PART OF SPEECH TAGGING EN LEMMATISERING VAN HET D-COI CORPUS

Frank Van Eynde

juli 2005

Centrum voor Computerlinguïstiek
K.U.Leuven

VOORWOORD

De in het D-coi corpus opgenomen teksten worden verrijkt met diverse lagen van annotatie. Van die annotatielagen worden er in deze tekst twee behandeld, namelijk de lemmatisering en de part of speech tagging. Voor het aanbrengen van deze annotaties wordt gebruik gemaakt van dezelfde richtlijnen en dezelfde tagset als voor het Corpus Gesproken Nederlands (1998-2004). Het enige verschil betreft de toevoeging van twee tags voor speciale tokens, d.w.z. SPEC(afkorting) en SPEC(symbool), zie 2.12.

Het eerste hoofdstuk stelt de belangrijkste kenmerken van de twee annotatielagen voor en introduceert de gebruikte notatie. Het tweede hoofdstuk biedt een volledig en gedetailleerd overzicht van de tagset, met inbegrip van de lemmatiseringsrichtlijnen. Het derde hoofdstuk plaatst de tagset in het ruimere perspectief van de EAGLES aanbevelingen. Het vierde hoofdstuk biedt een formele samenvatting van de tagset, met vermelding van voorbeelden.

Voor hun mondelinge en schriftelijke commentaar op de vorige versies van deze tekst dank ik Hans van Halteren, Walter Daelemans, Jakub Zavrel, Ineke Schuurman, Lisanne Teunissen, Richard Piepenbrock, Nelleke Oostdijk, Ard Sprenger, Gosse Bouma, Antal van den Bosch, Ton van der Wouden, Truus Kruyt, Peter-Arno Coppen, Vincent Vandeghinste en Peter Dirix.

Frank Van Eynde

Leuven, juli 2005

1 INLEIDING

De eerste stap in de taalkundige ontsluiting van het corpus behelst de toekenning van tags en lemmata aan de orthografische eenheden, d.w.z. woorden, leestekens en bijzondere symbolen. Hierbij is ernaar gestreefd om aan de volgende vereisten te voldoen.

- De tags moeten die informatie bevatten die traditioneel wordt geassocieerd met wat men onder 'woordontleding' verstaat; er moet dus aansluiting worden gezocht bij de indelingen van voor algemeen gebruik bedoelde grammatica's, zoals de *Algemene Nederlandse Spraakkunst* (ANS 1997). Wat hiermee bedoeld wordt, is in detail uitgewerkt in het stuk over de tagset, zie hoofdstuk 2.

- De tagset moet zoveel mogelijk aansluiten bij de heersende internationale standaarden. In de praktijk wordt vooral aansluiting gezocht bij de *EAGLES* standaard, zie hoofdstuk 3.

- Aan elk token wordt één en slechts één tag toegekend. Dit impliceert dat er *hoogstens* één tag mag worden toegekend en dat er dus gedisambigueerd moet worden. Dit impliceert ook dat er *minstens* één tag moet worden toegekend, en dat er dus voorzieningen moeten getroffen worden voor minder gebruikelijke tokens, zoals afgebroken en onverstaanbare 'woorden' en woorden uit dialecten of vreemde talen.

- Aan elk token wordt één en slechts één lemma toegekend. Ook hier dient er dus waar nodig gedisambigueerd te worden.

- De tags moeten een geschikte basis bieden voor de hogere niveaus van taalkundige annotatie, zoals de syntactische analyse en de identificatie van meerwoordcombinaties (lexicologische koppeling).

- De notatie moet overzichtelijk, compact en makkelijk te lezen zijn.

Gelet op de omvang van het corpus is ervoor gekozen om zowel de tagging als de lemmatisering zoveel mogelijk te automatiseren. Op basis van een in 1999 uitgevoerde evaluatie (Zavrel en Daelemans 1999) en rekening houdend met beschikbaarheidsbeperkingen is besloten om voor de lemmatisering gebruik te maken van MBLEM (Memory-based Lemmatizer, Tilburg) en om voor de tagging gebruik te maken van de TIMBL combi-tagger. Typisch voor een combi-tagger is dat hij door systematische vergelijking van de resultaten

van een aantal afzonderlijk werkende taggers tot een resultaat komt dat accurater is dan elk van de afzonderlijke taggers kan bereiken. In het geval van de TIMBL combi-tagger gaat het om een combinatie van de resultaten van TnT (Trigram 'n Tags, Saarbrücken), MBT (Memory-based Tagging, Tilburg), een maximum entropy tagger en een Brill tagger.

Om een hoge kwaliteit te garanderen wordt het resultaat van de automatische tagging en lemmatisering manueel gecontroleerd en waar nodig gecorrigeerd.

1.1 DE LEMMATISERING

De lemmatisering betreft de herleiding van geflecteerde vormen tot een basisvorm; in het geval van de werkwoorden is dat de infinitief, in het geval van de andere woordsoorten de stam. Indien de stam geen bestaand woord is, wordt de basisvorm geïdentificeerd met een geflecteerde vorm; dat is o.m. het geval bij inherent meervoudige substantieven (*hersenen, mazelen*) en bij substantieven die alleen een diminutiefvorm hebben (*akkefietje, ootje*). Hoe de lemmatisering precies is gebeurd, wordt in hoofdstuk 2 voor elke woordsoort afzonderlijk besproken.

De lemmatisering geschiedt woord voor woord. De delen van een scheidbaar samengesteld werkwoord, zoals in *ze geeft prachtig verluchte boeken uit*, krijgen dus elk afzonderlijk een lemma. De herleiding van *geven* en *uit* tot *uitgeven* geschiedt op het niveau van de lexicologische koppeling. Dat is ook het niveau waarop meerledige eigennamen zoals *Den Haag* als één geheel worden herkend.

1.2 DE TAGSET

De tagging betreft de toekenning van lexicale en morfo-syntactische kenmerken aan woordvormen in een specifieke context. Tot de relevante kenmerken behoort minimaal de **woordsoort**. Welke andere kenmerken ertoe behoren is vastgelegd in de tagset.

1.2.1 Tagsets voor het Nederlands

Voor de tagging van Nederlandse teksten zijn reeds verschillende tagsets ontwikkeld; voor een overzicht, zie Schuurman (1998) en de appendix van Zavrel en Daelemans (1999).

De meeste van die tagsets worden gekenmerkt door een beperkte **granulariteit**, in de zin dat ze minder dan 50 verschillende tags onderscheiden. De INL-tagset bijv., die o.m. door de CORRie tagger gebruikt wordt, beperkt zich tot de onderscheiding van de woordsoorten en bevat dus geen bijkomende morfo-syntactische kenmerken. Ook de tagsets van KEPER (24 tags), D-Tale (45 tags) en Xerox (49 tags) maken relatief weinig onderscheidingen. Zulke weinig gedifferentieerde tagsets kunnen voor sommige toepassingen volstaan, maar voor de annotatie van het CGN corpus is gekozen voor een hogere graad van granulariteit, omdat de tags voor een groot deel van het corpus de enige vorm van taalkundige ontleding zijn (syntactische analyse gebeurt slechts voor 10 %van het corpus). Tagsets die wel een groot aantal onderscheidingen maken en in dat opzicht aantrekkelijker zijn voor CGN, zijn die van WOTAN-2 en PAROLE.

Naast een hoge granulariteit is de mate van aansluiting bij bestaande **standaarden** een belangrijk criterium. Dat is enerzijds belangrijk omwille van de herkenbaarheid in een internationale context en anderzijds omwille van de toegankelijkheid van de annotaties voor de eindgebruikers. Om het eerste te garanderen is aansluiting gezocht bij de EAGLES-aanbevelingen; om het tweede te garanderen is aansluiting gezocht bij de ANS-97. Dat laatste heeft twee voordelen: (1) voor verdere toelichting kunnen we doorverwijzen naar de relevante secties in de ANS, en (2) aangezien de ANS voor een breed publiek is geschreven en zoveel mogelijk gebruik maakt van vertrouwde begrippen en onderscheidingen, vergroot de kans dat de CGN-tags ook voor niet-taalkundigen interpreteerbaar zijn.

Een derde selectiecriterium heeft betrekking op de begeleidende **documentatie**. Een tagset die zich beperkt tot een opsomming van de gebruikte woordsoorten en onderscheidingen is weinig geschikt voor CGN, omdat die geen garantie biedt voor een eenduidige interpretatie van de gegevens. Om bruikbaar te zijn moet de tagset ook aanwijzingen en richtlijnen bevatten die een eenduidige toekenning en interpretatie van de verschillende tags mogelijk maken. De noodzaak van zulke documentatie wordt bovendien groter naarmate de granulariteit van de tagset toeneemt.

Rekening houdend met deze criteria leken de tagsets van WOTAN-2 en PAROLE het meest in aanmerking te komen voor gebruik in CGN. Het probleem was echter dat ze bij de aanvang van het CGN-project nog volop in ontwikkeling waren. De WOTAN-2 tagset bestond slechts in een voorlopige versie, die tijdens de startfase van het CGN-project nog verschillende keren is bijgesteld, en voor de PAROLE tagset was geen begeleidende documentatie beschikbaar. Om die reden is er in de loop van 1998-1999 een nieuwe

tagset ontworpen voor gebruik in het CGN-project. De belangrijkste bronnen van inspiratie hierbij waren EAGLES, ANS-97 en de voorlopige versies van WOTAN-2. Vanwege de verwantschap van de taal en vooral vanwege de gelijkaardigheid van de uitgangspunten is ook de Stuttgart-Tübingen Tagset voor het Duits (STTS) een interessante bron gebleken.

Om de tagset op zijn praktische bruikbaarheid te testen is in de lente van 1999 een manueel annotatie-experiment uitgevoerd. De resultaten van dit experiment zijn uitvoerig beschreven in Zavrel (1999) en hebben de inrichting van de CGN-tagset mede bepaald.

1.2.2 De inrichting van de CGN tagset

In de CGN tagset worden de woordsoorten geassocieerd met twee groepen van morfo-syntactische kenmerken. De eerste groep bestaat uit **lexicale kenmerken**, zoals de indeling in nevenschikkende en onderschikkende voegwoorden of het onderscheid tussen bepaalde en onbepaalde lidwoorden. De tweede groep bestaat uit kenmerken die morfologische variatie coderen, zoals het getal bij substantieven of de trappen van vergelijking bij adjectieven. Tot de **morfologische kenmerken** behoren minimaal die welke na lemmatisering niet in het lemma zelf gereflecteerd zijn. Het substantief *tafels* bijv. wordt met het lemma *tafel* geassocieerd en de informatie i.v.m. de getalswaarde moet bijgevolg in een afzonderlijk feature worden opgenomen. De in de CGN-tagset opgenomen morfologische kenmerken zijn die welke inflectionele variatie coderen (getal, werkwoordstijden, naamval, e.d.), aangevuld met een aantal kenmerken die woordsoortbehoudende derivatie coderen, zoals de diminutiefvorming bij substantieven. Of een kenmerk lexicaal of morfologisch is, hangt soms van de woordsoort af. Getal bijv. is een morfologisch kenmerk bij de substantieven, maar een lexicaal kenmerk bij de voornaamwoorden. Welke kenmerken de tags precies bevatten, wordt in hoofdstuk 2 voor elke woordsoort afzonderlijk uitgespeld. Wat niet is opgenomen zijn semantische kenmerken. Het onderscheid tussen concrete en abstracte substantieven bijv. wordt niet gemaakt.

De tagging geschiedt net zoals de lemmatisering woord voor woord. Een vaste verbinding zoals *te goeder trouw* wordt dus behandeld als een sequentie van een voorzetsel, een adjectief en een substantief. Die werkwijze heeft gevolgen voor de inrichting van de tagset. Zo moet bijv. bij de substantieven en de adjectieven voorzien worden dat ze datiefvormen kunnen hebben, aangezien zulke vormen relatief vaak voorkomen in vaste verbindingen, zoals *ter plaatse, van harte, in feite*. Een ander gevolg is de noodzaak van

een bijzondere tag voor delen van eigennamen. In *Den Haag* bijvoorbeeld wordt aan beide woorden de tag SPEC(deeleigen) toegekend, zie 2.12. De identificatie van de eigennaam als één geheel geschiedt op het niveau van de lexicologische koppeling. Om diezelfde reden zijn er geen valentiepatronen in de tagset opgenomen. Een scheidbaar samengesteld werkwoord heeft immers niet noodzakelijk dezelfde valentie als het werkwoord dat er de kern van is: *uitgeven* bijv. is transitief, terwijl *geven* ditransitief is, en *uitlachen* is transitief, terwijl *lachen* intransitief is. Het toekennen van valentiepatronen heeft dus pas zin na de lexicologische koppeling.

Bij de toekenning van de tags is systematisch gekozen voor een morfo-syntactisch perspectief. De getalswaarden bijv. worden niet geïnterpreteerd in conceptuele maar in morfo-syntactische termen; substantieven als *boel* en *aantal* zijn dus enkelvoudig. De keuze van het morfo-syntactische perspectief is ook van belang bij de interpretatie van de woordsoorten. Een woord als *maandag* bijv. wordt vaak in bijwoordelijke functies gebruikt, zoals in *ik heb hem maandag nog gesproken*, maar is qua woordsoort een substantief, en wordt bij de tagging dan ook niet als bijwoord maar als substantief behandeld.

Aan de dialectwoorden, d.w.z. de woorden die bij de orthografische transcriptie voorzien zijn van het symbool '*d', worden geen volledige tags toegekend, maar alleen de woordsoort en de voor die woordsoort relevante TYPE features, zoals NTYPE voor substantieven en CONJTYPE voor voegwoorden. De reden voor het weglaten van de andere features is dat die vooral dienen voor het coderen van morfologische informatie en dat juist de morfologie van de dialecten sterk kan verschillen van die van de standaardtaal. De voornaamwoorden in combinaties als *dieën boek* en *dieë vent* bijvoorbeeld laten zich niet beschrijven in termen van de voor de standaardtaal gemaakte onderscheidingen. Om diezelfde reden vindt er bij de lemmatisering van dialectwoorden geen reductie tot een basisvorm plaats, maar wordt het lemma gelijkgesteld aan de woordvorm zelf.

1.2.3 Ambiguïteit en onderspecificatie

Ambiguïteit is een relatieve notie: een woordvorm is niet ambigu op zichzelf, maar alleen met betrekking tot een systeem van onderscheidingen. De vorm *snel* bijv. is POS-ambigu als men verschillende woordsoorten toekent aan het attributieve gebruik in *een snel paard* en het adverbiale in *dat paard loopt snel*. Het is daarentegen niet POS-ambigu als men dit ziet als verschillende gebruiken van hetzelfde adjectief. Het aantal en het soort van

ambiguïteiten waarmee de tagger geconfronteerd wordt is dus in belangrijke mate een functie van de keuzes die worden gemaakt bij de inrichting van de tagset.

Bij die inrichting is een onderscheid gemaakt tussen occasionele en systematische ambiguïteiten. De eerste betreffen individuele woorden: het woord *bij* bv. kan zowel een substantief als een voorzetsel zijn. De tweede betreffen groepen van woorden en hebben meestal betrekking op verschillen in gebruiksmogelijkheden: het feit dat het adjectief *snel* zowel attributief als adverbiaal gebruikt wordt is een eigenschap die het met vele andere adjectieven gemeen heeft. De tagset is zo gedefinieerd dat occasionele ambiguïteiten worden erkend en opgelost, terwijl de postulering van systematische ambiguïteiten zoveel mogelijk vermeden wordt.

De CGN tagset laat een zekere mate van **onderspecificatie** toe. Wat hiermee bedoeld is, kan worden toegelicht aan de hand van het feature NAAMVAL.

```
NAAMVAL = nominatief, oblique, genitief, datief.
```

Voorbeelden van deze waarden zijn resp. *ik, jij, hij, wij* voor de nominatief, *mij, jou, hem, ons* voor oblique, *'s avonds, wiens, elkaars* voor de genitief en *ter plaatse, in koelen bloede, in der minne* voor de datief. Zoals bekend is het onderscheid tussen nominatief en oblique alleen relevant voor de voornaamwoorden: bij de substantieven en de adjectieven wordt het systematisch geneutraliseerd. Om te vermijden dat in zulke gevallen een systematische ambiguïteit ontstaat, postuleren we een intermediaire waarde ('standaard') die generaliseert over nominatief en oblique. Iets soortgelijks doen we voor de genitief en de datief: omdat bij de adjectieven de genitief- en datiefvormen systematisch dezelfde vorm hebben, introduceren we een intermediaire waarde die over beide generaliseert ('bijzonder'). De partitie ziet er dan als volgt uit.

```
NAAMVAL = standaard (nominatief, oblique), bijzonder (genitief,
datief).
```

In deze hiërarchie staat 'standaard' voor de vormen zonder en 'bijzonder' voor de vormen met een naamvalssuffix. Per woordsoort wordt dan bepaald tot op welk niveau men de onderscheidingen wil maken.

In sommige gevallen is het ook nuttig om over een waarde te beschikken die over alle specifieke waarden generaliseert. De substantieven bijv. hebben een GENUS feature met de waarden 'zijdig' en 'onzijdig'. Die onderscheiding kan in de meeste gevallen eenduidig gemaakt worden, maar er is een klein aantal substantieven die beide genuswaarden kunnen hebben (*de/het riool, de/het*

filter); wanneer de context het niet toelaat om aan zulke substantieven een specifieke GENUS waarde toe te kennen, laten we het gebruik toe van de waarde 'genus'.

```
GENUS = genus (zijdig, onzijdig).
```

Hoewel het gebruik van onderspecificatie voordelen biedt, is er bij het bepalen van de tagset spaarzaam gebruik van gemaakt. Daar zijn twee redenen voor. Primo, bij overdadig gebruik verliezen de tags aan informativiteit. Secundo, als men voor eenzelfde feature zowel atomaire als intermediaire waarden in de tags toelaat, neemt het aantal mogelijke tags drastisch toe.

1.2.4 De definitie van de CGN tagset

Formeel gesproken is de CGN tagset een sixtupel <A,W,P,D,I,T>, waarin A een verzameling is van attributen, W van waarden, P van partities, D van declaraties, I van implicaties en T van tags. Features zijn die combinaties van attributen en waarden die voldoen aan de door de partities opgelegde beperkingen. Tags zijn lijsten van features die voldoen aan de door de declaraties en de implicaties opgelegde beperkingen.

Attributen en waarden. Een feature bestaat uit een attribuut en een waarde. Voor de eerste gebruiken we hoofdletters en voor de tweede kleine letters of cijfers; de twee worden gescheiden d.m.v. een gelijkheidsteken.

```
GENUS = onzijdig
```

De waarden zijn atomair, d.w.z. ze bestaan niet op hun beurt uit attribuut-waarde paren.

Zowel voor de attributen als voor de waarden worden Nederlandse termen gebruikt. Attributen die alleen relevant zijn voor een specifieke woordsoort krijgen een prefix; LWTYPE bijv. wordt alleen toegekend aan lidwoorden en NTYPE alleen aan substantieven.

Partities. Voor elk attribuut wordt bepaald welke de mogelijke waarden zijn. Dat gebeurt in de vorm van een partitie. Die kan de vorm hebben van een simpele opsomming, maar ook van een hiërarchie met intermediaire waarden.

```
[P03] NTYPE = soortnaam, eigennaam.
[P07] NAAMVAL = standaard (nominatief, oblique),
      bijzonder (genitief, datief).
```

Voor het gemak van de verwijzing worden de partities genummerd. Als een feature voor diverse woordsoorten relevant is, krijgt ze bij die verschillende woordsoorten hetzelfde nummer.

Declaraties. Een tag is een lijst van lexicale en morfo-syntactische kenmerken. Welke kenmerken relevant zijn voor welke woordsoorten wordt bepaald in declaraties als

[D08] <POS = werkwoord> $\Longrightarrow$ <WVORM>

Kenmerken die slechts voor sommige leden van een categorie relevant zijn, worden niet met de woordsoort als geheel geassocieerd maar met specifiekere kenmerken. Bij de werkwoorden bijv. zijn persoon, tijd en wijze wel voor de persoonsvormen relevant, maar niet voor de infinitieven en de deelwoorden.

[D09] <WVORM = persoonsvorm> $\Longrightarrow$ <PVTIJD, PVAGR>

De declaraties zijn net als de partities genummerd.

Als een waarde gepartitioneerd is in één of meer sub-waarden, dan erven de sub-waarden de morfo-syntactische kenmerken die geassocieerd zijn met de hogere waarde. Als de voornaamwoorden het feature NAAMVAL hebben bijv. dan hebben ook de determiners dat feature. Complexere vormen van inheritance, zoals multiple inheritance en default inheritance, zijn niet gebruikt.

Tags. Tags zijn geordende lijsten van features. Ze bevatten een POS feature en alle morfo-syntactische features die voor de betreffende woordsoort gedeclareerd zijn.

<POS = voegwoord, CONJTYPE = nevenschikkend>

De CGN-tags zijn dus intern gestructureerd en verschillen in dat opzicht van de monolitische tags die o.m. in het Brown corpus gebruikt zijn.

Naast de volledig uitgespelde notatie wordt ook gebruik gemaakt van een compacter formaat, waarin de namen van de attributen zijn weggelaten en waarin de namen van de waarden zijn afgekort. Voor de nevenschikkende voegwoorden ziet die er als volgt uit.

[T801] VG(neven)

De tags worden net als de declaraties en de partities genummerd. Het eerste cijfer van het nummer verwijst naar de woordsoort (T8). Tags die een ondergespecificeerde waarde bevatten, zoals 'genus', krijgen een U-nummer i.p.v. een T-nummer.

[U117] `N(soort,ev,basis,genus,stan)`

De tags die alleen voor de dialectwoorden bestemd zijn, krijgen een R-nummer (voor regionaal).

[R5xx] `VNW(aanw,det)`

Implicaties. Binnen een tag bestaan soms afhankelijkheden tussen de waarden van de features. Diminutieve substantieven bijv. zijn altijd onzijdig. Dat kan worden uitgedrukt in termen van een implicatie als

<POS = substantief, GRAAD = diminutief> $\Longrightarrow$ <GENUS = onzijdig>

2 DE D-COI-TAGSET

De orthografisch getranscribeerde spraak wordt gesegmenteerd in tokens. Dat zijn woorden, leestekens of speciale symbolen.

[P01] `TOKENTYPE = woord, speciaal, leesteken.`

01. TOKENTYPE. De tokens van het type 'woord' zijn de woordvormen zoals ze in de orthografische transcriptie voorkomen; die vormen kunnen gelijk zijn aan de stam, maar zullen—in vele gevallen—geflecteerd zijn. Met de woorden wordt een POS (Part of Speech) feature geassocieerd.

[D00] <TOKENTYPE = woord> $\Longrightarrow$ <POS>
[P02] `POS = substantief, adjectief, werkwoord, telwoord, voornaamwoord, lidwoord, voorzetsel, voegwoord, bijwoord, tussenwerpsel.`

02. POS. We maken een onderscheid tussen vier open klassen (substantief, adjectief, werkwoord, telwoord) en zes (min of meer) gesloten klassen (voornaamwoord, lidwoord, voorzetsel, voegwoord, bijwoord, tussenwerpsel). Deze indeling in tien woordsoorten is identiek aan die van de ANS-97. Er zijn wel enkele verschillen m.b.t. de classificatie van specifieke woorden: *er* bijv. wordt niet behandeld als een bijwoord, maar als een voornaamwoord, en *veel* en *weinig* worden niet behandeld als telwoorden maar als onbepaalde voornaamwoorden. Zulke verschillen worden bij de bespreking van de diverse woordsoorten vermeld en gemotiveerd.

Voor elk van de tien woordsoorten wordt in de volgende secties vermeld hoe ze worden onderscheiden van andere woordsoorten (Afgrenzing), welke features ermee geassocieerd worden (Declaraties) en welke waarden die features kunnen hebben (Partities). Alle features worden toegelicht met voorbeelden en voorzien van criteria voor de bepaling van de relevante waarden. Per woordsoort wordt ook vermeld of er afhankelijkheden bestaan tussen de waarden van de features (Implicaties) en welke de mogelijke tags zijn. Bij die laatste wordt een onderscheid gemaakt tussen de maximaal specifieke combinaties en de ondergespecificeerde combinaties; de eerste krijgen een T-nummer, de laatste een U-nummer. Tenslotte worden telkens richtlijnen gegeven voor de lemmatisering.

Voor de dialectwoorden is een afzonderlijke tagset met lagere granulariteit ingevoerd (zie 2.11). De speciale tokens en de leestekens worden afzonderlijk behandeld in de secties 2.12 en 2.13.

2.1 SUBSTANTIEVEN

2.1.1 Afgrenzing

Bij de identificatie van de substantieven dient speciale aanacht te worden besteed aan het onderscheid met nominaal gebruikte adjectieven (zie 2.2), deelwoorden en infinitieven (zie 2.3), telwoorden (zie 2.4) en speciale tokens (zie 2.12). Criteria voor het maken van die onderscheidingen worden gegeven in de secties over die andere woordsoorten.

2.1.2 Declaraties

Substantieven zijn gemarkeerd voor type (soortnaam of eigennaam), getal (enkelvoud of meervoud) en graad (diminutief of niet). Naamvalswaarden (standaard, genitief of datief) worden alleen aan de enkelvoudige substantieven toegekend en genuswaarden (zijdig of onzijdig) aan de enkelvoudige substantieven zonder naamvalssuffix.

[D01] <POS = substantief> $\Longrightarrow$ <NTYPE, GETAL, GRAAD>
[D02] <POS = substantief, GETAL = enkelvoud> $\Longrightarrow$ <NAAMVAL>
[D03] <POS = substantief, GETAL = enkelvoud, NAAMVAL = standaard>
$\Longrightarrow$ <GENUS>

NTYPE en GENUS zijn lexicale kenmerken; GRAAD, GETAL en NAAMVAL zijn morfologische kenmerken. Het diminutiefsuffix gaat steeds vooraf aan de suffixen van het meervoud (*hond-je-s*) en de genitief (*Jan-tje-s hond*).

2.1.3 Partities

[P03] NTYPE = soortnaam, eigennaam.
[P04] GETAL = enkelvoud, meervoud.
[P05] GRAAD = basis, diminutief.
[P06] GENUS = genus (zijdig, onzijdig).
[P07] NAAMVAL = standaard, genitief, datief.

03. NTYPE. Substantieven die bij de orthografische transcriptie zonder hoofdletter gespeld zijn, worden tot de soortnamen gerekend, behalve wanneer ze in de lexicografische praktijk algemeen als eigennaam behandeld

worden; dat geldt o.m. voor de namen van de maanden (*april*) en van de dagen van de week (*zondag*).

