inheems taalmateriaal. Dat zou betekenen dat

we alle leenwoorden in het lexicon op dezelfde manier analyseren als inheemse woorden. Ongeacht de vraag of een bepaald woorddeel al bestaansrecht heeft verworven in de taal, analyseren we het als zodanig.

Deze werkwijze is echter om verschillende redenen onwenselijk. Het belangrijkste bezwaar is dat het lexicon door deze werkwijze te veel 'vervuild' zou kunnen raken met uitheems taalmateriaal dat niet relevant is. Neem een Frans woord als *existant* ('bestaande'), aangetroffen in een 19de-eeuwse Nederlandse tekst. Volgens de hierboven geschetste werkwijze moeten we de analyse *exist+ant* opnemen in de database. Problematisch hieraan is dat het Nederlands het basiswoord *exist* noch het suffix *-ant* kent of heeft gekend. Het is daarmee voor een diachroon lexicon van het *Nederlands* volstrekt irrelevant om een morfologische analyse van de delen van *existant* te geven.

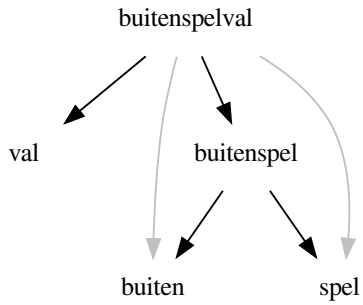
Binnen GiGaNT hebben we gekozen voor een tussenoplossing. We geven niet structureel een morfologische analyse van een uitheems woord, maar voeren ook geen detailstudie per uitheems woord uit. Het criterium om een (mogelijk) uitheems woord te analyseren is dat tenminste één van de samenstellende delen behoort of behoorde tot het Nederlands. Daarbij wordt geen onderscheid gemaakt tussen woorden die direct ontleend zijn aan een andere taal en woorden die binnen het Nederlands gevormd zijn. Het is immers niet de taak van GiGaNT om te bepalen welk affix wel en welk affix niet op een bepaald moment in de taalgeschiedenis geïdentificeerd werd in een woord. Wel is het de taak van GiGaNT om de verschillende woorddelen terugvindbaar te maken. Pas als een suffix is opgenomen in de morfologische analyse van een woord kan de mate van 'ingeburgerdheid' ervan worden onderzocht.

Wat betekent dit nu concreet? In zijn studie naar de ontstaansgeschiedenis van het suffix *-erij*, rekent Hüning (1999, 63) woorden als *ribauderie* ('buitensporig gedrag') en *seigneurie* ('heerschappij, bestuur') als leenwoorden (uit het Frans) en niet als afleidingen binnen het Nederlands. De reden hiervoor is dat de woorden in het Nederlands met dezelfde vorm en betekenis gebruikt worden als in het Frans. Hoewel dergelijke woorden ontleend zijn aan het Frans, worden ze binnen GiGaNT wel van een morfologische analyse voorzien, omdat het suffix *-erie* (*-erij*) geattesteerd is binnen het inheemse deel van het Nederlandse lexicon.

6. Onder de kap: de databasestructuur

In deze paragraaf zal ik kort de databasestructuur beschrijven van de morfologische module in GiGaNT.[8] Derivationele en compositionele morfologische gegevens worden gekoppeld aan het lemmaniveau in GiGaNT. De woordvormen

[8] Voor een uitvoerige beschrijving van de databasestructuur, zie Kenter (2010).



Figuur 5: Voorbeeld Directed Acyclic Graph voor *buitenspelval*. De zwarte pijlen staan voor 'moeder-kind'-relaties, de grijze pijlen geven de *transitive closures* weer.

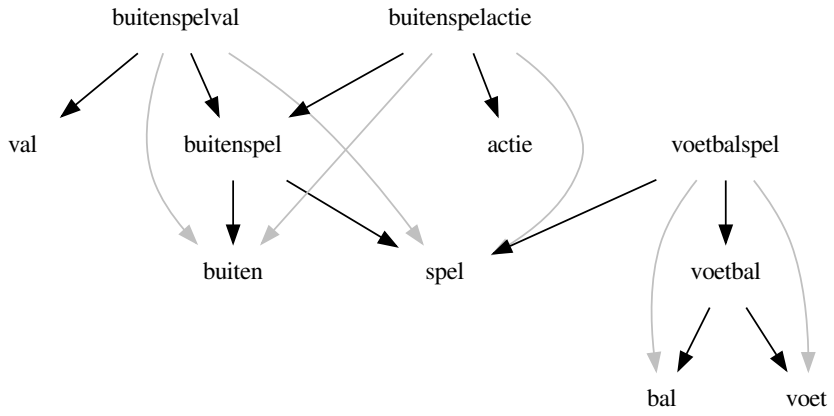
die aan de lemmata gekoppeld zijn erven deze informatie. Flexie wordt gekoppeld aan de woordvormen en is niet verbonden met de lemmata.

