INL SCHATKAMER VAN DE NEDERLANDSE TAAL

De morfosyntactische module van het GiGaNT-lexicon

Tilly Ruitenberg en Katrien Van pellicom
I.s.m. Jesse de Does en Katrien Depuydt
INL Working Papers - Taalbank Nederlands 3
Leiden, 2012

www.inl.nl

Inhoudsopgave

1 Inleiding: het GiGaNT-lexicon

De afdeling Taalbank van het Instituut voor Nederlandse Lexicologie (INL) heeft als opdracht de toegang tot de Nederlandse taalschat te waarborgen, onder andere door het bouwen van lexica en corpora. De lexica worden ingezet voor het ontsluiten en verrijken van teksten en corpora en voor het koppelen van corpora aan elkaar en aan andere teksten. Daarbij spelen bestaande bronnen, zoals de historische woordenboeken van het INL, een belangrijke rol.

Op dit moment wordt er bij de Taalbank gewerkt aan een nieuw, allesomvattend infrastructureel lexicon: het GiGaNT-lexicon.

GiGaNT staat voor 'Groot Geïntegreerd Lexicon van de Nederlandse Taal' en is een computationeel lexicon van het Nederlands van de zesde eeuw tot het hedendaags Nederlands. GiGaNT is eigenlijk een verzameling woorden, woordgroepen en woorddelen, telkens gekoppeld aan het bijhorende moderne Nederlandse lemma.

De GiGaNT-populatie is diachroon: GiGaNT omvat de taal van 500 tot heden.

GiGaNT is een functioneel en generiek lexicon. We gaan in principe uit van de morfosyntaxis zoals beschreven in de ANS, uiteraard aangevuld of uitgebreid waar dat vereist is, bijvoorbeeld voor historische vormen.

GiGaNT is een corpusgebaseerd lexicon: alle woordvormen in het lexicon zijn gekoppeld aan een of meer bewijsplaatsen in bestaand taalmateriaal. Ook worden alle aangetroffen woordvormen van de lemmata opgenomen in het lexicon. Spelling- en morfologische varianten[1] spelen immers een belangrijke rol bij de automatische herkenning en lemmatisering van historische woorden binnen Text Recognitionprogramma's.

GiGaNT is modulair opgebouwd: de combinatie woordvorm, lemma en woordsoort, aangeduid als type – lemma – PoS, vormt de kern, bijkomende informatie kan in aanvullende modules worden toegevoegd (zie bijvoorbeeld de morfologische module).

Tot slot willen we nog benadrukken dat het bij GiGaNT niet de bedoeling is om uiteindelijk één groot product af te leveren dat 'af' is, maar wel om een verzameling informatie bijeen te brengen die telkens kan worden aangevuld. Materiaal en werkwijze kunnen op die manier telkens, waar nodig, worden bijgesteld en uitgebreid.

GiGaNT is centraal onderdeel van een infrastructuur voor de beschrijving van de Nederlandse woordenschat die op het INL in ontwikkeling is.

GiGaNT zal op verschillende manieren toegankelijk worden gemaakt, waaronder als webservice.

1.1 Tools en thesaurus

Bij de inrichting van GiGaNT spelen de twee hoofdfuncties van het lexicon een grote rol: *tools* en *thesaurus*.

1.1.1 Computationeel-linguïstische toepassing van het lexicon in tools

[1] Een voorbeeld: de varianten aangetroffen bij uiterlijk.
uytterlijcste, uyterlijkste, d'uyterlijke, uiterlyke, uyterlijcke, uiterlijke, uyterlijck, uiterlyken, uiterlijkste, uiterlicke, wterlicke, wterlijcke, ulterlijk, uiterlyk, uiterlijk, uyterlick, wterlicken, d'uyterlijcke, uiterlijken, uiterlijks, wterlijck, uytterlicke, uitterlijke, ujterlijke, uytterlijk, uyterlycke, uyterlicken, uijterlicke, d'uiterlijcke, wtterlijcke, wterlyke, wtterlijk, uiterlijke, uuterlick, uuterlic, uyterlijke, uyterlijcken, uyterlicke, d'uiterlyke, wterlijke, vuyterlijcke, uuterlycke, uuterlicke, wterlijken, uyterlijcksten, uuyterlicke, uuyterlick, uuyterlycke, uytterlijcke, uytterlycke, uytterlick, vuytterlicke, uiterlijker, uyterlyck, uterliek, wterlijcken, uiterlijkst, uitterlijk, uytterlijcken, uyterlyk, uiterlijk, wterlick, uutterlijck, uuyterlicken, uyttelijck, uijterlijk, uytterlijck, uuterlijck, uiterlick, uitterlyk, uuyterlic, uuyterlyck, uuyterlijck, uiterlijck, uytterlyck, uterlyc, wterlijk.

De PoS-aanduiding zorgt ervoor dat niet alle vormen onoverzichtelijk in het lexicon worden opgeslagen, maar dat ze door middel van een gedetailleerde grammaticale karakterisering een vaste plek in het paradigma hebben.

Zo kan bijvoorbeeld direct worden opgemerkt dat de comparatiefvorm niet of nauwelijks voorkomt en dat het zelfstandig naamwoord vooral (of uitsluitend) in het enkelvoud voorkomt.

Op het gebied van morfosyntaxis willen we dat het lexicon kan worden ingezet bij onder andere de volgende taken:

- digitaal materiaal toegankelijk maken door middel van verrijkte indexering (het toevoegen van moderne lemmavormen aan zoekindexen) en query-expansie (het toevoegen van flexievormen en (spelling)varianten aan een zoekopdracht;
- koppelen van tekstmateriaal aan de wetenschappelijke woordenboeken van het Nederlands;
- automatische PoS-tagging en lemmatisering van grote hoeveelheden corpusmateriaal;
- automatische morfologische analyse van nieuwe types (samenstellings- en afleidingsmorfologie);
- Verbeteren van de kwaliteit van OCR, zowel door het gebruik tijdens het OCR-proces als door gebruik binnen postcorrectiesystemen;
- datamining en datasuppletie: opsporen van nieuw lexicon-, woordenboek- en grammaticamateriaal (zowel ouder als recent materiaal); belangrijk hierbij is het opsporen van neologismen, waarvoor het web als corpus gebruikt zal worden;
- expansie: genereren van paradigmatische woordvormen op basis van een set regels, afgeleid uit reeds bestaand lexiconmateriaal;
- Named Entity Recognition: herkenning van eigennamen (persoonsnamen, plaatsnamen en namen van organisaties)[2];
- In een volgende fase zullen ook tools ontwikkeld worden voor de ontsluiting van syntactische en semantische[3] informatie.

1.1.2 Thesaurus

GiGaNT fungeert ook als thesaurus: het wil de opslagplaats van taalkundige gegevens over de totale Nederlandse woordenschat zijn, met informatie over woordfamilies, semantische netwerken, woordreeksen, frequenties, chronologische gegevens, morfosyntaxis, tekstgegevens enz.
Alle informatie binnen GiGaNT is elektronisch beschikbaar en raadpleegbaar en kan dus direct worden ingezet voor allerlei toepassingen en bij onderzoek, zowel binnen het INL als in het wetenschapsveld daarbuiten.

1.2 Ontwikkelingen naar aanleiding van concrete projecten

De Taalbank participeerde van 1 januari 2008 tot 31 juni 2012 in het Europese IMPACT-project, Binnen dat project was het de taak van de Taalbank een lexicon met historische woordvormen te ontwikkelen, met als doel de OCR en de toegankelijkheid van historische teksten te verbeteren.

Intussen zijn ook de historische wetenschappelijke woordenboeken digitaal en uniform ontsloten, zodat van daaruit lexiconmateriaal voor GiGaNT gedestilleerd kan worden.

Deze twee projecten, IMPACT en de woordenboekapplicatie, hebben op natuurlijk wijze de prioriteit van het GiGaNT-ontwerp bepaald, nl. dat te allen tijde flexibel ingespeeld kan worden op belangrijke externe en interne ontwikkelingen, en hebben ook geleid tot verdere uitwerking van het lexiconontwerp. In eerste instantie was bijvoorbeeld een combinatie van functionele en vormelijke kenmerken voorzien als beschrijving van de woordsoort, zowel voor corpustagging als voor opslag in de thesaurus. In de loop van 2008 werd echter gekozen voor een expliciete tweedeling:

- een eenvoudige vormgestuurde woordsoortbeschrijving voor corpustagging (met alleen flexieaanduiding als eerste optie);

- een veel rijkere (want meer gedetailleerde) functionele paradigmatische woordsoortbeschrijving, die vooral geschikt is voor opslag van het materiaal in de thesaurus.

[2] Een namenlexicon maakt deel uit van het lexicon

[3] GiGaNT neemt op dit moment geen betekenisomschrijvingen op. De semantische informatie is beperkt tot korte glossen, om bepaalde ambigue lemma's van elkaar te kunnen onderscheiden of om aan te geven dat het om een nieuw (gebruik van een) woord gaat.

1.3 Bronnen

Het GiGaNT-lexicon zal allereerst gevuld worden met de lemmata en de daarbij gezochte woordvormen uit de historische woordenboeken van het INL: het Oudnederlands Woordenboek, het Vroegmiddelnederlands Woordenboek, het Middelnederlandsch Woordenboek en het Woordenboek der Nederlandsche Taal. Hieraan zal ook het materiaal van het Algemeen Nederlands Woordenboek (ANW) en van de Woordenlijst Nederlandse Taal (Groene Boekje) worden toegevoegd. GiGaNT vormt namelijk de kern van de nieuwe centrale data-infrastructuur van het INL, waarin alle informatie over de Nederlandse woordenschat wordt opgeslagen.

Daarnaast zullen nieuwe lemmata en ontbrekende woord(vorm)en worden gehaald uit nieuwe bronnen[4]: corpora/tekstverzamelingen uit interne en externe projecten. Enerzijds gaat het dus om nieuwe woorden die in de teksten worden aangetroffen en vervolgens met hun attestatie-informatie worden opgenomen, anderzijds om (door synthese of expansie) gegenereerde woord(vorm)en waarvan wordt bekeken of ze werkelijk voorkomen, waarna ze gekoppeld worden aan attestatie-informatie. Taalkundige informatie bij die woordvormen zal dan moeten worden ingepast in het GiGaNT-schema.

1.4 Verrijking

Het GiGaNT-lexicon wordt gevuld met vele miljoenen woordvormen. Zonder enige bijkomende informatie is zo'n gigantische opslaglijst van woordvormen van weinig tot geen nut, daarom wordt het lexicon verrijkt, waarbij aan alle woordvormen bepaalde informatie wordt gekoppeld.

1.4.1 Basisverrijking: type – lemma – PoS

De combinatie type (woordvorm) – lemma – PoS (Part of Speech; woordsoort) is als het ware de 'kapstok' waaraan alle verder toe te kennen verrijking, wordt opgehangen. Zie als voorbeeld tabel 1[5].

Het type is de abstractie over alle voorkomens van eenzelfde woordvorm in een lexicon. De woordvorm ('token') zelf is immers, als element van spraak- of schrijfuitingen, een tekst- of corpuselement.

Het lemma, bekend als ingang van woordenboeken, verbindt de woordvormen die bij dat lemma horen. In GiGaNT gaat het om een modern lemma[6].

De woordsoort is een puur grammaticaal element, dat iets zegt over het gedrag of de functie van een woord in een zin (morfosyntaxis). Binnen GiGaNT moeten de woordvormen die bij een lemma horen altijd tot één woordsoort behoren.

Tabel 1: voorbeeld van de GiGaNT-kapstokstructuur: type – lemma – PoS

TYPE	LEMMA	POS
aap	AAP	NOU-C
apen	AAP	NOU-C
leest	LEZEN	VRB
las	LEZEN	VRB
las	LAS	NOU-C

[4] Een voorbeeld van zo'n nieuwe bron is de "*Brieven als buit*"-collectie van prof. dr. M. Van der Wal: deze zal verrijkt worden en de resultaten zullen worden opgeslagen in GiGaNT.

[5] Het type *apen* heeft het lemma *aap*, dat als woordsoort NOU-C (zelfstandig naamwoord: soortnaam) heeft. Het lemma *aap* is de abstractie over alle woordvormen (types) van *aap* en is het hoofd van het raamwerk (of paradigma) waarin al deze *aap*-woordvormen hun plaats hebben. Door een lemma toe te kennen aan een woordvorm (type), wordt die als het ware verankerd in zijn paradigma.

[6] Vergelijk met de lemmata van de historische wetenschappelijke woordenboeken, die ook allemaal voorzien zijn van een modern lemma en daardoor met elkaar verbonden zijn.

klein	KLEIN	AA
kleine	KLEIN	ADJ
kleine	KLEINE	NOU-C
kleintje	KLEINTJE	NOU-C

1.4.2 Transcategorisatie-informatie

In GiGaNT is woordsoort een onderscheidend lemmakenmerk: *fietsen* (VRB) is een ander lemma dan *(het) fietsen* (NOU-C)[7]. We willen de relatie tussen zulke lemmaparen echter wel vastleggen. De **transcategorisaties** worden ingedeeld in types. In het lexicon slaan we lemmaparen samen met het type transcategorisatie op. Voorts bevat het lexicon informatie over de regelmatig terugkerende transcategorisatietypes. Wanneer (vrijwel) alle lemmata van een bepaalde woordsoort een pendant met een bepaalde andere woordsoort hebben, spreken we van *regelmatige* transcategorisatie. Anders is er sprake van *incidentele* transcategorisatie.