Substantieven die met een hoofdletter gespeld zijn, worden tot de eigennamen gerekend, behalve wanneer het gaat om (1) afkortingen van soortnamen, zoals *CD, LP, WC, TV, PC*; (2) samenstellingen waarvan de kern een soortnaam is en die als geheel ook als soortnaam functioneren (*het Ardennenoffensief, een typisch Randstad-probleem, NAVO-bombardementen*); woorden als *Noordzee*, *Waasland* en *Bondgenotenlaan* worden daarentegen wel tot de eigennamen gerekend, o.m. omdat de toevoeging van een meervoudsuitgang of een onbepaald lidwoord gemarkeerd is, vgl. *een Randstadprobleem* met *? een Noordzee*; (3) substantieven die deel uitmaken van een titel, zoals in *De Naam Van De Roos* en *Man Bijt Hond*; substantieven in titels worden alleen als eigennamen getagd als ze ook buiten de context van de titel alsdusdanig gebruikt worden; in *Het Verdriet Van België* is het eerste substantief dus een soortnaam en het tweede een eigennaam. Merk wel dat het hier om titels gaat (van boeken, TV- en radioprogramma's, platen, films, e.d.) en niet om namen van personen, plaatsen, kranten e.d. In de persoonsnaam *Roos De Hond* en de plaatsnaam *Den Haag* krijgen de substantieven niet de tag voor soortnamen, maar SPEC(deeleigen). Dat is de tag die wordt toegekend aan de delen van een eigennaam die uit verschillende woorden bestaat. In combinaties als *Den Haag*, *Freek De Jonge*, *Brusselse Steenweg* en *De Standaard*, krijgen de afzonderlijke woorden dus de tag SPEC(deeleigen), zie 2.12. De identificatie van de eigennaam als één geheel geschiedt op het niveau van de lexicologische koppeling. Let wel: in combinaties als *het Van Mierlo-effect* wordt het laatste woord als een soortnaam behandeld (vgl. *Randstad-probleem*) en heeft alleen *Van* de tag SPEC(deeleigen).

Aangezien deze indeling in soortnamen en eigennamen van lexicaalorthografische aard is, is ze niet afhankelijk van specifieke contexten. Een substantief als *Stella* wordt dus steeds als eigennaam getagd, ook in contexten waarin het niet individualiserend gebruikt is, zoals in *waar mijn Stella staat* en *drie Stella's a.u.b.*.

Bemerk ten slotte dat alleen substantieven de tag N(eigen, ...) kunnen krijgen. De adjectieven in *de Belgische regering* en *een Italiaanse wijn* zijn dus geen eigennamen, maar adjectieven. Hetzelfde geldt voor het tussenwerpsel AUB en de vreemdtalige SVP en CQ. Die behoren tot de speciale tokens.

04. GETAL. Het meervoud is gemarkeerd door een suffix (*-s, -en, ...*) of—uitzonderlijk—door suppletie (*zeelui, timmerlieden*). Substantieven zonder meervoudssuffix krijgen de waarde 'enkelvoud', ook wanneer ze naar de betekenis meervoudig zijn. Soortnamen met een collectiverende betekenis zoals *boel, aantal, hoop* krijgen dus de waarde 'enkelvoud'; dit geldt ook voor de maataanduidende substantieven in combinaties van het type *vijf jaar* en *drie liter*. Substantieven met een meervoudssuffix krijgen de waarde 'meervoud'. Dat geldt ook voor de pluralia tantum zoals *hersenen* en *ingewanden*.

Bij de eigennamen is de aanwezigheid van een meervoudssuffix niet altijd een teken van meervoudigheid. Een vorm als *Enkhuizen* bijvoorbeeld krijgt de waarde 'enkelvoud' omdat hij de typische kenmerken van een enkelvoudig substantief vertoont: zo combineert hij met *het*, als in *het/*de Enkhuizen van weleer*, en vereist hij een enkelvoudige persoonsvorm wanneer hij als onderwerp fungeert, als in *Enkhuizen ligt/*liggen ergens in Nederland*. Een vorm als *Ardennen* daarentegen is wel meervoudig, zoals blijkt uit *de/*het Ardennen liggen/*ligt in België*.

05. GRAAD. Diminutiefvormen zijn gemarkeerd door een suffix (*-je, -tje, -pje, -ke, ...*). Substantieven die dit suffix niet hebben krijgen de waarde 'basis'; die waarde wordt ook toegekend aan de substantieven die geen diminutiefvorm kunnen hebben, zoals *gebergte, vee*. Substantieven met een diminutiefsuffix krijgen de waarde 'diminutief'; die waarde wordt ook toegekend aan de substantieven die alleen een diminutiefvorm hebben zoals *ootje, meisje, nippertje*. Typisch voor de diminutieve substantieven is hun onzijdigheid: in het enkelvoud combineren ze met *het/dat/dit/ons* en niet met *de/die/deze/onze*. Bij de eigennamen wijst de aanwezigheid van een diminutiefsuffix niet altijd op diminutiviteit; een naam als *Nelleke* bijvoorbeeld combineert met typisch zijdige determiners (*die/?dat Nelleke toch*) en is dus niet diminutief.

06. GENUS. Bij POS tagging maken we alleen het onderscheid tussen zijdige en onzijdige substantieven; de verdere differentiatie van de zijdige substantieven in masculiene en feminiene is niet gemaakt. De zijdige substantieven zijn die welke in het enkelvoud determiners nemen als *de/die/deze/onze*, terwijl de onzijdige determiners nemen als *het/dat/dit/ons*.

GENUS betreft, net als GETAL, een morfo-syntactische onderscheiding, en geen semantische; *meisje* en *mannetje* zijn dus onzijdig, ook al verwijzen ze naar personen. Het is ook van belang om zich te realiseren dat de GE-

NUS waarden betrekking hebben op woordvormen en niet op lemma's; zo is *mannetje* morfo-syntactisch onzijdig, ook al is het corresponderende lemma zijdig (*de man*).

Een beperkt aantal soortnamen wordt zowel zijdig als onzijdig gebruikt. Als dit met een betekenisverschil gepaard gaat (*de/het bal, de/het blik*), onderscheiden we twee lemma's met een verschillend genus. Als dit niet met een betekenisverschil gepaard gaat (*de/het riool, de/het filter, de/het soort*), spreken we van één lemma en kennen we de ondergespecifeerde waarde ('genus') toe, wanneer de lokale context het niet mogelijk maakt om te disambigueren, zoals in *een filter* (voor een lijst van zulke substantieven, zie de ANS-97, p. 159); als de lokale context het toelaat om te disambigueren, zoals in *de filter*, wordt wel een specifieke waarde toegekend.

In combinaties als *de laatste drie jaar* en *om de vier uur* krijgen *jaar* en *uur* ondanks de aanwezigheid van *de* toch hun gebruikelijke waarde *onzijdig*. Het lidwoord bepaalt hier namelijk niet het maataanduidende substantief maar het telwoord.

Het genusonderscheid is ook voor de eigennamen relevant. Zo zijn persoonsnamen meestal zijdig (*de/*het Karel*) en namen van steden, landen en talen meestal onzijdig (*het/*de Brussel van toen, het/*de Frankrijk van de eeuwwisseling, het/*de Spaans*), zie de ANS-97, p. 285. Samengestelde eigennamen met een soortnaam als kern erven meestal het genus van die soortnaam (*het/*de Waasland*, *het/*de Zilvermeer*, *de/*het Noordzee, de/*het Kemmelberg*); een uitzondering is *de Bijlmermeer*. Als de keuze van het lidwoord vrij is, kiest men de generische waarde; dat geldt o.m. voor merknamen als *Linux* en *Esselte*.

Omdat het morfo-syntactische genus in meervoudsvormen en in vormen met een naamvalssuffix systematisch geneutraliseerd wordt, kennen we het alleen toe aan de enkelvoudige standaardvormen.

07. NAAMVAL. De genitief is gemarkeerd door het suffix *-s* en komt vooral voor bij substantieven die naar personen of tijdstippen verwijzen (*Otto's, vaders, 's avonds*); een beperkt aantal substantieven neemt het suffix *-en* (*des Heren, des mensen*). De datief is gemarkeerd door het suffix *-e* en komt vrijwel alleen in vaste verbindingen voor (*ter plaatse, te berde, in der minne*); zeer zeldzaam is de datief met een *-en* suffix, zoals in *ten voeten uit*. Niet elk substantief dat op *-e* eindigt, is een datiefvorm: in *aanvrage* en *proeve*, bijvoorbeeld, is de sjwa deel van de stam, net als in *bode* en *smeekbede*. Bij afwezigheid van een naamvalssuffix wordt de waarde 'standaard'

toegekend. Omdat het naamvalsonderscheid bij meervoudige substantieven systematisch geneutraliseerd wordt (*deze dagen, één dezer dagen*), wordt het feature er niet aan toegekend.

2.1.4 Implicaties

[I01] <POS = substantief, GRAAD = diminutief, GETAL = enkelvoud>
$\Longrightarrow$ <NAAMVAL $\neq$ datief>
[I02] <POS = substantief, GRAAD = diminutief, GETAL = enkelvoud, NAAMVAL = standaard> $\Longrightarrow$ <GENUS = onzijdig>

Diminutieve substantieven hebben geen datiefvorm, wellicht omdat ze al op een sjwa eindigen en de toevoeging van het datiefsuffix dus geen verschil zou maken.

Diminutieve substantieven zijn ook steeds onzijdig. Bemerk dat het hierbij gaat om het genus van de woordvorm (het morfo-syntactische genus) en niet om het genus van het lemma (het lexicale genus); die laatste kan immers zijdig zijn, ook al is de eerste onzijdig.

2.1.5 De tags

Het aantal maximaal specifieke tags bedraagt acht voor de soortnamen.

[T101]	N(soort,ev,basis,zijd,stan)	die stoel, elke avond, deze muziek, de filter
[T102]	N(soort,ev,basis,onz,stan)	het kind, ons huis, dat brood, dit land, het filter
[T103]	N(soort,ev,dim,onz,stan)	dit stoeltje, ons huisje, op 't nippertje
[T104]	N(soort,ev,basis,gen)	's avonds, de heer des huizes, des mensen
[T105]	N(soort,ev,dim,gen)	vadertjes pijp
[T106]	N(soort,ev,basis,dat)	ter plaatse, heden ten dage, te berde brengen
[T107]	N(soort,mv,basis)	stoelen, kinderen, hersenen
[T108]	N(soort,mv,dim)	stoeltjes, huisjes, hersentjes

Voor de eigennamen zijn er evenveel.

[T109]	`N(eigen,ev,basis,zijd,stan)`	de Noordzee, de Kemmelberg
[T110]	`N(eigen,ev,basis,onz,stan)`	het Hageland, het Albertkanaal, het Latijn
[T111]	`N(eigen,ev,dim,onz,stan)`	het slimme Kareltje
[T112]	`N(eigen,ev,basis,gen)`	des Heren, Hagelands trots, de Aa's bovenloop
[T113]	`N(eigen,ev,dim,gen)`	Kareltjes fiets
[T114]	`N(eigen,ev,basis,dat)`	wat den Here toekomt
[T115]	`N(eigen,mv,basis)`	de Ardennen, de Middeleeuwen, de Kempen
[T116]	`N(eigen,mv,dim)`	de Maatjes (een natuurreservaat in de Kempen)

De toekenning van ondergespecificeerde waarden is relevant voor GENUS. In gevallen waarin de context het niet toelaat om tussen de twee genuswaarden te kiezen wordt de generische waarde toegekend, zowel bij soortnamen als bij eigennamen.

[U117]	`N(soort,ev,basis,`*`genus`*`,stan)`	een riool, geen filter
[U118]	`N(eigen,ev,basis,`*`genus`*`,stan)`	Linux, Esselte

2.1.6 Lemmatisering

Voor de lemmatisering van de soortnamen gebruiken we de vorm zonder affixen, d.w.z. de basisvorm, in het enkelvoud, zonder naamvalssuffix. Als het substantief echter alleen een meervoudsvorm heeft (*mazelen, hersenen*), dan wordt die meervoudsvorm als lemma gebruikt. Hetzelfde geldt voor substantieven die alleen een diminutiefvorm hebben (*meisje, ootje, akkefietje*); zulke substantieven komen vooral in vaste verbindingen voor (*een robbertje vechten, zijn hachje redden, in het ootje nemen, op het nippertje*). Substantieven die alleen in de genitief- of de datiefvorm voorkomen, krijgen die verbogen vorm als lemma; een zeldzaam voorbeeld is de datiefvorm in *ten behoeve van*.

Dezelfde principes gelden voor de lemmatisering van de eigennamen: die worden tot een enkelvoudige basisvorm herleid (*(een) Fiatje* $\Longrightarrow$ *Fiat*, *(drie) Stella's* $\Longrightarrow$ *Stella*), behalve wanneer de vorm zonder affix niet bestaat.

Dat geldt o.m. voor vormen als *Ardennen, Antillen, Marollen, Enkhuizen, Middeleeuwen* en voor vormen als *Nelleke, Sneeuwwitje, Doornroosje*.

Een klein aantal substantieven heeft twee basisvormen, zoals *aanvraag/aanvrage* en *proef/proeve*; in dat geval onderscheiden we ook twee lemma's. Voor vormen als *proeven* en *aanvragen* wordt de keuze door de context bepaald: als ze voorkomen in een context waarin de enkelvoudsvorm een sjwa heeft, dan heeft ook het lemma van de meervoudsvorm een sjwa; anders kiest men voor het lemma zonder sjwa.

Afkortingen, zoals TV en CD, en soortnamen die omwille van het gebruik in een titel met een hoofdletter gespeld zijn, zoals de substantieven in *De Naam van de Roos*, krijgen een lemma zonder hoofdletter.

2.2 ADJECTIEVEN

2.2.1 Afgrenzing

Tot de adjectieven rekenen we niet alleen de prenominaal en de predikatief gebruikte bijvoeglijke voornaamwoorden, maar ook de zelfstandig (of nominaal) gebruikte en de adverbiaal gebruikte. Hoe die onderscheiden worden van resp. de substantieven en de bijwoorden wordt uitgelegd in 2.2.3. Het onderscheid tussen deelwoord en adjectief wordt in de sectie over de werkwoorden toegelicht, zie 2.3.1.

2.2.2 Declaraties

Alle adjectieven hebben de features POSITIE, GRAAD en BUIGING. De nominaal gebruikte zijn ook gemarkeerd voor GETAL en—als ze enkelvoudig en verbogen zijn—voor NAAMVAL. Dat laatste feature wordt ook toegekend aan de verbogen prenominale adjectieven.

[D04] <POS = adjectief> $\Longrightarrow$ <POSITIE, GRAAD, BUIGING>
[D05] <POSITIE = nominaal> $\Longrightarrow$ <GETAL-N>
[D06] <POS = adjectief, POSITIE = nominaal, BUIGING = met-e, GETAL-N = zonder-n> $\Longrightarrow$ <NAAMVAL>
[D07] <POS = adjectief, POSITIE = prenominaal, BUIGING = met-e> $\Longrightarrow$ <NAAMVAL>

Met uitzondering van POSITIE gaat het om morfologische kenmerken. Het graadssuffix gaat steeds vooraf aan het buigingssuffix, dat op zijn beurt gevolgd kan worden door een getal- of naamvalssuffix (*de oud-er-e-n*).

2.2.3 Partities

```
[P08] POSITIE = prenominaal, nominaal, postnominaal, vrij.
[P05] GRAAD = basis, comparatief, superlatief, diminutief.
[P09] BUIGING = zonder, met-e, met-s.
[P10] GETAL-N = zonder-n, meervoud-n.
[P07] NAAMVAL = standaard, bijzonder.
```

08. POSITIE. Adjectieven komen in diverse soorten van posities voor: prenominale (*een mooie tuin*), nominale (*het/de mooie*), postnominale (*iets moois*) en niet-nominale of vrije; tot die laatste rekenen we zowel de predikatieve (*dit is mooi, de tuin mooi maken*) als de adverbiale (*hij praat zo mooi*).

Bij prenominaal gebruik gaan de adjectieven vooraf aan het substantief dat ze bepalen. In deze positie is er variatie tussen vormen met een buigings *-e* en vormen zonder buigingsuitgang (*kleine* vs. *klein*). Tot het prenominale gebruik rekenen we ook het elliptische, als in de tweede conjunct van *Hij heeft een wit bord en ik een groen*; het antecedent hoeft niet noodzakelijk in dezelfde zin te staan: *Ik heb gisteren een witte telefoon gekocht. Hij past beter in het interieur dan die groene.* Een belangrijk argument voor de behandeling van deze adjectieven als prenominaal is dat ze bij elliptisch gebruik dezelfde variatie vertonen tussen vormen met en zonder buigingsuitgang als bij prenominaal gebruik. Bij nominaal en vrij gebruik vinden we resp. meer variatie en geen variatie.

Nominaal (of zelfstandig) gebruikte adjectieven worden niet als substantieven behandeld, maar als adjectieven. Daarvoor pleiten o.m. het bestaan van comparatief- en superlatiefvormen (*de ouderen, de rijksten*), de compatibiliteit met adverbiale graadaanduiders (*de zeer rijken*) en het feit dat de meervoudsvorming anders gebeurt dan bij de substantieven, zie GETAL-N. Wanneer ze naar personen verwijzen eindigen de nominaal gebruikte adjectieven steeds op een sjwa en kunnen ze een meervouds *-n* nemen: *de arme(n)* en *de blinde(n)* zijn dus adjectieven, maar *de liberaal* en *de conservatief* zijn substantieven. Wanneer ze niet naar personen verwijzen kunnen de nominaal gebruikte adjectieven ook zonder buigingsuitgang voorkomen, zoals in *iets in het groen schilderen*.

Het postnominale gebruik is stilistisch gemarkeerd en komt weinig voor (*kindeke klein*), behalve wanneer het adjectief eigen bepalingen heeft (*alle rivieren bevaarbaar in de winter*) of wanneer het een bepaling is bij een quantor (*niets bijzonders, iets groters*). In dat laatste geval is het adjectief meestal van een buigings-*s* voorzien. We spreken alleen van postnominaal gebruik wanneer het adjectief een nabepaling is bij een substantief; in combinaties als *drie maand lang* en *twee meter breed* is het adjectief geen nabepaling bij het substantief, maar is het adjectief de kern van een adjectivale groep en is de NP een bepaling bij het adjectief. De POSITIE waarde van het adjectief is dan afhankelijk van de ruimere context: in *een twee meter brede stoep* is het adjectief prenominaal en in *die stoep is twee meter breed* is het vrij.

Het niet-nominale of vrije gebruik omvat naast het predikatieve ook het adverbiale gebruik. Adverbiaal gebruikte adjectieven worden dus niet als bijwoorden behandeld maar als adjectieven, net als in de ANS-97, CELEX, RBN, WOTAN-2 en het Duitse STTS-95. Om de bijwoorden te onderscheiden van de adverbiaal gebruikte adjectieven hanteren we het volgende criterium: als het betreffende woord in dezelfde betekenis ook gebruikt wordt in prenominale posities, dan is het geen bijwoord maar een adjectief. Zo is *vrij* in *de vogels vrij laten rondvliegen* een adjectief in vrije positie, omdat het in deze context dezelfde betekenis heeft als het prenominale adjectief in *een vrije vogel*, terwijl het in *dat is vrij hoog* om het bijwoord *vrij* gaat, zie 2.9. Omdat er geen morfologische verschillen bestaan tussen predikatief en adverbiaal gebruikte adjectieven, en omdat het maken van het onderscheid slechts mogelijk is op basis van een volledige syntactische analyse, wordt het bij de POS tagging niet gemaakt.

Niet alle adjectieven komen in elk van de vier posities voor. Zo zijn er die alleen prenominaal gebruikt worden (*houten, ijzeren*) en andere die alleen vrij gebruikt worden (*jammer, beu*). Voor zulke adjectieven kunnen lexicale specificaties uitsluitsel bieden over de POSITIE waarde. Voor adjectieven die in alle vier de posities voorkomen, zoals *mooi*, moet bij de POS tagging de contextueel relevante waarde toegekend worden.

Vormen van het type *een-/het-/de-/dat-/diezelfde* zijn prenominaal wanneer ze een nominale kern bepalen, als in *hetzelfde paard*, en nominaal wanneer ze zelf de kern zijn van een nominale groep, als in *het zijn altijd dezelfden die niet opdagen*.

05. GRAAD. De comparatief wordt gemarkeerd door het suffix *-er*, de superlatief door *-st* en de diminutief door *-jes*, als in (*zachtjes, fijntjes, warm-*

pjes). De reden waarom de diminutief op dezelfde lijn wordt geplaatst als de comparatief en de superlatief is tweeledig: (1) ze zijn in complementaire distributie (een vorm kan niet tegelijk comparatief en diminutief zijn); (2) ze combineren met hetzelfde soort van adjectieven (de gradeerbare). Bij afwezigheid van een graadssuffix wordt de waarde 'basis' toegekend, ook in het geval van adjectieven die geen graadsaanduiding kunnen hebben (*houten, elektrisch*). Die waarde wordt ook toegekend aan gelede adjectieven waarin het graadssuffix niet aan het eind van het woord komt, zoals in *verderaf, dichterbij* en *hoogstgelegen, laatstleden*.

09. BUIGING. De meeste adjectieven hebben drie verschijningsvormen (*klein, kleine, (iets) kleins*). Dat die laatste vorm verschillend is van de genitief, blijkt uit de vorm van het suffix; een minimaal paar is de buigingsvorm in *iets zaligs* vs. de genitief in *zaliger gedachtenis*. Adjectieven waarvan de stam op een sis-klank eindigt (*grijs, boos, theoretisch*) kunnen geen *-s* nemen en hebben dus nooit de waarde 'met-s'. Iets soortgelijks geldt voor de adjectieven die geen sjwa kunnen nemen (*beige, timide, bescheiden, houten, lila, kaki*); die hebben nooit de waarde 'met-e'. De sjwa in *beige, timide, onderste* is deel van de stam. Hetzelfde geldt voor de sjwa in het postnominale *(1 juli) aanstaande*; postnominale adjectieven hebben overigens nooit een buigings *-e*.

10. GETAL-N. Bij nominaal gebruik kunnen de adjectieven een meervoudssuffix nemen (*de blinden, de zieken, de rijken*). Dat deze vormen wel degelijk adjectieven zijn en geen substantieven is hierboven al uitgelegd (zie POSITIE); aan de daar vermelde argumenten kan hier worden toegevoegd dat de meervoudsvorming bij nominale adjectieven in twee opzichten verschilt van die bij substantieven. Primo, het gebruikte suffix is steeds *-(e)n*, terwijl bij substantieven die op een sjwa eindigen ook *-s* gebruikt wordt (*dames, bodes, gewoontes, (on)voldoendes*). Secundo, de vorm met het meervoudssuffix kan alleen naar personen verwijzen; voor de substantieven geldt die restrictie niet.

Dit heeft twee gevolgen: (1) nominaal gebruikte adjectieven zonder *-n* zijn niet noodzakelijk enkelvoudig (*de grootste is/zijn al verkocht*); om die reden staat 'meervoud-n' niet tegenover 'enkelvoud-n' maar gewoon tegenover 'zonder-n'; (2) meervoudsvormen op *-n* zijn alleen adjectieven wanneer ze naar personen verwijzen; wiskundige termen als *kromme(n), variabele(n)* en *gemiddelde(n)* zijn dus substantieven, en geen zelfstandig gebruikte adjectieven; hetzelfde geldt voor vormen als *uiterste(n), groteske(n)* en *vereiste(n)*.

Omwille van die verschillen maken we ook in de notatie een onderscheid tussen het feature voor de getalsdistinctie bij substantieven (GETAL) en het feature voor de getalsdistinctie bij adjectieven (GETAL-N). Een minimaal paar is het substantief *ouders* vs. het nominaal gebruikte adjectief *ouderen*. Een bijkomende reden voor het onderscheiden van GETAL en GETAL-N is dat sommige woorden, meer bepaald de nominaal gebruikte bezittelijke voornaamwoorden, voor beide features gemarkeerd zijn; *de zijnen* bijv. heeft de waarde 'enkelvoud' voor GETAL en de waarde 'meervoud-n' voor GETAL-N, zie 2.5.

07. NAAMVAL. Het naamvalsonderscheid wordt alleen gemaakt bij de (pre)nominaal gebruikte adjectieven met een buigings *-e*. Bij aanwezigheid van een naamvalssuffix (*-er* of *-en*) krijgen die de waarde 'bijzonder'. Het onderscheid tussen genitief en datief wordt niet gemaakt, omdat het niet met verschillende vormen correspondeert: zo is de *-er* vorm een genitief in *zaliger gedachtenis* en een datief in *te goeder trouw*, en is de *-en* vorm een genitief in *des Allerhoogsten* en een datief in *in koelen bloede*. Bij afwezigheid van een naamvalssuffix wordt de waarde 'standaard' toegekend.

2.2.4 Implicaties

[I03] <POS = adjectief, GRAAD = superlatief> $\Longrightarrow$
<POSITIE $\neq$ postnominaal>
[I04] <POS = adjectief, GRAAD = diminutief> $\Longrightarrow$
<POSITIE = vrij>
[I05] <BUIGING = met-s> $\Longrightarrow$ <POSITIE = postnominaal>
[I06] <BUIGING = met-e> $\Longrightarrow$ <POSITIE = (pre)nominaal>

De eerste twee implicaties betreffen verbanden tussen GRAAD en POSITIE: de superlatiefvormen kunnen niet postnominaal gebruikt worden en de diminutiefvormen komen uitsluitend in predikatieve of adverbiale posities voor. De diminutiefvorm in *de kleintjes* is dus geen adjectief, maar de meervoudsvorm van het substantief *kleintje*; idem dito voor *de oudjes*.

De laatste twee implicaties leggen een verband tussen BUIGING en POSITIE. De *-s* vormen komen alleen in postnominale positie voor; de s-vormen in *van jongs af aan, er is nieuws, lekkers krijgen* beschouwen we niet als verbogen vormen van adjectieven, maar als substantieven met aparte lemmata. De *-e* vormen komen alleen in nominale en prenominale posities voor; de vorm *hele* in *een hele mooie tuin* wordt dus als prenominaal be-

handeld en niet als vrij. Bemerk dat deze twee implicaties geldig zijn voor alle woordsoorten met een BUIGINGsfeature, dus ook voor deelwoorden en determiners.

2.2.5 De tags

De interactie van de declaraties met de implicaties geeft voor de prenominale adjectieven negen maximaal specifieke combinaties: drie voor de basisvormen, drie voor de comparatieven en drie voor de superlatieven.