De in paragraaf 4.1 voorgestelde structuur kan het best vertaald worden in een hiërarchische databasestructuur. De hiërarchische relaties in de database zijn gelegd met behulp van een *Directed Acyclic Graph*. Figuur 5 geeft een voorbeeld van deze structuur voor *buitenspelval*.

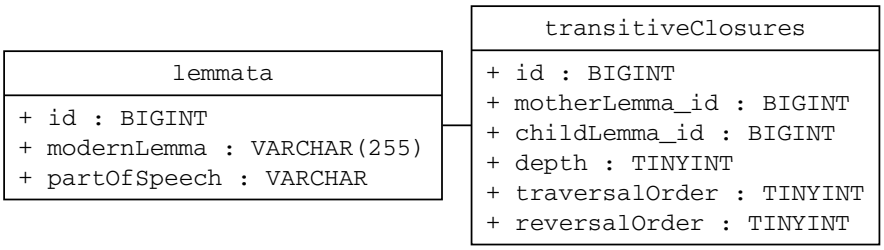
De bovenste knoop in de grafiek (*buitenspelval*) is de zogenoemde moederknoop. Deze knoop heeft als directe dochters de nomina *buitenspel* en *val*. Deze dochterknopen zijn op hun beurt weer verbonden met hun eigen kinderen (*buiten* en *spel*). Naast directe 'moeder-kind'-relaties, zijn ook de zogenoemde *transitive closures* gespecificeerd. Een *transitive closure* kunnen we opvatten als een 'grootmoeder-kleinkind'-relatie, waarbij de grootmoeder (*buitenspelval*) verbonden is met het kleinkind (*buiten*) via de moeder van het kleinkind (*buitenspel*).

In Figuur 6 zijn de analyses van *voetbalspel* en *buitenspelactie* toegevoegd aan de structuur. Figuur 6 maakt mooi zichtbaar dat syntagmatische en paradigmatische verbanden tussen woorden en woorddelen een natuurlijke plaats hebben in de databasestructuur.

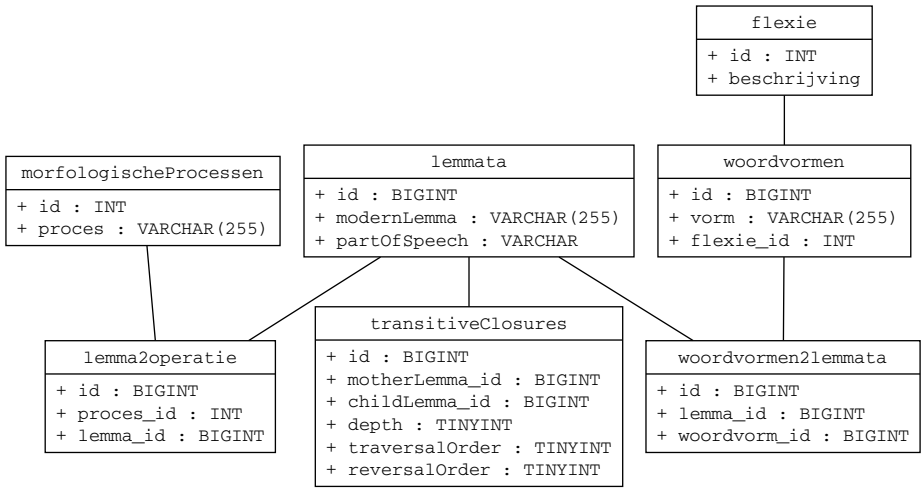
Dankzij de databasestructuur waarmee we de hiërarchische relaties in het lexicon kunnen leggen, hebben we voor de uiteindelijke database slechts twee tabellen nodig voor de basisonderdelen van de morfologische module. Figuur



Figuur 6: Voorbeeld Directed Acyclic Graph voor *buitenspelval*, *voetbalspel* en *buitenspelactie*. De zwarte pijlen staan voor 'moeder-kind'-relaties, de grijze pijlen geven de *transitive closures* weer.



Figuur 7: Databasemodel van de basisonderdelen in de morfologische module van GiGaNT.



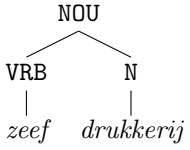
Figuur 8: Databasemodel van de morfologische module in GiGaNT.

7 geeft deze databasestructuur weer. De tabel LEMMATA is gekoppeld aan de tabel TRANSITIVECLOSURES waardoor het mogelijk wordt om de waardes van de moderne lemmata op te halen in de morfologische tabel.

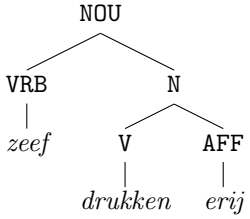
Nu we de databasestructuur voor de basisonderdelen van de morfologische module hebben bepaald kunnen we de overige componenten eraan toevoegen. In figuur 8 is de totale databasestructuur van de morfologische module weergegeven.