1.4.2.1 Regelmatige transcategorisatie

Binnen het lexicon wordt vastgelegd welke regelmatige woordsoortovergangen (nulconversies) mogelijk zijn. Dit gebeurt aan de hand van conversieregels. Een voorbeeld: bijvoeglijke naamwoorden kunnen functioneren als zelfstandige naamwoorden. Elk lemma krijgt dus één of meerdere nevenlemmata met een nevenwoordsoort toegekend.

1.4.2.2 Incidentele transcategorisatie

Een incidentele transcategorisatierelatie tussen 2 lemmata wordt op precies dezelfde manier in het lexicon opgenomen als een regelmatige transcategorisatie. Het enige verschil is dat de aanname dat in principe ALLE lemmata van een bepaalde woordsoort deze transcategorisatie kunnen ondergaan ontbreekt. Het wordt echter op een ander niveau vastgelegd en gebeurt niet aan de hand van een automatische procedure.

Een voorbeeld: deelwoorden kunnen zich gedragen en gebruikt worden als voorzetsel (uitgezonderd, hangende,...) Dit wordt niet op woordsoortniveau vastgelegd. Enkel bij een beperkte subgroep wordt een transcategorisatieregel toegevoegd, waardoor enkel dat type een nevenwoordsoort krijgt.

1.4.3 Andere verrijking

Naast de reeds behandelde elementen lemma en PoS zal bijkomende **taalkundige informatie** worden opgeslagen: spellingvariatiegegevens en morfologische informatie (analyse, flexieklasse, bouwvorm enz.). Daarnaast wordt ook **brontekst- en andere corpusinformatie** voorzien, omdat de voorgaande aspecten enkel betekenis krijgen als we ook weten of de woordvormen effectief voorkomen en zo ja, in welke teksten en in welke tijd. Het gaat dan ook om attestatiegegevens, tekst- en broninformatie, taallabels, teksttype en tekstkenmerken, datering, lokalisering, lexicale bron, verificatiegegevens, frequentiegegevens en eventuele betekenisglossen[8]. Er wordt gewerkt aan een semantische module voor het lexicon.

[7] Hierin volgen we het Vroegmiddelnederlands en het Oudnederlands woordenboek; de overige historische woordenboeken zijn in dit opzicht niet consequent.

[8] Een glosse is een korte betekenisomschrijving van een woord in een bepaalde tekst. Deze kan van belang zijn bij vb. homonymie, nieuwe betekenissen of nieuwe woorden.

2 Part of Speech in GiGaNT

Zoals gezegd is de woordsoortaanduiding een essentieel aspect van de inhoudelijke verrijking van het GiGaNT-lexicon. Aangezien het concept 'woordsoort' kan worden beschreven vanuit zeer verschillende invalshoeken, zoals die van de lexicografie, de (morfo-) syntaxis of de (afleidings- en samenstellings-)morfologie, is het belangrijk de juiste combinatie te vinden die aansluit bij de andere gemaakte keuzes met betrekking tot de inrichting van dit lexicon.

2.1 Functies van PoS in GiGaNT

In GiGaNT heeft PoS meerdere functies:
- als onderscheidend lemmakenmerk;
- ter grammaticale karakterisering van woordvormen in een tekst of corpus;
- als organiserend principe bij de opslag van de woordvorm en zijn varianten;
- als paradigma-aanvulling: bij gaten in het paradigma moeten woordvormen met die ontbrekende PoS worden bijgezocht of gegenereerd;
- als onderdeel van de morfologische analyse van het lemma.

2.1.1 Lemma-onderscheidend lemmakenmerk

In GiGaNT is woordsoort een lemma-onderscheidend lemmakenmerk: *fietsen* (VRB) is een ander lemma dan *(het) fietsen* (NOU-C).

2.1.2 Grammaticale karakterisering van woordvormen in een tekst of corpus

De grammaticale karakterisering van een woordvorm zegt in de eerste plaats iets over het gebruik in de zin.

Een voorbeeld: het werkwoord (VRB) *talen* heeft in de zin *Zij talen niet naar nog meer invloed* een andere functie dan het zelfstandig naamwoord (NOU-C) *talen* in de zin *Buitenlandse talen spreken is een groot goed*.

De grammaticale karakterisering zorgt er hier voor dat de ambiguïteit van de woordvorm vermindert, wanneer deze multi-interpretabel is.

Tabel 2: PoS en lemma bij woordvorm ambigu voor PoS

	WOORDVORM	**POS**	**LEMMA**
	talen	VRB|NOU-C	talen|taal
1		VRB	talen
2		NOU-C	taal

Een ander voorbeeld: het werkwoord (VRB) *bronzen* en het bijvoeglijk naamwoord (ADJ) *bronzen*.

Tabel 3: PoS en lemma bij woordvorm ambigu voor PoS

	WOORDVORM	**POS**	**LEMMA**
	talen	VRB|ADJ	bronzen
1		VRB	bronzen
2		ADJ	bronzen

Als een meer gedetailleerde grammaticale karakterisering nodig is, zou het lemma alleen bovendien niet meer voldoen als desambiguerende factor, maar moet de woordsoortaanduiding specifieker zijn, zoals uit onderstaande tabel blijkt.

Tabel 4: identiek lemma en gedetailleerd PoS (binnen een paradigma)

	WOORDVORM	**POS**	**LEMMA**
	talen	VRB	talen
1		VRB (fin, ind, pl, 1, 2, 3)	talen
2		VRB (inf)	talen

De infinitief en de indicatiefvormen hebben hetzelfde lemma en worden verder onderscheiden op basis van een analyse van de context.

Om een bepaald type te kunnen koppelen aan een lemma en zijn plaats in het paradigma te bepalen, is een uitgebreidere woordsoortbeschrijving nodig dan alleen de kale woordsoort die bij het lemma hoort. PoS wordt dan ook opgesplitst in *lemmawoordsoort* en *typewoordsoort* (zie verder onder 2.2).

De woordsoortbeschrijving zelf kan gemaakt worden vanuit de vorm (vormgestuurd), vanuit de functie (paradigmatisch). Binnen GiGaNT is gekozen om zowel de *vormgestuurde* als de *functionele of paradigmatische* beschrijving te gebruiken.

Tabel 5: overzicht woordsoortbeschrijvingen

POS		
Lemma_PoS	**Type_PoS**	
kale woordsoort	vormgestuurde beschrijving	functionele / paradigmatische beschrijving

Vormgestuurde beschrijving

Vormelijke woordsoortbeschrijvingen van de woordvorm kunnen grotendeels automatisch gegenereerd worden, omdat ze weinig specifiek zijn en daardoor bij toepassingen weinig fouten opleveren. Ook controle gaat betrekkelijk snel. Dit soort beschrijvingen gaat wel met flink wat ambiguïteit gepaard, maar die wordt duidelijk in kaart gebracht: de relatie tussen de vormgestuurde beschrijving en de paradigmatische tagset wordt vastgelegd in mappingregels.

Tabel 6: voorbeelden

TYPE	**LEMMA + POS**	**TYPE-POS, VORMGESTUURD**
aap	AAP_N	NOU-C (infl=o)
apen	AAP_N	NOU-C (infl=en)
leest	LEZEN_V	VRB (infl=t)
las	LEZEN_V	VRB (infl=klankwisseling+o)
las	LAS_N	NOU-C (infl=o)
klein	KLEIN_AA	AA (infl=o)
kleine	KLEIN_ADJ	ADJ (infl=e)
kleine	KLEINE_N	NOU-C (infl=o)

Functionele of paradigmatische beschrijving

Functionele of paradigmatische woordsoortbeschrijvingen sommen de functionele eigenschappen op van de woordvorm: getal, geslacht, persoon, naamval, tijd, wijze. Deze functionele eigenschappen worden traditioneel ondergebracht in een paradigma: een soort raamwerk waarvan de invulling wordt bepaald door de flexieklasse[9] van het lemma. Dit raamwerk speelt een belangrijke rol wanneer een nog (gedeeltelijk) leeg lexicon systematisch moet worden gevuld. Dat kan gebeuren met onder meer bestaand woordenboekmateriaal of met gegenereerde, "bijgebakken" vormen. Paradigma's zijn dus essentiële bouwstenen binnen het lexicon en zorgen er mee voor dat het gecontroleerd kan groeien.

Tabel 7: voorbeelden

TYPE	LEMMA + POS	TYPE-POS, PARADIGMATISCH
aap	AAP_N	NOU-C (sg)
apen	AAP_N	NOU-C (pl)
leest	LEZEN_V	VRB (fin, pres, sg, 2, 3)
las	LEZEN_V	VRB (fin, past, sg, 1, 2, 3)
las	LAS_N	NOU-C (sg)
klein	KLEIN_AA	AA
kleine	KLEIN_ADJ	ADJ (attr)
kleine	KLEINE_N	NOU-C (sg)

PoS-mappingregels en -records

Door deze twee verschillende woordvormbeschrijvingen samen is het mogelijk om woordvormen snel te herkennen en te duiden (=vormgestuurde aanpak), en deze vervolgens precies op hun plek in de database op te slaan (=paradigmatische aanpak).

Hiertoe leggen we de relatie tussen deze twee beschrijvingen in regels vast; we houden we een mapping bij die beide beschrijvingen van elke paradigmaplaats aan elkaar koppelt. De mappingrecords ondersteunen de semiautomatische toekenning van paradigmatische kenmerken en kunnen tevens het zoeken op paradigmatische eigenschappen binnen een slechts met formele eigenschappen verrijkt corpus ondersteunen. Dit onderdeel is nog niet uitgewerkt, mede vanwege de nauwe samenhang met de hiervoor noodzakelijke indeling van historische woordenschat in flexieklassen. Hieronder enige mogelijke voorbeelden:

PoS	Flexieklasse	Periode	T_f	T_p
NOU	-	Modern	Infl=en	Number=pl
NOU	Flexieklasse I[10], mannelijk en onzijdig	Middelnederlands	Infl=en	Case=dat,number=pl
NOU	Flexieklasse II, mannelijk en onzijdig	Middelnederlands	Infl=en	Case=dat,number=sg
NOU	Flexieklasse I, mannelijk en onzijdig	Middelnederlands	Infl=s	Case=gen,number=sg
NOU	s-meervoud	Modern	Infl=s	Number=pl
VRB	-	Modern	Infl=t	Number=sg,tense=pres,person=2
VRB	-	Modern	Infl=t	Number=sg,tense=pres,person=3

[9] Hiermee wordt de declinatie- of conjugatieklasse van het lemma bedoeld.

[10] Indeling vooralsnog volgens van Loey, Middelnederlandse Spraakkunst.

2.1.3 Organiserend principe bij de opslag van de woordvorm en zijn varianten

Met GiGaNT als achtergrondlexicon zal een woordvorm eerst door de toekenning van een lemma en een (primaire) woordsoort naar het goede paradigma worden geleid en vervolgens door een meer specifieke karakterisering of woordsoortaanduiding op zijn juiste plek in het paradigma[11] worden geplaatst.

De opslag van woordvormen gebeurt in de eerste plaats in vormlijsten / woordvormtabellen, aan de hand van een lemma en een grammaticale tag. Niet al het materiaal kan paradigmatisch beschreven worden: daarom worden ook woordvormlijsten met alleen vormgestuurde beschrijvingen opgeslagen.

Het is echter beter en vollediger om woordvormen op te slaan in paradigma's. GiGaNT als opslagplaats zal gaandeweg bestaan uit een groeiend aantal woordvormen, gekoppeld aan lemmata, waaruit een groeiend aantal steeds completere paradigma's kan worden samengesteld.

2.1.4 Paradigma-aanvulling

Wanneer wordt vastgesteld dat bepaalde paradigma's nog gaten vertonen of dat bepaalde lemmata nog geen woordvormen hebben, zullen deze worden aangevuld met nieuw corpusmateriaal of met gegenereerde woordvormen met als eigenschap die ontbrekende type-PoS. Voor een dergelijke expansie moeten regels worden afgeleid om ontbrekende vormen te genereren. Dit kan alleen op basis van een zo groot mogelijk aantal bekende lemmata met bijhorende woordvormen. Op basis van de woordsoortbeschrijving van de paradigmaposities kan men nl. voorspellen of bepaalde woordvormen mogelijk zijn. Gegenereerde (en verondersteld mogelijke) woordvormen worden dan voorlopig gemarkeerd met een vlaggetje om aan te geven dat ze geëxpandeerd zijn en dat werkelijke attestaties moeten worden opgespoord.

