

# 1. Inleiding

## 1.1 Achtergrond

Het Referentiebestand Nederlands (RBN) is een multifunctioneel lexicon voor het Nederlands dat ontwikkeld is op initiatief van de Commissie Lexicografische Vertaalvoorzieningen (CLVV). De CLVV is een door de toenmalige Nederlandse en de Vlaamse Ministers van Onderwijs ingestelde commissie die gedurende de periode 1993-2003 zorgde voor de totstandkoming van tweetalige woordenboeken en andere vertaalvoorzieningen tussen het Nederlands en andere talen.<sup>1</sup> Het betreft talen waarvoor er nog geen goede voorzieningen zijn en waarvoor er zonder overheidssteun ook geen zouden komen.

Het RBN is ingezet voor zowel lexicografische toepassingen (bv. woordenboekprojecten Nederlands-Arabisch, -Deens en -Portugees) als niet-lexicografische (Corpus Gesproken Nederlands, NL-Translex, Databank Overheidsterminologie). Voor het maken van een woordenboek is de informatie uit het RBN in feite te rijk. De uitvoerende teams die het bestand ter beschikking gesteld kregen, konden zelf uit het RBN schrappen, toevoegen en wijzigen. Vaak reduceerden zij de macrostructuur van 45.000 naar 25.000 tot 35.000 items.

In de loop van de tijd kreeg de CLVV verschillende verzoeken binnen om een uitgangsbestand ter beschikking te stellen voor het samenstellen van vertaalwoordenboeken voor talen zoals het Rifberber, Georgisch en Slowaaks. Voor dergelijke kleinere talen zou een bestand van 25.000 items of meer te groot zijn. Er bleek dus duidelijk behoefte te zijn aan een klein referentiebestand Nederlands, een selectie van rond de 10.000 woorden, maar wel met dezelfde rijke microstructuur als het grote RBN.

## 1.2 Doel en werkwijze RBN-klein

Het doel van het RBN-klein is om op een beredeneerde en systematische wijze een semantisch verfijnd basislexicon samen te stellen. In eerste instantie is vanuit lexicografisch oogpunt te werk gegaan, dus met de bedoeling een lexicon samen te stellen dat als basis voor een woordenboek gebruikt kan worden. Een andere benadering zou zich op taaltechnologisch gebruik kunnen richten, bv. information retrieval of semantic tagging. Bij een dergelijke benadering wordt de semantische verdieping van het lexicon echter sterk bepaald door de beoogde toepassing van het lexicon. Hoe verfijnder de semantiek, hoe meer mogelijkheden het biedt. Zonder een specifieke toepassing voor ogen te hebben, is het moeilijk iets te zeggen over de benodigde graad en het formaat van de verdieping. Het gebruik voor een taaltechnologische toepassing wordt dan ook beschouwd als een neveneffect, maar is niet het primaire doel van het RBN-klein.

Het RBN-klein heeft een kern van ongeveer 7.600 woorden, gebaseerd op 'objectieve' frequentie en op consistente wijze beschreven. Op basis van die beschrijving/onderverdeling in semantische klassen kan vastgesteld worden welke klassen lacunes kennen of juist oververtegenwoordigd zijn en kan vervolgens de kernlaag worden uitgebreid of ingedikt. Door beheersing van de aldus tot stand gekomen structuur is het lexicon naar behoefte verder eenvoudig uit te breiden of in te krimpen. Het RBN-klein bevat door deze werkwijze niet zomaar de 10.000 meest frequente woorden van het Nederlands, maar wel

---

<sup>1</sup> Voor meer informatie over de CLVV zie 'Woordenboekenbeleid in Nederland en Vlaanderen' eindrapport van de CLVV, A. Hoorntje en W. Martin (eds.) (2004, Den Haag).

zeer frequente woorden Nederlands, op systematische wijze aangevuld met de woorden die noodzakelijk zijn om tot een (minimale) samenhangende communicatieve kern te komen. Als selectie van de meest frequente/meest nuttige woorden voor een taalleerder én een rijke microstructuur is het RBN-klein een goed uitgangspunt voor woordenboeken voor kleinere talen.