[T201]	ADJ(prenom,basis,zonder)	een mooi huis, een houten pot
[T202]	ADJ(prenom,basis,met-e,stan)	mooie huizen, een grote pot
[T203]	ADJ(prenom,basis,met-e,bijz)	zaliger gedachtenis, van goeden huize
[T204]	ADJ(prenom,comp,zonder)	een mooier huis
[T205]	ADJ(prenom,comp,met-e,stan)	mooiere huizen, een grotere pot
[T206]	ADJ(prenom,comp,met-e,bijz)	van beteren huize
[T207]	ADJ(prenom,sup,zonder)	een alleraardigst mens
[T208]	ADJ(prenom,sup,met-e,stan)	de mooiste keuken, het grootste paard
[T209]	ADJ(prenom,sup,met-e,bijz)	bester kwaliteit

Bij de nominale adjectieven speelt ook het getalsonderscheid een rol, zodat er een aantal combinaties meer zijn.

[T210]	ADJ(nom,basis,zonder,zonder-n)	in het groot, in het bijzonder, het groen
[T211]	ADJ(nom,basis,zonder,mv-n)	de timiden, dezelfden
[T212]	ADJ(nom,basis,met-e,zonder-n,stan)	het leuke is dat ..., geef mij maar een grote met tartaar
[T213]	ADJ(nom,basis,met-e,zonder-n,bijz)	hosanna in den hogen
[T214]	ADJ(nom,basis,met-e,mv-n)	de rijken
[T215]	ADJ(nom,comp,zonder,zonder-n)	
[T216]	ADJ(nom,comp,met-e,zonder-n,stan)	een betere
[T217]	ADJ(nom,comp,met-e,zonder-n,bijz)	
[T218]	ADJ(nom,comp,met-e,mv-n)	de ouderen

[T219]	ADJ(nom,sup,zonder,zonder-n)	op z'n best, om ter snelst
[T220]	ADJ(nom,sup,met-e,zonder-n,stan)	het leukste is dat, het langste blijven
[T221]	ADJ(nom,sup,met-e,zonder-n,bijz)	des Allerhoogsten
[T222]	ADJ(nom,sup,met-e,mv-n)	de slimsten

Bij de postnominale adjectieven spelen naamval en getal geen rol en is de variatie in graad beperkt tot twee waarden (basis en comparatief).

[T223]	ADJ(postnom,basis,zonder)	rivieren bevaarbaar in de winter
[T224]	ADJ(postnom,basis,met-s)	iets moois
[T225]	ADJ(postnom,comp,zonder)	een getal groter dan drie
[T226]	ADJ(postnom,comp,met-s)	iets gekkers kon ik niet bedenken

Bij vrij gebruik ten slotte spelen naamval en getal evenmin een rol en is de waarde van het buigingsfeature invariabel.

[T227]	ADJ(vrij,basis,zonder)	die stok is lang, lang slapen
[T228]	ADJ(vrij,comp,zonder)	deze stok is langer, langer slapen
[T229]	ADJ(vrij,sup,zonder)	welke stok is het langst, het langst slapen, de verst afgelegen dorpen
[T230]	ADJ(vrij,dim,zonder)	het is hier stilletjes, stilletjes wegsluipen

2.2.6 Lemmatisering

Voor de specificatie van het lemma gebruiken we de vorm zonder affixen, d.w.z. de basisvorm, zonder buigings-, getals- of naamvalssuffix. Als het adjectief geen basisvorm heeft, zoals de comparatief in *eerdere pogingen*, dan nemen we de comparatief als lemma. Dat doen we ook wanneer de comparatiefvorm verzelfstandigd is; zo is het lemma van de comparatieven in *verder denk ik dat dit niet klopt* en *het verdere verloop van de procedure* gelijk aan *verder*, en niet aan *ver*; soortgelijke gevallen zijn de comparatieven in *vroeger was alles beter*, *vroegere pogingen*, *later zal je dat wel begrijpen* en *latere successen*. Typisch voor de verzelfstandigde comparatiefvormen is dat ze niet gecombineerd kunnen worden met een *dan* bepaling en dat ze

geen adverbiale graadsbepaling, zoals *veel*, kunnen nemen. In *zij springt veel verder dan ik* en *het was later/vroeger dan ik dacht* is het lemma dus wel gelijk aan de basisvormen *ver, laat, vroeg.*

Bij het afstrippen van de buigingsuitgang dient men er rekening mee te houden dat niet elk adjectief dat op een sjwa eindigt verbogen is; in *beige, timide, morbide, onderste* bijv. is de sjwa deel van de stam (en dus van het lemma), aangezien vormen als *beig, timid* en *onderst* niet bestaan. Adjectieven die alleen met een naamvalssuffix voorkomen, zoals *arren* in *in arren moede*, worden ook als lemma gebruikt.

2.3 WERKWOORDEN

2.3.1 Afgrenzing

Voor de persoonsvormen stellen zich nauwelijks problemen voor de afgrenzing t.o.v. andere woordsoorten; dat *vliegen* een werkwoord is in *we vliegen* en een substantief in *wat een vervelende vliegen* is zonder meer duidelijk.

Infinitieven behandelen we als werkwoorden, ook wanneer ze in nominale posities voorkomen (*het vallen van de bladeren, het polsstokspringen*). Er zijn wel enkele infinitiefvormen die een substantief homoniem hebben, zoals *leven.* In tegenstelling tot de infinitief, heeft het substantief een meervoudsvorm (*levens*) en een diminutiefvorm (*wat een leventje*) en wordt het courant gebruikt met een onbepaald lidwoord (*een leven*).

Ook deelwoorden behandelen we als werkwoorden, behalve wanneer ze duidelijk adjectivische eigenschappen vertonen. Dat is het geval voor vormen (1) met een niet-werkwoordelijke stam, zoals *getand, geaderd, gevleugeld*; (2) met een typisch adjectivaal prefix (*on-gehoord, on-deugend, aartsingewikkeld*); (3) met een graadssuffix, meestal een comparatief of superlatief (*opgewekter, spannendste*), zie ANS-97, p. 388-389; (4) met een buigings *-s* (*iets uitdagends, iets gezochts*), zie ANS-97, p. 412; (5) met een naamvalssuffix, meestal een datief (*te gepasten tijde, te bestemder plaatse, met voorbedachten rade*), zie ANS-97, p. 413. Deze criteria zijn makkelijk toe te passen omdat ze de vorm zelf betreffen, onafhankelijk van de context waarin hij optreedt. Voor de andere vormen wordt de keuze van de woordsoort door de context bepaald. Ook hier geldt dat de keuze voor een werkwoordelijke analyse de default is. Voor een adjectivische analyse wordt echter gekozen wanneer het deelwoord (6) gecombineerd is met een graadsaanduidende bepaling, zoals *zo opgewekt, zeer beperkte voorraad, te ingewikkeld, heel spannende film, hoogst opwindende lingerie, ...*; (7) een valentie heeft

die niet overeenstemt met die van het corresponderende werkwoord; vergelijk bijv. het werkwoordelijke gebruik van *bedacht* in *dat heb je zeker zelf bedacht* en *dat is bedacht (door ...)* met het adjectivale gebruik in *daar was ze niet op bedacht*. In de eerste twee voorbeelden gaat het om een vorm van het transitieve werkwoord *iets bedenken*, maar in het laatste voorbeeld is er geen werkwoord met een corresponderende valentie: * *op iets bedenken*; in de plaats daarvan gaat het om het adjectivische *op iets bedacht zijn*; (8) in de combinatie 'onvoltooid deelwoord + substantief' niet geparafraseerd kan worden als 'substantief + rel.pron. + persoonsvorm (ott,actief)' met behoud van lexicale betekenis. Zo is het deelwoord in *een spannende film* adjectivisch, omdat het niet geparafraseerd kan worden als *een film die spant*, terwijl het deelwoord in *een doodlopende straat* een werkwoord is, aangezien het geparafraseerd kan worden als *een straat die doodloopt*; (9) in de combinatie 'voltooid deelwoord + *zijn*' voorop moet staan in bijzinnen; vgl. het adjectivische deelwoord in * *dat hij na de lange strijd was uitgeput* met het verbale in *dat hij toen al was vertrokken*.

2.3.2 Declaraties

De werkwoorden worden met verschillende features geassocieerd, afhankelijk van of ze als persoonsvorm gebruikt zijn of als buigbare vorm.

[D08] <POS = werkwoord> $\Longrightarrow$ <WVORM>
[D09] <WVORM = persoonsvorm> $\Longrightarrow$ <PVTIJD, PVAGR>
[D10] <WVORM = buigbaar> $\Longrightarrow$ <POSITIE, BUIGING>
[D05] <POSITIE = nominaal> $\Longrightarrow$ <GETAL-N>

De persoonsvormen hebben dus drie features en de niet-finiete (of buigbare) vormen eveneens, behalve bij nominaal gebruik; dan hebben ze er vier.

2.3.3 Partities

[P11] WVORM = persoonsvorm, buigbaar (infinitief, onvdw, voltdw).
[P12] PVTIJD = tegenwoordig, verleden, conjunctief.
[P13] PVAGR = enkelvoud, meervoud, met-t.
[P08] POSITIE = prenominaal, nominaal, vrij.
[P09] BUIGING = zonder, met-e.
[P10] GETAL-N = zonder-n, meervoud-n.

11. WVORM. De buigbare werkwoordsvormen zijn de infinitief en de deelwoorden. De infinitieven hebben het suffix *-en* of in enkele gevallen *-n* (*zijn, gaan, slaan, staan, doen, zien*) en de onvoltooide deelwoorden worden gevormd door toevoeging van *-d* of *-de* aan de infinitief. De voltooide deelwoorden hebben het suffix *-d*, *-t* of *-en* en in vele gevallen ook het prefix *ge-*. Het onderscheiden van voltooid en passief deelwoord is geen taak voor de tagger.

12. PVTIJD. De tegenwoordige tijd is niet morfologisch gemarkeerd, de verleden tijd heeft het suffix *-de* of *-te* (suppletie voor de onregelmatige werkwoorden) en de conjunctief het suffix *-e* (*moge, leve, kome*), behalve wanneer de stam op een klinker eindigt (*het zij zo, het ga je goed*). Enkele werkwoorden hebben meer dan één grondvorm voor de tegenwoordige tijd (*kun/kan*, *zul/zal* en *ben/is/wees*). Dit fenomeen komt ook voor in de verleden tijd (*joeg/jaagde*, *wou/wilde*) en in de conjunctief (*zij, weze, ware*). De imperatief wordt niet onderscheiden van de tegenwoordige tijd.

13. PVAGR. Bij afwezigheid van een AGR suffix wordt de waarde 'enkelvoud' toegekend, omdat zulke vormen steeds enkelvoudig zijn; de persoonswaarde is variabel: eerste persoon (*ik kom, ik kwam, ik speelde, ik maakte*), tweede persoon (*kom je, je mag, je kwam, je speelde, je maakte, kom*) of derde persoon (*hij mag, ze kwam, het speelde, hij maakte, leve de koning*). In combinatie met de PVTIJD distinctie levert dit voor de meeste werkwoorden drie vormen op: één voor de tegenwoordige tijd en de imperatief (*kom, leef*), één voor de verleden tijd (*kwam, leefde*) en één voor de conjunctief (*kome, leve*). Als een werkwoord meer dan één vorm heeft voor een bepaalde PVTIJD waarde, dan zijn er uiteraard ook meer combinaties mogelijk met PVAGR. Zo is niet alleen *kun* de enkelvoudsvorm van de tegenwoordige tijd van *kunnen*, maar ook *kan*; idem dito voor *ben* en *is* in het geval van *zijn*.

De persoonsvormen met een *-(e)n* suffix zijn steeds meervoudig en krijgen dus de waarde 'meervoud'; ook hier is de persoonswaarde variabel (*wij/jullie/zij komen, wij/jullie/zij kwamen, wij/jullie/zij zijn*). In combinatie met de PVTIJD distinctie zijn er voor de meeste werkwoorden twee vormen: één voor de tegenwoordige tijd (*komen, leven, maken*) en één voor de verleden tijd (*kwamen, leefden, maakten*). De imperatief- en conjunctiefvormen nemen dit suffix niet.

De vormen met een *-t* suffix kunnen zowel enkelvoudig als meervoudig zijn, en krijgen de waarde 'met-t'; ze kunnen van de tweede persoon zijn (*je komt,*

je bent, gij waart, gaat u zitten, geeft acht) of van de derde (*hij komt*). De meeste werkwoorden hebben twee vormen met dit suffix: één voor de tegenwoordige tijd en de imperatief (*komt, geeft, gaat*) en één voor de verleden tijd (*kwaamt, gaaft*); die laatste bestaan wel alleen voor de onregelmatige werkwoorden in combinatie met een *gij*-subject, en komen dus vooral in het Vlaams voor. Zoals te verwachten is, beschikt *zijn* ook hier over een grotere variatie: in plaats van één vorm voor de tegenwoordige tijd en de imperatief heeft het er drie (*bent, zijt, weest*). De conjunctiefvormen nemen dit suffix niet. De waarde 'met-t' wordt alleen toegekend, als *-t* een AGR suffix is; als het deel is van de stam, zoals in *hij zit*, dan wordt de waarde 'enkelvoud' toegekend.

08. POSITIE. Net als de adjectieven komen de deelwoorden in diverse soorten van posities voor: prenominale (*een fraai versierde kerstboom, een slapend lid*), nominale (*de gedupeerde, het geschrevene, de wachtenden*) en vrije (*we zijn opgelicht, wat is hier gaande*, *achternagestaard door de menigte reed hij langzaam weg, hij liep luid lachend weg*). Er worden geen aparte tags voorzien voor postnominaal gebruik, omdat de deelwoorden in die positie geen variatie vertonen en dus systematisch gelijk zijn aan vrij gebruikte deelwoorden (*een boom versierd met slingers, alle burgers residerend in Brussel*). Tot het vrije gebruik behoren ook de gevallen waarin het deelwoord het verbale complement is van een hulpwerkwoord (*ze hebben ons niets gezegd, we worden morgen ontslagen*). Bij vrij gebruik wordt het onvoltooid deelwoord soms ingeleid door *al*, als in *al doende leert men.*

Infinitieven worden vooral gebruikt als complementen van andere werkwoorden, meestal hulpwerkwoorden. Naast dit vrije gebruik is er het nominale (*het schaatsen*) en het prenominale (*de nog te lezen post*). Bij prenominaal gebruik wordt de infinitief steeds ingeleid door *te*; bij nominaal gebruik wordt de infinitief vaak maar niet altijd ingeleid door *het.* Bij vrij gebruik treft men zowel infinitieven aan met *te* als zonder *te.* Een voordeel van deze behandeling is dat ze geen systematische ambiguïteiten introduceert: het is niet nodig om verschillende woordsoorten toe te kennen aan de infinitieven in *ze kan lezen, dat was te verwachten, lang wachten is vervelend, het lossen van de duiven* en *de nog te lezen stukken.*

09. BUIGING. In prenominale posities vertonen de deelwoorden—net als de adjectieven—variatie tussen vormen met en vormen zonder *-e* (*een slapend kind, een slapende man*; *een getemd paard, een getemde feeks*). In nominale posities nemen de deelwoorden steeds een buigings *-e* (*het geschrevene, het*

vervelende). In vrije (predikatieve en adverbiale) posities neemt het voltooid deelwoord—net als de adjectieven—geen buigings *-e*; het onvoltooid deelwoord komt in vrije posities zowel met als zonder sjwa voor (*al doende leert men, hij liep (al) zingend de trap af*), maar die sjwa beschouwen we als een optioneel deel van het deelwoordaffix, en niet als een buigingsuitgang, aangezien het voorkomen ervan in het geheel niet wordt bepaald door de regels die ten grondslag liggen aan het gebruik van de buigings *-e*. Bij de infinitieven is er minder variatie, omdat die bijna allemaal op *-en* eindigen, maar er zijn enkele uitzonderingen (*niet mis te verstane bewoordingen, een niet te weerstane verleiding*).

In tegenstelling tot de adjectieven hebben de buigbare werkwoorden geen *-s* vormen: deelwoorden met een *-s* suffix worden immers tot de adjectieven gerekend (zie A.) en infinitieven zijn niet compatibel met dit suffix (*iets te eten(*s), niets te melden(*s)*). Het suffix in combinaties als *nog vier uur gaans, wetens en willens, tot ziens* en *tot bloedens toe*, is geen buigingsuitgang, maar een derivationeel suffix dat bijwoorden afleidt van infinitieven of substantieven (vgl. *deel-s, daag-s*), zie de ANS-97, p. 738.

10. GETAL-N. Bij nominaal gebruik kunnen de deelwoorden een meervoudssuffix nemen (*de gedupeerden, de wachtenden*); net als bij de adjectieven is dit alleen mogelijk wanneer naar personen wordt verwezen. De nominaal gebruikte infinitief is steeds enkelvoudig (*het wachten valt me zwaar*).

2.3.4 Implicaties

[I06] <BUIGING = met-e> $\Longrightarrow$ <POSITIE = (pre)nominaal>
[I07] <WVORM = infinitief, POSITIE = nominaal> $\Longrightarrow$
<BUIGING = zonder, GETAL-N = zonder-n>
[I08] <PVTIJD = conjunctief> $\Longrightarrow$ <PVAGR = enkelvoud>

De eerste implicatie betreft de buigbare vormen en geldt ook voor de adjectieven en de determiners. De tweede stelt dat nominaal gebruikte infinitieven geen buigings- of meervoudsaffix kunnen nemen. De derde stelt dat de conjunctiefvormen steeds enkelvoudig zijn; de meervoudsvormen zijn immers systematisch gelijk aan die van de tegenwoordige tijd.

2.3.5 De tags

Als we ons tot de maximaal specifieke combinaties beperken, dan zijn er voor de persoonsvormen negen.

[T301]	WW(pv,tgw,ev)	kom, speel
[T302]	WW(pv,tgw,mv)	komen, spelen
[T303]	WW(pv,tgw,met-t)	komt, speelt
[T304]	WW(pv,verl,ev)	kwam, speelde
[T305]	WW(pv,verl,mv)	kwamen, speelden
[T306]	WW(pv,verl,met-t)	kwaamt, gingt
[T309]	WW(pv,conj,ev)	kome, leve de koning

Voor de infinitief zijn er vier combinaties en voor de deelwoorden telkens vijf.

[T310]	WW(inf,prenom,zonder)	de nog te lezen post
[T311]	WW(inf,prenom,met-e)	een niet te weerstane verleiding
[T312]	WW(inf,nom,zonder,zonder-n)	(het) spelen, (het) schaatsen
[T314]	WW(inf,vrij,zonder)	zal komen
[T315]	WW(vd,prenom,zonder)	een verwittigd man, een gekregen paard
[T316]	WW(vd,prenom,met-e)	een getemde feeks
[T317]	WW(vd,nom,met-e,zonder-n)	het geschrevene, een gekwetste
[T318]	WW(vd,nom,met-e,mv-n)	gekwetsten, gedupeerden
[T320]	WW(vd,vrij,zonder)	is gekomen, een boom versierd met slingers
[T321]	WW(od,prenom,zonder)	een slapend kind
[T322]	WW(od,prenom,met-e)	een piano spelende aap, slapende kinderen
[T323]	WW(od,nom,met-e,zonder-n)	het resterende, een klagende
[T324]	WW(od,nom,met-e,mv-n)	de wachtenden
[T326]	WW(od,vrij,zonder)	liep lachend weg, al doende leert men

De intermediaire waarde 'buigbaar' voor WVORM is gebruikt voor het formuleren van de declaraties (D10), maar is niet bedoeld voor onderspecificatie in de tags. Vormen die zowel infinitief als voltooid deelwoord kunnen zijn (*bekomen*, *vergaan*), worden dus gedisambigueerd.

2.3.6 Lemmatisering

Als lemma gebruiken we niet de stam, maar de infinitiefvorm.

2.4 TELWOORDEN

Om tal van redenen zou het taalkundig gesproken de voorkeur verdienen om de hoofdtelwoorden als substantieven te behandelen, meer bepaald als soortnamen, en de rangtelwoorden als adjectieven. Ter wille van de herkenbaarheid en de daaruit voortvloeiende eis van conformiteit aan de ANS-praktijk en de EAGLES-aanbevelingen is daar niet voor gekozen. Het gevolg is wel dat de criteria voor het onderscheiden van hoofdtelwoorden en substantieven enigszins kunstmatig zijn; hetzelfde geldt voor de toepassing van het POSITIE onderscheid op de hoofdtelwoorden.

2.4.1 Afgrenzing

Tot de telwoorden rekenen we alle woorden die aparte vormen hebben voor hoofd- en rangtelwoord. Daartoe behoren uiteraard de namen van de getallen, zoals *twee(de), dertig(ste), zeshonderd(ste), ...*, maar ook de woorden *elfendertig(ste), tig(ste), hoeveel(ste), evenveel(ste)* en *zoveel(ste)*. Woorden als *beide*, *veel* en *weinig* horen op basis van dit criterium niet bij de telwoorden; in plaats daarvan worden ze als onbepaalde voornaamwoorden behandeld, zie 2.5. Eveneens tot de onbepaalde voornaamwoorden rekenen we het gebruik van *één* in verbindingen als *het één en ander* en *één en al aandacht*; kenmerkend voor deze verbindingen is dat *één* er niet in complementaire distributie staat met andere telwoorden.

Telwoorden dienen te worden onderscheiden van substantieven. Voor de vormen met een meervoudssuffix kan het onderscheid tussen telwoord en substantief worden toegelicht aan de hand van een minimaal paar als *met z'n zevenen* vs. *hij heeft twee zevens*. In het eerste voorbeeld is *zeven* als een nominaal gebruikt telwoord behandeld (GETAL-N = meervoud-n) en in het

tweede als een substantief (GETAL = meervoud). Op vergelijkbare wijze worden de meervoudsvormen in *met z'n tweetjes* en *met z'n honderden tegelijk* als nominaal gebruikte telwoorden getagd, terwijl de meervoudsvormen in *ik heb nog twee tientjes* en *honderden kisten* als substantieven worden behandeld. De vormen zonder meervoudssuffix rekenen we tot de substantieven als ze de kern vormen van een enkelvoudige NP, als in *ze heeft een zes*, *deze vijf is mooier dan die* en *zes is/*zijn deelbaar door drie*; als de resulterende NP daarentegen meervoudig is, dan gaat het om een telwoord, als in *deze vijf zijn mooier dan die* en *er zijn/*is er zes ontsnapt*.

Telwoorden dienen ook te worden onderscheiden van bijwoorden. De tijdsaanduidende vormen in *tegen énen* en *na zessen* bijv. zijn geen meervoudige telwoorden met een GETAL-N uitgang, zoals in *met z'n zessen*, maar bijwoorden. Het suffix *-en* kan hier dus gezien worden als een derivationeel affix voor de afleiding van bijwoorden. Het komt ook voor in vormen als *voren, achteren, onderen*, zie 2.7. Eveneens tot de bijwoorden rekenen we vormen als *halfzeven*; telwoorden kunnen dat immers niet zijn, aangezien er geen corresponderend rangtelwoord is.

Breuken als *vijf achtste* worden behandeld als combinaties van resp. een hoofdtelwoord en een rangtelwoord. Als ze in één woord geschreven zijn, als in *een tweederde meederheid, een viertiende baan, ik werk nu viertiende*, dan rekenen we ze tot de adjectieven; op die manier kunnen we het adjectief *viertiende* ook onderscheiden van het rangtelwoord *veertiende*. Combinaties als *anderhalve, zesenhalve, driekwart* rekenen we eveneens tot de adjectieven; ze hebben immers geen corresponderend rangtelwoord.

2.4.2 Declaraties

[D11] <POS = telwoord> $\Longrightarrow$ <NUMTYPE, POSITIE>
[D05] <POSITIE = nominaal> $\Longrightarrow$ <GETAL-N>
[D12] <NUMTYPE = hoofdtelwoord, POSITIE = nominaal> $\Longrightarrow$ <GRAAD>
[D13] <POS = telwoord, POSITIE = prenominaal> $\Longrightarrow$ <NAAMVAL>

In tegenstelling tot de adjectieven hebben de telwoorden geen buigingsfeature, aangezien die distinctie systematisch geneutraliseerd is: de hoofdtelwoorden nemen immers nooit een buigings-*e* en de rangtelwoorden eindigen steeds op een sjwa. De enige vorm waarvoor het onderscheid relevant zou kunnen zijn, is *één* (*die éne keer*), maar omdat die vorm ook als onbepaald voornaamwoord gebruikt wordt en in die functie sowieso een buigingsfea-

ture heeft, behandelen we de vorm *éne* als een onbepaald voornaamwoord, en niet als een telwoord.

2.4.3 Partities

[P14] `NUMTYPE = hoofdtelwoord, rangtelwoord.`
[P08] `POSITIE = prenominaal, nominaal, vrij.`
[P10] `GETAL-N = zonder-n, meervoud-n.`
[P05] `GRAAD = basis, diminutief.`
[P07] `NAAMVAL = standaard, bijzonder.`

14. NUMTYPE. Tot de hoofdtelwoorden behoren o.m. *één, twee, drie, zoveel*. De rangtelwoorden zijn de corresponderende vormen met het suffix *-ste* of *-de*, cf. *eerste, tweede, derde, zoveelste*.

08. POSITIE. De hoofdtelwoorden worden als prenominaal behandeld wanneer ze voorafgaan aan een substantief en bepalen hoeveel exemplaren er van het door het substantief aangeduide bedoeld zijn, zoals in *één hond* en *vijf kinderen*. In verbindingen als *vijf juli* is het telwoord dus niet prenominaal, aangezien het niet bepaalt om hoeveel juli's het gaat. Net als de adjectieven kunnen de prenominale telwoorden elliptisch gebruikt worden, zoals *twee* in *hij krijgt vijf rode knikkers en jij twee groene*. De rangtelwoorden worden als prenominaal behandeld wanneer ze voorafgaan aan een substantief en bepalen welke plaats het door het substantief aangeduide heeft, zoals in *het vijfde kind*.

Telwoorden met een meervoudsaffix (zie GETAL-N) of een diminutiefsuffix (zie GRAAD) zijn steeds nominaal. Vormen zonder die affixen zijn nominaal wanneer ze in dezelfde posities voorkomen als de vormen met zulke affixen, zoals in zinnen met een kwantitatief *er*, cf. *er is er één(tje) ontsnapt*. Dat dit laatste gebruik niet prenominaal is blijkt o.m. uit het feit dat de diminutiefvormen niet prenominaal gebruikt kunnen worden. Voorbeelden van nominaal gebruikte rangtelwoorden zijn *Lodewijk de Veertiende, de dertiende van elke maand, zij was (de) vierde* en *altijd (de) tweede eindigen*.

De hoofdtelwoorden die niet (pre)nominaal gebruikt zijn, beschouwen we als vrij. Hiertoe behoort niet alleen het predikatieve gebruik, als in *hij wordt zestig*, en het adverbiale, als in *hij reed minstens honderd*, maar ook het gebruik in NPs waarin het telwoord niet het aantal aanduidt van datgene wat door het nominale hoofd van de NP wordt aangeduid; relevante voorbeelden

betreffen het gebruik van *twintig* in *twintig juli, de jaren twintig, pagina twintig* en *twee euro twintig*. In die laatste drie voorbeelden gaat het dus niet om postnominaal gebruik: in *de jaren twintig* bijv. gaat het niet om een periode van twintig jaar, maar om een periode van tien jaar (1920 t.e.m. 1929), en in *pagina twintig* gaat het niet om twintig pagina's, maar om 1 pagina.