Hoe completer de paradigma's zullen zijn, hoe beter met behulp van tools gebruik kan worden gemaakt van het in het lexicon aanwezige materiaal: vb. corpustagging, woordvormexpansie en lemmatisering van woordvormen. De aanwezigheid van meer en completere paradigma's in het lexicon zorgt dus voor het sneller en doeltreffender werken met het lexicon.

2.1.5 Onderdeel van de morfologische analyse van het lemma

PoS functioneert ook in de morfologische module: in GiGaNT krijgt elke lemma een morfologische oppervlakteanalyse, die bestaat uit de woordsoorten van de samenstellende delen, en, bij afleidingen, het affix.

Samengestelde en afgeleide woorden worden gekoppeld door de morfologische analyse aan de lemma's waaruit ze samengesteld zijn of het lemma waarvan ze afgeleid zijn. Op die manier is de plaats van dat deel binnen de samenstelling of afleiding zichtbaar en duidelijk.

Uitgebreide informatie is te vinden in het aparte stuk over de morfologische analyse in GiGaNT.

2.2 Soorten PoS in GiGaNT

Het multifunctionele karakter van GiGaNT en de complexiteit van de woordsoorttoekenning vragen om een compleet overzicht van de verschijningsvormen van PoS in GiGaNT.

[11] Iedere type-lemma-PoS-trits maakt deel uit van een paradigma. Elk paradigma bevat minimaal één zo'n trits.

2.2.1 Lexicale en functionele woordsoort

De betekenis en het gebruik van woorden zijn constant in beweging. Elk woord heeft enerzijds een (of meerdere) gelexicaliseerde woordsoort(en), die we in woordenboeken e.d. kunnen terugvinden, maar kan anderzijds functioneel als andere, niet noodzakelijk in het woordenboek vermelde woordsoort(en) voorkomen. Dit is het onderscheid tussen de lexicale en functionele woordsoort. Zodra een bepaald gebruik vaak genoeg voorkomt in de praktijk, kan de bijhorende functionele woordsoort gelexicaliseerd worden: het woord krijgt er dan nog een woordsoort bij in het woordenboek of naslagwerk. De geboekstaafde woordsoort (=lexicaal) is dus eigenlijk steeds de resultante van het gebruik (=functioneel).

Een algemeen voorbeeld: het gebruik van infinitieven (VRB) als zelfstandig naamwoord (NOU-C) is een productief verschijnsel. Iedere infinitief kan als zelfstandig naamwoord gebruikt worden, maar niet iedere infinitief wordt ook daadwerkelijk als zelfstandig naamwoord gelexicaliseerd en in woordenboeken opgenomen. Woordenboeken maken immers keuzes op basis van attestatiefrequentie en op basis van de betekenisontwikkeling en het afleidings- en samenstellingsgedrag van de 'nieuwe' woordsoort. Als elke infinitief zich als een zelfstandig naamwoord kan gedragen, hoeft dat nl. niet meer in het woordenboek vermeld te worden, behalve als daaruit een nieuwe betekenisontwikkeling ontstaat.

Concreet zal binnen GiGaNT bij elke woordsoort worden aangegeven naar welke andere woordsoorten systematische woordsoortverschuiving optreedt. Elk lemma met een bepaalde 'primaire' woordsoort krijgt dan een of meerdere nevenlemmata met een eigen nevenwoordsoort en een daarbij aansluitend paradigma. In GiGaNT wordt dus geen onderscheid gemaakt tussen gelexicaliseerde en niet-gelexicaliseerde gevallen: vb. de lemma's *leven* en *fluisteren* (beide VRB) hebben een transcategorisatierelatie met een lemma met woordsoort NOU-C, hoewel *leven* (NOU-C) wel gelexicaliseerd is en *fluisteren* (NOU-C) niet. Uit GiGaNT kan men dus niet afleiden of het om een lexicale of een functionele woordsoort gaat[12].

2.2.2 Lemmawoordsoort

De lemmawoordsoort is de woordsoort die in GiGaNT aan het lemma wordt gehangen. Het gaat altijd om een korte (kale) woordsoortbeschrijving, eventueel gevolgd door een lemmatype, vb. VRB, ADJ, NOU-C (zelfstandig naamwoord: soortnaam), NOU-P (zelfstandig naamwoord: eigennaam).

In de diverse bronnen binnen het INL worden tot nog toe verschillende woordsoortaanduidingen gebruikt. Uiteraard willen we dit uniformeren: de GiGaNT-lemmawoordsoort is speciaal ontwikkeld om in alle bronnen één en dezelfde naamgeving te gebruiken. Daarnaast is in de thesaurus nog een apart blok woordsoortinformatie te vinden: de GTB-woordsoort, met daarin informatie uit WNT, VMNW, MNW en ONW. De woordsoorten van de GTB-applicatie zijn qua naamgeving geüniformeerd, maar niet gewijzigd of verbeterd. Ze kunnen wel aangepast en gecorrigeerd worden bij de toekenning in GiGaNT, omdat bijvoorbeeld gevallen waarbij een functionele woordsoort toch gelexicaliseerd lijkt, zullen worden opgelost aan de hand van de reeds vermelde transcategorisatieregels.

[12] In woordenboeken als ONW en VMNW, en ook in GiGaNT, heeft de oppositie lexicaal-functioneel (en het gebruik van frequentiegegevens) geen zin: zodra er een andere woordsoort geconstateerd is – ook al is er maar één enkele bewijsplaats – dan krijgt zo'n woord een ander lemma-PoS.

2.2.3 Typewoordsoort

De typewoordsoort (of woordvormwoordsoort) is de woordsoort die in GiGaNT aan de geanalyseerde woordvorm wordt gehangen.

Zoals reeds eerder vermeld, kan de woordsoortbeschrijving gemaakt worden vanuit de vorm (vormgestuurd) of vanuit de functie (paradigmatisch). De relatie tussen beide beschrijvingen wordt vastgelegd in mappingregels.

Bij de paradigmatische typewoordsoort wordt de functie gedetailleerd beschreven met kenmerken die de woordsoortparadigma's bepalen.

Bij de vormgestuurde typewoordsoort gaat het louter om een beschrijving van de vorm, een andere weergave van de woordvorm, aan de hand van globale, maar wel karakteristieke verschillen met zijn lemma (vb. deelwoordprefix, klinkerwisselingen enz.). Die verschillen kunnen veelal met een programma worden vastgesteld (corpustagging) of met een programma uit bestaande paradigma's worden afgeleid.

Paradigmatische en vormgebaseerde woordsoortbeschrijvingen zijn altijd gerelateerd. Mede door het verlies aan flexie zal een vormgebaseerde beschrijving minder precieze informatie bevatten, zodat er vaak sprake zal zijn van ' many-to-many'-mapping: één vormelijke beschrijving leidt tot meerdere paradigmatische beschrijvingen en omgekeerd.

2.3 Het vastleggen van PoS-kenmerken in tagsets

Om de verschillende lexicale en grammaticale aspecten van de woordvormbeschrijving zo eenduidig en helder mogelijk vast te leggen, wordt gebruik gemaakt van tagsets.

Zo'n tagset wordt altijd ontwikkeld met het oog op een bepaald doel en kan niet los worden gezien van bepaalde aspecten van het lexicon: de opzet, de inrichting, het model, de doelen, de populatie, de bouw, de groei enz.

Een PoS-tagset is een combinatie van woordsoortaanduidingen met *kenmerken* (of *features*, vb. Person, Number) en *waardes* (of *values*, vb. 1, 2, 3, singular, plural). Het doel is tweevoudig: enerzijds 'woordvormen in context' van PoS-informatie voorzien, anderzijds alle woordvormen een eigen opslagplaats binnen het lexicon geven.

2.3.1 Oppositie corpustagging - lexiconopslag

Het multifunctionele karakter van het GiGaNT-lexicon zorgt ervoor dat het veel omvattender en complexer is dan een simpel corpuslexicon.

In de thesaurus wordt een compleet stelsel kenmerken opgeslagen. De toegepaste corpustagset van een corpusgerelateerd lexicon is daarvan slechts een mogelijke afgeleide. Zo zullen bepaalde kenmerken wel in de paradigmabouw van belang zijn, teneinde alle woordvormen netjes op hun plaats in de thesaurus op te slaan, maar niet per se noodzakelijk/haalbaar zijn voor corpustagging (vb. of het om een 1e, 2e of 3e persoon meervoud gaat)[13].

[13] Andere voorbeelden zijn kenmerken als gender en case en onderscheidingen als indicatief, conjunctief en imperatief. Deze zijn echter minder belangrijk en bovendien bij corpustagging achterwege gelaten omwille van haalbaarheid.

2.3.2 Oppositie diachronie – synchronie (flexieverlies)

Het zou inefficiënt zijn om bijvoorbeeld een 20e-eeuwse tekst te taggen met een PoS-tagset met Oud- en Middelnederlandse kenmerken: in oudere fases komen nl. veel meer verschillende uitgangen voor dan nu.

Hoewel het de bedoeling is om een diachroon toepasbare tagset te ontwikkelen, hebben we toch besloten een 'periodemarkering' in te voeren, zodat bepaalde kenmerken van de tagset alleen toegepast worden als ze aansluiten bij het grammaticale systeem van (een tekst van) een bepaalde periode. Deze 'periodemarkering' wordt uitgevoerd door middel van de toepassingsrestrictie 'mandatory when applicable'. Ten behoeve van deze restrictie worden teksten eerst gescreend, o.a. om vast te stellen welke kenmerken van de tagset van toepassing zijn. Zo kunnen tekst en tagset en tools volledig op elkaar afgestemd worden.

Het is dus niet mogelijk één set paradigmakenmerken samen te stellen, waarin alle diachrone eigenschappen van een lemma en zijn woordvormen tegelijk een plaats krijgen. Wel kunnen we door de koppeling van woordvormen aan bepaalde gedateerde attestaties synchrone of periodespecifieke paradigma's of overzichten extraheren.

2.4Inrichting van de Master-PoS-tagset

Het diachrone uitgangspunt en de multifunctionele toepassingsmogelijkheden van GiGaNT maken de beschikbaarheid van een flexibel tagsetsysteem noodzakelijk. Voor GiGaNT is er dan ook een *mastertagset* ontwikkeld, waaruit diverse soorten tagsets en paradigma's kunnen worden afgeleid en waarbinnen de reeds vermelde opposities tot hun recht komen.

Daartoe is de tagset als volgt ingericht:

2.4.1 Beschrijving van de kenmerken

Bij het taggen wordt bij elk kenmerk aangegeven

- binnen welk onderdeel van de tagset (lemma, paradigma, vorm) of binnen welke module (morfologie, transcategorisatie) het functioneert;
- wat de status is (optioneel, mandatory, mandatory when applicable);
- of het voorkomt in de core of in de extensie;
- of het intern ("onder de motorkap") of extern (bij corpustagging) van toepassing is;
- op welk onderdeel van de woordsoort het van toepassing is (restricties).

Tabel 8: voorbeeld van restrictie

VRB (FINITENESS=FINITE)	**PARADIGMA: MANDATORY**		
MOOD	conjunctive	imperative	indicative
	conj	imp	ind

In de linker kopcel is vastgelegd dat binnen de thesaurus het kenmerk MOOD (met mogelijke waardes conjunctief, imperatief en indicatief) enkel van toepassing is op de finiete vormen van het werkwoord. De rechter kopcel geeft aan dat voor opname in het paradigma (de thesaurus), de waardes altijd toegekend moeten worden.

2.4.2 Filters

Een gefilterde tagset voor een specifieke periode of toepassing kan eenvoudig worden afgeleid uit de master PoS-tagset.

Mogelijke filters zijn:

- PoS- en kenmerkfilter ter verkrijging van een vereenvoudigde tagset voor een specifiek doel (retrieval, interne of externe verrijking) bijvoorbeeld alleen de hoofdwoordsoort
- locatie of periode filter; verwijder bijvoorbeeld voor een bepaalde periode niet relevante features uit de tagset (naamval voor modern Nederlands)

2.4.3 Haalbaarheid: onderspecificatie en neutralisatie

Omdat we in toepassingen kunnen werken met gefilterde tagsets hoeven we ons bij de vaststelling van de PoS-kenmerken voor de master-PoS-tagset niet te beperken tot kenmerken die voor alle toepassingen haalbaar zijn.

Daarnaast kan haalbaarheid van de invulling van een bepaald kenmerk ook bevorderd worden door onderspecificatie en neutralisatie.

2.4.3.1Onderspecificatie

Onderspecificatie houdt in: het bij elkaar nemen van posities die geen vormelijk verschil hebben. Doel van onderspecificatie is tijdswinst: als een betekenisverschil niet leidt tot een vormverschil, maar wel tot een unieke paradigmapositie, kan (maar moet niet) met onderspecificatie gewerkt worden.

Een voorbeeld: bij de 1e, 2e en 3e persoon meervoud van het werkwoord kunnen de 3 posities met een combiwaarde benoemd worden: person=1|2|3.

Meestal gaat het bij onderspecificatie om een algemener paradigmaniveau: vb. alle verledentijdsvormen enkelvoud of meervoud van een werkwoord kunnen samengenomen worden.