Naast een systematischer macrostructuur biedt een verfijndere semantische typologie ook meer greep op de informatie uit de microstructuur van het RBN. Leden van dezelfde subcategorie (bv. lichaamsdelen, beroepsnamen, fruit) kunnen op dezelfde wijze beschreven, vertaald en van collocaties voorzien worden.

Het RBN-klein is geïntegreerd in het RBN. De microstructuur van het RBN-klein is die van het RBN. De semantische typologie van het RBN is in het RBN-klein voor de nomina verder uitgebreid met de categorie *semantisch type*.

## 2. Samenstelling RBN- klein

### 2.1 Kernlaag RBN-klein

#### 2.1.1 Basis

De kernlaag van het RBN-klein wordt gevormd door ongeveer 7.600 woorden, gebaseerd deels op frequentie en deels op distributie over verschillende bronnen. De basis hiervoor wordt gevormd door een macroselectie die reeds bij de Vrije Universiteit bestond, bestemd was voor leerwoordenboeken (het zgn. VU-basisvocabularium) en geacht wordt een geschikte basis te vormen voor het RBN-klein. Deze selectie bestaat uit ca. 20.000 lemmata, de centrale woordenschat van de gemiddelde moedertaalspreker. Voor een selectie uit deze selectie is gebruik gemaakt van de volgende bronnen:

- Schrooten, W. en A. Vermeer, *Woorden in het basisonderwijs* (1994). Dit boek bevat een inventarisatie van het taalaanbod dat leerlingen in het basisonderwijs krijgen. Uit dat aanbod zijn bijna 2 miljoen woorden verzameld (wat ruim 25.000 verschillende lemmata opleverde). In het boek is een lijst opgenomen van de 15.000 woorden die meer dan één keer voorkomen in dit aanbod.
- frequente woorden uit het corpus Instituut voor Nederlandse Lexicologie (5 miljoen-corpus);
- bestand Vrije Universiteit (1990) (= VU-basisvocabularium);
- bestand functiewoorden KU Leuven.

Er werd destijds vanuit gegaan dat de uiteindelijke macro van het leerwoordenboek uit ca. 15.000 woorden zou bestaan. De selectie van 20.000 items bestaat daarom uit twee categorieën:

- 'woorden die zeker worden opgenomen' (6.925 stuks), bestaande uit:
  - woorden met een frequentie van > 4 op het corpus van 2 miljoen van Schrooten (6.806 stuks)
  - functiewoorden (niet adverbialen en voor zover nog niet opgenomen) (119 stuks).
- 'twijfelgevallen' (13.020 stuks). Deze zouden nog verder bekeken worden en ongeveer de helft geschrapt moeten worden.

Voor de kernlaag van het RBN-klein zijn de 6.925 meest nuttige, zeer frequente woorden van de VU-selectie voor het leerwoordenboek aangevuld met 1.083 woorden die frequent, maar niet in deze selectie opgenomen zijn. Deze toegevoegde woorden zijn geselecteerd op basis van een INL-frequentie hoger dan 103 op 5 miljoen. De woorden die opgenomen zijn in de basis van de kernlaag van het RBN-klein zijn dus opgenomen op grond van de volgende criteria:

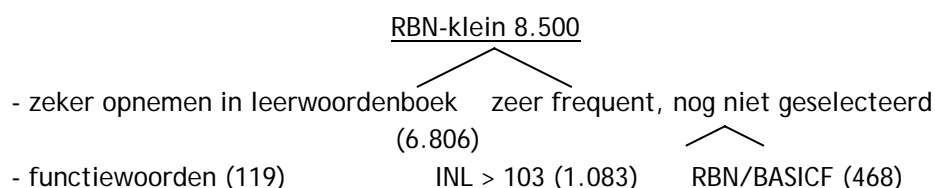
#### RBN-klein 8.000

Frequentie > 4 in Schrooten	6.806
Functiewoorden (niet adverbia)	119
Frequentie INL > 103	1.083
	<b>8.008</b>

#### 2.1.2 Uitbreiding van de basis

Al gauw bleek echter dat een aantal frequente woorden (bv. *slapen, fluiten, kruipen*) niet tot deze selectie behoorde. Minder frequent geachte woorden zoals *bolderkar, fakir, stencilen* en *benoorden* zaten er wel bij. Dit komt door het soort bronnen dat gebruikt is voor de selectie. De lijst van Schrooten is gebaseerd op materiaal voor de basisschool en bevat dus relatief meer woorden uit sprookjes, kinderboeken, schoolboeken en 'schooltaal' etc.