De hiërarchische representatie van de morfologische structuur hoeft met de gekozen databasestructuur niet te worden opgeslagen, maar kan 'on the fly' gegenereerd worden. Dit heeft als voordeel dat er op verschillende niveaus een structuurrepresentatie gevormd kan worden: De gebruiker kan een analyse vragen van een woord op het eerste niveau van analyse (analyse 35a), of op een dieper niveau (analyse 35b):

(35) (a)



(b)



De familierelaties tussen de woorden zijn eveneens niet apart opgenomen in een tabel aangezien deze gemakkelijk gegenereerd kunnen worden met queries in de database. Het voordeel van deze manier van data genereren is dat er minder opslagruimte nodig is. Zeker met het oog op het beoogde formaat van GiGaNT is dit van groot belang.

Referenties

Baayen, R. H., 1991. De CELEX lexicale databank. Forum der Letteren 32 (3), 221–231.

Baayen, R. H., Piepenbrock, R., Gulikers, L., 1995. The celex lexical database (release 2). CD-ROM, Linguistic Data Consortium, University of Pennsylvania, Philidelphia, U.S.A.

Booij, G., Van Santen, A., 1998. Morfologie: de woordstructuur van het Nederlands. Amsterdam University Press.

De Haas, W., Trommelen, M., 1993. Morfologisch Handboek van het Nederlands. Een overzicht van de woordvorming. 's-Gravenhage, SDU Uitgeverij.

De Jong, N. H., Schreuder, R., Baayen, R. H., 2000. The morphological family size effect and morphology. Language and Cognitive Processes 15 (4/5), 329–365.

Hüning, M., 1999. Woordensmederij. De geschiedenis van het suffix *-erij*. No. 19 in LOT Dissertation series. Holland Academics Graphics, Den Haag.

Hüning, M., 2000. Monica en andere gates. het ontstaan van een morfologisch procédé. Nederlandse Taalkunde 5 (2), 121–132.

Hüning, M., Van Santen, A., 1994. Produktiviteitsveranderingen: de adjectieven op -lijk en -baar. Leuvense Bijdragen 83 (1), 1–29.

Karsdorp, F., 2010. Evaluatie morfologische annotaties in GiGaNT, intern verslag, INL.

Karsdorp, F., Beekhuizen, B., 2010. Regelmaat in een regelloos systeem: De nederlandse superlatief, to be published in *Voortgang: Jaarboek van de Neerlandistiek*.

Kenter, T., 2010. Hiërarchishce structuren in relationele databases, intern verslag, INL.

Kestemont, M., Daelemans, W., de Pauw, G., 2010. Weigh your words – memory-based lemmatization for middle dutch. Literary and Linguistic Computing 25 (3), 287–301.

Krott, A., Baayen, R. H., Schreuder, R., 2001. Analogy in morphology: modeling the choice of linking morphemes in dutch. Linguistics 39 (1), 51–93.

Laureys, T., De Pauw, G., Van Hamme, H., Daelemans, W., Van Compernolle, Dirk, 2004. Evaluation and adaptation of the celex dutch morphological database. In: Proceedings 4th international conference on language resources and evaluation. Vol. IV. Lissabon, Portugal, pp. 1247–1250.

Meesters, G., 2004. Marginale morfologie in het Nederlands: paradigmatische samenstellingen, neoklassieke composita en splintercomposita. Studie op het gebied van de Nederlandse taalkunde. Koninklijke Academie voor Nederlandse Taal- en Letterkunde, Gent.

Neijt, A., Schreuder, R., 2009. De rol van en en s in samenstellingen. veranderend nederlands verandert gedachten over het taalsysteem. In: Boogaart, R., Lalleman, J., Mooijaart, M., van der Wal, M. (Eds.), Woorden wisselen. No. 20 in SNLreeks. Stichting Neerlandistiek Leiden, Leiden, pp. 93–104.

Schultink, H., 1962. De morfologische valentie van het ongelede adjectief. Utrecht: HES publishers.

Tálasi, Z., 2009. Het nederlandse prefix *ge-* in historisch perspectief. 'ge-+werkwoordstam'-afleidingen in grammatica's, woordenboeken en teksten. Ph.D. thesis, Leiden University.

Van Bree, C., 1996. Historische taalkunde, tweede, herziene druk Edition. Acco, Leuven, Amersfoort.

Van den Toorn, M., 1975. Nederlandse Grammatica. H.D. Tjeenk Willink Groningen.

Van der Sijs, N., 1996. Leenwoordenboek: De invloed van andere talen op het Nederlands. Sdu Uitgevers, Den Haag.
Van der Sijs, N., Van Santen, A., 2006. Een historisch-morfologische database. Taal en Tongval 19, 153–168.
Van Loey, A., 1964. Schönfelds Historische Grammatica van het Nederlands. Thieme, Zutphen.