Een voorbeeld: bij de tegenwoordigetijdsvormen van werkwoorden met stam op –t kan je nooit aan de vorm zien of het om de 1e, 2e of 3e persoon enkelvoud gaat: ik|jij|hij eet.

Het is echter ook mogelijk om onderspecificatie op een subgroep, of zelfs op een individueel lemmaniveau toe te passen.

Het is mogelijk onderspecificatie alleen in een bepaalde fase toe te passen, waarbij dan op een later moment het ondergespecificeerde materiaal nauwkeurig benoemd kan worden.

Ten slotte zijn ook de dubbele woordsoortaanduidingen in GiGaNT een voorbeeld van onderspecificatie: AA (bijvoeglijk naamwoord / bijwoord), PD (pronoun / determiner).

2.4.3.2 Neutralisatie

Sommige verschillen worden systematisch geneutraliseerd, bijvoorbeeld de overeenkomst tussen geslacht en getal bij bijvoeglijke naamwoorden, die op een bepaald moment enkel nog van toepassing is op het enkelvoud, omdat de verschillen in het meervoud niet meer bestaan (= systematisch worden geneutraliseerd).

In GiGaNT wordt dit opgelost met de toekenning ‘mandatory when applicable’, waarmee wordt aangegeven dat als een kenmerk niet meer van toepassing is, dat dan het verschil niet meer herkenbaar en dus geneutraliseerd is.

2.4.3.3 Notatie van neutralisatie, onderspecificatie en onzekerheid

Deze worden gebruikt om niet-gespecificeerdheid en niet-toegekendheid eenduidig aan te geven.

waardes:

unspecified (unspec.) wordt toegevoegd om aan te geven dat een kenmerk niet nader gespecificeerd kan worden in de aangetroffen context

other geeft aan dat een van de bestaande waardes niet toereikend is, maar dat het niet nodig is een nieuwe waarde vast te leggen

tekens:

? kan toegevoegd worden achter een waarde waarover men onzeker is

/ geeft aan dat het niet uit te maken is welke van de waardes van toepassing is

+ geeft aan dat er nog geen waarde is toegekend maar dat dit wel moet

- geeft aan dat het niet mogelijk is een waarde toe te kennen

, 2 of meer waardes gescheiden door een komma wil zeggen dat 2 of meer waardes mogelijk zijn

Ø geeft aan dat een kenmerk niet van toepassing is

| geeft aan dat beide (of alle) waardes van toepassing zijn en dat we (bewust) niet verder willen specificeren

Haalbaarheid is een zeer krachtig criterium. Daarom is, naast deze mechanismes en tekens, ook de indeling van de tagset hierop gericht. Dit komt later nog aan bod.

2.4.4 Tagsetstandaarden en terminologie

Aangezien het lexicon bruikbaar moet zijn voor onderzoekers uit alle theorieën, is het belangrijk dat de PoS-kenmerknamen en -labels zoveel mogelijk theorieneutraal zijn en aansluiten bij gangbare opvattingen. Idealiter moet een tagset voldoen aan internationale standaarden, zoals EAGLES. Bij het INL zijn in de loop der projecten al enkele zeer specifieke tagsets ontwikkeld voor corpustagging en lexiconbouw (PAROLE, GTB) die aan deze standaard voldoen. GiGaNT is echter meer dan een corpustagset en zal dus ook nog andere kenmerken opnemen. Het is de bedoeling om de complete GiGaNT-tagset in ISOCAT te registreren om hem zo te gebruiken binnen de hele CLARIN-community.

Naast theoretische verschillen zijn er ook doelafhankelijke verschillen tussen PoS-stelsels, bijvoorbeeld voor gebruik in woordenboeken, in lexica en in tag- en retrievalomgevingen. De PoS-benamingen en -labels in GiGaNT zullen voor alle doelen dezelfde zijn of ten minste corresponderen.

Ook zullen veel types die in het lexicon opgenomen worden, afkomstig zijn uit zeer diverse bronnen: historische en moderne woordenboeken, corpora, lexica, datamining. Veel van deze types zullen (deels) al voorzien zijn van PoS-informatie. Het is dus nodig een vertaaltabel te maken voor de PoS-tags van alle bronnen van GiGaNT, zoals vb. de woordenboekbronnen: ONW, VMNW, MNW, WNT en ANW.

Ook voorzien we een protocol voor lexiconinvoer, waarin systematisch beschreven staat hoe te handelen bij invoer van extern materiaal.

2.4.5 Formaat

De GiGaNT-database is een relationele database met een XML-export naar TEI P5. Er wordt hierbij gebruik gemaakt van standaardbenamingen voor de velden en standaardwaardes bij de invulling van de velden. Voorwaarde voor het gebruik is transparantie en goede herkenbaarheid.

2.5 Indeling van de Master-PoS-tagset

We onderscheiden 3 hoofdgroepen van woordsoortkenmerken binnen de Master-PoS-tagset: lemmakenmerken, paradigmakenmerken en vormkenmerken.

2.5.1 Lemmakenmerken

Onder lemmakenmerken verstaan we de eigenschappen zoals die meestal in woordenboeken worden aangetroffen: woordsoortaanduiding; functionele kenmerken als transitiviteit, scheidbaarheid; enz.

Deze eigenschappen worden vanuit het oogpunt van haalbaarheid onderverdeeld in drie toepassingscategorieën: core, extension en expansion.

2.5.1.1 Core

Core- of kerneigenschappen worden voor de totale GiGaNT-populatie bijgehouden.

Het gaat om de volgende kenmerken:

Tabel 9: Corekenmerken

LEMMA_CORE	VAN TOEPASSING OP (woordsoort)
Conjugation type	VRB
Type	VRB, ADV, NUM, PD, CONJ, RES
Person	PD (pers/poss/refl)
Number	PD (pers/poss/refl)
Subtype	PD (d-p), CONJ

2.5.1.2 Expansion

Expansie-eigenschappen zijn kenmerken die te maken hebben met het aanmaken ("bijbakken") van vormen voor paradigma-aanvulling. Expansie wordt alleen toegepast op de vervoegbare en verbuigbare woordsoorten[14] van een open klasse[15].

Tabel 10: Expansionkenmerken

LEMMA_EXPANSION	VAN TOEPASSING OP (woordsoort)
Separability	VRB

2.5.1.3 Extension

Verrijkte data uit externe bronnen wordt overgenomen. Indien we vinden dat de eigenschappen van belang kunnen zijn voor GiGaNT, maar ze niet in het eigen materiaal zitten (of niet overeenkomen met eigen eigenschappen), worden ze gemarkeerd als extensie. Uiteraard worden alle originele kenmerken ook bewaard. Op die manier gaat er niets relevants verloren en blijven de gegevens beschikbaar voor retrieval.

Het gaat om de volgende kenmerken:

[14] Sommige verbuigbare woordsoorten komen toch voor met een onvolledig paradigma: vb. bijvoeglijke naamwoorden die wel attributief voorkomen maar toch niet verbuigbaar zijn.

[15] PD zijn bijvoorbeeld ook verbuigbaar, maar vormen een gesloten klasse. De meeste vormen zijn dus al bekend, zodat geen expansie nodig is.

Tabel 11: Extensionkenmerken

LEMMA_EXTENSION	VAN TOEPASSING OP (woordsoort)
Conjugation type - x	VRB
Verb class	VRB (main, strong)
Valency	VRB
Government	VRB, ADP (prep)
Gender	NOU-C, PD (pers/poss/pron)
Adverb subtype	ADV
Type	PD, ADP

2.5.2 Paradigmakenmerken

Niet alle woordsoorten hebben paradigmatische eigenschappen. Paradigmatische kenmerken komen voor bij open woordklassen met verbuigings-/vervoegingskenmerken: werkwoorden, zelfstandige naamwoorden en bijvoeglijke naamwoorden. Ook de van bijvoeglijke naamwoorden afgeleide bijwoorden en de voornaamwoorden en telwoorden hebben verbogen vormen.

Paradigmakenmerken worden vanuit het oogpunt van haalbaarheid, efficiency en inzichtelijkheid onderverdeeld in twee categorieën: *mandatory* en *mandatory when applicable* (MWA). Hiermee wordt de mate van verplichtheid aangegeven.

2.5.2.1Mandatory

Mandatory eigenschappen zijn kenmerken die verplicht moeten worden toegekend.
Het gaat om volgende kenmerken:

Tabel 12: Mandatory kenmerken

PARADIGMA - MANDATORY	VAN TOEPASSING OP (woordsoort)
Finiteness	VRB
Mood	VRB (fin)
Tense	VRB (fin/part)
Number agreement	VRB (fin)
Person agreement	VRB (fin)
Number	NOU-C
Degree	AA

2.5.2.2 Mandatory when applicable

MWA-eigenschappen zijn kenmerken die enkel toegekend moeten worden als dat mogelijk is, afhankelijk van het grammaticale systeem van de gescreende tekst.
Het gaat om volgende kenmerken:

Tabel 13: MWA-kenmerken

PARADIGMA - MANDATORY WHEN APPLICABLE	VAN TOEPASSING OP (woordsoort)
Case	VRB (inf/ger), AA, NUM, PD
Position	AA, NUM, PD

Number agreement	AA, NUM, PD (det)
Gender agreement	AA, NUM, PD (det)

2.5.3 Vormkenmerken

Vormtags worden alleen voorzien voor woordsoorten met een paradigma. De andere woordsoorten (onverbuigbare bijwoorden, voorzetsels, voegwoorden en residuals) zonder paradigma, krijgen als woordsoorttag een kale PoS-tag.

Vooralsnog worden de volgende vormkenmerken onderscheiden:

Tabel 14: vormkenmerken

VORMKENMERKEN	VAN TOEPASSING OP (woordsoort)
Prefix	VRB *(ge-* aan begin woord*)*
Infix	VRB *(ge-* na ander prefix voor stam*)*
Stem change (stamverandering)	VRB (sterke wwn.), NOU-C
Inflexion (inflectie)	VRB, NOU-C, ADJ, NUM, PD

2.5.4 Schrijfvorm

Schrijfwijze wordt binnen GiGaNT toegevoegd om aan te geven of de aangetroffen woordvorm een weergave is van het woord in overeenstemming met de gangbare schrijfwijze voor volledig uitgeschreven woorden, dan wel op andere wijze genoteerd is, zij het als afkorting, getal in cijfers of anderszins.

Tabel 15: Extra paradigmakenmerken

SCHRIJFVORM	VAN TOEPASSING OP (woordsoort)
Written form[16]	All PoS: VRB, NOU-C, NOU-P, AA, ADV, NUM, PD, ADP, CONJ, INT, RES, AFF
Written form – number system	NUM

2.6 Toepassing van de PoS-tagset

Het verrijken van een tekst gebeurt, zoals reeds eerder vermeld, aan de hand van 3 procedés: lemmatoekenning, PoS-tagging en koppelen van bibliografische gegevens[17] aan een woordvorm. Deze procedés kunnen gescheiden plaatsvinden, maar het is interessanter om ze, waar mogelijk en nodig, te combineren.

2.6.1 Attesteren

Voor het toevoegen van verrijking aan een tekst wordt een tool voor corpusgebaseerde lexiconbouw[18] gebruikt, waarin alle benodigde onderdelen samen aanwezig zijn:

- de tekst en de metadata bij die tekst (met lemma's en PoS-tags per token en met bibliografische gegevens)
- een meelopende lijst met alle (voor dat onderdeel) toegestane PoS-tags

[16] 'Geschreven vorm' is zowel een paradigmakenmerk als een vormkenmerk en verhuist mee binnen de mappingregel.

[17] Het koppelen van bibliografische gegevens aan een woordvorm gebeurt automatisch en hoeft dus niet te gebeuren tijdens de (semi-)handmatige lemma-en PoS-toekenning.

[18] http://www.digitisation.eu/tools/toolbox-for-lexicon-building/corpus-based-lexicon-tool-cobalt/

- een lijst van bestaande lemmata (een selectie of van alle woordenboeken samen)

2.6.1.1 Controle van automatische lemmatisering

Het koppelen van lemma en PoS-tags aan woordvormen gebeurt mogelijk al eerder automatisch door een tagger-lemmatiseerder, waardoor de taak van de attesteerder zich beperkt tot het controleren en eventueel verbeteren van het voorgestelde, en het definitief vastleggen.

2.6.1.2 Attestatie in historische woordenboeken

Soms zijn woordvormen in een woordenboekartikel fysiek gekoppeld aan het lemma waaronder ze te vinden zijn. Dit is vaak het geval in de zeer uitgebreide artikelen van de historische woordenboeken[19]. Dit is ook een vorm van attestatie, en ook hier moet de attesteerder controleren of het programma de juiste en alle woordvormen in het woordenboekartikel heeft gevonden en of alle gegevens correct zijn. De huidige inhoud van het GiGaNT lexicon is op deze wijze verkregen.

2.6.1.3 Attestatiecriteria: haalbaarheid en betrouwbaarheid

Ieder type met lemma en woordsoordtag of paradigmaplaats moet gekoppeld worden aan een niet-ambigu token in een tekst (bewijsplaats, attestatie). Het is duidelijk dat haalbaarheid hier dan afhankelijk is van de mate van verfijndheid van een paradigma: hoe meer verfijning, hoe meer attestaties er nodig zijn. Voor paradigma-attestaties gelden echter andere criteria[20].