In het RBN wordt de frequentie van de woorden uit het VU-basisvocabularium aangegeven met de waarden BASIC (neutraal), BASICF (hoogfrequent) en RARE. Bij de latere uitbreiding van het RBN tot 45.000 woorden zijn de waarden NFREQ/NEUTRAL (neutraal; default) en RARE uit het 5 miljoen corpus van het INL gebruikt (Deze waarden vormen slechts een indicatie van de frequentie omdat tijdens de redactie geen vaste referentiebron beschikbaar was). De RBN-klein basislijst van 8.008 woorden is vergeleken met de woorden met frequentieaanduiding BASICF (2.289 stuks), waarna de in de basislijst ontbrekende woorden zijn toegevoegd. In totaal waren dat 468 stuks, zodat de uiteindelijke selectie voor de kernlaag van het RBN-klein uit 8.476 woorden bestaat. Uitgaande van een basisvocabulaire van 20.000 items kan de samenstelling als volgt worden voorgesteld:



De samenstelling van de kernlaag RBN-klein ingedeeld naar woordklasse:

	ja*	ja-extra**	BASICF	totaal
nomina	3.742	633	286	4.661
adjectiva	910	262	109	1.281
verba	1.479	188	73	1.740
	6.131	1.083	468	<b>7.682***</b>

\* de categorie 'zeker opnemen' uit het VU-basisvocabularium

- \*\* de categorie 'twijfelgevallen' uit het VU-basisvocabularium  
 \*\*\* adverbia, functiewoorden en interjecties zijn buiten beschouwing gelaten. Hiervan is later een nieuwe selectie gemaakt.

Voor het RBN-klein zijn de nomina, adjectiva, verba, functiewoorden, adverbia en interjecties gereduceerd. Voor de geografica zie 2.4.

## 2.2 Semantische subcategorisering

### 2.2.1 Nomina

De semantische subcategorisering van de nomina vindt plaats op basis van de indeling in semantische typen uit het RBN (zie verder documentatie RBN). Per semantisch type is een verdere onderverdeling gemaakt in subcategorieën. Dit geldt voornamelijk voor de concreta. Van de abstracta is slechts een klein gedeelte gesubcategoriseerd.

De gevormde subcategorieën zijn semantisch bepaalde, duidelijk afgebakende, vrij algemene betekenisgroepen. In totaal worden er binnen de nomina 108 (incl. de 'deel van-categorieën') semantische subcategorieën onderscheiden. Bij het samenstellen van de categorieën is zeer pragmatisch te werk gegaan. Om de verschillende categorieën te vormen is gekeken welke (duidelijke) klassen van woorden in de RBN-klein-selectie voorkwamen. Het was niet de bedoeling om honderden, kleine subcategorieën over te houden, dat zou voor het doel van de subcategorisering niet werkbaar zijn. In principe is een subcategorie een subcategorie van één bepaald semantisch type uit het RBN, al vindt er hier en daar enige overlap tussen semantische typen plaats (bv. *product* komt voor bij zowel ARTEFACT als SUBSTANCE). Voor de woorden die niet bij een categorie onder te brengen zijn is er een categorie *overig*. Bij deze restcategorie vindt verdere onderverdeling plaats door middel van de domeinlabels uit het RBN.

Allereerst is er per semantisch type uit het RBN een onderverdeling gemaakt tussen de woorden die een algemene en deze die een specifieke betekenis hebben. Die met een specifieke betekenis zijn daarna verder opgedeeld in subcategorieën, bv.:

ARTEFACT > artefact algemeen > *apparaat*

ARTEFACT > specifiek > apparaat/machine > *camera*.