Als het telwoord in verschillende woorden gesplitst is, dan wordt elk van die delen op dezelfde manier behandeld. In *tweehonderd veertien pagina's* zijn de telwoorden dus allebei prenominaal en in *pagina tweehonderd veertien* zijn ze allebei vrij.

10. GETAL-N. Bij nominaal gebruik kunnen de telwoorden het meervoudssuffix *-en* nemen, cf. *met z'n vieren*. Van hetzelfde type zijn de vormen in *wij tweeën* en *zij vijven* (vergelijkbaar met *wij fietsers*). Ook rangtelwoorden kunnen een meervoudssuffix nemen in nominale posities, cf. *de eersten*. Bemerk de semantische beperking tot verwijzing naar personen.

Van een heel andere aard zijn vormen als *na vieren* in de betekenis van *na vier uur*. In dit geval gaat het niet om een meervoudsvorm die naar personen verwijst, maar om een bijwoord. Dat dit geen meervoudsvorm kan zijn, blijkt o.m. uit de combinatie *tegen énen*. Een motivering voor de classificatie als bijwoord is dat de toevoeging van *-en* wel vaker een bijwoord oplevert: toepassing op het voorzetsel *voor* bijv. geeft het bijwoord *voren*, zie 2.7.

05. GRAAD. Typisch Nederlands is het bestaan van diminutiefvormen voor een aantal hoofdtelwoorden, cf. *op z'n eentje, met z'n tweetjes*. Bemerk dat de getalsonderscheiding ook hier relevant is.

07. NAAMVAL. Datiefvormen hebben het suffix *-en* of *-er* en komen vooral in vaste verbindingen voor (*te elfder ure, te enen male*). Genitiefvormen zijn zeer zeldzaam; een mogelijk voorbeeld is het Bijbelse *eens geestes zijn*. Omwille van die zeldzaamheid maken we—net als bij de adjectieven—alleen een onderscheid tussen vormen met een naamvalssuffix ('bijzonder') en vormen zonder naamvalssuffix ('standaard').

2.4.4 Implicaties

[I09] <NUMTYPE = rangtelwoord> $\Longrightarrow$ <POSITIE $\neq$ vrij>

De rangtelwoorden worden alleen (pre)nominaal gebruikt.

2.4.5 De tags

Voor de hoofdtelwoorden zijn er zeven maximaal specifieke combinaties.

[T401]	TW(hoofd,prenom,stan)	vier cijfers
[T402]	TW(hoofd,prenom,bijz)	eens geestes zijn, ten enen male
[T403]	TW(hoofd,nom,zonder-n,basis)	
[T404]	TW(hoofd,nom,mv-n,basis)	met z'n vieren
[T405]	TW(hoofd,nom,zonder-n,dim)	er is er eentje ontsnapt, op z'n eentje
[T406]	TW(hoofd,nom,mv-n,dim)	met z'n tweetjes
[T407]	TW(hoofd,vrij)	veertig worden, zoveel sneller, pagina vijf, de jaren zestig, zes juli

Voor de rangtelwoorden zijn er vanwege de afwezigheid van GRAAD en het ontbreken van vrij gebruik slechts vier.

[T408]	TW(rang,prenom,stan)	de vierde man
[T409]	TW(rang,prenom,bijz)	te elfder ure
[T410]	TW(rang,nom,zonder-n)	het eerste, (de) vierde zijn
[T411]	TW(rang,nom,mv-n)	de eersten, iets aan derden verkopen

2.4.6 Lemmatisering

Als lemma gebruiken we de basisvorm van het hoofdtelwoord. Enkele voorbeelden zijn *drie, zestig, zoveel, tig*.

2.5 VOORNAAMWOORDEN

De voornaamwoorden vormen een heterogene klasse en worden daarom in sommige tagsets, zoals PAROLE, in twee aparte woordsoorten ondergebracht, nl. de pronomina, die doorgaans zelf kern zijn van een NP, en de determiners, die meer gebruikt worden als bepaling bij een substantief. Voor CGN is evenwel besloten om beide tot dezelfde woordsoort te rekenen en om het pronomen/determiner onderscheid in termen van een afzonderlijk

feature weer te geven (PDTYPE). Die keuze is deels gemotiveerd door het feit dat zowel de ANS-97 als EAGLES geen verschillende woordsoorten voor pronomina en determiners gebruiken, en deels door het feit dat de pronomina en determiners naast hun verschillen ook een aantal kenmerken gemeen hebben (VWTYPE, NAAMVAL).

2.5.1 Afgrenzing

Voor de identificatie van de voornaamwoorden is in grote mate de ANS-97 gevolgd. Op enkele punten is er echter van afgeweken.

Een eerste punt van divergentie betreft de zgn. voornaamwoordelijke bijwoorden *hier, daar, waar, ergens, nergens* en *overal*. De classificatie ervan als bijwoorden in de ANS-97 is wellicht ingegeven door het feit dat ze in een combinatie als *ik woon hier/daar al zestien jaar* de kern zijn van een locatieve bepaling. In het kader van de CGN-tagset is die overweging echter niet doorslaggevend, omdat de woordsoortindeling geheel op vormelijke en niet op functionele criteria gebaseerd is; zo zijn *maandag* en *jaren* in *ze komen maandag* en *ik heb jaren in Portugal gewoond* wel de kern van een temporele bepaling, maar qua woordsoort zijn het allebei substantieven. Op soortgelijke wijze kan men stellen dat de voornaamwoordelijke bijwoorden pronomina zijn die o.m. als plaatsbepaling gebruikt worden. Een voordeel van deze analyse is dat ze ook van toepassing is op hun gebruik in PPs: in combinatie met een voorzetsel nemen ze immers de plaats in van de voornaamwoorden *dit, dat, wat, iets, niets* en *alles*, vgl. het ongrammaticale *op wat* met het grammaticale *waarop*. Orthografisch vormt deze combinatie weliswaar één geheel, zodat de tagger het ook als één woord zal behandelen, maar dat is niet het geval wanneer het voornaamwoord van het voorzetsel gescheiden is, als in *daar wacht ik niet op*. In zulke gevallen moet *daar* een eigen woordsoort krijgen en de beste kandidaat daarvoor is 'voornaamwoord', niet alleen omwille van de complementaire distributie met de voornaamwoorden, maar ook omwille van de relevantie van het feature VWTYPE. Zo zijn *hier* en *daar* aanwijzend, *waar* vragend of betrekkelijk, en *ergens, nergens* en *overal* onbepaald.

Een tweede punt van divergentie betreft het veel besproken woord *er*. De ANS-97 behandelt het als een bijwoord en onderscheidt er vier gebruiken voor. Twee ervan komen overeen met die van de voornaamwoordelijke bijwoorden, m.n. het locatieve, als in *ik kom er niet graag*, en het PP-gebruik, als in *ze wacht er al maanden op*, waarin *er* in complementaire distributie is met het persoonlijke voornaamwoord *het*. Naar analogie met de behande-

ling van *daar* en cs. rekenen we *er* dan ook tot de voornaamwoorden. Die behandeling is overigens ook toepasselijk op de twee andere gebruiken. Een ervan is wat de ANS-97 het presentatieve noemt. Hiermee is het voorlopige of expletieve onderwerp bedoeld in zinnen als *er staat een man voor de deur*; bemerk dat er ook in dit gebruik sprake is van complementaire distributie met *het*, cf. *er wordt gezegd dat* vs. *het wordt betreurd dat*. Tenslotte is er het kwantitatieve gebruik, als in *ik heb er vijf*, dat wellicht nog duidelijker dan de drie andere van pronominale aard is. Historisch gaat het overigens om de genitiefvorm van een persoonlijk voornaamwoord, zie de ANS-97, 464. Kortom, er zijn diverse argumenten voor de classificatie van *er* als een voornaamwoord, eerder dan als een bijwoord.

Een derde punt van divergentie betreft *veel/meer/meest, weinig/minder/minst* en *beide*. De ANS-97 rekent die tot de telwoorden, maar omdat die woorden i.t.t. de echte telwoorden geen corresponderende rangtelwoorden hebben, volgen we dat gebruik niet. In plaats daarvan rekenen we ze tot de voornaamwoorden, meer bepaald tot de onbepaalde determiners. Dat is overigens ook meer in overeenstemming met de classificatie van hun vertaal-equivalenten in andere talen.

Daarnaast zijn er een aantal woorden die de ANS-97 tot de aanwijzende of onbepaalde voornaamwoorden rekent (of althans in die secties behandelt), maar die in het CGN bij andere woordsoorten worden ondergebracht. Het gaat o.m. om de adjectieven *dergelijk(e), soortgelijk(e), dusdanig(e), zodanig(e), -zelfde* en om de bijwoorden *zelf, genoeg, zat*. Verder vermeldt de ANS-97 bij de onbepaalde voornaamwoorden een aantal meer-woord-combinaties zoals *deze of gene, dit of dat, een en ander, een paar, ...* (p. 356). Voor CGN gaat het in zulke gevallen om combinaties van woorden die elk hun eigen woordsoort krijgen, zoals voegwoord, voornaamwoord, lidwoord, substantief, e.d.

Om uitsluitsel te bieden in geval van twijfel is een lexicon samengesteld waarin alle voornaamwoorden zijn opgenomen, met inbegrip van hun tag(s) en lemma. Dat lexicon is integraal opgenomen in het CGN-lexicon.

2.5.2 Declaraties

De heterogeniteit van deze klasse wordt weerspiegeld door het grote aantal declaraties dat er nodig is om tot passende feature-combinaties te komen.

[D14] <POS = voornaamwoord> $\Longrightarrow$ <VWTYPE, PDTYPE, NAAMVAL>
[D15] <PDTYPE = pronomen> $\Longrightarrow$ <STATUS, PERSOON, GETAL>
[D16] <VWTYPE = persoonlijk, NAAMVAL = standaard, PERSOON = 3,
GETAL = enkelvoud> $\Longrightarrow$ <GENUS>
[D17] <PDTYPE = determiner> $\Longrightarrow$ <POSITIE, BUIGING>
[D05] <POSITIE = nominaal> $\Longrightarrow$ <GETAL-N>
[D18] <PDTYPE = determiner, POSITIE = prenominaal> $\Longrightarrow$ <NPAGR>
[D19] <PDTYPE = gradeerbaar> $\Longrightarrow$ <GRAAD>
[D20] <VWTYPE = bezittelijk> $\Longrightarrow$ <STATUS, PERSOON, GETAL>

Alle voornaamwoorden hebben de features VWTYPE, PDTYPE en NAAMVAL. De pronomina hebben daarnaast ook STATUS, PERSOON en GETAL, desgevallend aangevuld met GENUS. De determiners hebben naast de gemeenschappelijke drie ook POSITIE en BUIGING, desgevallend aangevuld met GETAL-N, NPAGR en/of GRAAD. Voor de bezittelijke voornaamwoorden is er een aparte declaratie toegevoegd; als determiners hebben die uiteraard alle features die met de determiners geassocieerd zijn, maar daarnaast ook de features STATUS, PERSOON en GETAL.

2.5.3 Partities

[P15] VWTYPE = pr (persoonlijk, reflexief), reciprook, bezittelijk,
vb (vragend, betrekkelijk), exclamatief, aanwijzend, onbepaald.
[P16] PDTYPE = pronomen (adv-pronomen), determiner (gradeerbaar).
[P07] NAAMVAL = standaard (nominatief, oblique), genitief, datief.
[P17] STATUS = vol, gereduceerd, nadruk.
[P18] PERSOON = persoon (1, 2 (2v, 2b), 3 (3p (3m, 3v), 3o)).
[P04] GETAL = getal (enkelvoud, meervoud).
[P06] GENUS = masculien, feminien, onzijdig.
[P08] POSITIE = prenominaal, nominaal, vrij.
[P09] BUIGING = zonder, met-e.
[P19] NPAGR = agr (evon, rest (evz, mv)), agr3 (evmo,
rest3 (evf, mv)).
[P10] GETAL-N = zonder-n, meervoud-n.
[P05] GRAAD = basis, comparatief, superlatief, diminutief.

15. VWTYPE. De indeling in negen types stemt overeen met die van de ANS-97. Net als in EAGLES zijn er gemeenschappelijke intermediaire waarden voor de persoonlijke en reflexieve voornaamwoorden enerzijds, en voor

de vragende en betrekkelijke voornaamwoorden anderzijds; dat bespaart ons het postuleren van een systematische ambiguïteit voor voornaamwoorden die beide rollen kunnen vervullen (resp. *me, mij, ons, je* en *wat, welke*).

16. PDTYPE. De indeling in pronomina en determiners is onafhankelijk van de VWTYPE classificatie, maar dat neemt niet weg dat er verbanden zijn. Kort gezegd, (1) de persoonlijke, reflexieve en reciproke voornaamwoorden zijn pronomina; (2) de bezittelijke voornaamwoorden zijn determiners; (3) de vragende, betrekkelijke en exclamatieve voornaamwoorden zijn—met uitzondering van *welk(e)* en *hetgeen*—pronomina; (4) de aanwijzende en onbepaalde voornaamwoorden zijn deels determiners en deels pronomina. Binnen de pronomina onderscheiden we een apart subtype voor de voornaamwoordelijke bijwoorden en *er* (adverbiaal-pronomen) en binnen de determiners onderscheiden we een apart subtype voor de onbepaalde determiners met trappen van vergelijking, zoals *veel* en *weinig* (gradeerbaar). In het lexicon is voor elk voornaamwoord aangegeven wat z'n PDTYPE waarde is. De criteria die bij de toekenning van die waarden gebruikt zijn, worden hier kort toegelicht.

Het onderscheid tussen pronomina en determiners vertoont gelijkenis met het onderscheid dat de ANS-97 maakt tussen zelfstandige en niet-zelfstandige voornaamwoorden, maar het valt er niet mee samen. Daar zijn twee redenen voor: (1) determiners worden—net als de adjectieven—ook zelfstandig gebruikt, met name in nominale posities; (2) de genitiefvormen van de pronomina worden—net als die van de substantieven—ook niet-zelfstandig gebruikt, met name in pre- of postnominale posities. Het is dus niet voldoende om de pronomina te onderscheiden van de niet-zelfstandige of prenominale determiners (A), ze moeten ook worden onderscheiden van de zelfstandig of nominaal gebruikte determiners (B).

A. Een eerste verschil tussen prenominaal gebruikte determiners en pronomina is dat de eerste in naamval moeten overeenstemmen met het substantief dat ze bepalen, terwijl de pronomina die als bepaling bij een substantief gebruikt zijn, steeds de genitiefvorm hebben, ook wanneer het gemodificeerde substantief een standaardvorm is. Het bezittelijke *mijn* is dus een determiner, omdat zijn naamval niet verschillend mag zijn van die van het substantief, vgl. *mijn/*mijns hoofd* en *mijns/*mijn inziens*; idem dito voor het onbepaalde *alle* in *alle/*allen leden* en *te allen/*alle prijze*. De voornaamwoorden in *wiens huis* en *mijns gelijke* daarentegen zijn pronomina, aangezien het contrast in naamval (genitief vs. standaard) niet tot ongrammaticaliteit leidt; hetzelfde geldt voor pronomina die volgen op het woord

dat ze bepalen: in *wie uwer* (= wie van u) en *één hunner* (= één van hun) zijn de nabepalingen pronomina en geen determiners, omdat het verschil in naamval met het voorafgaande woord geen incompatibiliteit veroorzaakt.

Een tweede verschil is dat de prenominaal gebruikte determiners verplicht overeenkomst vertonen in genus en getal met het substantief dat ze bepalen, terwijl de pronomina die als bepaling bij een substantief gebruikt zijn, zulke overeenkomst niet hoeven te vertonen. Het demonstratieve voornaamwoord in *deze tafel/boeken* is dus een determiner, omdat het alleen compatibel is met enkelvoudig zijdige of meervoudige substantieven, cf. **deze boek*. Idem dito voor het onbepaalde *elke*, dat alleen met enkelvoudige zijdige substantieven combineert *elke gans/*boek/*ganzen*. De voornaamwoorden in *wiens paarden* en *diens hemden* daarentegen zijn pronomina, want hoewel ze allebei enkelvoudig en zijdig zijn, zijn ze toch compatibel met meervoudige en onzijdige substantieven.

Een derde verschil blijkt bij de volgende parafrasetest: als bij vervanging van de genitief door een PP, zowel het voornaamwoord als het substantief in de PP verschijnen, dan gaat het om een determiner; als daarentegen alleen het voornaamwoord in de PP verschijnt, dan gaat het om een pronomen. Deze test laat duidelijk het verschil zien tussen het bezittelijke *mijns*, dat een determiner is, en het persoonlijke *mijns*, dat een pronomen is.

mijns inziens	===	naar mijn inzien	DET
mijns gelijke	===	de gelijke van mij	PRO

B. Voor het onderscheiden van pronomina en nominaal gebruikte determiners zijn andere testen vereist, aangezien m.b.t. de net vermelde criteria de nominaal gebruikte determiners zich net als de pronomina gedragen. In *met aller instemming* bijv. heeft alleen het voornaamwoord de genitiefvorm, is er geen overeenstemming met het substantief in genus en getal (*aller* is meervoud en *instemming* enkelvoud) en bevat de PP-parafrase alleen het voornaamwoord ('met de instemming van allen'); toch is *aller* geen pronomen maar een nominaal gebruikte determiner. Om dit onderscheid te maken is de morfologische structuur van het voornaamwoord de beslissende factor: (1) nominaal gebruikte determiners bevatten namelijk steeds een buigings -*e*, en (2) als ze een meervoudsvorm hebben, dan is die gemarkeerd door het suffix -*n*, met de restrictie dat de resulterende vorm alleen naar personen kan verwijzen: *aller* voldoet aan beide criteria. De pronomina daarentegen bevatten geen buigings -*e*, en als ze een aparte meervoudsvorm hebben, dan wordt die niet door toevoeging van -*n* gevormd maar door andere vormen van suffigering of suppletie. Zo is het voornaamwoord in *met*

wier instemming een pronomen, aangezien het geen buigings *-e* bevat en de meervoudsvorm niet door *-n* wordt gevormd.

Kortom, het gaat hier niet om een tweeledig onderscheid tussen zelfstandig en niet-zelfstandig, zoals in de ANS-97, maar om een drieledig onderscheid tussen prenominale (bijvoeglijke) determiners, nominale (zelfstandige) determiners en pronomina. Ter illustratie en verdere verduidelijking passen we de criteria toe op een concreet voorbeeld, *schrijver dezes*. Uit de parafrase ('de schrijver van dit') blijkt dat alleen *dezes* een genitief is; het is dus zeker geen bijvoeglijke determiner, maar ofwel een pronomen ofwel een nominale determiner. De keuze wordt bepaald door de morfologische structuur: aangezien *deze* de buigings *-e* bevat en een meervoud heeft op *-n*, dat uitsluitend naar personen verwijst (*dezen*), is het geen pronomen, maar een nominale determiner. Hetzelfde geldt voor de genitief in *de 20ste dezer*.

In sommige gevallen geven de criteria enigszins onverwachte resultaten. Zo is het aanwijzende *deze* een determiner, zowel bij nominaal als bij prenominaal gebruik, terwijl *dit, dat* en *die* determiners zijn bij prenominaal gebruik, maar pronomina bij zelfstandig gebruik: ze bevatten immers geen buigings *-e* en hebben geen meervoud op *-n*. Dat kan op het eerste gezicht willekeurig lijken, maar bij nader inzien blijkt dat de zelfstandig gebruikte *dit, dat, die* (i.t.t. *deze*) nog een aantal andere kenmerken vertonen van pronomina. Zo hebben de zelfstandig gebruikte *dit* en *dat* aparte nadruksvormen (*ditte, datte*), net als een aantal andere pronomina (*ikke, watte*), terwijl zulke vormen voor de nominale determiners niet bestaan. Voor *die* geldt bovendien dat het (i.t.t. *deze*) ook als betrekkelijk voornaamwoord gebruikt wordt en in die functie ontegenzeggelijk een pronomen is.

07. NAAMVAL. De standaardvormen hebben geen naamvalssuffix en worden verder verdeeld in nominatief en oblique. Nominatief zijn de vormen die als het subject van persoonsvormen gebruikt worden (*ik, jij, men*); oblique zijn de vormen die o.m. als object van werkwoorden en voorzetsels gebruikt worden (*mij, jou, hem*). Pronomina die zowel oblique als nominatief kunnen zijn (*je, wie, iemand*) krijgen de waarde 'standaard'. Bij de determiners is het onderscheid tussen nominatief en oblique systematisch geneutraliseerd.

De genitief wordt gemarkeerd door *-s* (*mijns inziens, zijns gelijke, wiens hoed, elkaars fiets, iemands auto*) of *-(e)r* (*dezer dagen, aller landen, wier hoed, wie uwer zonder zonde is, tot veler verbazing*). De datief wordt gemarkeerd door *-(e)n* (*met dien verstande, te allen prijze*) of *-(e)r* (*te eniger tijd, te mijner ere*).

Bemerk dat de naamvalsdistinctie ook voor de adverbiale pronomina relevant is. Zo hebben *hier, daar, waar, ergens, nergens* en *overal* de waarde 'oblique', en heeft *er* diezelfde waarde bij locatief en PP-gebruik, maar 'nominatief' bij het presentatieve gebruik en 'genitief' bij het kwantitatieve gebruik. Bij de POS-tagging onderscheiden we het kwantitatieve *er* (NAAMVAL = genitief) van de drie andere gebruiken, maar binnen die laatste maken we geen verder onderscheid tussen nominatief en oblique (NAAMVAL = standaard).

Zoals uit dit overzicht blijkt, speelt naamval een veel belangrijker rol in de analyse van de voornaamwoorden dan in de analyse van de substantieven, adjectieven en telwoorden. Dat wordt ook in de declaraties weerspiegeld: terwijl voor die laatste het feature alleen in specifieke gevallen wordt toegekend (zie D02, D06, D07 en D13), wordt het bij de voornaamwoorden uniform toegekend (D14), dus ook aan de meervoudige en de onverbogen vormen.

17. STATUS. Dit feature wordt alleen toegekend aan de pronomina en de bezittelijke determiners. Tot de gereduceerde vormen rekenen we alle vormen zonder klinker (*'t, z'n, d'r, ...*), de monosyllabische vormen met sjwa (*het, ze, er,...*) en de vormen *zich, ie*. Tot de vormen met nadruk rekenen we de voornaamwoorden met geïncorporeerd *-zelf* of *-lie(den)* en de vormen *ikke, ditte, datte, watte*. Alle andere vormen krijgen de waarde 'vol'.

18. PERSOON. Net als STATUS wordt dit feature toegekend aan de pronomina en de bezittelijke determiners. Dat kan op het eerste gezicht te ruim lijken, omdat de bekende 1-2-3 distinctie alleen relevant is voor de persoonlijke, reflexieve en bezittelijke voornaamwoorden. De reden waarom het feature hier toch een ruimere toepassing krijgt, is dat er binnen de derde persoon een aantal verdere onderscheidingen worden gemaakt die ook voor de andere pronomina relevant zijn. Zo is het onderscheid tussen pronomina die een persoonlijke referent vereisen (3p) en die welke een onpersoonlijke referent vereisen (3o) ook van belang voor de vragende, betrekkelijke en onbepaalde pronomina: *wie, (n)iemand, iedereen* bijvoorbeeld vereisen een persoonlijke referent en *wat,(n)iets, alles* een onpersoonlijke. Bij de pronomina met een persoonlijke referent wordt bovendien een verder onderscheid gemaakt tussen mannelijke (3m) en vrouwelijke referenten (3vr), en ook dat onderscheid is van belang voor de vragende en betrekkelijke pronomina: de genitiefvormen *wiens* en *wier* bijvoorbeeld vereisen —in het enkelvoud— respectievelijk een mannelijke en een vrouwelijke referent. Bemerk dat het hier

niet om het morfo-syntactische genus van het woord gaat, maar om het natuurlijke geslacht van de referent. Het onderscheid kan worden geïllustreerd aan de hand van het contrast tussen *hij* en *hijzelf*. De eerste vorm is masculien en kan zowel een persoonlijke als een onpersoonlijke referent hebben; de tweede vorm is eveneens masculien, maar kan bovendien uitsluitend naar manspersonen verwijzen. Aangezien deze indeling volgens de aard van de referent eerder van semantisch-pragmatische dan van morfo-syntactische aard is, zou men kunnen stellen dat ze niet thuishoort in de CGN tagset; de reden waarom ze toch is opgenomen is dat ze in het pronominale systeem zo'n belangrijke rol speelt dat zowel de ANS-97 als EAGLES soortgelijke distincties maken, c.q. aanbevelen. Dezelfde opmerking geldt voor de verdere differentiatie binnen de tweede persoon, waar we het bekende onderscheid maken tussen vertrouwelijke vormen (*je, jij, jou, jullie*) en beleefdheidsvormen (*u*); die krijgen resp. de waarden '2v' en '2b'. Bij de vooral in Vlaanderen gebruikte *gij*-vormen (*ge, gij, u*) is het onderscheid geneutraliseerd en wordt dus de waarde '2' toegekend. Voor de eerste persoon worden geen verdere subtypes onderscheiden. De generische waarde 'persoon', tenslotte, wordt toegekend aan de reciproke voornaamwoorden (*elkaar, mekaar, ...*), omdat hun antecedent van eender welke persoon kan zijn.

04. GETAL. Het gaat hierbij om het getal van de referent, m.a.w. om het onderscheid tussen verwijzing naar een enkelvoudige entiteit (*hij, iemand, jouw*) en verwijzing naar een groep van entiteiten (*wij, hun, jullie*). De onderscheiding is relevant voor de pronomina en de bezittelijke voornaamwoorden, maar niet voor de andere determiners. De meervoudsvorming wordt niet gekenmerkt door suffigering, zoals bij de substantieven, maar door suppletie. Bij een aantal pronomina is het onderscheid geneutraliseerd (*zich, wie*); die krijgen de generische waarde 'getal'. Het verschil tussen GETAL en GETAL-N wordt bij dit laatste feature toegelicht.

06. GENUS. Naast het natuurlijke geslacht, dat als een subpartitie van de derde persoon wordt behandeld, is er het morfo-syntactische genus. Dit feature wordt alleen toegekend aan de enkelvoudige standaardvormen van de persoonlijke voornaamwoorden van de derde persoon. Het onderscheid tussen masculien en feminien, dat bij de substantieven geïgnoreerd is, wordt hier wel gemaakt.