2.6.1.3.1 Oudste en jongste vindplaats

Uit het oogpunt van haalbaarheid worden in eerste instantie de oudste en de jongste bewijsplaatsen gekoppeld, uit alle woordenboeken. De eerste lexiconpopulatie bestaat uit woordvormen van het WNT (attestaties van tussen 1500 en 2000), daarbij worden telkens de oudste en jongste bewijsplaatsen van de andere woordenboeken toegevoegd. Reeds gekoppelde bewijsplaatsen blijven in het lexicon behouden, waardoor het uiteindelijke totaal een goed diachroon beeld zal geven van de voorkomens van een woordvorm.

Bovendien kunnen na analyse zo nodig bijkomende bewijsplaatsen worden toegevoegd, bijvoorbeeld 1 per eeuw.

Daarnaast zal ook worden nagegaan welke paradigmaplaatsen gaten vertonen, zodat meer specifiek naar attestaties kan worden gezocht in bepaalde bronnen uit bepaalde periodes.

2.6.1.3.2 Betrouwbaarheid

Opgenomen bewijsplaatsen moeten aan een bepaalde betrouwbaarheid voldoen. Dit betekent dat we een voorkeur hebben voor betrouwbare diplomatische edities van historisch materiaal. Wanneer we toch minder betrouwbare, kritische edities gebruiken (zoals die bijvoorbeeld ten grondslag liggen aan het Middelnederlandsch Woordenboek), dient dat te worden vastgelegd. Dergelijke kwalificaties met betrekking tot betrouwbaarheid worden vastgelegd in de metadata.

[19] In het ONW en het VNMW vinden we bijvoorbeeld zo'n lijstjes met woordvormen voorin het artikel.
[20] Dit in tegenstelling tot bijvoorbeeld corpustagging, waarbij elke woordvorm behandeld moet worden.

2.6.2 behandeling van ambiguïteit

PoS en ambiguïteit: een overzicht van de problematiek

Ambiguïteit van de toegekende woordsoort komt voor als gevolg van verschillende verschijnselen. Deze worden hieronder opgesomd:

A1. Regelmatige transcategorisaties of woordsoortverschuivingen met een lemmaverschuiving als resultaat.
A2. Incidentele transcategorisaties met een lemmaverschuiving als gevolg.
A3. Regelmatige transcategorisatie zonder lemmaverschuiving (zie ook nulafleidingen: E1).

B1. Toevallige ambiguïteit tussen homografe lemma's met identieke woordsoort.
B2. Toevallige ambiguïteit tussen homografe lemma's met verschillende woordsoort.

C. Verschillen in lexicalisatie tussen woordenboeken en woordenlijsten.

D. Verschillen in woordsoortbenoeming en de opvatting daarover tussen woordenboeken en grammatica's.

E1. Ambiguïteiten op woordvormniveau, waar geen systematische transcategorisatierelatie bestaat tussen de woordsoorten (zoals in A1 of A2 wel het geval is), maar wel vaak een morfologische relatie (nulafleidingen).
E2. Interne woordvormambiguïteit binnen woordsoorten.

Workflowoplossingen

Ambiguïteiten worden gefaseerd aangepakt aan de hand van tijdelijke toekenning van PoS en/of lemma[21]. Hiervoor zijn verschillende mogelijkheden:

a) onderspecificatie (al dan niet tijdelijk)

Een woordvorm kan soms tijdelijk, soms blijvend een ambigu label krijgen. We gebruiken hiervoor de rechte streep[22] bij twee (of meer) lemmata, woordsoorten of een combinatie van beide.
Een speciaal geval van onderspecificatie zijn de koepellemmata: homonieme lemmata (lemmata die dezelfde lemmavorm en woordsoort dragen, maar op een andere manier van elkaar verschillen, vb. in betekenis, flexieklasse, geslacht enz.) die bij elkaar worden gezet onder één lemma[23]. Deze koepellemmata zijn voorbeelden van blijvende onderspecificatie en bieden vele mogelijkheden. Zo kan vb. een variant die slechts bij een van de homoniemen voorkomt direct bij het specifieke lemma worden ondergebracht[24], terwijl bij bepaalde samenstellingen niet direct een keuze tussen de lemmata hoeft te worden gemaakt[25].

b) ambiguïteit in de vormgestuurde tags en mappingrecords

De vormgestuurde woordsoortbeschrijving is in feite een hybride label, omdat morfologische kenmerken (vb. flexie) hier morfosyntactische kenmerken (vb. person, number, mood) vervangen. Het is mogelijk dat een vormgebaseerde beschrijving minder precieze informatie bevat, waardoor

[21] Dit leidt echter niet noodzakelijk tot het oplossen van alle ambiguïteiten.
[22] Cf. onderspecificatie bij Master-PoS-tagset.
[23] Voorbeeld: koepellemma *plegen*: *plegen1* (*placht (*VRB (fin, past, sg))) vs. *plegen2* (*pleegde (*VRB (fin, past, sg))).
[24] Een voorbeeld: de oude vorm *har* kan alleen voorkomen bij *heer* 'vorst' en niet bij *heer* 'leger'.
[25] Een voorbeeld: het al dan niet koppelen van *tuinbank* en *handelsbank* aan een bepaald lemma *bank*.

er vaak sprake zal zijn van 'many-to-many'-mapping: één vormgestuurde beschrijving leidt tot meerdere paradigmatische beschrijvingen[26] en omgekeerd.

Hoe ziet de oplossing van onderspecificatievormen en tijdelijke ambigue toekenningen er nu concreet uit?

In eerste instantie krijgt een nieuwe woordvorm verschillende mogelijke lemma- en PoS-toekenningen. In de vervolgprocedures (automatisch dan wel handmatig) worden dan een aantal toekenningen of relaties weggenomen: het lemma blijft op hetzelfde niveau behouden, maar de relatie met de woordvorm wordt als het ware 'weggehapt'.

2.6.3 behandeling van varianten

2.6.3.1Opzet

GiGaNT neemt zoals gezegd alle morfosyntactische varianten op. Daardoor kan men een overzicht krijgen van welke vormen in welke periode of taalvariëteit mogelijk zijn en/of daadwerkelijk gebruikt worden.

GiGaNT behandelt alle varianten gelijk, vb. met betrekking tot opname in het lexicon en attesteringsrecht[27]. Wel worden criteria vastgesteld om te bepalen of bepaalde woordvormen een eigen paradigmaplaats krijgen, dan wel of ze als variant gekoppeld worden aan een reeds bestaande paradigmapositie.

Het zal uiteindelijk mogelijk zijn om bij een bepaald lemma, per periode en per taalvariant, een apart subparadigma af te leiden.

2.6.3.2 Plaatsing van varianten in de structuur

Reguliere varianten worden wel gemapt.

Incidentele morfosyntactische gevallen van variatie vallen uiteraard buiten de mappingregels, maar krijgen wel een plaats in het paradigma.

2.6.3.3 Criteria

Hoe wordt bepaald wanneer een woordvorm een morfosyntactische variant (variant van een bepaalde positie in het paradigma) is, dan wel een eigen paradigmapositie krijgt?

1. In normale omstandigheden wordt een unieke positie van een woordvorm in een paradigma bepaald door een vormverschil dat met een betekenis-/functieverschil gepaard gaat, op basis van paradigmatische eigenschappen. Elke unieke positie in een paradigma wordt dus weerspiegeld in een unieke beschrijving, die op minimaal één waarde verschilt van een andere beschrijving.

[26] Een voorbeeld: de woordvorm fietsen word vormelijk geanalyseerd als 'fiets plus –en'. Paradigmatisch zijn er echter verschillende mogelijkheden: NOU-C(pl), VRB(fin,ind,pl,1,2,3) of VRB(inf). Zie ook sectie 3) bij 2.2 Soorten PoS in GiGaNT. [Zo was het niet bedoeld. De vormtag kiest altijd wel de 'primaire' woordsoort, maar niet noodzakelijkerwijs de positie in het paradigma].

[27] Echter, ten behoeve van vb. spelling, moet er echter wel een canonieke vorm voor een bepaalde periode worden aangeduid.

2. Als een vormverschil echter niet leidt tot een betekenis-/functieverschil, wordt de woordvorm als een variant beschouwd en dus gekoppeld aan een reeds bestaande paradigmapositie[28].
3. In GiGaNT is het incidenteel echter mogelijk dat een vormverschil wel leidt tot een betekenis-/functieverschil, maar toch niet tot een verschillende paradigmapositie. Een voorbeeld daarvan is *goed(e)* in 'het goede kind' vs. 'een goed kind'.

De definitie van een variant binnen GiGaNT is dus: een woordvorm is een variant van een andere woordvorm als een vormverschil niet leidt tot een verschil in paradigma(positie).

3 Woordsoorten: algemeen

Woorden en hun gebruik zijn continu in beweging. Het benoemen van de woordsoorten (categorisering) is een belangrijke schakel in de desambiguering van woordvormen.

De meeste woorden hebben door de eeuwen heen een 'primaire' woordsoort gekregen: de lexicale woordsoort. Daarnaast wordt echter ook functioneel gebruik herkend en benoemd in woordenboeken en grammatica's. Als dergelijk functioneel gebruik genoeg voorkomt, wordt de functionele woordsoort 'gelexicaliseerd': het woord krijgt er dan nog een extra (neven)woordsoort bij in de beschrijving in grammatica of woordenboek.

Een voorbeeld: een infinitief kan ook voorkomen als zelfstandig naamwoord: de woordsoort is dan VRB en NOU-C. Dit is bovendien een productief verschijnsel: elke infinitief kan namelijk gebruikt worden als zelfstandig naamwoord. Niet elke infinitief wordt echter ook als zelfstandig naamwoord in de verschillende materialen teruggevonden.

In GiGaNT kiezen we ervoor om zowel de lexicale als de functionele woordsoort op te nemen.

We onderscheiden in GiGaNT de volgende woordsoorten[29] (lemma_PoS):

Tabel 26: overzichtstabel woordsoorten

WOORDSOORT	**afkorting**
Werkwoord	VRB
Zelfstandig naamwoord: soortnaam	NOU-C
Zelfstandig naamwoord: eigennaam	NOU-P
Bijvoeglijk naamwoord / Bijwoord	AA
Bijwoord	ADV
Telwoord	NUM
Pronoun / Determiner	PD
Voorzetsel	ADP
Voegwoord	CONJ
Tussenwerpsel (interjectie)	INT
Residual	RES
Affix (en andere gebonden morfemen)	AFF

Enkele opmerkingen hierbij:

- Elke woordsoort kan zowel als 'primaire' woordsoort als als nevenwoordsoort voorkomen. Als nevenwoordsoort gaat het dan om de resultante van een transcategorisatieregel of van een morfologisch procedé.

[28] Een voorbeeld: *loopt* in 'ik loopt' is een variant van *loop* in 'ik loop' en krijgt dus dezelfde paradigmatische tag: VRB (pres, ind, sg, 1).

[29] Zie voor de subcategorieën van de 'primaire' woordsoorten het kenmerk TYPE bij de verschillende woordsoorten en de tabellen 9 en 11.

- Vanwege de ambiguïteit en/of het gebrek aan vormelijk onderscheid tussen bijvoeglijk en bijwoordelijk gebruik van de basisvormen van bijvoeglijke naamwoorden hebben we de koepelwoordsoort AA in het leven geroepen. Om vergelijkbare redenen worden ook de lidwoorden behandeld bij de voornaamwoorden.

Hierna worden behandeld per woordsoort (hoofdstuk 4 e.v.):
- afbakening & indeling
- lemmakenmerken
- paradigmakenmerken
- varianten
- transcategorisatieregels

3.1 Kenmerken

Elk kenmerk word in de volgende hoofdstukken per woordsoort behandeld, ook als het van toepassing is op meer dan één woordsoort. De waardes voor elk kenmerk kunnen nl. voor iedere woordsoort anders zijn (vb. TYPE).

Enige uitzondering hierop vormt het kenmerk schrijfvorm (zie 2.5.4), dat voor op één na alle woordsoorten dezelfde waardes heeft.

Tabel 17: mogelijke waardes voor schrijfvorm voor alle woordsoorten behalve het telwoord

POS					
WRITTEN FORM	abbreviated	truncated	form variant	misspelled	other
afkorting	abbr	trunc	fv	mis	oth

Tabel 18: mogelijke extra waardes voor schrijfvorm voor woordsoort telwoord

POS					
WRITTEN FORM	letters	digit	roman	mixed-digit	mixed-roman
afkorting	let	dig	rom	mix-dig	mix-rom

3.2 Transcategorisatie

Bij elke woordsoort worden zowel de regelmatige als de incidentele transcategorisaties vermeld. De regelmatige staan in witte cellen, de incidentele in grijze.