Bij ANIMATE (= HUMAN en NONHUMAN) is daarnaast ook onderscheid gemaakt tussen groep/geen groep (default), bv.:

HUMAN > specifiek > actor > *klant*

HUMAN > specifiek > groep > *publiek*

NONHUMAN is verder onderverdeeld in flora en fauna:

NONHUMAN > specifiek > fauna > zoogdier > *schaap*

NONHUMAN > specifiek > fauna > groep > *kudde*

Bij ARTEFACT/CONCROTHER is onderscheid gemaakt tussen deel/geheel (default), bv.:

ARTEFACT > specifiek > vervoermiddel > *vrachtwagen*

ARTEFACT > specifiek > vervoermiddel > deel v. > *laadbak*

De semantische typen INSTITUTION, PLACE en DYNAMIC zijn niet in subcategorieën onderverdeeld; verdere specificatie vindt alleen door middel van domeinlabels plaats. Voor

de NONDYNAMIC worden enkele subcategorieën onderscheiden en verloopt de onderverdeling verder ook via de domeinlabels.

INSTITUTION > [pol] > *VN*

DYNAMIC > [med] > *abortus*

Domeinlabels kunnen ook bij de andere semantische typen voor een verdere onderverdeling van de subcategorieën zorgen, bv.:

HUMAN > beroepsnaam/beoefenaar > [techn] > *monteur*

Wel verschilt de rol die de domeinlabels bij de verdere onderverdeling spelen, nogal per subcategorie. In sommige gevallen vallen de subcategorieën min of meer samen met een bepaald domein en zijn de domeinlabels niet van belang, bv.: de subcategorie *lichaamsdeel* en [anatomie], *plant* en [plantkunde]/[biologie], *artefact mbt. (water)weg* en [verkeer] en [weg/waterb].

Een aantal subcategorieën komt voor bij twee of meer semantische typen. Zo staan bv. *kaas* (SUBSTANCE [cul]), *brood* (ARTEFACT [cul]) en *beleg* (CONCROTHER [cul]) onder *product*. Het verschil tussen de semantische typen ARTEFACT en CONCROTHER of tussen ARTEFACT en PLACE is in enkele gevallen niet zo duidelijk. Sport- en spelattributen, een vrij duidelijk afgebakende betekenisgroep, staan verdeeld over ARTEFACT en CONCROTHER, eventueel voorzien van hun eigen domeinlabel (*spel/ wielrennen/paardensport etc.*).

bv. *wandelstok* CONCROTHER en *tennisbal* ARTEFACT. In het RBN-klein zijn deze nu samengebracht onder de categorie *artefact mbt. sport/spel*.

bv: *eetkamer* (PLACE), *woonkamer* (ARTEFACT > PLACE) en *kinderkamer* (CONCROTHER) staan in het RBN-klein samen onder

*bouwwerk/ruimte (deel v.)*.

## 2.2.2 Opmerkingen/problemen subcategorisering

### 1. afbakening/begrenzing van semantisch veld

Gesloten klassen (bv. *windrichting*, *zintuig*, *maand*) en relatief gesloten klassen (bv. *meubelstuk*, *lichaamsdeel*, *vrucht*, *religie*) zijn eenvoudig te bepalen en te definiëren. Moeilijker wordt het om van allerlei soorten artefacten en concrothers niet teveel en niet te weinig bruikbare, samenhangende semantische groepen te vormen. Nu eenmaal niet ieder woord kan bij een bepaalde categorie ondergebracht worden. Soms zijn er ook meerdere mogelijkheden voor een subcategorie. bv. *helling* [schuin aflopende werf] is een *structuur*, maar ook een *artefact mbt. water/weg*. Een *wolk* [massa water- of ijsdeeltjes] is een *landschappelijk verschijnsel natuurlijk* maar past ook onder *hoeveelheid/vorm*. Categorieën zoals bv. *groente* en *drank* ontbreken in het RBN-klein, hoewel dit toch duidelijk afgebakende betekenisgroepen zijn. Groenten staan in het RBN onder NONHUMAN (de 'plantachtige' groente bv. *sla*) en CONCROTHER (de vruchten van planten, bv. *courgette*). Omwille van de overzichtelijkheid (niet teveel subgroepen of overlappende categorieën) is ervoor gekozen in het RBN-klein alleen de categorieën *plant* en *vrucht* (zie bijlage 1 voor definities van de categorieën) op te nemen. Om dezelfde reden is er geen aparte categorie dranken en zijn deze te vinden onder *product* en eventueel *natuurlijke stof*. Wellicht kunnen dergelijke subcategorieën in de toekomst nog toegevoegd worden.