08. POSITIE. Determiners worden vooral prenominaal gebruikt (*deze tafel, welke kast*) en nominaal (*heb je deze al, welke bedoel je, de zijne*). Post-

nominaal gebruik, als in *kindeke mijn*, is zo uitzonderlijk dat we er geen voorzieningen voor treffen. Vrij gebruik komt vooral voor bij de onbepaalde determiners *elk* en *ieder*, als in *ze hebben elk/ieder een appel gekregen*, en bij determiners die aan een hele NP voorafgaan, als in *al de mensen, zulk een ellende, welk een dwaasheid*. Bemerk dat die laatste positie niet prenominaal is, omdat de determiner hier geen morfologische variatie (en geen overeenkomst met het substantief) vertoont. De behandeling als vrije determiner is geschikter, omdat de pre-NP positie typisch is voor bijwoordelijk gebruikte elementen, cf. *zelfs een boswachter, ook de mannen, alleen de kinderen, precies die vlinder*.

09. BUIGING. Vele determiners hebben twee verschijningsvormen, cf. *welk(e), ieder(e), zulk(e), al(le), ons/onze*. Er zijn net als bij de buigbare werkwoorden geen *-s* vormen.

10. GETAL-N. Bij nominaal gebruik kunnen sommige determiners een meervoudssuffix nemen (*dezen, sommigen, enkelen, allen, de zijnen*). Net als bij de adjectieven geldt de beperking tot persoonlijke referenten. Aangezien het feature alleen aan de nominaal gebruikte determiners wordt toegekend, en dus niet aan de pronomina, is er slechts één type van voornaamwoorden dat zowel een GETAL als een GETAL-N feature heeft, m.n. de nominaal gebruikte bezittelijke voornaamwoorden. Dat het nuttig is om beide features te hebben kan worden aangetoond aan de hand van een vorm als *de zijnen*: die heeft 'enkelvoud' als waarde voor GETAL en 'meervoud-n' als waarde voor GETAL-N.

19. NPAGR. Bij prenominaal gebruik geldt dat sommige determiners een enkelvoudig onzijdig substantief vereisen (*dit, dat, welk, elk, ieder*), andere een enkelvoudig zijdig substantief (*elke, iedere*) en nog andere een enkelvoudig zijdig of een meervoudig substantief (*deze, die, welke*). NPAGR heeft dus net als GETAL en GETAL-N met getalsonderscheidingen te maken. Dit hoeft niet tot verwarring te leiden, aangezien NPAGR en GETAL-N nooit samen voorkomen; de enige voornaamwoorden die zowel een GETAL als een NPAGR feature hebben zijn de prenominaal gebruikte bezittelijke voornaamwoorden. Het nut hiervan kan worden aangetoond aan de hand van een combinatie als *ons huis*, waarin *ons* de waarde 'meervoud' heeft voor GETAL en de waarde 'enkelvoud onzijdig' voor NPAGR.

Een complicerende factor is de aanwezigheid van sporen van een drie-genera systeem. Zulke sporen treft men vooral aan in vaste verbindingen en archaïsch taalgebruik. De datiefvorm in *met dien verstande* bijv. vereist een masculien of onzijdig enkelvoud, terwijl het voornaamwoord in *in dier voege* een feminien enkelvoudig of een meervoudig substantief vereist. In het algemeen geldt dat de genitief- en datiefvormen het oude 3 genera systeem volgen, terwijl de standaardvormen het hedendaagse 2 genera systeem volgen.

05. GRAAD. Dit feature wordt alleen toegekend aan de gradeerbare determiners. Naast de basisvormen *veel, min, weinig*, zijn er de comparatieven *meer, minder*, de superlatieven *meest(e), minst(e)* en de diminutief *minnetjes*.

2.5.4 Implicaties

[I10] <VWTYPE = persoonlijk> $\Longrightarrow$ <PDTYPE = pronomen>
[I11] <VWTYPE = reflexief> $\Longrightarrow$ <PDTYPE = pronomen, NAAMVAL = oblique>
[I12] <VWTYPE = reciprook> $\Longrightarrow$ <PDTYPE = pronomen,
NAAMVAL $\neq$ nominatief, STATUS = vol, GETAL = meervoud>
[I13] <VWTYPE = bezittelijk> $\Longrightarrow$
<PDTYPE = determiner, POSITIE $\neq$ vrij>
[I14] <VWTYPE = vragend, PDTYPE = pronomen> $\Longrightarrow$ <STATUS $\neq$ gereduceerd>
[I15] <VWTYPE = betrekkelijk, PDTYPE = pronomen> $\Longrightarrow$ <STATUS = vol>
[I16] <VWTYPE = exclamatief, PDTYPE = pronomen> $\Longrightarrow$ <STATUS = vol>
[I06] <BUIGING = met-e> $\Longrightarrow$ <POSITIE = (pre)nominaal>
[I17] <PDTYPE = determiner, POSITIE = prenominaal, NAAMVAL = standaard>
$\Longrightarrow$ <NPAGR = agr>
[I18] <PDTYPE = determiner, POSITIE = prenominaal, NAAMVAL = bijzonder>
$\Longrightarrow$ <NPAGR = agr3>

De eerste zeven implicaties hebben betrekking op specifieke klassen van voornaamwoorden. De achtste is ook van toepassing op adjectieven en deelwoorden. De laatste twee gelden voor prenominale determiners, en in feite ook voor lidwoorden, maar omdat die tot een andere woordsoort gerekend worden, hebben ze hun eigen implicaties, zie 2.6.

2.5.5 De tags

Omdat het aantal combinaties bijzonder groot is, beperken we ons hier tot een overzicht waarin alleen de waarden voor VWTYPE, PDTYPE, NAAMVAL en POSITIE zijn vastgelegd. De andere features geven we een generische waarde, behalve wanneer de implicaties de toekenning van een specifiekere waarde mogelijk maken. Op die manier komen we tot een classificatie van types, die elk één of meer specifieke tags subsumeren. Bij elk type wordt aangegeven hoeveel tags erdoor gesubsumeerd worden. Een volledig overzicht van de 188 individuele tags is te vinden in de annex.

Aangezien het in deze sectie om types van tags gaat, heeft het geen zin om een onderscheid te maken tussen T- en U-tags. Aan het eind van de sectie wordt wel per feature aangegeven wanneer het gebruik van ondergespecificeerde waarden geoorloofd is.

De persoonlijke, reflexieve en reciproke voornaamwoorden. Deze voornaamwoorden zijn steeds pronomina. Hun tags bevatten dus naast VWTYPE, PDTYPE en NAAMVAL ook de features STATUS, PERSOON en GETAL; voor de standaardvormen van de persoonlijke voornaamwoorden van de derde persoon enkelvoud komt daar bovendien nog GENUS bij. Tot de persoonlijke voornaamwoorden van de derde persoon rekenen we ook 'men' en 'het'. Het kwantitatieve 'er' daarentegen rekenen we tot de onbepaalde voornaamwoorden en de andere gebruiken van 'er' tot de aanwijzende. We onderscheiden 8 types die in totaal 54 tags subsumeren.

[501-22] VNW(pers,pron,nomin,status,persoon,getal(,genus))	ik, we, wijzelf, ikke, jij, gij, ge, men, hij, ie
[502-9] VNW(pers,pron,obl,status,persoon,getal(,genus))	jou, hen, hem, 'm, haar
[503-4] VNW(pers,pron,stan,status,persoon,getal(,genus))	het, ze, jullie
[504-6] VNW(pers,pron,gen,vol,persoon,getal)	wie uwer, mijns gelijke
[505-9] VNW(pr,pron,obl,status,persoon,getal)	me, mij, ons, mezelf
[506-2] VNW(refl,pron,obl,status,3,getal)	zich, zichzelf
[507-1] VNW(recip,pron,obl,vol,persoon,mv)	elkaar, mekaar, elkander

[508-1]	VNW(recip,pron,gen,vol,persoon,mv)	elkaars, mekaars, elkanders

De bezittelijke voornaamwoorden. De bezittelijke voornaamwoorden zijn determiners, en hebben dus naast VWTYPE, PDTYPE en NAAMVAL ook de features POSITIE en BUIGING. De prenominale hebben bovendien ook NPAGR en de nominale GETAL-N. Typisch voor de bezittelijke determiners is dat ze daarnaast ook de features hebben die kenmerkend zijn voor de pronomina, d.w.z. STATUS, PERSOON en GETAL. We onderscheiden 5 types (drie prenominale en twee nominale) die in totaal 63 tags subsumeren; postnominaal gebruik, als in *kindeke mijn*, en vrij gebruik, als in *dat is mijn*, zijn zo uitzonderlijk dat er geen tags voor voorzien worden.

[509-17]	VNW(bez,det,stan,status,persoon,getal,prenom,buiging,agr)	mijn paard, mijne heren, m'n kapsel, je hoed
[510-13]	VNW(bez,det,gen,vol,persoon,getal,prenom,buiging,agr)	mijns inziens, een mijner vrienden
[511-13]	VNW(bez,det,dat,vol,persoon,getal,prenom,met-e,agr)	te mijnen huize, te mijner ere
[512-14]	VNW(bez,det,stan,vol,persoon,getal,nom,met-e,getal-n)	de mijne, de zijnen
[513-6]	VNW(bez,det,dat,vol,persoon,getal,nom,met-e,getal-n)	ten onzent

De vragende, betrekkelijke en exclamatieve voornaamwoorden. De meeste van deze voornaamwoorden zijn pronomina. De VWTYPE waarde is in vele gevallen het intermediaire 'vb': er is immers slechts één vragend voornaamwoord dat niet tevens als betrekkelijk voornaamwoord wordt gebruikt (*watte*). De waarde voor NAAMVAL is het intermediaire 'standaard', behalve bij het inherent oblique adverbiaal-pronomen (*waar*). Omdat deze voornaamwoorden getopicaliseerd worden en gereduceerde vormen niet getopicaliseerd kunnen worden, is de STATUS waarde 'vol' of 'nadruk'. De waarde voor PERSOON is '3' of een subtype daarvan, behalve voor het betrekkelijke *die*, cf. *ik die hier zo lang gewerkt heb, ...*

Gemeenschappelijk aan deze voornaamwoorden is dat ze getopicaliseerd worden. De vragende vnw. komen zowel in hoofd- als bijzinnen voor, de betrekkelijke alleen in bijzinnen en de exclamatieve alleen in hoofdzinnen. De STATUS waarde van de vragende pronomina is 'vol' of 'nadruk', maar niet

'gereduceerd'; die van de betrekkelijke en exclamatieve pronomina is steeds 'vol'.

[514-1]	VNW(vrag,pron,stan,nadr,3o,ev)	watte
[515-2]	VNW(betr,pron,stan,vol,persoon,getal)	de man die daar staat, het kind dat je daar ziet
[516-2]	VNW(betr,pron,gen,vol,3o,getal)	het warenhuis welks directeur hem een baan had aangeboden
[517-2]	VNW(vb,pron,stan,vol,3,getal)	wie gaat er mee, wat ik niet begrijp is ...
[518-3]	VNW(vb,pron,gen,vol,3p,getal)	wiens hoed is dit, de vrouw wier hoed daar hangt
[519-1]	VNW(vb,adv-pron,obl,vol,3o,getal)	waar ga je naartoe, de trein waar we op staan te wachten
[520-1]	VNW(excl,pron,stan,vol,3,getal)	wat een dwaasheid, wat kan jij liegen zeg

De enige determiner in deze klasse is *welk(e)*. Hij wordt meestal als vragend voornaamwoord gebruikt, maar sporadisch ook als betrekkelijk of exclamatief voornaamwoord.

[521-2]	VNW(vb,det,stan,prenom,buiging,agr)	welke stoel, welk kind
[522-1]	VNW(vb,det,stan,nom,met-e,getal-n)	welke vind jij de mooiste, de procedures welke bij zo'n gelegenheid gevolgd worden
[515-2]	VNW(betr,det,stan,nom,buiging,getal-n)	hetgeen ik wil zeggen
[523-1]	VNW(excl,det,stan,vrij,zonder)	welk een dwaasheid

We onderscheiden dus 11 types die tezamen 18 tags subsumeren.

De aanwijzende voornaamwoorden. Deze klasse omvat zowel pronomina als determiners. We onderscheiden 11 types die in totaal 19 tags subsumeren. De determiners worden meestal prenominaal of nominaal gebruikt, maar er zijn ook gevallen van vrij gebruik. De pronomina, waartoe

we ook de adverbiale pronomina *hier, daar, d'r* en het niet-kwantitatieve *er* rekenen, zijn steeds van de derde persoon. Over het onderscheid tussen pronomina en nominale determiners, zie sectie B, onder PDTYPE.

[524-3]	VNW(aanw,pron,stan,status,3,getal)	dat(te), dit(te), die
[525-2]	VNW(aanw,pron,gen,vol,3,ev)	diens voorkeur, en dies meer
[526-1]	VNW(aanw,adv-pron,obl,status,3o,getal)	hier, daar, d'r
[527-1]	VNW(aanw,adv-pron,stan,red,3,getal)	het niet-kwantitatieve 'er'
[528-4]	VNW(aanw,det,stan,prenom,buiging,npagr)	dat boek
[529-1]	VNW(aanw,det,gen,prenom,met-e,rest3)	een dezer dagen, de notulen dier vergadering
[530-2]	VNW(aanw,det,dat,prenom,met-e,agr3)	te dien tijde, in dier voege
[531-2]	VNW(aanw,det,stan,nom,met-e,getal-n)	deze(n), gene(n), degene
[532-1]	VNW(aanw,det,gen,nom,met-e,getal-n)	schrijver dezes, de twintigste dezer
[533-1]	VNW(aanw,det,dat,nom,met-e,getal-n)	dat is dan bij dezen beslist
[534-1]	VNW(aanw,det,stan,vrij,zonder)	zulk een vreemde gedachte

Net als de ANS-97 rekenen we *soortgelijk(e), dergelijk(e), zodanig(e)* en *dusdanig(e)* niet tot de aanwijzende voornaamwoorden, maar tot de adjectieven. Hetzelfde geldt voor de samenstellingen met *-zelfde(n).*

De onbepaalde voornaamwoorden. Ook deze klasse omvat zowel pronomina als determiners. Er zijn 14 types die tezamen 34 tags subsumeren. De pronomina, die ook de adverbiaal-pronomina *(n)ergens, overal* en het kwantitatieve *er* omvatten, zijn alle van de derde persoon. De onbepaalde voornaamwoorden vertonen dezelfde variatie als de aanwijzende. Er zijn enerzijds de prenominale determiners, de nominale en de vrije, en anderzijds de pronomina, m.i.v. de adverbiale pronomina *overal* en *(n)ergens.* Tot die laatste rekenen we ook het kwantitatieve *er.* De verschillende vormen van *veel, weinig, beide* rekenen we niet tot de telwoorden, maar tot de gradeerbare determiners. Bemerk echter dat *zoveel, evenveel, hoeveel* wel telwoorden zijn aangezien ze corresponderende rangtelwoorden hebben.

[535-2]	VNW(onbep,pron,stan,vol,3,ev)	(n)iets, (n)iemand, iedereen, alles, wat snoep
[536-1]	VNW(onbep,pron,gen,vol,3p,ev)	andermans, (n)iemands, ieders
[537-1]	VNW(onbep,adv-pron,obl,vol,3o,getal)	(n)ergens, overal
[538-1]	VNW(onbep,adv-pron,gen,red,3,getal)	het kwantitatieve 'er'

De onbepaalde determiners omvatten ook de gradeerbare (PDTYPE = graad), die naast de typische determiner features ook een GRAAD feature hebben.

[539-6]	VNW(onbep,det,stan,prenom,buiging,agr)	beide mannen, elk kind, iedere keer
[540-1]	VNW(onbep,det,gen,prenom,met-e,mv)	proletariërs aller landen
[541-2]	VNW(onbep,det,dat,prenom,met-e,agr3)	te allen prijze, te eniger tijd
[542-6]	VNW(onbep,grad,stan,prenom,buiging,agr,graad)	veel plezier, vele uren, minder werk, de meeste mensen
[543-3]	VNW(onbep,det,stan,nom,met-e,getal-n)	allen zijn tevreden, sommigen zijn gevlucht
[544-1]	VNW(onbep,det,gen,nom,met-e,getal-n)	met aller instemming, tot beider verbazing
[545-5]	VNW(onbep,grad,stan,nom,met-e,getal-n,graad)	velen zijn geroepen, weinigen zijn uitverkoren
[546-1]	VNW(onbep,grad,gen,nom,met-e,getal-n,graad)	tot veler verbazing
[547-1]	VNW(onbep,det,stan,vrij,zonder)	ze kregen elk/ieder/allebei een bal, al die mensen

[548-3]	VNW(onbep,grad,stan,vrij,zonder,graad)	minder werken, meer slapen, dat is te weinig, veel groter

Genoeg, zat, allerhande rekenen we niet tot de voornaamwoorden, maar tot de bijwoorden (vnw. worden immers niet postnominaal gebruikt) en *ietsje, (een) weinigje* rekenen we tot de substantieven (de diminutiefvormen van vnw. eindigen immers op *-jes*, niet op *-je*).

Ondergespecificeerde combinaties. Net als bij de substantieven laten we een beperkte mate van onderspecificatie toe. De features die daarvoor in aanmerking komen zijn de volgende.

[P15] VWTYPE = pr (persoonlijk, reflexief), reciprook, bezittelijk, vb (vragend, betrekkelijk), exclamatief, aanwijzend, onbepaald.
[P07] NAAMVAL = standaard (nominatief, oblique), genitief, datief.
[P18] PERSOON = persoon (1, 2 (2v, 2b), 3 (3p (3m, 3v), 3o)).
[P04] GETAL = getal (enkelvoud, meervoud).
[P19] NPAGR = agr (evon, rest (evz, mv)), agr3 (evmo, rest3 (evf, mv)).

15. VWTYPE. De intermediaire waarden 'pr' en 'vb' worden toegekend wanneer aan de vorm van het voornaamwoord niet kan worden afgelezen of het persoonlijk of reflexief dan wel vragend of betrekkelijk is. De vormen die daarvoor in aanmerking komen zijn die welke in het functiewoordenlexicon een intermediaire waarde hebben gekregen.

07. NAAMVAL. De intermediaire waarde 'standaard' wordt gebruikt wanneer aan de vorm van het voornaamwoord niet kan worden afgelezen of het nominatief of oblique is. Dat geldt voor alle determiners en voor die pronomina waarvoor het nominatief/oblique onderscheid geneutraliseerd wordt. Voor een lijst, zie het lexicon.

18. PERSOON. De intermediaire waarden '2', '3' en '3p' worden toegekend aan de voornaamwoorden waarvoor het lexicon geen specifiekere waarden vermeldt.

04. GETAL. De generische waarde 'getal' wordt toegekend aan de voornaamwoorden waarvoor in het lexicon geen specifieke waarde is gegeven. Het gaat o.m. om de adverbiaal-pronomina, het reflexieve *zich(zelf)* en de *u*-vormen.

19. NPAGR. De intermediaire waarden 'agr', 'agr3', 'rest' en 'rest3' worden toegekend aan de prenominale determiners die in het lexicon geen specifiekere waarde hebben gekregen.

2.5.6 Lemmatisering

Het lemma wordt geïdentificeerd met de onverbogen standaardvorm. Inflectionele affixen worden afgestript, behalve wanneer de grondvorm niet als zelfstandig woord gebruikt wordt, zoals bij de genitiefvorm *andermans*. Bij suppletie, zoals nominatief *ik* vs. oblique *mij*, krijgen beide vormen een verschillend lemma. De gereduceerde vormen zonder klinker, zoals *'t* en *z'n*, krijgen de corresponderende volle vorm als lemma; de gereduceerde vormen met klinker daarentegen krijgen hun eigen lemma: de lemmawaarde van *me* is dus niet 'mij', maar 'me'. Ingeval van twijfel, raadpleeg het lexicon.

2.6 LIDWOORDEN

De lidwoorden zijn eigenlijk determiners, en kunnen in termen van dezelfde features beschreven worden. Omdat het in Nederlandse grammatica's echter gebruikelijk is om de lidwoorden als een aparte woordsoort te behandelen (zie o.m. de ANS-97), sluit het CGN zich bij die traditie aan.

2.6.1 Afgrenzing

Het bepaalde lidwoord *het* moet worden onderscheiden van het homonieme persoonlijke voornaamwoord, en het onbepaalde *een* moet worden onderscheiden van het telwoord en de onbepaalde determiner; dat laatste is eenvoudig omdat er een duidelijk verschil in uitspraak is (sjwa vs. heldere ee), waarvan we mogen aannemen dat het bij transcriptie wordt aangegeven door de aan-of afwezigheid van accenttekens. De genitiefvorm *des* moet worden onderscheiden van het homonieme bijwoord in constructies van het type *des te beter*.

2.6.2 Declaraties

[D21] <POS = lidwoord> $\Longrightarrow$ <LWTYPE, NAAMVAL, NPAGR>

Er is geen feature voor BUIGING, omdat er—in de standaardtaal—geen onderscheid wordt gemaakt tussen verbogen en onverbogen lidwoorden. Voor de behandeling van de dialectische vorm in *ne vent*, zie 2.11.

2.6.3 Partities

[P20] LWTYPE = bepaald, onbepaald.
[P07] NAAMVAL = standaard, genitief, datief.
[P19] NPAGR = agr (evon, rest), agr3 (evmo, rest3 (evf, mv)).

20. LWTYPE. Het exclamatieve *een* wordt net als in de ANS-97 als 'onbepaald' beschouwd.

07. NAAMVAL. Het bepaalde lidwoord *de* heeft naast de standaardvorm ook de genitiefvormen *des, der* en *'s* (*des duivels, der Nederlandse taal, 's avonds*), evenals de datiefvormen *der* en *den*; die laatste komen alleen in vaste verbindingen voor, zoals *in der minne, op den duur, uit den boze*.

19. NPAGR. *Het* vereist een enkelvoudig onzijdig substantief ('evon') en *de* een enkelvoudig zijdig of meervoudig substantief ('rest'). Het onbepaalde lidwoord legt geen beperkingen op, cf. *een paard, een man, een mensen dat er waren!*

Een complicerende factor is de aanwezigheid van sporen van een drie-genera systeem ('agr3'). Zulke sporen treft men vooral aan in vaste verbindingen en bij archaïsch taalgebruik. De genitiefvorm *des* bijv. vereist een masculien of onzijdig enkelvoud ('evmo'), terwijl *der* een feminien enkelvoudig of een meervoudig substantief vereist ('rest3').

2.6.4 Implicaties

[I19] <POS = lidwoord, NAAMVAL = standaard> $\Longrightarrow$ <NPAGR = agr>
[I20] <POS = lidwoord, NAAMVAL = bijzonder> $\Longrightarrow$ <NPAGR = agr3>

De genitief en de datief volgen het oude drie-genera systeem, terwijl de standaardvormen het hedendaagse twee-genera systeem volgen.

2.6.5 De tags

Voor de lidwoorden geldt net als voor de voornaamwoorden dat een volledige lijst van de combinaties nauwelijks verschilt van wat in het lexicon wordt geboden. Omdat het aantal lidwoorden echter zo klein is, kunnen we hier wel de volledige lijst geven. Er zijn in totaal negen combinaties, waaronder twee ondergespecificeerde.

[T601]	LID(bep,stan,evon)	het kind, in 't geniep
[T602]	LID(bep,stan,rest)	de hond(en), de kinderen
[T603]	LID(bep,gen,evmo)	des duivels, de plaats des onheils, 's avonds, 's maandags
[U604]	LID(bep,gen,*rest3*)	der Nederlandse taal, der Belgen
[T605]	LID(bep,dat,evmo)	op den duur, om den brode, uit den boze

[T606]	LID(bep,dat,evf)	in der minne
[T607]	LID(bep,dat,mv)	die in den hemelen zijt
[U608]	LID(onbep,stan,*agr*)	een kind, een mensen dat er waren
[T609]	LID(onbep,gen,evf)	de kracht ener vrouw

Er zijn ons geen voorbeelden bekend van masculiene of onzijdige genitiefvormen voor het onbepaalde lidwoord. De ANS-97 vermeldt weliswaar *eens geestes zijn*, maar dat is een genitief van het telwoord *één*.

2.6.6 Lemmatisering

De lemmata voor de lidwoorden zijn 'de', 'het' en 'een'. Het onbepaalde ''n' ontbreekt, omdat het protocol voor ortografische transcriptie de spelling 'een' voorschrijft; de vorm wordt alleen toegestaan in de verbinding *zo'n*, maar die wordt in z'n geheel als aanwijzend voornaamwoord behandeld.

2.7 VOORZETSELS

2.7.1 Afgrenzing

Voorzetsels nemen meestal een complement. Dat kan niet alleen een NP zijn, maar ook een PP, een bijwoord, een adjectief, een telwoord of een verbale projectie (V, VP, S). In dat laatste geval worden ze vaak tot de voegwoorden gerekend, waardoor een systematische ambiguïteit ontstaat voor woorden als *tot, sedert, sinds, voor, na, naar, zonder, met, door, om*. CGN volgt dit gebruik niet en behandelt zulke woorden steeds als voorzetsels. Ook de *te* die een infinitief inleidt, rekenen we tot de voorzetsels, evenals de *aan* in *aan het vissen zijn*, de *op* in *op springen staan* en de *uit* in *uit vissen gaan*.

Als het complement een NP of een PP is, kan het aan het voorzetsel voorafgaan, als in *loopt overal tegen, rijdt de berg op, de hele dag door, onder de brug door*. In zulke gevallen kan het complement ook geëxtraheerd worden, zodat het voorzetsel alleen achter blijft, als in *waar denk je aan*, *rijdt die berg alleen op* en *door die muur kunnen we niet heen*. We rekenen alleen die woorden tot de postpositionele voorzetsels die voorafgegaan kunnen worden door een adverbiaal pronomen; *heen* en *af* worden dus tot de voorzetsels gerekend, maar *terug, weg* en *geleden* niet (**erterug, *hierweg, *waargeleden*); die laatste rekenen we tot de bijwoorden.

Voorzetsels kunnen ook zonder complement gebruikt worden, net zoals vele transitieve werkwoorden ook intransitief gebruikt worden. Dat is bijv. het geval voor predikatief gebruikte voorzetsels (*het bier is op, het licht is aan, hij is vroeg op*) en voor bijwoordelijk gebruikte voorzetsels, zoals *boven* in *naar boven gaan*. Tot de intransitief gebruikte voorzetsels rekenen we ook de scheidbare delen van werkwoorden die dezelfde vorm hebben als een voorzetsel (*belt ... op, geeft ... uit*). Dat

vermijdt de creatie van een systematische ambiguïteit tussen het voorzetsel *op* en het homonieme partikel of bijwoord. Noot: het niet-verbale deel van een scheidbaar samengesteld werkwoord kan ook een substantief zijn (*haal diep adem*), een adjectief (*doe de glazen nog eens vol*) of een bijwoord (*we komen morgen samen*).