Tabel 19: overzichtstabel transcategorisatie

MAIN POS	**TARGET POS**	**VOORBEELD**
VRB	NOU-C (n, sg)	het geven, het schrijven

VRB(part)	AA [ADJ]	het gedeeld genoegen, een kirrend kind
VRB(part)	AA [ADV]	zij liep huppelend de trap op
VRB(part, past)	CONJ (sub)	(aan)gezien, uitgezonderd
VRB(part, pres)	ADP	aangaande, betreffende, hangende
VRB(part, past)	ADP	gezien
NOU-C	ADV	plankgas, retour; maandag, donderdagavond; begin, eind
NOU-C	ADJ (infl=o)	een formica tafel, een hardboard kast, een tricot trui
NOU-C	INT	hemel!
NOU-C	ADP	richting (“naar”)
AA	NOU-C	mijn lief
AA	NOU-C	het rood, het geel; het Vlaams, het Amsterdams; de glossy
AA	ADP	inwendig, overeenkomstig, relatief
ADV	NOU-C	achteruit; buitenspel; andante, allegro; oost, west; alias; het hier en nu, het waarom, het hoe
ADV	CONJ	telkens
ADV	INT	toe (nou)!
ADV	ADJ	het vake gebruik, een dakpansgewijze opslag; apart; een (on)affe zin; de ane televisie, een uite kachel
NUM(card)	NOU-C	een zes, een zeven, een acht
NUM(ord)	NOU-C	de eerste, de tweede, de derde
NUM(ord)	ADV	eerst
PD	NOU-C	een hij, een zij, de grote onbekende men; de haren, de zijnen, de meesten
PD	ADV	al, allemaal, alles, geen, zodanig
PD	ADJ	niet, zelf
PD	CON	noweder (“geen van beiden > noch”)
PD	ADP	wes (“van wie; wat ook” > tot ”(tijdsbep.)”)
ADP	NOU-C	(de) voor(s) en tegen(s), (de) pro(‘s) en contra(‘s)
ADP	ADV	achter, boven; buiten, te, achterin, achterop, binnenin
ADP	CONJ	met, na, naar, niettegenstaande, om (te), tot, zonder, voor
ADP	ADV	ook
CONJ	NOU-C	de althansen van de dagvaarding, een maar weten te vinden
CONJ	ADV	althans, naargelang
CONJ	ADP	mits, tenzij
CONJ	ADJ	aangezien (“in aanzien”)
INT	NOU-C	ach, aha, adieu, hallo, kiekeboe, vaarwel
INT	ADV	koest

INT	ADJ	oké
AFF	NOU-C	een isme, de ultra

4 Werkwoord

tag: VRB

4.1 Afbakening

Op lemmaniveau onderscheiden we:
- zwakke en niet-zwakke werkwoorden
- hoofd-, hulp- en koppelwerkwoorden
- transitieve, intransitieve, onpersoonlijke en reflexieve werkwoorden
- al dan niet scheidbare werkwoorden

Daarnaast zijn er twee typeringen specifiek in gebruik voor ONW-trefwoorden: het onderscheid tussen verschillende werkwoordklassen en het gebruik van naamvallen bij specifieke voorzetsels.

Op paradigmatisch niveau kunnen werkwoorden worden gedefinieerd op het vlak van finietheid, wijs, tijd, getal, persoon en naamval.

4.2 Lemmakenmerken

We geven een overzicht van de kenmerken die kunnen voorkomen in core, extension en expansion, met hun mogelijke waardes.

4.2.1 Conjugation type

tag: Conj type

Dit geeft aan om welk soort werkwoord het gaat en welke vormen daarbij kan verwachten.

Volgende waardes[30] zijn mogelijk:

VRB	**LEMMA-CORE**	
CONJ TYPE	weak	non-weak
	wk	nwk

VRB	**LEMMA-EXTENSION**		
CONJ TYPE-X	weak	strong	irregular
	wk	st	irreg
	leven	sterven	bakken, zijn, hebben, kunnen, durven

[30] Er zijn verschillende classificaties mogelijk; dit is degene die voor GiGaNT werd gekozen.

4.2.2 Verb class

tag: VRB class

Dit geeft aan tot welke klasse het sterke werkwoord behoort. Werkwoordklasse wordt alleen onderscheiden voor de ONW-periode, omdat we die beschikbaar hebben in het woordenboek.

Volgende waardes zijn mogelijk:

VRB (main, strong)	**LEMMA-EXTENSION**								
VRB class	1	2a	2b	3a	3b	4	5	6	7
	grijpen	liegen	druipen	beginnen	belgen	nemen	bidden	dragen	lopen

4.2.3 Type

tag: Type

Dit geeft aan om welk type werkwoord het gaat.

Volgende waardes zijn mogelijk:

VRB	**LEMMA-CORE**		
TYPE	main	auxiliary	copular
	mai	aux	cop
	lezen, kopen, fietsen	hebben, zijn, worden	zijn, worden, blijven

4.2.4 Valency

tag: Val

Dit geeft de valentie van het werkwoord aan.

Volgende waardes zijn mogelijk:

VRB (TYPE=MAIN)	**LEMMA-EXTENSION**			
VAL	transitive	intransitive	impersonal	reflexive
	trans	intr	imp	refl
	maken	zitten	regenen	zich vergissen

4.2.5 Government

tag: Gov

Dit geeft aan of het werkwoord een voorzetsel of een bepaalde naamval vereist. Government wordt alleen onderscheiden voor de ONW-periode. Dit kenmerk lijkt echter zinvol toe te passen voor werkwoorden uit ander woordenboeken, vb. MNW.

Volgende waardes zijn mogelijk:

VRB (TYPE=MAIN)	LEMMA-EXTENSION			
GOV	+ preposition	+ genitive	+ dative	+ accusative
	prep	gen	dat	acc
	bigān mit *(begaan met)*	*wunderon *(zich verbazen over)*	antduon *(deur openen voor iem.)*	antfān *(ontvangen)*

4.2.6 Separability

tag: Sep

Dit geeft aan of het werkwoord al dan niet scheidbaar[31] is.

Volgende waardes zijn mogelijk:

VRB	LEMMA-EXPANSION	
SEP	yes	no
	y	n

4.3 Paradigmakenmerken

We geven een overzicht van de kenmerken die kunnen voorkomen (zowel 'mandatory' als 'mandatory when applicable').

4.3.1 Finiteness

tag: Finiteness

Dit geeft aan of het om een al dan niet finiete vorm van het werkwoord gaat.

Volgende waardes zijn mogelijk:

VRB	PARADIGMA: MANDATORY		
FINITENESS	finite	infinitive/gerund	participle
	fin	inf/ger	part

4.3.2 Mood

tag: Mood

Dit geeft om welke wijs van het werkwoord het gaat. Dit kenmerk komt alleen voor bij finiete vormen.

[31] Scheidbaarheid heeft in GiGaNT geen invloed op het paradigma: de los geschreven vormen worden gezien als variant van de aaneengeschreven vormen.

Volgende waardes zijn mogelijk:

VRB (FINITENESS=FIN)	PARADIGMA: MANDATORY		
MOOD	conjunctive	imperative	indicative
	conj	imp	ind

4.3.3 Tense

tag: Tense

Dit geeft aan in welke tijd het werkwoord staat. Dit kenmerk geldt alleen voor finiete vormen en deelwoorden.

Volgende waardes zijn mogelijk:

VRB (FINITENESS=FIN, PART)	PARADIGMA: MANDATORY	
TENSE	present	past
	pres	past

4.3.4 Number agreement

tag: NA

Dit geeft de overeenkomst in getal met het onderwerp aan die het werkwoord uitdrukt. Dit kenmerk geldt alleen voor finiete vormen.

Volgende waardes zijn mogelijk:

VRB (FINITENESS=FIN)	PARADIGMA: MANDATORY	
NA	singular	plural
	sg	pl

4.3.5 Person agreement

tag: PA

Dit geeft de overeenkomst in persoon met het onderwerp aan. Dit kenmerk geldt alleen voor finiete vormen.

Volgende waardes zijn mogelijk:

VRB (FINITENESS=FIN)	PARADIGMA: MANDATORY		
PA	1	2	3

4.3.6 Case

tag: Case

Dit geeft aan of het om een specifieke naamval van het werkwoord als zelfstandig naamwoord gaat. Case wordt speciaal voor de verbogen vorm van infinitief/gerundium, die niet transcategoriseert naar een zelfstandig naamwoord, bij het werkwoord opgenomen. Van het gerundium komen alleen de genitief en de datief voor (vb. in het (vroeg)Middelnederlands;

vergelijk met hedendaagse uitdrukkingen als: een uur gaans, het is menens, prijzenswaardig, tot ziens[32]).

Volgende waardes zijn mogelijk:

VRB (FINITENESS=INF/GER)	PARADIGMA: MANDATORY WHEN APPLICABLE	
CASE	genitive	dative
	gen	dat
	slapens	te gevene

4.4 Transcategorisatieregels

MAIN POS: VRB	TARGET POS	VOORBEELD
VRB	NOU-C (n, sg)	het geven, het schrijven
VRB(part)	AA [ADJ]	het gedeeld genoegen, een kirrend kind
VRB(part)	AA [ADV]	zij liep huppelend de trap op
VRB(part, past)	CONJ (sub)	(aan)gezien, uitgezonderd
VRB(part, pres)	ADP	aangaande, betreffende, hangende
VRB(part, past)	ADP	gezien

- De (onverbogen vorm van de) infinitief kan voorkomen als zelfstandig naamwoord, al dan niet voorafgegaan door een lidwoord, *vb. (het) leven, (het) zijn*.
- Verbogen deelwoordvormen vallen onder de koepelwoordsoort AA (bijwoord/bijvoeglijk naamwoord). Attributieve verbogen deelwoordvormen zijn daarbij altijd bijvoeglijke naamwoorden.
- Onverbogen deelwoordvormen worden verwerkt als werkwoorden, met uitzondering van de attributieve, die als bijvoeglijke naamwoorden worden verwerkt[33].

5 Zelfstandig naamwoord: soortnaam

tag: NOU-C

5.1 Afbakening

Op lemmaniveau onderscheiden we:

- mannelijke, vrouwelijke en onzijdige woorden

Op paradigmatisch niveau kunnen zelfstandige naamwoorden worden gedefinieerd op het vlak van getal en naamval.

[32] A.M. Duinhoven, Middelnederlandse syntaxis dl 2, 193 [1997]

[33] Er zijn gevallen van voltooide deelwoorden die onverbogen zijn, predicatief gebruikt worden en waar de link met het werkwoord niet altijd (meer) duidelijk is; deze categorie behoeft nog onderzoek; is het bijvoorbeeld haalbaar om ze te onderscheiden?

Diminutieven krijgen een eigen ingang (lemma en woordsoort) en worden behandeld als reguliere afleidingen, *vb. etentje, leventje, eentje.*

Verkortingen krijgen een eigen ingang, *vb. homo, airco.*

5.2 Lemmakenmerken

We geven een overzicht van de kenmerken die kunnen voorkomen in core, extension en expansion, met hun mogelijke waardes.

5.2.1 Gender

tag: Gender

Uiteraard is het mogelijk enkel het onderscheid te maken tussen de-woorden en het-woorden. In GiGaNT is echter gekozen voor de driedeling in geslacht: mannelijk – vrouwelijk – onzijdig (en combinaties daarvan). Slecht één voorbeeld van een voordeel van deze keuze: er ontstaan zo geen problemen bij afgeleide verkleinwoorden met een geslacht dat verschilt van het hoofdwoord, *vb. het mannetje (NOU-C(neu)) – de man (NOU-C(masc)).*

Bij zelfstandige naamwoorden heeft geslacht geen paradigmatische gevolgen[34].

Volgende waardes zijn mogelijk:

NOU-C	**LEMMA-EXTENSION**						
GENDER	masculine	feminine	neuter	masculine/ neuter	masculine/ feminine[35]	feminine/ neuter	unknown
	m	f	n	m/n	m/f	f/n	m/f/n

5.3 Paradigmakenmerken

We geven een overzicht van de kenmerken die kunnen voorkomen (zowel 'mandatory' als 'mandatory when applicable').

5.3.1 Number

tag: Number

Dit geeft aan wat het getal van het zelfstandig naamwoord is.

Volgende waardes zijn mogelijk:

NOU-C	**PARADIGMA: MANDATORY**	
NUMBER	singular	plural
	sg	pl

[34] Als er toch paradigmatische verschillen terug te voeren zijn op geslacht, dan worden die behandeld als variant: zie hiervoor bij 'varianten'.

[35] In sommige woordenboeken wordt hiervoor de term 'gemeenslachtig' gebruikt, vb. WNT.

5.3.2 Case

tag: Case

Dit geeft aan in welke naamval het zelfstandig naamwoord staat. Hoewel er in het ONW en het VMNW een vocativus wordt opgenomen in de beschrijving van de flexie, wordt dit niet overgenomen en behandelen we de vormen onder de nominatief zolang er geen overtuigende voorbeelden worden aangetroffen die een vormverschil aangeven.