### 2. structuur

De indeling in subcategorieën is niet hiërarchisch, maar vrij plat. Een 'tweede niveau' wordt alleen gevormd door de *deel van-* categorieën. De andere subcategorieën liggen op

hetzelfde niveau, terwijl ze in sommige gevallen in feite een subgroep van elkaar zijn. Bv. *muziekinstrument* is eigenlijk een subgroep van *instrument/gereedschap*, maar vormt door haar duidelijk afgebakende betekenis een aparte subcategorie van ARTEFACT en ligt op hetzelfde niveau als *instrument/gereedschap*.

Binnen een subcategorie kunnen ook verschillende niveaus worden onderscheiden waar verder geen rekening mee gehouden wordt, bv. *vader* en *stiefvader* zitten beide in de groep *familie/verwantschapsnaam*.

### 3. deel-geheel

Dit probleem doet zich vooral voor bij CONCROTHER. Een vertrek kan gezien worden als een ruimte op zich, maar ook als een deel van een grotere ruimte. Bij *bouwwerk/vertrek (deel v.)* zijn deel en geheel samengenomen. Bij de overige subcategorieën is een aparte categorie *deel van...* gecreëerd.

### 4. abstracta

Van de abstracta (NONDYNAMIC en DYNAMIC) is maar ongeveer 5% gesubcategoriseerd. Alleen enkele, duidelijk herkenbare groepen of gesloten klassen zijn eruit gehaald zoals kleuren, getallen, talen en windstreken. Een aantal abstracta paste onder gevormde subcategorieën van andere semantische typen, bv.: *kudde*  
NONDYNAMIC > *fauna groep*

## 2.2.3 Adjectiva

Voor de indeling van de adjectiva uit het RBN-klein voldoet de semantische typering uit het RBN (zie verder documentatie RBN).

## 2.2.4 Verba

Voor de indeling van de verba uit het RBN-klein is uitgegaan van de onderverdeling uit het RBN in Action-, State- en Process-verba en de gedeeltelijk uitgevoerde indeling in subklassen (zie verder documentatie RBN).

## 2.3 Paradigmatische uitbreiding

### 2.3.1 werkwijze

Op basis van de onderverdeling van het RBN-klein in semantische subcategorieën kan vastgesteld worden welke woorden er binnen een bepaalde subcategorie voorkomen en welke er ontbreken. De uitbreiding van de ca. 7.600 kernwoorden (zie 2.1.2) tot in totaal ca. 10.000 woorden is op basis van de semantische indeling tot stand gekomen. Opgemerkt dient te worden dat het bij het vaststellen en opvullen van hiaten niet gaat om woorden, maar om woordbetekenissen. Een woord is in het RBN-klein op grond van zijn betekenis ingedeeld in een bepaalde subcategorie. Bij het toevoegen van een woord aan een subcategorie wordt een woord in een bepaalde betekenis toegevoegd. Indien dit woord nog andere betekenissen heeft, worden deze ondergebracht bij de subcategorieën waartoe zij behoren en vullen zij op hun beurt die subcategorieën aan. Minder frequente betekenissen van een woord zijn dus niet geschrapt.

Bv. naast *sneeuwbal* [bal van samengepakte sneeuw] ARTEFACT > hoeveelheid/vorm] komen verder voor:

[borrel van jenever of brandewijn] ARTEFACT > product [cul]  
 [bepaalde heestersoort] NONHUMAN > plant]  
 [bepaalde methode van verkoop] NONDYNAMIC > [handel].

Het schrappen van niet-frequente betekenissen zou wellicht later nog uitgevoerd kunnen worden.

In totaal zijn toegevoegd aan de kernlaag van het RBN-klein:

nomina	977
adjectiva	167
verba	215
adverbia	352
functiewoorden	277
interjecties	47
	<b>2.035</b>

	kernlaag	uitbreiding	totaal
nomina	4661	977	5638
adjectiva	1281	167	1448
verba	1740	215	1955
adverbia	-	352	352
functiewoorden	-	277	277
interjecties	-	47	47
	<b>7.682</b>	<b>2.035</b>	<b>9.717</b>