Bij de voorzetsels die aan vreemde talen ontleend zijn, maken we een onderscheid tussen die welke ook met een Nederlands complement gecombineerd worden, zoals *per trein* en *drie à vier glazen*, en die welke uitsluitend met een complement uit diezelfde vreemde taal gecombineerd worden, zoals *ad hoc* en *en profil*. De eerste worden gewoon tot de voorzetsels gerekend; de laatste krijgen de tag SPEC(vreemd). Het gaat hierbij vooral om ontleningen uit het Latijn (*ab ovo, ad fundum, cum laude, ex machina, ex voto, intra muros, inter alia, post mortem, pro deo, pro domo, salva veritate*) en uit het Italiaans (*con amore, con brio, sotto voce*).

2.7.2 Declaraties

[D22] <POS = voorzetsel> $\Longrightarrow$ <VZTYPE>

2.7.3 Partities

[P21] VZTYPE = initieel (versmolten), finaal.

21. VZTYPE. Voorzetsels die voorafgaan aan hun complement krijgen de waarde 'initieel'. Dat complement kan zowel een NP zijn als een PP, een bijwoord, een adjectief, een telwoord of een verbale projectie (V, VP of S). Voorzetsels die op hun complement volgen krijgen de waarde 'finaal'; het voorafgaande complement is vaak een adverbiaal pronomen (*hij zit er net naast*), maar het kan ook een NP zijn (*de trap af*) of een PP (*van het dak af, door de eeuwen heen*). Een belangrijk verschil tussen initieel en finaal gebruik is dat in het eerste geval het complement onmiddellijk op het voorzetsel moet volgen (modulo parenthese), terwijl in het tweede geval het complement van het voorzetsel gescheiden kan zijn door andere zinsdelen: dit komt vooral voor bij adverbiale pronomina ([*daar*] *denk ik niet eens* [*aan*]), maar ook bij PPs ([*door die muur*] *kom je met dat boortje niet* [*heen*]); voor meer voorbeelden van die laatste, zie de ANS-97, 509-510.

Voorzetsels zonder complement of intransitieve voorzetsels worden o.m. gebruikt als het niet-verbale lid van een scheidbaar samengesteld werkwoord (*belt zijn dochter op*), als het complement van een ander voorzetsel (*naar boven/binnen gaan*), als kern van een predikaat (*op zijn, binnen zijn, in zijn*) of als kern van een VP-bepaler (*binnen spelen, niet zonder kunnen*). Omdat de groep van voorzetsels die intransitief gebruikt kunnen worden een subset is van de voorzetsels die finaal gebruikt kunnen worden, krijgen ze bij tagging de waarde 'finaal'.

In termen van het VZTYPE onderscheid kunnen we drie klassen van voorzetsels onderscheiden: die welke uitsluitend initieel gebruikt worden, zoals *per, sinds, se-*

dert, te; die welke uitsluitend finaal gebruikt worden, zoals *af, heen, vandaan*; en die welke in beide soorten van posities voorkomen: *aan* bijvoorbeeld is initieel in *aan de wand, aan het vissen zijn, aan het zeuren gaan* en finaal in *achter de stoet aan, tegen de veertig aan, kondigt een vertraging aan, heeft een zwarte rok aan.* Enkele voorzetsels hebben verschillende vormen voor het initiële en het finale gebruik, zoals *met/me(d)e* en *tot/toe.*

Binnen de inherent initiële voorzetsels onderscheiden we een aparte klasse van versmolten voorzetsels, d.w.z. voorzetsels die één geheel vormen met een bepaald lidwoord. In het Duits, het Frans en het Italiaans zijn er verschillende van zulke voorzetsels; het Nederlands heeft alleen de vormen *ter* en *ten.*

Met uitzondering van de twee versmolten voorzetsels zijn de Nederlandse voorzetsels morfologisch invariabel. Vormen als *voorste, achterste, onderste, bovenste, benedenste, binnenste* en *buitenste* zijn geen superlatiefvormen van voorzetsels maar adjectieven. Iets soortgelijks geldt voor de vormen *onderen, voren, achteren*; dat zijn geen verbogen voorzetsels maar bijwoorden.

2.7.4 Implicaties

Implicaties zijn er niet te melden.

2.7.5 De tags

Er zijn drie mogelijke combinaties.

```
[T701] VZ(init)   met een lepeltje, met Jan in het hospitaal,
                  met zo te roepen
[T702] VZ(fin)    liep de trap af, bij de beesten af, speelt
                  het bandje af, kletsen flink wat af
[T703] VZ(versm)  ten strijde, ten hoogste, ter plaatse
```

Er is geen ruimte voor onderspecificatie. Bij de tagging dient steeds een specifieke waarde te worden toegekend voor VZTYPE.

2.7.6 Lemmatisering

Het lemma is voor de voorzetsels identiek aan de woordvorm, behalve in het geval van de twee versmolten voorzetsels: *ter, ten* krijgen 'te' als lemma.

2.8 VOEGWOORDEN

2.8.1 Afgrenzing

Tot de voegwoorden behoren twee klassen van woorden met sterk verschillende eigenschappen: de nevenschikkende en de onderschikkende.

Nevenschikkende voegwoorden kunnen zowel zinnen inleiden als kleinere woordgroepen en zelfs delen van woorden. Wanneer ze een zin inleiden dan kan die eender welke volgorde vertonen: V-1, V-2 of V-finaal. De nevenschikkende voegwoorden vormen een kleine gesloten klasse: tot de leden ervan rekenen we o.m. *en, of, ofwel, noch, maar, want, hetzij.* Zie het lexicon voor een volledige lijst.

De onderschikkende voegwoorden kunnen in drie groepen verdeeld worden: (1) de complementeerders *dat, of, als* en *dan*; (2) de combinaties van een voorzetsel en een complementeerder (*alsof, doordat, nadat, omdat, opdat, totdat, voordat*) of van een bijwoord en een complementeerder (*eerdat, zodat*); (3) een restgroep (*(ter)wijl, (al)hoewel, (voor)aleer, alvorens, tenzij, zodra, ...*); zie het lexicon voor een langere lijst.

Een typische eigenschap van de complementeerders is dat ze voorafgegaan kunnen worden door een vooropgeplaatst zinsdeel, als in *leuk dat het was*, *wie of er gebeld heeft* en *rijk als ze was.* Bemerk dat de complementeerders alle vier ambigu zijn: *dat* is ook een aanwijzend of betrekkelijk voornaamwoord, *of* is ook een nevenschikkend voegwoord, *als* een voorzetsel en *dan* een bijwoord. Wat het voegwoord *dat* onderscheidt van het betrekkelijke voornaamwoord is dat het in de bijzin geen argumentsfunctie heeft: het is noch een onderwerp noch een voorwerp, maar gewoon een inleider van de bijzin. Vgl. bijv. het voornaamwoord in *het feit dat vaak over het hoofd wordt gezien* met het voegwoord in *het feit dat dit vaak over het hoofd gezien wordt.* Aangezien het voornaamwoord met zijn antecedent moet overeenstemmen in genus en getal, treedt het alleen op in combinatie met enkelvoudige onzijdige NPs, terwijl het voegwoord niet aan die restrictie onderhevig is, cf. *de geruchten dat de volgende Paus een Italiaan moet zijn* en *de verwachting dat ze wel zal komen.*

Onderschikkende voegwoorden leiden meestal een bijzin in. In dat geval moet die zin de typische bijzinsvolgorde vertonen (V-finaal). Dit volgorde-criterium is bruikbaar om de voegwoorden te onderscheiden van de bijwoorden. Vergelijk bijv. het voegwoord in *wanneer we naar Milaan gaan* met het bijwoord in *wanneer gaan we naar Milaan?*. Op dezelfde manier kunnen we een onderscheid maken tussen de voegwoorden *toen, nu* en *dan* en de homonieme bijwoorden; *dan* bijv. is een bijwoord in *dan gaan we naar Milaan* en in *als dit een droom is, dan word ik liefst niet wakker*, aangezien de deelzinnen die erdoor ingeleid worden niet de typische bijzinsvolgorde vertonen. Om diezelfde reden is de concessieve *al* in *al is de leugen nog zo snel* geen voegwoord meer een bijwoord. Het volgorde-criterium is ook bruikbaar om het redengevende voegwoord *daar* te onderscheiden van het homonieme adverbiale voornaamwoord.

Niet alle woorden die een verbale projectie met bijzinsvolgorde inleiden, rekenen we tot de voegwoorden. Als het woord qua vorm identiek is aan een voorzetsel, zoals *voor, na, naar, om, tot, sedert, sinds, zonder*, dan beschouwen we het niet als ambigu (voorzetsel of voegwoord) maar steeds als voorzetsel, zie 2.7. Een zeldzaam geval van ambiguïteit tussen voegwoord en voorzetsel is *als*: dat is een voegwoord wanneer het een conditionele bijzin of het tweede lid van een vergelijking inleidt, maar niet wanneer het een irrealis inleidt, zoals in *als was het een droom.*

Onderschikkende voegwoorden kunnen—net als de nevenschikkende—ook kleinere woordgroepen inleiden; dat geldt o.m. voor het gebruik van *als* en *dan* als inleiders van het tweede lid van een vergelijking, cf. *rijker dan Bill* en *zo groot als jij*; andere voorbeelden zijn *behalve* en *hoewel* in een combinatie als *hoewel vlijtig en verstandig vond ie geen baan.*

2.8.2 Declaraties

[D23] <POS = voegwoord> $\Longrightarrow$ <CONJTYPE>

2.8.3 Partities

[P22] CONJTYPE = nevenschikkend, onderschikkend.

22. CONJTYPE. Het onderscheid tussen nevenschikkende en onderschikkende voegwoorden is geen semantisch maar een puur syntactisch onderscheid. Het causale *omdat* bijv. behoort tot de onderschikkende voegwoorden, omdat het bijzinsvolgode vereist, terwijl het quasi-synonieme *want* tot de nevenschikkers behoort, aangezien het de typische hoofdzinsvolgorde vereist (V-2). Het enige voegwoord dat zowel nevenschikkend als onderschikkend kan zijn is *of.*

2.8.4 Implicaties

Implicaties zijn er niet te melden.

2.8.5 De tags

Het aantal combinaties is beperkt tot twee.

[T801] VG(neven) Jan en Peter; en toen gebeurde het
[T802] VG(onder) ze komt niet, omdat ze zich niet goed voelt

2.8.6 Lemmatisering

Het lemma is identiek aan de woordvorm.

2.9 BIJWOORDEN

Nog meer dan de voornaamwoorden vormen de bijwoorden een heterogene klasse. Om te bepalen welke woorden ertoe behoren gebruiken we vormelijke criteria en geen functionele. Zo is een woord dat de kern vormt van een bijwoordelijke bepaling, niet noodzakelijk een bijwoord. De tijdsbepaling in *we gaan zondag naar Milaan* bijvoorbeeld is geen bijwoord maar een substantief, en de bepaling van wijze in *hij praat snel* behandelen we evenmin als een bijwoord, maar als een adverbiaal gebruikt adjectief (<POS = adjectief, POSITIE = vrij>, zie 2.2). Om adverbiaal gebruikte adjectieven te onderscheiden van bijwoorden gebruiken we het volgende criterium: wanneer de adverbiale vorm ook prenominaal gebruikt kan worden in eenzelfde of gelijkaardige betekenis, dan gaat het om een adjectief; anders gaat het om een bijwoord. Zo is de vorm *vrij* in *je kan hier vrij rondlopen* een adverbiaal gebruikt adjectief, terwijl diezelfde vorm in *een vrij warme dag* een bijwoord is.

Aangezien de CGN tagset niet alleen voor de adjectieven voorziet dat ze een adverbiaal gebruik kunnen hebben, maar ook voor de deelwoorden, de telwoorden en de determiners, wordt de heterogeniteit van de bijwoordenklasse enigszins in de perken gehouden, maar zelfs dan blijven hun aantal en hun diversiteit groot.

Om die diversiteit in kaart te brengen zijn we inductief tewerk gegaan: vertrekkend van een lijst van woorden die in CELEX als bijwoord geklasseerd zijn (ong. 850 woorden) zijn we tot een morfologisch gebaseerde classificatie gekomen. Naast de vrij grote groep van (1) ongelede bijwoorden (*nu, niet, nog, al, hoe, ...*) onderscheiden we diverse types van—naar de vorm—gelede bijwoorden: (2) die met een adverbiaal suffix (*stomweg, beroepshalve, derwaarts*), (3) die met een adverbiale kern (*welnu, hierzo*), (4) die met een prepositionele kern (*tussenin, bovenaan*), (5) die met een geïncorporeerd voornaamwoordelijk complement (*daarin, erop, waarover, desondanks, bovendien*), (6) die met een nominale kern (*uitermate, binnenshuis, bergaf*), (7) die met een adjectivale kern (*stilaan, voluit*), en (8) die met een verbale kern (*ongetwijfeld, welteverstaan*). Tenslotte is er de aparte groep van de (9) leenwoorden (*incognito, sowieso, normaliter*). Voor een volledige overzicht en verdere subclassificaties, zie het functiewoordenlexicon.

Van de bijwoorden die in het lexicon opgenomen zijn, is er slechts een zeer klein deel dat morfologische variatie vertoont: er zijn de diminutiefvormen *strakjes, saampjes, eventjes, ...* en er zijn de comparatief- en superlatiefvormen van o.m. *graag*. Men zou daaruit kunnen besluiten om ook voor de bijwoorden het GRAAD feature te declareren, maar of dat de moeite waard is valt te betwijfelen, want naast het feit dat de onderscheiding slechts voor een zeer klein aantal van de bijwoorden relevantie heeft, is er het feit dat die niet-basisvormen gelexicaliseerd zijn. Zo is *liever* een compositioneel interpreteerbare comparatiefvorm van het adjectief *lief*, maar als comparatief van het bijwoord *graag* is het eigenlijk een vorm sui-generis. Hetzelfde

geldt voor de superlatiefvormen in *ze komt liefst alleen*, *hoogst verleidelijk*, *eerst doe je dit en dan dat* en *laatst zag ik een merkwaardig tafereel*. Om die reden kiezen we ervoor om de (schaarse) diminutief-, comparatief- en superlatiefvormen van de bijwoorden als aparte lemmata te behandelen.

Een andere potentiële uitzondering op de morfologische invariabiliteit van de bijwoorden zijn de buigingsvormen van de intensifiers *heel* en *erg* in combinaties als *een hele lange tafel* en *een erge leuke vakantie*. Het introduceren van een BUIGINGsfeature voor de bijwoorden om deze twee geïsoleerde en bovendien niet geheel grammaticale vormen te behandelen lijkt ons niet opportuun. In de plaats daarvan behandelen we beide vormen als prenominaal gebruikte adjectieven. Bemerk overigens dat ze—in enigszins andere betekenissen—ook alsdusdanig gebruikt worden, cf. *een hele dag* en *erge pijnen*.

Als gevolg van deze twee ingrepen kunnen we morfologische invariabiliteit inderdaad tot de typische kenmerken van de bijwoorden rekenen.

Aangezien de bijwoorden alleen een POS feature krijgen, zijn er geen implicaties te melden en is er slechts één combinatie.

```
[T901] BIJW()      gisteren, nu, niet, nog, al, hoe
```

Het lemma is identiek aan de woordvorm, behalve in het geval van getronceerde vormen als *'ns* en *d'rover*; die worden, zoals alle getronceerde vormen, herleid tot een vorm zonder weglatingsteken, i.c. *eens* en *erover*.

2.10 TUSSENWERPSELS

Tot de tussenwerpsels rekenen we die woorden die gewoonlijk als een zelfstandige taaluiting gebruikt worden. In navolging van de ANS-97 onderscheiden we drie types: (1) de klanknabootsingen, zoals *kukeleku*; (2) de expressies van emoties van de spreker, zoals uitdrukkingen van pijn, verwondering, frustratie e.d.; hiertoe rekenen we ook krachttermen en vloeken; (3) de formules voor sociaal verkeer, zoals begroetingen, bedankingen, verontschuldigingen e.d.

Niet elk woord dat als zelfstandige taaluiting gebruikt wordt, is een tussenwerpsel. Imperatieven als *bijt* en bijwoorden als *weg* bijvoorbeeld kunnen wel als zelfstandige taaluiting gebruikt worden, maar worden in het CGN corpus tot de werkwoorden, resp. bijwoorden gerekend. Dit volgt uit het algemene principe dat de toekenning van POS tags op vormelijke criteria gebaseerd is en niet op functionele. Om diezelfde reden rekenen we woorden als *munt, kruis* en *Christus* tot de substantieven, ook wanneer ze als uitroep gebruikt zijn. Voor een lijst van de woorden die we tot de tussenwerpsels rekenen, zie het CGN lexicon.

De tussenwerpsels krijgen net als de bijwoorden alleen een POS feature. Implicaties zijn er dus niet te melden, en er is slechts één combinatie.

[T001] TSW() oei, amai, uh, hoera, AUB

Het lemma is identiek aan de woordvorm.

2.11 DIALECTWOORDEN

Tot de dialectwoorden rekenen we alle woorden die bij de orthografische transcriptie voorzien zijn van de aanduiding '*d'. Omdat de onderscheidingen die voor de standaardtaal gelden niet altijd van toepasssing zijn op de dialectwoorden kennen we aan deze woorden alleen de POS en XTYPE features toe. Ter onderscheiding van de andere woorden, nemen we in hun tags ook de waarde 'dial(ectisch)' op.

Er zijn in totaal 28 combinaties. we geven ze een R-nummer (voor regionaal), omdat de D-nummers al voor de declaraties gebruikt worden.

[R101] N(soort,dial) bompa*d, ne*d lange *frak*d*
[R102] N(eigen,dial)

[R201] ADJ(dial) ne*d *langen*d* toot*d

[R301] WW(dial) 'k *zen*d* nie*d thuis, 'k *hem*d* gee*d geld

[R401] TW(hoofd,dial)
[R402] TW(rang,dial) den*d *elfste*d*

[R501] VNW(pers,pron,dial) kom *de*d* gij mee, 'k heb *ulie*d* gezien
[R502] VNW(refl,pron,dial)
[R503] VNW(recip,pron,dial) we zien *malkanderen*d* niet veel
[R504] VNW(bez,det,dial) hij heeft *z'ne*d* frak*d vergeten
[R505] VNW(vrag,pron,dial)
[R506] VNW(vrag,det,dial)
[R507] VNW(betr,pron,dial)
[R508] VNW(betr,det,dial)
[R509] VNW(excl,pron,dial)
[R510] VNW(excl,det,dial)
[R511] VNW(aanw,pron,dial)
[R512] VNW(aanw,det,dial) *diejen*d* boek, *dees*d* week
[R513] VNW(onbep,pron,dial) z' hebben *iet*d* gezien
[R514] VNW(onbep,det,dial) ze kan *elken*d* dag vertrekken

[R601] LID(bep,dial) het gevecht met *den*d* beer

[R602] LID(onbep,dial) *nen*d* toffe gast, *ne*d* vieze vent

[R701] VZ(init,dial) *me*d* veel geduld
[R702] VZ(fin,dial)

[R801] VG(neven,dial)
[R802] VG(onder,dial) 't schijnt *da*d* we mogen komen

[R901] BW(dial) *efkes*d*, *nie*d*

[R001] TSW(dial) *neeje*d*, *wablieft*d*

Het lemma is identiek aan de woordvorm.

2.12 SPECIALE TOKENS

Tot deze groep behoren de tokens die niet bij de normale woordsoorten ondergebracht kunnen worden. Ze hebben geen POS feature, maar wel een SPECTYPE feature.

[D24] <TOKENTYPE = speciaal> $\Longrightarrow$ <SPECTYPE>

[P23] SPECTYPE = afgebroken, onverstaanbaar, vreemd, deeleigen, meta, commentaar, achtergrond, afkorting, symbool.

23. SPECTYPE. Tot de afgebroken tokens rekenen we onvolledige woorden. Ze worden vaak door een streepje gevolgd of voorafgegaan, zoals het eerste token in *binnen- en buitenland* en het laatste in *regeringsvoorstellen en -beslissingen*. Deze tag wordt ook toegekend aan tokens waarvan de orthografische transcriptie in het CGN eindigt op *a, zoals *uitge*a*. Het equivalent hiervan in D-coi zijn tokens die eindigen op een beletselteken, zoals *uitge...*.

Tot de afkortingen rekenen we tokens als *d.w.z.* en *enz.*, ook wanneer de dots zijn weggelaten, zoals in *dwz*. Woorden als TV, WC en DVD worden niet als afkortingen getagd maar als soortnamen, o.m. omdat ze de morfologie van soortnamen vertonen: bemerk bijv. de meervoudsuitgang in TVs en het diminutiefaffix in PC-tje (zie 2.1.3). Hetzelfde geldt voor de soortnamen in *fig. 6* en *hfdst. 8*. Acroniemen als IBM en EU worden evenmin als afkortingen getagd maar als eigennamen. De tag 'afkorting' wordt dus vooral toegekend aan tokens die bij volledige uitspelling woordgroepen of niet-nominale woorden zouden zijn.

Tot de onverstaanbare tokens rekenen we die welke in de orthografische transcriptie van het CGN zijn gemarkeerd als ggg (niet-talige uiting), xxx (niet verstaan) of Xxx (niet verstane naam). Het equivalent ervan in D-coi zijn onleesbare tokens. Denk hierbij bijv. aan handgeschreven (al dan niet gedigitaliseerde) teksten.

Tot de vreemde tokens behoren die welke in de orthografische transcriptie van het CGN van het teken *v voorzien zijn. Daarnaast wordt de waarde ook toegekend aan die leenwoorden die morfo-syntactisch niet in het Nederlandse systeem geïntegreerd zijn. Hiermee wordt bedoeld dat ze niet in termen van de tags voor de tien basiswoordsoorten beschreven kunnen worden, bijvoorbeeld omdat ze onderscheidingen maken die niet tot het Nederlandse systeem behoren, zoals de Latijnse ablatief *anno*. Het criterium voor de toekenning van SPEC(vreemd) is dus niet gebaseerd op gebruiksfrequentie, vertrouwdheid of conformiteit aan de Nederlandse uitspraak, maar enkel en alleen op morfo-syntactische integratie. SPEC(vreemd) wordt daarom ook toegekend aan de delen van meerwoordsuitdrukkingen die in z'n geheel wel ingeburgerd zijn, maar waarvan de afzonderlijke woorden niet in termen van de Nederlandse CGN tags beschreven kunnen worden, zoals *ad hoc*, *wishful thinking*, *al dente*, *en profil*, e.d.

De tag 'symbool' wordt toegekend aan niet-alfabetische tekens, zoals het procentteken (%), het dollarteken, de apestaart (@), enz. Hij wordt ook toegekend aan wiskundige, natuurkundige, scheikundige e.a. symbolen, zoals het gelijkheidsteken (=) en NaCl, evenals aan de emoticons in chat- en e-mailberichten, zoals (:(). Leestekens worden niet als SPEC(symbool) getagd maar als LET(), zie 13.

De waarde 'deel van eigennaam' kennen we toe aan de delen van meerledige eigennamen, zowel Nederlandse (*Den Haag*, *Piet De Zager*, *Hans Van Halteren*) als vreemde (*Rio De Janeiro*, *Yom Kippur*, *Labour Day*, *Herald Tribune*).

De waarde 'meta' wordt toegekend aan woorden die in zelfnoemfunctie gebruikt zijn, zoals in de combinatie *het woord 'omtrent'*. De reden voor deze speciale behandeling is dat de toekenning van de normale tag voor het woord in kwestie (in dit geval 'voorzetsel') misleidend zou zijn, zowel voor de gebruiker als voor de tagger. Deze waarde wordt ook toegekend aan in zelfnoemfunctie gebruikte tokens die niet met een woord corresponderen, maar met een letter of een symbool, zoals in *hoofdletter A* en *het teken +*.

De waarde 'comment' wordt toegekend aan fragmenten die bij de orthografische transcriptie in het CGN als zodanig gemarkeerd zijn; het gaat om stukjes commentaar. Het equivalent ervan in D-coi zijn fragmenten die door de annotator zijn toegevoegd.

De waarde 'achtergrond' wordt toegekend aan fragmenten die bij de orthografische transcriptie in het CGN als zodanig gemarkeerd zijn; het gaat om achtergrondgeluiden.

De relevante tags zijn de volgende.

```
[T002] SPEC(afgebr)      uitge*a, uitge..., binnen-
[T003] SPEC(onverst)     ggg, xxx, Xxx
[T004] SPEC(vreemd)      whatever*v, ad, hoc, wishful
[T005] SPEC(deeleigen)   Den, Haag, New, York
```

```
[T006] SPEC(meta)      (het woord) omtrent
[T008] SPEC(comment)   voor commentaren
[T009] SPEC(achter)    voor achtergrondgeluid
[T010] SPEC(afk)       d.w.z., dwz, enz., EHBO
[T011] SPEC(symb)      @, %, NaCl, =, emoticons
```

Om problemen te vermijden bij de uitlijning op het scherm wordt ook aan de speciale tokens een lemma toegekend. De waarde ervan is gelijk aan het token ('_').

2.13 LEESTEKENS

In het Corpus Gesproken Nederlands werden bij de orthografische transcriptie slechts drie leestekens gebruikt, m.n. het punt, het beletselteken (...) en het vraagteken. In D-coi wordt de tag LET() ook toegekend aan de andere leestekens, zoals het uitroepteken, de komma, het dubbele punt, enz.

```
[T007] LET()      ., ?, !, ;, :, ,, ', ``,
```

Om problemen bij de uitlijning op het scherm te vermijden wordt ook aan de leestekens een lemma toegekend. De waarde ervan is gelijk aan het token.

3 VERGELIJKING MET EAGLES

In de EAGLES-standaard voor tagsets wordt een drievoudig onderscheid gemaakt tussen (1) verplichte attributen en waarden, (2) aanbevolen attributen en waarden, en (3) bijzondere extensies, die enerzijds bestaan uit optionele toevoegingen en anderzijds uit taalspecifieke toevoegingen. Aan elk ervan wordt een aparte sectie gewijd; in een vierde en laatste sectie worden de CGN features vermeld die niet in de EAGLES aanbevelingen voorkomen.

3.1 VERPLICHT

Tot deze categorie behoort slechts één attribuut, m.n. POS. Het heeft dertien waarden, waarvan er tien precies overeenkomen met de tien woordsoorten die in de CGN tagset onderscheiden worden. Van de overige drie correspondeert er een met de speciale tokens ('Residual') en een met de leestekens ('Punctuation'). De dertiende heet 'Unique' en is bedoeld voor 'categories with a unique or very small membership', zoals het Engelse 'infinitival *to*'. CGN heeft geen equivalent voor 'unique': het expletieve *er* en het infinitivale *te*, die er volgens EAGLES voor in aanmerking komen, zijn ondergebracht bij de tien bestaande woordsoorten, resp. bij de voornaamwoorden en de voorzetsels.