Volgende waardes zijn mogelijk:

NOU-C	**PARADIGMA: MANDATORY WHEN APPLICABLE**			
CASE	nominative	genitive	dative	accusative
	nom	gen	dat	acc

5.4 Varianten

Wanneer een vormelijk verschil tussen twee woorden niet tot enig verschil in betekenis of gebruik leidt, beschouwen we ze als varianten met identieke paradigmapositie. Sommige zelfstandige naamwoorden hebben verschillende mogelijke meervouden (vb. zoon – zonen - zoons). Al deze meervoudsvormen hebben echter een gelijke paradigmapositie en dus een identieke tag: NOU-C(pl).

Is er wel een met het vormverschil corresponderend verschil in betekenis of gebruik, dan is er meestal wel sprake van een andere paradigmatische tag. Echter, als een zelfstandig naamwoord met een dubbel geslacht (mannelijk en vrouwelijk of vrouwelijk en onzijdig), in het enkelvoud verschillende uitgangen krijgt verbonden aan het verschil in geslacht, worden deze als varianten behandeld. Een voorbeeld: Vroegmiddelnederlands bederve (‘bederf’), met vormen als bederve, bederf in de accusatief singularis.

5.5 Transcategorisatieregels

MAIN POS: NOU-C	**TARGET POS**	**VOORBEELD**
NOU-C	ADV	plankgas, retour; maandag, donderdagavond; begin, eind
NOU-C	ADJ (infl=o)	een formica tafel, een hardboard kast, een tricot trui
NOU-C	INT	hemel!
NOU-C	ADP	richting (=naar)

6 Zelfstandig naamwoord: eigennaam

tag: NOU-P

6.1 Afbakening

Aangezien we eigennamen willen onderscheiden in het gewoon lexicon, hebben we ze in een aparte ‘woordsoort’ ondergebracht.

In het Named Entity-lexicon krijgen eigennamen op dit moment twee woordsoorten, *vb. Sinterklaas* → woordsoorten NOU-C (soortnaam) en NOU-P (eigennaam).

Voor eigennamen worden persoonsnamen, plaatsnamen en organisaties onderscheiden. We baseren ons voor de toekenning hiervan op de "1999 Named Entity Recognition Task Definition." Version 1.4, August 27, 1999, van Nancy Chinchor, Erica et.al., de annotatierichtlijnen van de CoNLL shared tasks on Named Entity Recognition.

7 Bijvoeglijk naamwoord / Bijwoord

tag: AA

7.1 Afbakening

De dubbele woordsoort bijvoeglijk naamwoord/bijwoord werd in het leven geroepen voor bijvoeglijke naamwoorden die ook als bijwoord gebruikt (kunnen) worden.

Bijvoeglijke naamwoorden die nooit als bijwoord voorkomen, worden ook in dit hoofdstuk behandeld (zie 7.2).

Bijwoorden die niet afgeleid zijn van bijvoeglijke naamwoorden worden in het volgende hoofdstuk behandeld.

De verkleinwoorden van bijwoorden, *vb. kalmpjes, frisjes, liefjes, netjes* worden wel in dit hoofdstuk behandeld, omdat hiervan toch ook bijvoeglijk gebruik wordt vastgesteld (omdat de s-uitgang niet meer door alle gebruikers als adverbiaal wordt ervaren). Hetzelfde stellen we ook vast bij andere bijwoorden op -s (*vb. bloot(s)hoofdse, blootsvoetse, binnentijdse, binnen(s)mondse, binnenshuize, eerstdaagse*) en ook bij bijwoorden op -gewijze (*vb. dakpansgewijze*).

Het bijwoord wordt aanzien als een paradigmapositie binnen de woordsoort bijvoeglijk naamwoord/bijwoord (AA), zie 7.3.2.

7.2 Paradigmakenmerken

We geven een overzicht van de kenmerken die kunnen voorkomen (zowel 'mandatory' als 'mandatory when applicable').

7.2.1 Degree

tag: Degree

Volgende waardes zijn mogelijk:

AA	**PARADIGMA: MANDATORY**		
DEGREE	positive	comparative	superlative
	pos	comp	sup

Merk wel dat bijvoeglijke naamwoorden die ontstaan door transcategorisatie van voltooide deelwoorden op -en[36], in het moderne Nederlands altijd maar één vorm hebben (en dus geen trappen van vergelijking).

[36] Vb.*gevallen*.

7.2.2 Position

tag: Position

Bijvoeglijke naamwoorden kunnen prenominaal, postnominaal[37], predicatief en adverbiaal gebruikt voorkomen. In het vroege en ook het dialectische taalgebruik hebben ze op die posities (soms) wel flexie (buigings-e), hoewel dit laatste niet frequent optreedt.

Volgende waardes zijn mogelijk:

AA	**PARADIGMA: MANDATORY WHEN APPLICABLE**			
POSITION	prenominal	postnominal	predicative	adverbial
	prenom	postnom	pred	adv

7.2.3 Number agreement

tag: NA

Dit geeft de overeenkomst in getal aan. Attributieve bijvoeglijke naamwoorden in oude fases[38] erven het getal van het zelfstandig naamwoord dat ze bepalen.

Volgende waardes zijn mogelijk:

AA	**PARADIGMA: MANDATORY WHEN APPLICABLE**	
NA	singular	plural
	sg	pl

7.2.4 Gender agreement

tag: GA

Dit geeft de overeenkomst in geslacht[39] aan.

Volgende waardes zijn mogelijk:

AA	**PARADIGMA: MANDATORY WHEN APPLICABLE**					
GA	masculine	feminine	neuter	masculine/ neuter	masculine/ feminine	feminine/ neuter
	m	f	n	m/n	m/f	f/n

7.2.5 Case

tag: Case

Dit geeft aan in welke naamval[40] het bijvoeglijk naamwoord staat.

[37] In het Oud- en Vroegmiddelnederlands.

[38] Vb. ONW. Ondanks dat er geen verschil in vorm is in het meervoud, wordt meervoud wel toegekend in de woordenboeken: zowel bij attributief (pre- en postnominaal) als bij predicatief gebruik.

[39] In het ONW wordt gender toegepast ongeacht de positie van het bijvoeglijk naamwoord (prenominaal, postnominaal of predicatief) én ondanks dat postnominale en predicatieve bijvoeglijke naamwoorden daardoor niet een andere vorm hebben. Ook aan meervoudsvormen wordt gender toegekend, ook al is er geen verschil in vorm.

[40] In het ONW wordt ook van pluralisvormen altijd de naamval bepaald.

Volgende waardes zijn mogelijk:

AA	PARADIGMA: MANDATORY WHEN APPLICABLE			
CASE	nominative	genitive	dative	accusative
	nom	gen	dat	acc

Predicatieve bijvoeglijke naamwoorden staan in de regel[41] in de nominatief.

7.3 Varianten

Het onderscheid dat ontstaat door gebruik van een bijvoeglijk naamwoord in een onbepaalde woordgroep met een onzijdig zelfstandig naamwoord als kern (*vb. een groot huis* versus *het grote huis*) en dat terug te leiden is op het onderscheid sterk/zwak, wordt in GiGaNT behandeld als variant. Het gaat immers om varianten in dezelfde paradigmapositie.

Hetzelfde geldt voor het onderscheid bij *vb. een groot man - een grote man* of *een goed man - een goede man*: ook deze krijgen een identieke tag en worden behandeld als varianten. Het verschil in betekenis heeft hier dus geen belang.

7.4 Transcategorisatieregels

MAIN POS	TARGET POS	VOORBEELD
AA	NOU-C	mijn lief
AA	NOU-C	het rood, het geel; het Vlaams, het Amsterdams; de glossy
AA	ADP	inwendig, overeenkomstig, relatief

8 Bijwoord

tag: ADV

8.1 Afbakening

Er is een grote groep van bijwoorden die niet in een transcategorisatierelatie tot een bijvoeglijk naamwoord staan, en dus enkel als bijwoord voorkomen:

- de 'echte' bijwoorden: *vb. echter, nooit, achtereenvolgens*
- voorzetselcombinaties: *vb. onderuit, vanbuiten*
- de pronominale bijwoorden: *vb. daarover*

Op lemmaniveau maken we een onderscheid tussen verschillende types en subtypes van bijwoorden.

[41] In de volgende aanhaling vinden we **unbewollene** in de acc. sg. in predicatief gebruik: Scona bistu, friundina min, wanda thu thie candidam uestem, quam in baptismo accepisti, **unbewollene** behaldost. Mooi ben je, mijn vriendin, omdat je het witte kleed dat je bij je doop hebt aangenomen, onbevlekt houdt. LW 101,05 Egmond, Holland, ca. 1100.

8.2 Lemmakenmerken

We geven een overzicht van de kenmerken die kunnen voorkomen in core en extension, met hun mogelijke waardes.

8.2.1 Type

tag: Type

Dit geeft aan om welk type bijwoord het gaat.

Volgende waardes zijn mogelijk:

ADV	**LEMMA-CORE**		
TYPE	genuine	conjunctional	pronominal
	gen	conj	pron

De voegwoordelijke bijwoorden en de voorzetselbijwoorden, zullen in een aparte bewerking gesubcategoriseerd worden.

8.2.2 Adverb subtype

tag: Subtype

Dit geeft voor elk bijwoord het subtype aan.

Volgende waardes zijn mogelijk:

ADV (PRON)	**LEMMA-EXTENSION**		
SUBTYPE			
Nederlands	**Engels**	**afkorting**	**voorbeeld**
persoonlijke	personal	pers	ik, wij, mij, ons
aanwijzend	demonstrative	dem	dit, dat, deze, die
betrekkelijk	relative	rel	welk(e), dat, wat, die
onbepaald	indefinite	indef	wanen “waar vandaan ook maar“, nergens
vragend	interrogative	inter	welk(e), wie, wat

Onbepaalde voornaamwoordelijke bijwoorden worden als aparte delen behandeld (en dus niet als multiword expressions), omdat ze alleen gescheiden voorkomen.

8.3 Varianten

Zoals hierboven vermeld, zijn de echte bijwoorden onverbuigbaar. Toch treedt soms flexie op, vb. bij ‘heel’ : *een/de heel/hele mooie jongen.* ‘Hele’ wordt in die gevallen behandeld als variant van ‘heel’.

8.4 Transcategorisatieregels

MAIN POS: ADV	TARGET POS	VOORBEELD
ADV	NOU-C	achteruit; buitenspel; andante, allegro; oost, west; alias; het hier en nu, het waarom, het hoe
ADV	CONJ	telkens
ADV	INT	toe (nou)!
ADV	ADJ	het vake gebruik, een dakpansgewijze opslag; apart; een (on)affe zin; de ane televisie, een uite kachel

9 Telwoord

tag: NUM

9.1 Afbakening

Op lemmaniveau onderscheiden we hoofdtelwoorden en rangtelwoorden. Er wordt geen aparte categorie voorzien voor breukgetallen: deze worden in GiGaNT gewoon tot de telwoorden gerekend. Het gaat immers om een combinatie van een hoofdtelwoord met een rangtelwoord.

Op paradigmatisch niveau kunnen telwoorden worden gedefinieerd op het vlak van positie, overeenkomst in geslacht, getalsovereenkomst en naamval.

Woorden als *veel, weinig, alle, geen, enige, sommige* enz. worden soms wel onbepaalde telwoorden genoemd. In GiGaNT worden zij gecategoriseerd[42] als quantifiers binnen de categorie pronoun/determiner. Ze worden verder behandeld in hoofdstuk 10.

9.2 Lemmakenmerk

Enkel het kenmerk type komt voor in de core.

9.2.1 Type

tag: Type

Dit geeft aan of het om een hoofdtelwoord of een rangtelwoord gaat.

Volgende waardes zijn mogelijk:

NUM	LEMMA-CORE	
TYPE	cardinal	ordinal
	card	ord

9.3 Paradigmakenmerken

We geven een overzicht van de kenmerken die kunnen voorkomen (zowel 'mandatory' als 'mandatory when applicable').

[42] In aansluiting bij o.a. *Modern Grammar of Dutch en A Communicative Grammar of English* (Leech & Svartvik).

9.3.1 Position

tag: Position

Dit geeft aan of het telwoord prenominaal, postnominaal[43], predicatief voorkomt.

Volgende waardes zijn mogelijk:

NUM	PARADIGMA: MANDATORY WHEN APPLICABLE			
POSITION	prenominal	postnominal	predicative	other[44]
	prenom	postnom	pred	oth
			zij zijn twee	één, twee, drie; ik weeg tachtig

9.3.2 Gender agreement

tag: GA

Dit geeft de overeenkomst in geslacht aan.

Volgende waardes zijn mogelijk:

NUM	PARADIGMA: MANDATORY WHEN APPLICABLE		
GA	masculine	feminine	neuter
	m	f	n

9.3.3 Number agreement

tag: NA

Dit geeft de overeenkomst in getal aan.

Volgende waardes zijn mogelijk:

NUM	PARADIGMA: MANDATORY WHEN APPLICABLE	
NA	singular	plural
	sg	pl

9.3.4 Case

tag: Case

Dit geeft aan in welke naamval het telwoord staat.

Volgende waardes zijn mogelijk:

NUM	PARADIGMA: MANDATORY WHEN APPLICABLE

[43] In het Oud- en Vroegmiddelnederlands.