Voor subcategorieën die gesloten klassen vormen zoals dagen van de week, letters van het alfabet, muzieknoten is eenvoudig vast te stellen of de categorie compleet is. De categorie kent een beperkt aantal leden en hiaten kunnen gemakkelijk aangevuld worden. Voor relatief gesloten categorieën (bv. *kledingstuk*, *muziekinstrument*, *familie/verwantschapsnaam*, *lichaamsdeel*, de diverse flora- en faunacategorieën) is het opsporen van gaten door de structuur en overzichtelijkheid van de categorie te doen. Binnen deze categorieën kunnen woordvelden of subgroepen onderscheiden worden (bv. *kledingstuk* -> regenkleding, nachtkleding, sportkleding, schoeisel, hoofddeksele, accessoires etc.) die steun bieden om de subcategorie op compleetheid te onderzoeken. Voor deze relatief gesloten klassen geldt dat de set waaruit gekozen kan worden in principe eindig is. Voor volledig open klassen zoals *kwalificerende persoonsnaam*, *actor*, *structuur*, *natuurlijke stof* en ook de klassen *overig* zoals *relatiennaam overig*, *artefact overig* is het uiteraard veel moeilijker om te bepalen of er grote hiaten zijn. Domeinlabels kunnen daarbij nog enige steun bieden.

Om gaten in de subcategorieën op te sporen en aan te vullen/uit te breiden is gebruik gemaakt van:

- de betekenisrelaties antonymie, hyponymie en in mindere mate synonymie:
  - aanvulling met een complementair antoniem (vnl. bij de adjectiva)
  - voorbeeld: *drijfnat\** - *kurkdroog* +

een sterretje (\*) na een woord wil zeggen dat het woord tot de kernlaag van RBN-klein behoort. Een plusje (+) betekent dat het woord ontbrak en daaraan is toegevoegd.

- aanvulling met een convers antoniem

voorbeeld: *aqueduct\** - *viaduct* +

voorbeeld: *schoonzus\** - *zwager* +

- aanvulling met een hyperoniem

voorbeeld: *seriemoordenaar\** - *moordenaar* +

- aanvulling met een hyponiem of co-hyponiem

voorbeeld: *havenstad\** - *hoofdstad* +, *wereldstad* +

voorbeeld: *studio\**, *flat\** - *appartement* +

Hier is o.a gebruik gemaakt van de definities uit het RBN (zoeken op het hyponiem bv. *stad*) en eigen inzicht.

- aanvulling met een synoniem:

Dit is zo weinig mogelijk gedaan. In feite zou je er voor kunnen kiezen om helemaal geen synoniemen in een basislexicon op te nemen, aangezien het de bedoeling is om de gebruiker een minimale woordenschat aan te bieden.

• de woordveldmethode

een woordveld is een verzameling lexemen die onderlinge samenhang vertoont binnen een bepaald ervaringsdomein. Om meer grip te krijgen op de (relatief) open klassen van de subcategorieën zijn, indien mogelijk, woordvelden binnen de subcategorieën gevormd.

Bv. subcategorie *kledingstuk*: woordvelden *hoofddeksel*, *schoeisel*; subcategorie *meubelstuk*: woordveld *zitmeubel* etc.

Voorbeeld: *kledingstuk/hoofddeksel*: [*kap, tulband, sluier, helm, pet, muts, puntmuts, steek, hoed*]\* - *hoofddoek* +

• de domeinlabels uit het RBN Van  
de domeinlabels is gebruik gemaakt voor verdere onderverdeling van subcategorieën en van de semantische typen die niet in subcategorieën onderverdeeld zijn (INSTITUTION, DYNAMIC, NON DYNAMIC).

• realisatie van de onderlinge paradigma's tussen nomina, adjectiva en de verba.  
Voorbeeld: *slaap-slaperig-slapen*; *macht-machtig* etc.