3.2 AANBEVOLEN

De aanbevolen features worden per woordsoort gegeven.

SUBSTANTIEVEN

De vier aanbevolen features (Type, Gender, Number, Case) zijn ook in de CGN-tagset opgenomen. Ze corresponderen respectievelijk met NTYPE, GENUS, GETAL en NAAMVAL.

ADJECTIEVEN

Van de vier aanbevolen features (Degree, Gender, Number, Case) komen er drie voor in de CGN-tagset. Genus ontbreekt, omdat de Nederlandse adjectieven, i.t.t. bijv. de Franse, niet voor genus gemarkeerd zijn. Bovendien kent CGN alleen aan de nominaal gebruikte adjectieven een getalswaarde toe.

WERKWOORDEN

EAGLES vermeldt niet minder dan acht aanbevolen features. 'Finiteness' correspondeert met het onderscheid tussen persoonsvormen en buigbare vormen (WVORM)

en ‘Tense’ correspondeert met het onderscheid tussen tegenwoordige en verleden tijd (PVTIJD). Voor ‘Verb form/Mood’ is in CGN geen apart feature voorzien, maar de relevante distincties zijn verdeeld over WVORM (infinitief, deelwoord) en PVTIJD (conjunctief). ‘Number’ is deel van PVAGR voor de persoonsvormen en van GETAL-N voor de (nominaal gebruikte) buigbare vormen. ‘Person’ heeft geen equivalent in de CGN-tagset, maar is mee opgenomen in PVAGR.

Van de drie resterende features zijn er twee niet relevant voor het Nederlands: ‘Voice’ is niet relevant omdat het Nederlandse passief niet morfologisch gemarkeerd wordt (i.t.t. het Grieks en het Deens), en ‘Gender’ is niet relevant omdat de Nederlandse deelwoorden geen variatie in genus vertonen (i.t.t. de deelwoorden van de Romaanse talen).

‘Status’ tenslotte betreft de onderscheiding tussen hoofd- en hulpwerkwoorden. Een equivalent ervan voor CGN zou het onderscheid zijn tussen hoofd-, hulp- en koppelwerkwoorden, maar omdat het maken ervan een volledige syntactische analyse van de zin vereist, is het niet in de tagset opgenomen: morfologische kenmerken volstaan immers niet, omdat de hulp- en koppelwerkwoorden dezelfde morfologische variatie vertonen als de hoofdwerkwoorden, en lexicale kenmerken volstaan evenmin omdat alle hulp- en koppelwerkwoorden homoniemen hebben die tot de hoofdwerkwoorden gerekend worden.

TELWOORDEN

EAGLES vermeldt vijf aanbevolen features. ‘Type’ correspondeert met NUMTYPE, ‘Case’ met NAAMVAL en ‘Number’ met GETAL-N. ‘Gender’ is ook hier niet relevant. Het feature ‘Function’ met de waarden ‘pronoun’, ‘determiner’ en ‘adjective’ is bedoeld ‘to indicate the part-of-speech function of a word within the numeral category’. Wat daar precies mee bedoeld wordt, is echter niet uitgelegd.

VOORNAAMWOORDEN

EAGLES vermeldt acht aanbevolen features. Twee ervan betreffen de onderscheidingen die wij onder VWTYPE hebben gegroepeerd, nl. ‘Pronoun-Type’ (demonstrative, indefinite, possessive, int/rel, pers/refl) en ‘Determiner-Type’ (demonstrative, indefinite, possessive, int/rel, partitive). Een derde feature, ‘Category’ (pronoun, determiner, both) correspondeert met PDTYPE; wat EAGLES onder ‘both’ verstaat, wordt niet uitgelegd. Bemerk dat CGN door het loskoppelen van PDTYPE en VWTYPE de redundantie vermijdt van het EAGLES voorstel.

‘Case’ correspondeert met NAAMVAL en ‘Person’ met PERSOON. ‘Number’ en ‘Gender’ corresponderen met GETAL en GENUS voor de pronomina en met NPAGR voor de prenominale determiners; bij de nominale determiners correspondeert ‘Number’ bovendien met GETAL-N.

'Possessive', tenslotte, met de waarden 'singular' en 'plural', is relevant voor de bezittelijke voornaamwoorden: in *la nostra casa* bijv. is *nostra* tegelijk (1ste persoon) meervoud en enkelvoudig. In CGN wordt dit onderscheid gemaakt door aan de bezittelijke voornaamwoorden zowel GETAL als NPAGR (prenominaal) of GETAL-N (nominaal) toe te kennen.

LIDWOORDEN

De vier in EAGLES aanbevolen features komen ook terug in de CGN tagset: 'Article-Type' correspondeert met LWTYPE, 'Case' met NAAMVAL, en 'Number' en 'Gender' met NPAGR.

VOORZETSELS

EAGLES noemt de voorzetsels 'adpositions' en kent er slechts een feature aan toe ('Type'), met de waarde 'Preposition'. Dat correspondeert met de waarde 'initieel' van het VZTYPE feature.

VOEGWOORDEN

Het feature 'Type' met de waarden 'coordinating' en 'subordinating' correspondeert met CONJTYPE en de waarden 'nevenschikkend' en 'onderschikkend'.

BIJWOORDEN

EAGLES vermeldt 'Degree' (positive, comparative, superlative). Aangezien CGN de bijwoordelijk gebruikte adjectieven als adjectieven behandeld en niet als bijwoorden, zijn er zeer weinig bijwoorden waarvoor de GRAAD variatie relevant is. Bovendien zijn de comparatieven en superlatieven van zulke bijwoorden vaak gelexicaliseerd. Om die reden hebben we dit feature niet in de CGN-tagset opgenomen, zie ook 2.9.

SPECIALE TOKENS

EAGLES beveelt drie features aan. Het eerste (Type) heeft zes mogelijke waarden (foreign word, formula, symbol, acronym, abbreviation, unclassified); die laatste waarde wordt gebruikt voor onvolledige woorden en 'pause fillers'. Het corresponderende CGN feature is SPECTYPE met de waarden 'vreemd', 'afgebroken' en 'onverstaanbaar'; CGN heeft geen equivalenten voor 'symbol' en 'formula', omdat zulke tekens in de getranscribeerde gsproken taal niet voorkomen: mensen zeggen niet '$', maar 'dollar'; CGN heeft evenmin een equivalent voor 'acronym' omdat

het die tot de substantieven rekent. Daarmee vervalt ook de behoefte aan de twee andere EAGLES features (Gender en Number).

3.3 OPTIONEEL

Zoals vermeld in de inleiding wordt binnen deze groep nog een verder onderscheid gemaakt tussen 'application- or task specific extensions' en 'language specific extensions'. We noemen die in deze paragraaf resp. A- en L-extensies. Ze betreffen ofwel de toevoeging van nieuwe attributen ofwel de toevoeging van extra waarden aan de aanbevolen attributen.

De L-extensies worden in deze paragraaf alleen vermeld als ze voor het Nederlands relevant zijn; attributen of waarden waarvan expliciet gezegd wordt dat ze speciaal voor het Deens, het Engels of een andere specifieke Euro-taal bedoeld zijn worden dus niet besproken.

SUBSTANTIEVEN

Als enige A-extensie vermeldt EAGLES het feature 'Countability' met de waarden 'mass' en 'countable'. De CGN-tagset bevat geen equivalent van dit feature, omdat het een semantische onderscheiding betreft.

ADJECTIEVEN

Bij de A-extensies vermeldt EAGLES de features 'Use' met de waarden 'attributive' en 'predicative', en 'NP Function' met de waarden 'premodifying', 'postmodifying' en 'head-function'. Die corresponderen—ongeveer—met het POSITIE feature: de 'NP Function' waarden corresponderen resp. met 'prenominaal', 'postnominaal' en 'nominaal'. Het vierde POSITIE feature ('vrij') correspondeert met 'predicative', maar heeft een ruimere interpretatie omdat het ook het adverbiale gebruik dekt. Die verruiming is nuttig omdat in het Nederlands de adverbiaal gebruikte adjectieven precies dezelfde vorm hebben als de predikatief gebruikte; in dat opzicht verschilt het Nederlands van de meeste andere eurotalen, waarin adverbiaal gebruik door een suffix gemarkeerd wordt, cf. het Engelse *-ly*, het Franse *-ment* en het Italiaanse *-mente*.

Een derde A-extensie betreft het feature 'Inflection-type' met de waarden 'weak-flection', 'strong-flection' en 'mixed'. Het vertoont enige gelijkenis met het BUIGINGs feature, maar de waarden zijn zo verschillend dat het eigenlijk om een ander feature gaat.

WERKWOORDEN

Bij de A-extensies vermeldt EAGLES vier features. 'Aspect' is wel relevant voor het Grieks en de Romaanse talen, maar niet voor het Nederlands. 'Separability' is niet opgenomen omdat de toekenning van de waarden een volledige syntactische analyse vereist; *wacht* bijv. is 'separable' in *ik wacht voorlopig de resultaten van het experiment af* maar niet in *ik wacht al weken op de resultaten van het experiment.* 'Reflexivity' is niet opgenomen omdat het de valentie van de werkwoorden betreft. Evenmin opgenomen is het feature 'Auxiliary', dat aangeeft welk hulpwerkwoord wordt geselecteerd voor de voltooide tijd (*hebben, zijn*).

VOORNAAMWOORDEN

EAGLES vermeldt drie A-extensies. 'Special Pronoun Type' heeft drie waarden (personal, reflexive, reciprocal) en 'Wh-Type' eveneens (interrogative, relative, exclamatory); ze hebben betrekking op specifieke onderdelen van de VWTYPE indeling in CGN. 'Politeness' heeft twee waarden (polite, familiar) en is in CGN deel van de PERSOONspartitie.

De L-extensie 'Strength' met de waarden 'weak' en 'strong' correspondeert met STATUS.

VOORZETSELS

De enige A-extensie betreft geen nieuw feature maar de toevoeging van de waarde 'Fused prep-art' aan 'Type'. Het correspondeert met de waarde 'versmolten' in CGN. Bij de L-extensies worden nog twee extra-waarden vermeld, nl. 'Postposition' en 'Circumposition'. De eerste correspondeert met 'finaal', de tweede heeft geen equivalent in CGN, omdat circumposities worden behandeld als combinaties van een initieel voorzetsel, een complement en een finaal voorzetsel.

VOEGWOORDEN

Bij de A-extensies vermeldt EAGLES het feature 'Coord-Type' met de waarden 'simple', 'correlative', 'initial' en 'non-initial'. In de eerste versie van de CGN tagset was een corresponderend feature opgenomen, maar dat is omwille van interpretatiemoeilijkheden bij de toekenning verwijderd. Bovendien is het herkennen van discontinue voegwoorden als delen van een geheel (*hetzij ... hetzij ..., noch ... noch ...*) geen taak van de tagger, maar van syntactische analyse en/of van lexicologische koppeling.

Bij de A-extensies vermeldt EAGLES 'Adverb-Type', 'Polarity' en 'Wh-type'. 'Adverb-Type' heeft de waarden 'general', 'degree' en—bij L-extensie—'particle' en 'pronominal'. Die indeling komt in feite neer op de isolering van drie kleine subklassen van bijwoorden, terwijl alle andere naar een grote restgroep verwezen worden ('general'). Het leek ons niet opportuun om daarvoor een CGN-equivalent te voorzien, temeer omdat we de zgn. voornaamwoordelijke bijwoorden als voornaamwoorden behandelen en omdat we de meeste partikels tot de intransitieve voorzetsels rekenen, zie 2.5 en 2.7.

'Polarity' met de waarden 'wh-type' and 'non-wh-type' onderscheidt de 'wh-adverbs' van de andere en 'Wh-type' verdeelt de eerste verder in vragende, betrekkelijke en exclamatieve. In het Nederlands zijn de wh-adverbs beperkt tot *hoe, wanneer* en de combinaties van *waar* met een voorzetsel, zoals *waarop, waarvan, ...*; de andere wh-woorden (*waar* zelf inbegrepen) zijn voornaamwoorden. Het leek ons niet de moeite waard om voor zo'n klein deel van de bijwoorden twee extra features te introduceren.

3.4 NIET VERMELD DOOR EAGLES

Niet vermeld in EAGLES, maar wel opgenomen in de CGN tagset, is de onderscheiding tussen basisvormen en diminutiefvormen. Het ontbreken in EAGLES is wellicht te wijten aan het feit dat diminutiefvorming in de meeste eurotalen geen productief proces is, maar in het Nederlands (en het Italiaans) is het dat wel, overigens niet alleen bij de substantieven, maar ook bij de adjectieven, de telwoorden en de gradeerbare determiners. Opname in de tagset is dus wenselijk, omdat we anders aparte lemmata moeten postuleren voor basisvormen en diminutiefvormen, wat haaks zou staan op de lexicografische praktijk.

Een ander punt waarop CGN verder gaat dan de EAGLES aanbevelingen betreft het gebruik van de features POSITIE en BUIGING. Die zijn in EAGLES vermeld bij de A-extensies voor de adjectieven, maar worden in CGN ook gebruikt voor de buigbare werkwoorden en de determiners. Het ruimere gebruik van BUIGING heeft te maken met het feit dat de buigingsvariatie (zonder, met-e, met-s) in andere eurotalen correspondeert met genus- en getalsonderscheidingen.

4 SYNTHESE

Formeel gesproken is de CGN tagset een sixtupel <A,W,P,D,I,T>, waarin A een verzameling is van attributen, W van waarden, P van partities, D van declaraties, I van implicaties en T van tags. Features zijn die combinaties van attributen en waarden die voldoen aan de door de partities opgelegde beperkingen. Tags zijn lijsten van features die voldoen aan de door de declaraties en de implicaties opgelegde beperkingen.

4.1 DE PARTITIES

```
[P01] TOKENTYPE = woord, speciaal, leesteken.
[P02] POS = substantief, adjectief, werkwoord, telwoord, voornaamwoord,
      lidwoord, voorzetsel, voegwoord, bijwoord, tussenwerpsel.
[P03] NTYPE = soortnaam, eigennaam.
[P04] GETAL = getal (enkelvoud, meervoud).
[P05] GRAAD = basis, comparatief, superlatief, diminutief.
[P06] GENUS = genus (zijdig (masculien, feminien), onzijdig).
[P07] NAAMVAL = standaard (nominatief, oblique), bijzonder (genitief, datief).
[P08] POSITIE = prenominaal, nominaal, postnominaal, vrij.
[P09] BUIGING = zonder, met-e, met-s.
[P10] GETAL-N = zonder-n, meervoud-n.
[P11] WVORM = persoonsvorm, buigbaar (infinitief, onvdw, voltdw).
[P12] PVTIJD = tegenwoordig, verleden, conjunctief.
[P13] PVAGR = enkelvoud, meervoud, met-t.
[P14] NUMTYPE = hoofdtelwoord, rangtelwoord.
[P15] VWTYPE = pr (persoonlijk, reflexief), reciprook, bezittelijk,
      vb (vragend, betrekkelijk), exclamatief, aanwijzend, onbepaald.
[P16] PDTYPE = pronomen (adv-pronomen), determiner (gradeerbaar).
[P17] PERSOON = persoon (1, 2 (2v, 2b), 3 (3p (3m, 3v), 3o)).
[P18] STATUS = vol, gereduceerd, nadruk.
[P19] NPAGR = agr (evon, rest (evz, mv)), agr3 (evmo, rest3 (evf, mv)).
[P20] LWTYPE = bepaald, onbepaald.
[P21] VZTYPE = initieel (versmolten), finaal.
[P22] CONJTYPE = nevenschikkend, onderschikkend.
[P23] SPECTYPE = afgebroken, onverstaanbaar, vreemd, deeleigen, meta,
      commentaar, achtergrond, afkorting, symbool.
```

4.2 DE DECLARATIES

[D00] <TOKENTYPE = woord> ⟹ <POS>
[D01] <POS = substantief> ⟹ <NTYPE, GETAL, GRAAD>
[D02] <POS = substantief, GETAL = enkelvoud> ⟹ <NAAMVAL>
[D03] <POS = substantief, GETAL = enkelvoud, NAAMVAL = standaard> ⟹
<GENUS>
[D04] <POS = adjectief> ⟹ <POSITIE, GRAAD, BUIGING>
[D05] <POSITIE = nominaal> ⟹ <GETAL-N>
[D06] <POS = adjectief, POSITIE = nominaal, BUIGING = met-e,
GETAL-N = zonder-n> ⟹ <NAAMVAL>
[D07] <POS = adjectief, POSITIE = prenominaal, BUIGING = met-e> ⟹ <NAAMVAL>
[D08] <POS = werkwoord> ⟹ <WVORM>
[D09] <WVORM = persoonsvorm> ⟹ <PVTIJD, PVAGR>
[D10] <WVORM = buigbaar> ⟹ <POSITIE, BUIGING>
[D11] <POS = telwoord> ⟹ <NUMTYPE, POSITIE>
[D12] <NUMTYPE = hoofdtelwoord, POSITIE = nominaal> ⟹ <GRAAD>
[D13] <POS = telwoord, POSITIE = prenominaal> ⟹ <NAAMVAL>
[D14] <POS = voornaamwoord> ⟹ <VWTYPE, PDTYPE, NAAMVAL>
[D15] <PDTYPE = pronomen> ⟹ <STATUS, PERSOON, GETAL>
[D16] <VWTYPE = persoonlijk, NAAMVAL = standaard, PERSOON = 3,
GETAL = enkelvoud> ⟹ <GENUS>
[D17] <PDTYPE = determiner> ⟹ <POSITIE, BUIGING>
[D18] <PDTYPE = determiner, POSITIE = prenominaal> ⟹ <NPAGR>
[D19] <PDTYPE = gradeerbaar> ⟹ <GRAAD>
[D20] <VWTYPE = bezittelijk> ⟹ <STATUS, PERSOON, GETAL>
[D21] <POS = lidwoord> ⟹ <LWTYPE, NAAMVAL, NPAGR>
[D22] <POS = voorzetsel> ⟹ <VZTYPE>
[D23] <POS = voegwoord> ⟹ <CONJTYPE>
[D24] <TOKENTYPE = speciaal> ⟹ <SPECTYPE>

4.3 DE IMPLICATIES

[I01] <POS = substantief, GRAAD = diminutief, GETAL = enkelvoud>
$\Longrightarrow$ <NAAMVAL $\neq$ datief>
[I02] <POS = substantief, GRAAD = diminutief, GETAL = enkelvoud,
NAAMVAL = standaard> $\Longrightarrow$ <GENUS = onzijdig>
[I03] <POS = adjectief, GRAAD = superlatief> $\Longrightarrow$ <POSITIE $\neq$ postnominaal>
[I04] <POS = adjectief, GRAAD = diminutief> $\Longrightarrow$ <POSITIE = vrij>
[I05] <BUIGING = met-s> $\Longrightarrow$ <POSITIE = postnominaal>
[I06] <BUIGING = met-e> $\Longrightarrow$ <POSITIE = (pre)nominaal>
[I07] <WVORM = infinitief, POSITIE = nominaal> $\Longrightarrow$
<BUIGING = zonder, GETAL-N = zonder-n>
[I08] <PVTIJD = conjunctief> $\Longrightarrow$ <PVAGR = enkelvoud>
[I09] <NUMTYPE = rangtelwoord> $\Longrightarrow$ <POSITIE $\neq$ vrij>
[I10] <VWTYPE = persoonlijk> $\Longrightarrow$ <PDTYPE = pronomen>
[I11] <VWTYPE = reflexief> $\Longrightarrow$ <PDTYPE = pronomen, NAAMVAL = oblique>
[I12] <VWTYPE = reciprook> $\Longrightarrow$ <PDTYPE = pronomen,
NAAMVAL $\neq$ nominatief, STATUS = vol, GETAL = meervoud>
[I13] <VWTYPE = bezittelijk> $\Longrightarrow$ <PDTYPE = determiner, POSITIE $\neq$ vrij>
[I14] <VWTYPE = vragend, PDTYPE = pronomen> $\Longrightarrow$ <STATUS $\neq$ gereduceerd>
[I15] <VWTYPE = betrekkelijk, PDTYPE = pronomen> $\Longrightarrow$ <STATUS = vol>
[I16] <VWTYPE = exclamatief, PDTYPE = pronomen> $\Longrightarrow$ <STATUS = vol>
[I17] <PDTYPE = determiner, POSITIE = prenominaal, NAAMVAL = standaard>
$\Longrightarrow$ <NPAGR = agr>
[I18] <PDTYPE = determiner, POSITIE = prenominaal, NAAMVAL = bijzonder>
$\Longrightarrow$ <NPAGR = agr3>
[I19] <POS = lidwoord, NAAMVAL = standaard> $\Longrightarrow$ <NPAGR = agr>
[I20] <POS = lidwoord, NAAMVAL = bijzonder> $\Longrightarrow$ <NPAGR = agr3>

4.4 DE TAGS

POS	T	U	TOTAAL
substantieven	16	2	18
adjectieven	30		30
werkwoorden	21		21
telwoorden	11		11
voornaamwoorden	43	145	188
lidwoorden	7	2	9
voorzetsels	3		3
voegwoorden	2		2
bijwoorden	1		1
tussenwerpsels	1		1
dialectwoorden		28	28
speciale tokens	7		7
leestekens	1		1
TOTAAL	143	177	320

[T101] N(soort,ev,basis,zijd,stan)	die stoel, deze muziek, de filter
[T102] N(soort,ev,basis,onz,stan)	het kind, ons huis, het filter
[T103] N(soort,ev,dim,onz,stan)	dit stoeltje, op 't nippertje
[T104] N(soort,ev,basis,gen)	's avonds, de heer des huizes
[T105] N(soort,ev,dim,gen)	vadertjes pijp
[T106] N(soort,ev,basis,dat)	ter plaatse, heden ten dage
[T107] N(soort,mv,basis)	stoelen, kinderen, hersenen
[T108] N(soort,mv,dim)	stoeltjes, huisjes, hersentjes
[T109] N(eigen,ev,basis,zijd,stan)	de Noordzee, de Kemmelberg, Karel
[T110] N(eigen,ev,basis,onz,stan)	het Hageland, het Nederlands
[T111] N(eigen,ev,dim,onz,stan)	het slimme Kareltje

[T112]	N(eigen,ev,basis,gen)	des Heren, Hagelands trots
[T113]	N(eigen,ev,dim,gen)	Kareltjes fiets
[T114]	N(eigen,ev,basis,dat)	wat den Here toekomt
[T115]	N(eigen,mv,basis)	de Ardennen, de Middeleeuwen
[T116]	N(eigen,mv,dim)	de Maatjes
[U117]	N(soort,ev,basis,*genus*,stan)	een riool, geen filter
[U118]	N(eigen,ev,basis,*genus*,stan)	Linux, Esselte
[T201]	ADJ(prenom,basis,zonder)	een mooi huis, een houten pot
[T202]	ADJ(prenom,basis,met-e,stan)	mooie huizen, een grote pot
[T203]	ADJ(prenom,basis,met-e,bijz)	zaliger gedachtenis, van goeden huize
[T204]	ADJ(prenom,comp,zonder)	een mooier huis
[T205]	ADJ(prenom,comp,met-e,stan)	mooiere huizen, een grotere pot
[T206]	ADJ(prenom,comp,met-e,bijz)	van beteren huize
[T207]	ADJ(prenom,sup,zonder)	een alleraardigst mens
[T208]	ADJ(prenom,sup,met-e,stan)	de mooiste keuken, het grootste paard
[T209]	ADJ(prenom,sup,met-e,bijz)	bester kwaliteit
[T210]	ADJ(nom,basis,zonder,zonder-n)	in het groot, het groen
[T211]	ADJ(nom,basis,zonder,mv-n)	de timiden, dezelfden
[T212]	ADJ(nom,basis,met-e,zonder-n,stan)	het leuke is dat, een grote met tartaar
[T213]	ADJ(nom,basis,met-e,zonder-n,bijz)	hosanna in den hogen
[T214]	ADJ(nom,basis,met-e,mv-n)	de rijken
[T215]	ADJ(nom,comp,zonder,zonder-n)	
[T216]	ADJ(nom,comp,met-e,zonder-n,stan)	een betere
[T217]	ADJ(nom,comp,met-e,zonder-n,bijz)	
[T218]	ADJ(nom,comp,met-e,mv-n)	de ouderen
[T219]	ADJ(nom,sup,zonder,zonder-n)	op z'n best, om ter snelst
[T220]	ADJ(nom,sup,met-e,zonder-n,stan)	het leukste is dat, het langste blijven

[T221]	ADJ(nom,sup,met-e,zonder-n,bijz)	des Allerhoogsten
[T222]	ADJ(nom,sup,met-e,mv-n)	de slimsten
[T223]	ADJ(postnom,basis,zonder)	rivieren bevaarbaar in de winter
[T224]	ADJ(postnom,basis,met-s)	iets moois
[T225]	ADJ(postnom,comp,zonder)	een getal groter dan 3
[T226]	ADJ(postnom,comp,met-s)	iets gekkers kon ik niet bedenken
[T227]	ADJ(vrij,basis,zonder)	die stok is lang, lang slapen
[T228]	ADJ(vrij,comp,zonder)	deze stok is langer, langer slapen
[T229]	ADJ(vrij,sup,zonder)	die stok is het langst, het langst slapen
[T230]	ADJ(vrij,dim,zonder)	het is hier stilletjes, stilletjes weggaan
[T301]	WW(pv,tgw,ev)	ik kom, speel je, hij is, zwijg
[T302]	WW(pv,tgw,mv)	komen, spelen
[T303]	WW(pv,tgw,met-t)	jij komt, hij speelt, zwijgt
[T304]	WW(pv,verl,ev)	kwam, speelde
[T305]	WW(pv,verl,mv)	kwamen, speelden
[T306]	WW(pv,verl,met-t)	kwaamt, gingt
[T309]	WW(pv,conj,ev)	kome, leve de koning
[T310]	WW(inf,prenom,zonder)	de nog te lezen post
[T311]	WW(inf,prenom,met-e)	een niet te weerstane verleiding
[T312]	WW(inf,nom,zonder,zonder-n)	(het) spelen, (het) schaatsen
[T314]	WW(inf,vrij,zonder)	zal komen
[T315]	WW(vd,prenom,zonder)	een verwittigd man, een gekregen paard
[T316]	WW(vd,prenom,met-e)	een getemde feeks
[T317]	WW(vd,nom,met-e,zonder-n)	het geschrevene, een gekwetste