[44] Schijnbaar predicatief is het idiomatische gebruik van bepaalde hoofdtelwoorden ter aanduiding van iemands leeftijd, *vb. morgen wordt mijn opa tachtig* (=tachtig jaar oud). Dit gebruik wordt ook behandeld als 'other'.

CASE	nominative	genitive	dative	accusative
	nom	gen	dat	acc

9.3.5 Written form numeral system

tag: Wf_numsys

Dit is een extra paradigmakenmerk : het gaat nl. niet om een kenmerk dat standaard in de vormleer bij een paradigma hoort, maar om een kenmerk dat we binnen GiGaNT toevoegen om variatie goed te kunnen vastleggen. Specifiek gaat het hier om een schrijfmatig kenmerk.

Volgende waardes zijn mogelijk:

NUM	PARADIGMA-ADDITION AND FORM				
WF_NUMSYS	letters	digit	roman	mixed-digit	mixed-roman
	let	dig	rom	mix-dig	mix-rom

9.4Transcategorisatieregels

MAIN POS: NUM	TARGET POS	VOORBEELD
NUM(card)	NOU-C	een zes, een zeven, een acht
NUM(ord)	NOU-C	de eerste, de tweede, de derde
NUM(ord)	ADV	eerst

10 Pronoun / Determiner

tag: PD

10.1 Afbakening

In deze categorie zijn alle voornaamwoorden opgenomen. Daarnaast worden ook lidwoorden getagd als een subcategorie van PD.

Zoals reeds vermeld aan het begin van hoofdstuk 9 worden de zgn. 'onbepaalde telwoorden' (*vb. veel, weinig, alle, geen, enige, sommige*) eveneens opgenomen in deze categorie, en wel als quantifiers. Ze hebben ongeveer dezelfde functie als telwoorden, maar geven een vage(re) hoeveelheid of aantal aan.

Op lemmaniveau onderscheiden we in eerste instantie de kenmerken: type voornaamwoord, persoon, geslacht, getal en functie. Bijkomend wordt van lidwoorden gezegd of ze al dan niet bepaald zijn.

Op paradigmatisch niveau wordt in deze categorie gedefinieerd op het vlak van positie, naamval, overeenkomst in geslacht en getal.

10.2 Lemmakenmerken

We geven een overzicht van de kenmerken die kunnen voorkomen in core, extension en expansion, met hun mogelijke waardes.

10.2.1 Type

tag: Type

Dit geeft aan om welk type voornaamwoord het gaat.

Volgende waardes zijn mogelijk:

PD	**LEMMA-CORE**							
TYPE	personal	d-pronomina	indefinite	w-pronomina	reflexive	reciprocal	possessive	quantifier
	pers	d-p	indef	w-p	refl	recip	poss	quant

Enkele opmerkingen hierbij:
- d-pronomina: (aanwijzende) deiktische pronoun/determiners, *vb. de, die, dat, deze, dit, dusdanige, gene, zodanige*
- w-pronomina: vragend-betrekkelijke[45] pronoun/determiners, *vb. wie, welke, hoedanig*
- quantifiers: zgn. onbepaalde telwoorden, vb. *veel, weinig, alle, geen, enige, sommige*

10.2.2 Uitgebreid type

tag: Type

Volgende waardes zijn mogelijk:

PD	**LEMMA-EXTENSION**								
TYPE	personal	demon-strative	relative	in-definite	reflexive	reciprocal	possessive	inter-rogative	other
	pers	dem	rel	indef	refl	recip	poss	int	oth

10.2.3 Subtype lidwoord

tag: Subtype_Art

Dit geeft aan of het om een bepaald of onbepaald lidwoord gaat.

Volgende waardes zijn mogelijk:

PD (D-P)	**LEMMA-CORE**	
SUBTYPE: ART	indefinite	definite
	indef	def

10.2.4 Person

tag: Person

Enkel de persoonlijke, bezittelijke en reflexieve[46] voornaamwoorden kennen het kenmerk 'persoon'.

[45] Deze kunnen zowel vragend als betrekkelijk gebruikt worden.
[46] Alleen 'zich' en de vormen met '-zelf' (vb. *mezelf, jezelf, zichzelf*) zijn uitsluitend te taggen als 'reflexief' .

Volgende waardes zijn mogelijk:

PD (PERS, REFL, POSS)	LEMMA-CORE		
PERSON	1	2	3

10.2.5 Gender

tag: Gender

Bij voornaamwoorden duidt deze tag het natuurlijk geslacht[47] (zijnde dat van de referent) aan.

Volgende waardes zijn mogelijk:

PD (PRON, POSS, PERS)	LEMMA-EXTENSION			
GENDER	masculine	feminine	neuter	non applicable
	m	f	n	na

10.2.6 Number

tag: Number

Dit geeft aan wat het getal van het voornaamwoord is.

Volgende waardes zijn mogelijk[48]:

PD (PERS, REFL, POSS)	LEMMA-CORE	
NUMBER	singular	plural
	sg	pl

10.3 Paradigmakenmerken

We geven een overzicht van de kenmerken die kunnen voorkomen (zowel 'mandatory' als 'mandatory when applicable').

10.3.1 Position

tag: Position

Een aantal voornaamwoorden kunnen zowel zelfstandig als bijvoeglijk gebruikt worden. Enkel de persoonlijke en reflexieve voornaamwoorden kennen geen bijvoeglijk gebruik. Het bezittelijk voornaamwoord en het lidwoord komen alleen bijvoeglijk voor.

Determiners kunnen prenominaal, postnominaal en predicatief[49] voorkomen.

[47] d.i. het tegengestelde van het congruentiegeslacht.

[48] *zich, gij/ge* en *u* worden zijn zowel enkelvoud als meervoud (sg + pl).

[49] Zie bijvoorbeeld: '**Sulich** is min drut, ande her is ouch min friund - thaz wizzet, ir iuncfrouwon!': 'Zo is mijn geliefde en hij is ook mijn vriend; dat moeten jullie goed weten, jonge vrouwen!' - LW 097,01 Egmond, Holland, ca. 1100.

Volgende waardes zijn mogelijk:

PD	PARADIGMA: MANDATORY WHEN APPLICABLE			
POSITION	prenominal	postnominal	predicative	pronominal
	prenom	postnom	pred	pron

10.3.2 Case

tag: Case

Dit geeft aan of we te maken hebben met een bepaalde naamval.

Volgende waardes zijn mogelijk:

PD	PARADIGMA: MANDATORY WHEN APPLICABLE				
CASE	nominative	genitive	dative	accusative	other
	nom	gen	dat	acc	oth

Alle voornaamwoorden kunnen verbogen worden, behalve 'zich'.

10.3.3 Gender agreement

tag: GA

Determiners kunnen in geslacht[50] overeenkomen met het zelfstandig naamwoord waar ze bij staan.

Volgende waardes zijn mogelijk:

PD (FUNCTION=DET)	PARADIGMA: MANDATORY WHEN APPLICABLE		
GA	masculine	feminine	neuter
	m	f	n

10.3.4 Number agreement

tag: NA

Dit geeft de overeenkomst in getal aan van de determiners.

Volgende waardes zijn mogelijk:

PD (FUNCTION=DET)	PARADIGMA: MANDATORY WHEN APPLICABLE	
NA	singular	plural
	sg	pl

[50] o.a. in ONW

10.4 Transcategorisatieregels

MAIN POS: PD	TARGET POS	VOORBEELD
PD	NOU-C	een hij, een zij, de grote onbekende men; de haren, de zijnen, de meesten
PD	ADV	al, allemaal, alles, geen, zodanig
PD	ADJ	niet, zelf
PD	CON	noweder (“geen van beiden > noch”)
PD	ADP	wes (“van wie; wat ook” > tot ”(tijdsbep.)”)

11 Voorzetsel

tag: ADP

11.1 Afbakening

In deze categorie zijn alle voorzetsels opgenomen. Voorzetsels kennen geen verbuiging; er zijn dus geen paradigmatische kenmerken.

Naast enkelvoudige voorzetsels, zijn ook complexe of samengestelde voorzetsels in deze categorie opgenomen, *vb. niettegenstaande, dienaangaande, dientengevolge*.

11.2 Lemmakenmerken

We geven een overzicht van de kenmerken die kunnen voorkomen in core en extension, met hun mogelijke waardes.

11.2.1 Type

tag: Type

Dit geeft aan of het om een voor- of achterzetsel gaat.

Volgende waardes zijn mogelijk:

ADP	LEMMA-EXTENSION		
TYPE	preposition	postposition	circumposition[51]
	pre	post	circ
	in, uit	beneven	onder...door, tussen...door

11.2.2 Government

tag: Gov

Dit geeft aan of het voorzetsel een bepaalde naamval vereist.

[51] zgn. ‘onderbroken voorzetsel’.

Volgende waardes zijn mogelijk:

ADP (PREP)	LEMMA-EXTENSION		
GOV	+ genitive	+ dative + dative or accusative	+ accusative
	gen	dat dat/acc	acc

11.3 Varianten

Voorzetsels met schijnbare verbuiging worden behandeld als varianten, aangezien het niet gaat om systematische paradigmatische verschijnselen. Een voorbeeld is *ombe* 'om'.

11.4 Transcategorisatieregels

MAIN POS: ADP	TARGET POS	VOORBEELD
ADP	NOU-C	(de) voor(s) en tegen(s), (de) pro('s) en contra('s)
ADP	ADV	achter, boven; buiten, te, achterin, achterop, binnenin
ADP	CONJ	met, na, naar, niettegenstaande, om (te), tot, zonder, voor
ADP	ADV	ook

12 Voegwoord

tag: CONJ

12.1 Afbakening

In deze categorie zijn alle voegwoorden opgenomen. Voegwoorden kennen geen verbuiging; er zijn dus geen paradigmatische kenmerken.

12.2 Lemmakenmerken

We geven een overzicht van de kenmerken die kunnen voorkomen in core en extension, met hun mogelijke waardes.

12.2.1 Type

tag: Type

Dit geeft aan of het om een nevenschikkend of een onderschikkend voegwoord gaat.

Volgende waardes zijn mogelijk:

CONJ	LEMMA-CORE	
TYPE	coordinating	subordinating
	coor	sub

12.2.2 Subtype

tag: Subtype

Dit geeft voor elk onderschikkend voegwoord het subtype aan.

Volgende waardes zijn mogelijk:

CONJ	LEMMA-CORE		
SUBTYPE	negative (ontkennend/negatief)	neg	neware
	comparative (vergelijkend)	comp	alse
	explicative (explicatief)	expl	zoals
	relative (relatief)	rel	dan
	quality (van hoedanigheid)	qual	als

12.3 Varianten

Voegwoorden met schijnbare verbuiging worden behandeld als varianten, aangezien het niet gaat om systematische paradigmatische verschijnselen. Een voorbeeld is *danne* 'dan'.

12.4 Transcategorisatieregels

MAIN POS: CONJ	TARGET POS	VOORBEELD
CONJ	NOU-C	de althansen van de dagvaarding, een maar weten te vinden
CONJ	ADV	althans, naargelang
CONJ	ADP	mits, tenzij
CONJ	ADJ	aangezien (=in aanzien)

13 Tussenwerpsel (interjectie)

tag: INT

13.1 Afbakening

In deze categorie zijn alle tussenwerpsels opgenomen, *vb. ach, foei, jaja*. Deze vertonen geen specifieke lemma- of paradigmatische kenmerken.

13.2 Transcategorisatieregels

MAIN POS: INT	TARGET POS	VOORBEELD
INT	NOU-C	ach, aha, adieu, hallo, kiekeboe, vaarwel
INT	ADV	koest
INT	ADJ	oké

14 INT

15 Residual

tag: RES

15.1 Afbakening

Dit is geen traditionele woordsoort, maar In deze categorie vinden we de woordvormen terug die niet tot een van de andere woordsoorten behoren. Het gaat als het ware om een restcategorie voor vormen waarmee we anders niets kunnen of willen doen: woordvormen die in een corpus voorkomen en waarbij je dus een of andere codering nodig hebt, maar waarbij nog moet worden uitgemaakt wat als ingang in het lexicon zal worden opgenomen (aangezien het lexicon ook wordt gebruikt bij corpustagging).

Concreet vinden we hierin vooral formules, uitheemse woorden en symbolen terug.

Residuals vertonen geen specifieke paradigmatische kenmerken.

15.2 Lemmakenmerken

RES	LEMMA-CORE			
TYPE	formula	foreign	symbol	unknown
	form	for	sym	unkn
	E=mc²	virtually[52]	$, %, @	(vb. iedere corrupte woordvorm)

16 Affix (en andere gebonden morfemen)

tag: AFF

16.1 Afbakening

Een affix is een niet-zelfstandig element dat voor of achter een grondwoord (of een stam) wordt gevoegd om zo een nieuw woord te vormen.

Affixen worden verder besproken in de morfologische module.

[52] Ieder woord uit een vreemde taal dat niet als leenwoord kan worden opgevat.

16.2 Transcategorisatieregels

MAIN POS: AFF	TARGET POS	VOORBEELD
AFF	NOU-C	een isme, de ultra