• lexicale functies:

- aanvulling met het instrument bij de actor/actie

voorbeeld: *zagen\** - *zaag* +

- aanvulling met het instituut als actor

voorbeeld: *verzekering\**, *verzekeraar\** - *verzekeringsmaatschappij* +

- aanvulling met het collectief bij de singularisering

voorbeeld: *hagelsteen\** - *hagel* +

• Multi-Taal Beeldwoordenboek

Het Beeldwoordenboek is een leermethode die gebruikt wordt om de woordenschat van taalleerders van het Nederlands uit te breiden in de allereerste fase die komt na de eerste centrale opvang en na het leren van de eerste Nederlandse taalbeginselen. De woorden en begrippen (alleen nomina) die in het beeldwoordenboek staan komen voort



uit de directe praktische omgeving van de volwassen mens en zijn ondergebracht in 26 thema's (o.a. het lichaam, gezondheid, inschrijfformulier, de klok, de kalender, kleding, schoeisel, dorp/gemeente, levensmiddelen, het weer, huis binnen, huis buiten). Deze thema's zijn bekeken en vergeleken met de RBN-klein-selectie en waar nodig aangevuld. Woorden die o.a. opgenomen zijn: *huisnummer*, *postcode*, *nationaliteit* (thema: inschrijfformulier), *bestek* (thema: huis binnen).

- eigen inzicht.

### 2.3.2 Opmerkingen en problemen

- begrenzing van woordveld en van uitbreiding  
Aangezien het RBN-klein een (klein) basislexicon is, mogen de categorieën niet teveel (te diep) uitgebreid worden, maar echte hiaten dienen wel opgevuld te worden. De beslissing om een woord wel/niet op te nemen is behoorlijk subjectief.  
Bv. de subcategorie *familie/verwantschapsnaam* bevat in de kernlaag de naaste familieleden (bv. *vader*, *moeder*, *zoon* etc.) en is, naast het aanvullen met de ontbrekende delen van paren (bv. *kleindochter-kleinzoon*), vervolgens uitgebreid met schoonfamilie (*zwager*, *schoonzus*, *schoonouders* etc.). Maar de uitbreiding van de categorie gaat bv. niet zover dat achterkleinkinderen en betovergrootouders ook opgenomen zijn.  
Van bv. de subcategorie *fauna* komen in de kernlaag als hyperoniemen van soortnamen o.a. *pelsdier*, *huisdier*, *roofdier*, *weekdier*, *knaagdier*, *zoogdier*, *insect* voor. Hieraan is toegevoegd *schaaldier* (als hyperoniem van *krab*, *kreeft*, *garnaal*, *pissebed* die in RBN-klein staan). Maar niet opgenomen zijn minder relevant geachte hyperoniemen als *trekdier*, *buideldier*, *muieldier*, *kuddedier*, *offerdier* en *troeteldier*.
- Aangezien het RBN-klein een onderdeel is van het RBN kunnen alleen woorden die in het RBN voorkomen, aan de selectie toegevoegd worden.
- Het uitbreiden van concreta is gemakkelijker dan van abstracta, omdat concreta nu eenmaal veel grijpbaarder zijn. Het gevaar hiervan is dat er met relatief meer concreta dan abstracta uitgebreid is. Binnen de concreta zijn wellicht de meest overzichtelijke en meest gestructureerde subcategorieën ook meer uitgebreid dan die categorieën waarop minder greep en overzicht te krijgen is. Bij een meer gestructureerde categorie zijn hiaten nu eenmaal beter te ontdekken.
- Het uitbreiden van nomina is eenvoudiger dan van adjectiva en verba. Binnen de adjectiva is nog gekeken naar antoniemen. Voor de verba en gedeeltelijk ook voor de adjectiva geldt dat de uitbreiding plaats vond op basis van de aanwezige nomina (bv. werkwoord toevoegen op basis van de actor: *verraad\**, *verrader\**- *verraden* +)

### 2.4 Geografica

De geografica die in de kernlaag van RBN-klein voorkomen zijn, zijn vrij willekeurig. Het betreft de categorieën *inwoner/volk*, *gebergte*, *rivier*, *stad*, *land*, *provincie* en *taal*. Er is voor gekozen om in het RBN-klein op te nemen:

- de eigen en omringende landen van het Nederlandse taalgebied, aangevuld met de grotere EU-landen en enkele grote landen in de wereld;
- de talen die bij deze landen horen en een aantal 'speciale' talen (bv. gebarentaal)
- de bijbehorende inwoner en adjectief.

Voor meer geografica kunnen gebruikers van het RBN-klein een beroep doen op de geografica uit het RBN. Het RBN bevat twee bestanden van geografische namen (nomina en adjectiva), een basisbestand van 1.900 items en groter bestand van 3.320 items. Hieruit kan een eigen selectie gemaakt worden.