[T318]	WW(vd,nom,met-e,mv-n)	gekwetsten, gedupeerden
[T320]	WW(vd,vrij,zonder)	is gekomen
[T321]	WW(od,prenom,zonder)	een slapend kind
[T322]	WW(od,prenom,met-e)	een piano spelende aap, slapende kinderen
[T323]	WW(od,nom,met-e,zonder-n)	het resterende, een klagende
[T324]	WW(od,nom,met-e,mv-n)	de wachtenden
[T326]	WW(od,vrij,zonder)	liep lachend weg, al doende leert men
[T401]	TW(hoofd,prenom,stan)	vier cijfers
[T402]	TW(hoofd,prenom,bijz)	eens geestes zijn, te enen male
[T403]	TW(hoofd,nom,zonder-n,basis)	er is er een ontsnapt
[T404]	TW(hoofd,nom,mv-n,basis)	met z'n vieren
[T405]	TW(hoofd,nom,zonder-n,dim)	er is er eentje ontsnapt, op z'n eentje
[T406]	TW(hoofd,nom,mv-n,dim)	met z'n tweetjes
[T407]	TW(hoofd,vrij)	veertig worden, honderd rijden, hoeveel sneller
[T408]	TW(rang,prenom,stan)	de vierde man
[T409]	TW(rang,prenom,bijz)	te elfder ure
[T410]	TW(rang,nom,zonder-n)	het eerste, (de) vierde eindigen, Karel de Vijfde
[T411]	TW(rang,nom,mv-n)	de eersten, iets aan derden verkopen
[T501a]	VNW(pers,pron,nomin,vol,1,ev)	ik
[T501b]	VNW(pers,pron,nomin,nadr,1,ev)	ikzelf, ikke
[T501c]	VNW(pers,pron,nomin,red,1,ev)	'k
[T501d]	VNW(pers,pron,nomin,vol,1,mv)	wij
[T501e]	VNW(pers,pron,nomin,nadr,1,mv)	wijzelf
[T501f]	VNW(pers,pron,nomin,red,1,mv)	we
[T501g]	VNW(pers,pron,nomin,vol,2v,ev)	jij
[T501h]	VNW(pers,pron,nomin,nadr,2v,ev)	jijzelf

[T501i]	VNW(pers,pron,nomin,red,2v,ev)	je
[U501j]	VNW(pers,pron,nomin,vol,2b,*getal*)	u
[U501k]	VNW(pers,pron,nomin,nadr,2b,*getal*)	uzelf
[U501l]	VNW(pers,pron,nomin,vol,*2,getal*)	gij
[U501m]	VNW(pers,pron,nomin,nadr,*2,getal*)	gijzelf
[U501n]	VNW(pers,pron,nomin,red,*2,getal*)	ge
[U501o]	VNW(pers,pron,nomin,vol,*3*,ev,masc)	hij
[T501p]	VNW(pers,pron,nomin,nadr,3m,ev,masc)	hijzelf
[U501q]	VNW(pers,pron,nomin,red,*3*,ev,masc)	ie
[U501r]	VNW(pers,pron,nomin,red,*3p*,ev,masc)	men
[T501s]	VNW(pers,pron,nomin,vol,3v,ev,fem)	zij
[T501t]	VNW(pers,pron,nomin,nadr,3v,ev,fem)	zijzelf
[U501u]	VNW(pers,pron,nomin,vol,*3p*,mv)	zij
[U501v]	VNW(pers,pron,nomin,nadr,*3p*,mv)	zijzelf
[T502a]	VNW(pers,pron,obl,vol,2v,ev)	jou
[U502b]	VNW(pers,pron,obl,vol,*3*,ev,masc)	hem
[T502c]	VNW(pers,pron,obl,nadr,3m,ev,masc)	hemzelf
[U502d]	VNW(pers,pron,obl,red,*3*,ev,masc)	’m
[U502e]	VNW(pers,pron,obl,vol,*3,getal*,fem)	haar
[U502f]	VNW(pers,pron,obl,nadr,3v,*getal*,fem)	haarzelf
[U502g]	VNW(pers,pron,obl,red,3v,*getal*,fem)	’r, d’r
[U502h]	VNW(pers,pron,obl,vol,*3p*,mv)	hen, hun
[U502i]	VNW(pers,pron,obl,nadr,*3p*,mv)	henzelf, hunzelf
[U503a]	VNW(pers,pron,*stan*,nadr,2v,mv)	jullie
[U503b]	VNW(pers,pron,*stan*,red,*3*,ev,onz)	het, ’t
[U503c]	VNW(pers,pron,*stan*,red,*3*,ev,fem)	ze
[U503d]	VNW(pers,pron,*stan*,red,*3*,mv)	ze
[T504a]	VNW(pers,pron,gen,vol,1,ev)	mijns gelijke, gedenk mijner
[T504b]	VNW(pers,pron,gen,vol,1,mv)	ons gelijke, velen onzer
[U504c]	VNW(pers,pron,gen,vol,*2,getal*)	uws gelijke, wie uwer
[T504d]	VNW(pers,pron,gen,vol,3m,ev)	zijns gelijke, zijner
[U504e]	VNW(pers,pron,gen,vol,3v,*getal*)	haars gelijke, harer
[U504f]	VNW(pers,pron,gen,vol,*3p*,mv)	huns gelijke, een hunner
[U505a]	VNW(*pr*,pron,obl,vol,1,ev)	mij
[U505b]	VNW(*pr*,pron,obl,nadr,1,ev)	mezelf, mijzelf
[U505c]	VNW(*pr*,pron,obl,red,1,ev)	me
[U505d]	VNW(*pr*,pron,obl,vol,1,mv)	ons

[U505e]	VNW(*pr*,pron,obl,nadr,1,mv)	onszelf
[U505f]	VNW(*pr*,pron,obl,red,2v,*getal*)	je
[U505g]	VNW(*pr*,pron,obl,nadr,2v,*getal*)	jezelf
[U505h]	VNW(*pr*,pron,obl,vol,*2,getal*)	u
[U505i]	VNW(*pr*,pron,obl,nadr,*2,getal*)	uzelf
[U506a]	VNW(refl,pron,obl,red,*3,getal*)	zich
[U506b]	VNW(refl,pron,obl,nadr,*3,getal*)	zichzelf
[U507a]	VNW(recip,pron,obl,vol,*persoon*,mv)	elkaar, mekaar, elkander
[U508a]	VNW(recip,pron,gen,vol,*persoon*,mv)	elkaars, mekaars, elkanders
[U509a]	VNW(bez,det,*stan*,vol,1,ev,prenom,zonder,*agr*)	mijn paard(en)
[U509b]	VNW(bez,det,*stan*,vol,1,ev,prenom,met-e,rest)	mijne heren
[U509c]	VNW(bez,det,*stan*,red,1,ev,prenom,zonder,*agr*)	m'n paard(en)
[U509d]	VNW(bez,det,*stan*,vol,1,mv,prenom,zonder,evon)	ons paard
[U509e]	VNW(bez,det,*stan*,vol,1,mv,prenom,met-e,rest)	onze paarden
[U509f]	VNW(bez,det,*stan*,vol,*2,getal*,prenom,zonder,*agr*)	uw paard(en)
[U509g]	VNW(bez,det,*stan*,vol,*2,getal*,prenom,met-e,rest)	uwe heiligheid
[U509h]	VNW(bez,det,*stan*,vol,2v,ev,prenom,zonder,*agr*)	jouw paard(en)
[U509i]	VNW(bez,det,*stan*,red,2v,ev,prenom,zonder,*agr*)	je paard(en)
[U509j]	VNW(bez,det,*stan*,nadr,2v,mv,prenom,zonder,*agr*)	jullie paard(en)
[U509k]	VNW(bez,det,*stan*,vol,*3*,ev,prenom,zonder,*agr*)	zijn paard(en), haar kind
[U5091]	VNW(bez,det,*stan*,vol,3m,ev,prenom,met-e,rest)	zijne excellentie
[U509m]	VNW(bez,det,*stan*,vol,3v,ev,prenom,met-e,rest)	hare majesteit
[U509n]	VNW(bez,det,*stan*,red,*3*,ev,prenom,zonder,*agr*)	z'n paard
[U509o]	VNW(bez,det,*stan*,vol,*3*,mv,prenom,zonder,*agr*)	hun paarden
[U509p]	VNW(bez,det,*stan*,vol,*3p*,mv,prenom,met-e,rest)	hunne
[U509q]	VNW(bez,det,*stan*,red,*3,getal*,prenom,zonder,*agr*)	'r paard, d'r paard
[T510a]	VNW(bez,det,gen,vol,1,ev,prenom,zonder,evmo)	mijns inziens
[U510b]	VNW(bez,det,gen,vol,1,ev,prenom,met-e,*rest3*)	een mijner vrienden
[T510c]	VNW(bez,det,gen,vol,1,mv,prenom,met-e,evmo)	onzes inziens
[U510d]	VNW(bez,det,gen,vol,1,mv,prenom,met-e,*rest3*)	een onzer vrienden
[U510e]	VNW(bez,det,gen,vol,*2*,getal,prenom,zonder,evmo)	uws
[U510f]	VNW(bez,det,gen,vol,*2*,getal,prenom,met-e,*rest3*)	een uwer vrienden
[U510g]	VNW(bez,det,gen,vol,2v,ev,prenom,met-e,*rest3*)	een jouwer vrienden
[U510h]	VNW(bez,det,gen,vol,*3*,ev,prenom,zonder,evmo)	zijns inziens
[U510i]	VNW(bez,det,gen,vol,*3*,ev,prenom,met-e,*rest3*)	een zijner vrienden
[T510j]	VNW(bez,det,gen,vol,3v,ev,prenom,zonder,evmo)	haars inziens

[U510k]	VNW(bez,det,gen,vol,3v,ev,prenom,met-e,*rest3*)	een harer vrienden
[U510l]	VNW(bez,det,gen,vol,*3p*,mv,prenom,zonder,evmo)	huns inziens
[U510m]	VNW(bez,det,gen,vol,*3p*,mv,prenom,met-e,*rest3*)	een hunner vrienden
[T511a]	VNW(bez,det,dat,vol,1,ev,prenom,met-e,evmo)	te mijnen huize
[T511b]	VNW(bez,det,dat,vol,1,ev,prenom,met-e,evf)	te mijner ere
[T511c]	VNW(bez,det,dat,vol,1,mv,prenom,met-e,evmo)	te onzen behoeve
[T511d]	VNW(bez,det,dat,vol,1,mv,prenom,met-e,evf)	te onzer ere
[U511e]	VNW(bez,det,dat,vol,*2*,getal,prenom,met-e,evmo)	te uwen behoeve
[U511f]	VNW(bez,det,dat,vol,*2*,getal,prenom,met-e,evf)	te uwer ere
[T511g]	VNW(bez,det,dat,vol,2v,ev,prenom,met-e,evf)	te jouwer nagedachtenis
[U511h]	VNW(bez,det,dat,vol,*3*,ev,prenom,met-e,evmo)	zijnen
[U511i]	VNW(bez,det,dat,vol,*3*,ev,prenom,met-e,evf)	te zijner tijd
[T511j]	VNW(bez,det,dat,vol,3v,ev,prenom,met-e,evmo)	haren
[T511k]	VNW(bez,det,dat,vol,3v,ev,prenom,met-e,evf)	te harer ere
[U511l]	VNW(bez,det,dat,vol,*3p*,mv,prenom,met-e,evmo)	hunnen
[U511m]	VNW(bez,det,dat,vol,*3p*,mv,prenom,met-e,evf)	te hunner ere
[U512h]	VNW(bez,det,*stan*,vol,1,ev,nom,met-e,zonder-n)	het mijne
[U512i]	VNW(bez,det,*stan*,vol,1,mv,nom,met-e,zonder-n)	de onze
[U512j]	VNW(bez,det,*stan*,vol,*2,getal*,nom,met-e,zonder-n)	het uwe
[U512k]	VNW(bez,det,*stan*,vol,2v,ev,nom,met-e,zonder-n)	de jouwe
[U512l]	VNW(bez,det,*stan*,vol,3m,ev,nom,met-e,zonder-n)	het zijne
[U512m]	VNW(bez,det,*stan*,vol,3v,ev,nom,met-e,zonder-n)	de hare
[U512n]	VNW(bez,det,*stan*,vol,*3p*,mv,nom,met-e,zonder-n)	het hunne
[U512o]	VNW(bez,det,*stan*,vol,1,ev,nom,met-e,mv-n)	de mijnen
[U512p]	VNW(bez,det,*stan*,vol,1,mv,nom,met-e,mv-n)	de onzen
[U512q]	VNW(bez,det,*stan*,vol,*2,getal*,nom,met-e,mv-n)	de uwen
[U512r]	VNW(bez,det,*stan*,vol,2v,ev,nom,met-e,mv-n)	de jouwen
[U512s]	VNW(bez,det,*stan*,vol,3m,ev,nom,met-e,mv-n)	de zijnen
[U512t]	VNW(bez,det,*stan*,vol,3v,ev,nom,met-e,mv-n)	de haren
[U512u]	VNW(bez,det,*stan*,vol,*3p*,mv,nom,met-e,mv-n)	de hunnen
[T513a]	VNW(bez,det,dat,vol,1,ev,nom,met-e,zonder-n)	te mijnent
[T513b]	VNW(bez,det,dat,vol,1,mv,nom,met-e,zonder-n)	ten onzent
[U513c]	VNW(bez,det,dat,vol,*2*,getal,nom,met-e,zonder-n)	ten uwent
[T513d]	VNW(bez,det,dat,vol,3m,ev,nom,met-e,zonder-n)	te zijnent
[T513e]	VNW(bez,det,dat,vol,3v,ev,nom,met-e,zonder-n)	ten harent
[U513f]	VNW(bez,det,dat,vol,*3p*,mv,nom,met-e,zonder-n)	ten hunnent
[U514a]	VNW(vrag,pron,*stan*,nadr,3o,ev)	watte
[U515a]	VNW(betr,pron,*stan*,vol,*persoon,getal*)	de man die daar staat

[U515b]	VNW(betr,pron,*stan*,vol,*3*,ev)	het kind dat je daar ziet
[U515c]	VNW(betr,det,*stan*,nom,zonder,zonder-n)	hetgeen je daar ziet, het feest tijdens hetwelk
[U515d]	VNW(betr,det,*stan*,nom,met-e,zonder-n)	op hetgene de gemeente doet
[T516a]	VNW(betr,pron,gen,vol,3o,ev)	het warenhuis welks directeur hem een baan had aangeboden
[U516b]	VNW(betr,pron,gen,vol,3o,*getal*)	de kathedraal welker gewelven
[U517a]	VNW(*vb*,pron,*stan*,vol,*3p,getal*)	wie gaat er mee
[U517b]	VNW(*vb*,pron,*stan*,vol,3o,ev)	wat ik niet begrijp is
[U518a]	VNW(*vb*,pron,gen,vol,3m,ev)	wiens hoed is dit
[U518b]	VNW(*vb*,pron,gen,vol,3v,ev)	de vrouw wier hoed daar hangt
[U518c]	VNW(*vb*,pron,gen,vol,*3p*,mv)	de studenten tegen wier houding ...
[U519a]	VNW(*vb*,adv-pron,obl,vol,3o,*getal*)	waar ga je naartoe, de trein waar we op staan te wachten
[U520a]	VNW(excl,pron,*stan*,vol,*3,getal*)	wat een dwaasheid, wat kan jij liegen zeg
[U521a]	VNW(*vb*,det,*stan*,prenom,zonder,evon)	welk kind
[U521b]	VNW(*vb*,det,*stan*,prenom,met-e,rest)	welke kinderen
[U522a]	VNW(*vb*,det,*stan*,nom,met-e,zonder-n)	welke vind jij de mooiste
[U523a]	VNW(excl,det,*stan*,vrij,zonder)	welk een dwaasheid
[U524a]	VNW(aanw,pron,*stan*,vol,3o,ev)	dat, dit, zulks
[U524b]	VNW(aanw,pron,*stan*,nadr,3o,ev)	datte, ditte
[U524c]	VNW(aanw,pron,*stan*,vol,*3,getal*)	die
[T525a]	VNW(aanw,pron,gen,vol,3m,ev)	diens voorkeur
[T525b]	VNW(aanw,pron,gen,vol,3o,ev)	en dies meer
[U526a]	VNW(aanw,adv-pron,obl,vol,3o,*getal*)	hier, daar
[U527a]	VNW(aanw,adv-pron,*stan*,red,*3,getal*)	d'r, het niet-kwantitatieve 'er'

[U528a]	VNW(aanw,det,*stan*,prenom,zonder,evon)	dat boek, dit dier, ginds bos, zulk hout
[U528b]	VNW(aanw,det,*stan*,prenom,zonder,rest)	die stoel(en)
[U528c]	VNW(aanw,det,*stan*,prenom,zonder,*agr*)	zo'n boek(en)
[U528d]	VNW(aanw,det,*stan*,prenom,met-e,rest)	deze man, gene zijde, gindse heuvel, zulke balken
[U529a]	VNW(aanw,det,gen,prenom,met-e,*rest3*)	een dezer dagen, de notulen dier vergadering
[T530a]	VNW(aanw,det,dat,prenom,met-e,evmo)	te dien tijde
[T530b]	VNW(aanw,det,dat,prenom,met-e,evf)	in dier voege
[U531b]	VNW(aanw,det,*stan*,nom,met-e,zonder-n)	deze, gene, datgene, degene, diegene
[U531c]	VNW(aanw,det,*stan*,nom,met-e,mv-n)	dezen, genen, degenen, diegenen
[T532a]	VNW(aanw,det,gen,nom,met-e,zonder-n)	schrijver dezes, de twintigste dezer
[T533a]	VNW(aanw,det,dat,nom,met-e,zonder-n)	dat is dan bij dezen beslist
[U534a]	VNW(aanw,det,*stan*,vrij,zonder)	zulk een vreemde gedachte
[U535a]	VNW(onbep,pron,*stan*,vol,*3p*,ev)	alleman, (n)iemand, iedereen, elkeen, menigeen
[U535b]	VNW(onbep,pron,*stan*,vol,3o,ev)	alles, (n)iets, niks, wat, zoiets
[U536a]	VNW(onbep,pron,gen,vol,*3p*,ev)	allemans, andermans, (n)iemands, (een)ieders
[U537a]	VNW(onbep,adv-pron,obl,vol,3o,*getal*)	(n)ergens, overal
[U538a]	VNW(onbep,adv-pron,gen,red,*3,getal*)	het kwantitatieve 'er'
[U539a]	VNW(onbep,det,*stan*,prenom,zonder,evon)	elk huis, ieder kind, enig benul, een enkel woord, sommig bier
[U539b]	VNW(onbep,det,*stan*,prenom,zonder,*agr*)	geen kind(eren), menig politicus
[U539c]	VNW(onbep,det,*stan*,prenom,met-e,evz)	elke hond, iedere keer, ene mijnheer X, menige

[U539d]	VNW(onbep,det,*stan*,prenom,met-e,mv)	ettelijke
[U539e]	VNW(onbep,det,*stan*,prenom,met-e,rest)	sommige, enige, enkele
[U539f]	VNW(onbep,det,*stan*,prenom,met-e,*agr*)	alle mensen, hoop, vee
[T540a]	VNW(onbep,det,gen,prenom,met-e,mv)	proletariërs aller landen
[T541a]	VNW(onbep,det,dat,prenom,met-e,evmo)	te allen prijze
[T541b]	VNW(onbep,det,dat,prenom,met-e,evf)	te eniger tijd
[U542a]	VNW(onbep,grad,*stan*,prenom,zonder,*agr*,basis)	veel plezier, weinig geld
[U542b]	VNW(onbep,grad,*stan*,prenom,met-e,*agr*,basis)	het vele plezier, de weinige toeschouwers
[U542c]	VNW(onbep,grad,*stan*,prenom,met-e,mv,basis)	beide mannen
[U542d]	VNW(onbep,grad,*stan*,prenom,zonder,*agr*,comp)	meer tijd, minder werk
[U542e]	VNW(onbep,grad,*stan*,prenom,met-e,*agr*,sup)	de meeste mensen, het minste tijd
[U542f]	VNW(onbep,grad,*stan*,prenom,met-e,*agr*,comp)	in mindere mate
[U543a]	VNW(onbep,det,*stan*,nom,met-e,mv-n)	allen, sommigen, enkelen, de enen
[U543b]	VNW(onbep,det,*stan*,nom,met-e,zonder-n)	het éne ... het andere
[U543c]	VNW(onbep,det,*stan*,nom,zonder,zonder-n)	het één en ander
[T544a]	VNW(onbep,det,gen,nom,met-e,mv-n)	met aller instemming
[U545a]	VNW(onbep,grad,*stan*,nom,met-e,zonder-n,basis)	het weinige
[U545b]	VNW(onbep,grad,*stan*,nom,met-e,mv-n,basis)	velen, weinigen, beiden
[U545d]	VNW(onbep,grad,*stan*,nom,met-e,zonder-n,sup)	het minste wat je kan zeggen, de meeste
[U545e]	VNW(onbep,grad,*stan*,nom,met-e,mv-n,sup)	de minsten, de meesten
[U545f]	VNW(onbep,grad,*stan*,nom,zonder,mv-n,dim)	met z'n beidjes
[T546a]	VNW(onbep,grad,gen,nom,met-e,mv-n,basis)	tot veler verbazing, met beider instemming
[U547a]	VNW(onbep,det,*stan*,vrij,zonder)	ze kregen elk/ieder/allebei een bal, al die mensen
[U548a]	VNW(onbep,grad,*stan*,vrij,zonder,basis)	dat is te weinig, veel groter
[U548b]	VNW(onbep,grad,*stan*,vrij,zonder,sup)	de minst gevraagde, de meest gezochte

[U548c] VNW(onbep,grad,*stan*,vrij,zonder,comp)	minder werken, meer slapen
[T601] LID(bep,stan,evon)	het kind, in 't geniep
[T602] LID(bep,stan,rest)	de hond(en), de kinderen
[T603] LID(bep,gen,evmo)	des duivels, 's avonds,
[U604] LID(bep,gen,*rest3*)	der Nederlandse taal, der Belgen
[T605] LID(bep,dat,evmo)	op den duur, om den brode
[T606] LID(bep,dat,evf)	in der minne
[T607] LID(bep,dat,mv)	die in den hemelen zijt
[U608] LID(onbep,stan,*agr*)	een kind, een mensen dat er waren
[T609] LID(onbep,gen,evf)	de kracht ener vrouw
[T701] VZ(init)	met een lepeltje, met Jan in het hospitaal, met zo te roepen
[T702] VZ(fin)	liep de trap af, bij de beesten af, speelt het bandje af
[T703] VZ(versm)	ten strijde, ten hoogste, ter plaatse
[T801] VG(neven)	Jan en Peter; en toen gebeurde het
[T802] VG(onder)	omdat ze zich niet goed voelt
[T901] BW()	gisteren, nu, niet, nog, al, hoe
[T001] TSW()	oei, amai, uh, hoera
DIALECTWOORDEN	
[R101] N(soort,dial)	bompa*d

[R102]	N(eigen,dial)	
[R201]	ADJ(dial)	ne*d langen*d toot*d
[R301]	WW(dial)	'k zen*d nie*d thuis, 'k hem*d gee*d geld
[R401]	TW(hoofd,dial)	
[R402]	TW(rang,dial)	den*d elfste*d
[R501]	VNW(pers,pron,dial)	kom de*d gij mee, 'k heb ulie*d gezien
[R502]	VNW(refl,pron,dial)	
[R503]	VNW(recip,pron,dial)	we zien malkanderen*d niet veel
[R504]	VNW(bez,det,dial)	hij heeft z'ne*d frak*d vergeten
[R505]	VNW(vrag,pron,dial)	
[R506]	VNW(vrag,det,dial)	
[R507]	VNW(betr,pron,dial)	
[R508]	VNW(betr,det,dial)	
[R509]	VNW(excl,pron,dial)	
[R510]	VNW(excl,det,dial)	
[R511]	VNW(aanw,pron,dial)	
[R512]	VNW(aanw,det,dial)	diejen*d boek, dees*d week
[R513]	VNW(onbep,pron,dial)	z' hebben iet*d gezien
[R514]	VNW(onbep,det,dial)	ze kan elken*d dag vertrekken
[R601]	LID(bep,dial)	het gevecht met den*d beer
[R602]	LID(onbep,dial)	nen*d toffe gast, ne*d vieze vent
[R701]	VZ(init,dial)	me*d veel geduld
[R702]	VZ(fin,dial)	
[R801]	VG(neven,dial)	
[R802]	VG(onder,dial)	't schijnt da*d ze nie*d kunnen komen
[R901]	BW(dial)	efkes*d, nie*d
[R001]	TSW(dial)	neeje*d, wablieft*d
SPECIALE TOKENS		
[T002]	SPEC(afgebr)	uitge*a, binnen-
[T003]	SPEC(onverst)	ggg, xxx, Xxx
[T004]	SPEC(vreemd)	whatever*v, ad, hoc, wishful

[T005]	SPEC(deeleigen)	Den, Haag, New, York
[T006]	SPEC(meta)	(het woord) homosexueel
[T008]	SPEC(comment)	voor commentaren
[T009]	SPEC(achter)	voor achtergrondgeluid
[T010]	SPEC(afk)	d.w.z., dwz, enz., EHBO
[T011]	SPEC(symb)	@, %, NaCl, =, emoticons

LEESTEKENS

[T007]	LET()	., ..., ?

5 REFERENTIES

[EAGLES] Expert Advisory Group on Language Engineering Standards (1996). *Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG - TCWG - MAC/R. Version of March, 1996.

[ANS-97] W. Haeseryn, K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn (1997), *Algemene Nederlandse Spraakkunst*. 2de geheel herziene druk. Martinus Nijhoff, Groningen & Wolters Plantyn, Deurne.

[WOTAN-2] H. van Halteren (1999), The WOTAN2 Tagset Manual (under construction). Nijmegen, april 1999.

[STTS-95] A. Schiller, S. Teufel & C. Thielen (1995), Guidelines für das Tagging deutscher Textcorpora mit STTS (Stuttgart-Tübingen Tagset).

G. Booij & A. van Santen (1998), *Morfologie. De woordstructuur van het Nederlands*. 2de geheel herziene druk. Amsterdam University Press, 1998.

I. Schuurman (1998), POS taggers. Leuven, november 1998.

F. Van Eynde (2001), CGN-Functiewoordenlexicon. Leuven, juni 2001 (integraal opgenomen in het CGN-lexicon).

F. Van Eynde, J. Zavrel & W. Daelemans (2000), Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus. In M. Gavrilidou et al. (eds), *Proceedings of the Second International Conference on Language Resources and Evaluation*. Volume III. p. 1427-1433. Athens, 2000.

J. Zavrel (1999), Annotator-overeenstemming bij het manuele taggingexperiment. Tilburg, juni 1999.

J. Zavrel & W. Daelemans (1999), Evaluatie van Part-of-Speech taggers voor het Corpus Gesproken Nederlands. Tilburg, juli 1999.