Automatische indeling van Nederlandse woorden op basis van etymologische filters

VINCENT J. VAN HEUVEN, ANNEKE H. NEIJT
EN MAARTEN HIJZELENDOORN

1. Inleiding

Talen veranderen voortdurend. Een belangrijke oorzaak van taalverandering is beïnvloeding vanuit andere talen. Zo heeft de Nederlandse woordenschat zich in de loop van de eeuwen verrijkt (sommigen beweren verarmd) met woorden die afkomstig zijn uit andere talen. Er zijn woorden binnengekomen vanuit het Grieks en Latijn, soms rechtstreeks in de klassieke fase, soms getrapt via het Middeleeuws Latijn. Later is grootscheeps ontleend aan het Frans, en nog weer later aan het Engels, terwijl in alle perioden ook ontleend is, maar dan op kleinere schaal, aan nog weer andere talen zoals het Arabisch, Hebreeuws, Maleis, etc. Op het moment dat woorden voor het eerst binnenkomen in het Nederlands zullen ze in meerdere of mindere mate afwijken in uitspraak en structuur van wat in onze taal gebruikelijk is. De importwoorden ondergaan na verloop van tijd veranderingen waardoor ze zich allengs beter gaan voegen in het gareel van het Nederlands. In grove lijnen kunnen we zeggen dat woorden zich ingrijpender hebben aangepast aan de structuur van het Nederlands naar mate ze langer geleden in onze taal zijn binnengedrongen.

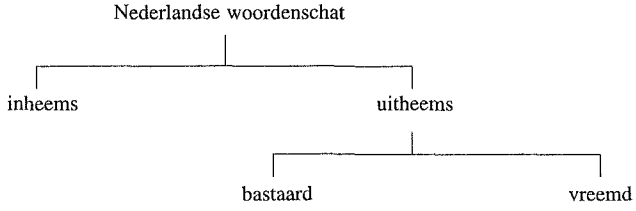
Onze woordenschat wordt traditioneel ingedeeld in drie categorieën van aangepastheid aan de Nederlandse norm.

1. Inheems: dit zijn de woorden van Germaanse oorsprong, die van oudsher tot onze taal behoren, de z.g. erfwoordenschat (b.v. *man, vrouw, kind*). Enkele uitzonderingen (*au/ou, ei/ij*) daargelaten worden deze woorden klankzuiver gespeld, d.w.z. dat hun schrijfwijze volledig voorspelbaar is, gegeven hun uitspraak.
2. Bastaard: dit zijn woorden van klassieke oorsprong, die inmiddels ingrijpend zijn aangepast aan het inheemse systeem (b.v. *consequent, extract, apathie*). Ze bevatten alleen nog maar inheemse klanken, maar kunnen uitheems aandoen door afwijkende klankcombinaties, b.v. *ps* aan het woordbegin zoals in *psycholoog*. In hun spelling verraadt dit soort woorden dikwijls nog zijn uitheemse herkomst door het gebruik van exotische grafieën, b.v. (*c, qu, x, th*).
3. Vreemd: dit zijn recente ontleningen, die zich (nog) niet of maar onvolledig hebben aangepast aan het inheemse systeem. Deze woorden bevatten dik-

Spektator, 23-4, 279-291

wijls on-Nederlandse klanken (b.v. *garage, douche, goal*). Vreemde woorden behouden meestal de spelling van de taal waaruit ze afkomstig zijn.

Deze indeling kan het best worden opgevat als een paar hiërarchisch geordende tweedelingen, volgens onderstaand schema:



Merk op dat, hoewel taalkundigen bij discussies over de mate van aangepastheid van leenwoorden aan de inheemse norm vooral acht slaan op de taalkundige (d.w.z. hoorbare) eigenschappen van woorden, men in spellingkwesties vooral let op de schrijfwijze van woorden. Doordat de Nederlandse spelling er traditioneel aan hecht de herkomst van leenwoorden herkenbaar te houden is er een (asymmetrische) relatie tussen spelling en etymologische status: woorden met ongebruikelijke grafieën zijn met zekerheid uitheems, maar lang niet alle uitheemse woorden bevatten exotische grafieën. In dit verband is het ons altijd als principieel onjuist voorgekomen om regels voor spellingwijziging te formuleren die inhaken op de bestaande of oudere schrijfpraktijk. Een regel als "schrijf woorden met <qu> voortaan met <kw>" heeft taalkundig geen status en is in het aanvankelijk schrijfonderwijs zinloos: de jonge generatie spellers weet niet hoe de oude spelling eruit zag. In de voorstellen van de jongste spellingcommissie wordt dit type formuleringen dan ook nergens gebruikt maar hebben spellingregels uitsluitend betrekking op hoorbare eigenschappen van woorden (Neijt & Zuidema 1994: 23).

De eerste twee auteurs van dit artikel waren lid van deze jongste spellingcommissie, die volgens zijn instellingsbeschikking (onder andere) van het Comité van Ministers van de Nederlands Taalunie de opdracht had gekregen een vergaand consequente spellingregeling te ontwerpen voor alleen de bastaardwoorden (Neijt & Zuidema 1994: 19). Bij implicatie werd de commissie niet geacht zich uit te spreken over de spelling van de vreemde woorden, noch over die van de inheemse woorden. Het is dus voor de werkwijze van de spellingcommissie van wezensbelang geweest om te komen tot een scherpe afbakening van de drie typen woorden binnen de Nederlandse woordenschat. Daarbij heeft de commissie niet volstaan met de gebruikelijke intuïtieve indeling, maar juist willen komen tot een automatisch toepasbaar stelsel van criteria waarmee deze indeling uitgevoerd kan worden. Over deze pogingen gaat dit artikel.

Als het mogelijk zou zijn om de driedeling van de Nederlandse woordenschat in inheemse, bastaard- en vreemde woorden met objectief toepasbare criteria tot stand te brengen, dan kan een aantal wensen in vervulling gaan. We kunnen dan

op basis van de groep inheemse woorden vaststellen wat de klankzuivere spelling is. Van de groep van de bastaardwoorden kunnen we proberen de eigen spellingsystematiek te doorgronden en in regels te vangen. Als alternatief kunnen we overwegen de bastaardwoorden onder het spellingregime van de inheemse woorden te brengen. Woorden, ten slotte, die volgens de objectieve criteria als vreemd moeten worden aangemerkt, behouden hun buitenlandse spelling. Recente ontleningen die niet als vreemd ontmaskerd kunnen worden door onze criteria, en dus ook niet als uitheems ervaren zullen worden door Nederlandse taalgebruikers, zouden dan – naar wij aannemen – zonder bezwaar omgespeld kunnen worden volgens de inheemse spellingsystematiek.

2. Vraagstelling en plan van aanpak

De eerste vraag die we met ons onderzoek willen beantwoorden is:

> *kunnen we, zonder acht te slaan op het spellingbeeld, een formeel onderscheid maken tussen de woordcategorieën inheems, bastaard en vreemd?*

Het gaat hier dus om de vraag hoe goed uitheemse woorden zich hebben aangepast aan het inheemse taalsysteem. De uiteindelijke juistheid van de indeling van een woord kan alleen worden bepaald door onderzoek te doen naar de intuïties van de Nederlandse taalgemeenschap, door te kijken naar de mate van vreemdheid die Nederlanders desgevraagd toekennen aan een woord. Dit intuïtie-onderzoek is – voor zover we hebben kunnen nagaan – nooit uitgevoerd, en zal in de praktijk ook onuitvoerbaar zijn, al was het alleen al omdat we dan vreemdheidsoordelen zouden moeten inwinnen over vele duizenden woorden. In onze aanpak leek het beter eerst een stelsel van regels te ontwerpen waarmee ieder willekeurig woord zou kunnen worden ingedeeld in de categorieën inheems, bastaard of vreemd, en daarna steekproefsgewijs na te gaan in hoeverre de indeling klopt. Wij hebben er bovendien voorshands van afgezien om de indeling te toetsen aan de intuïtie van naïeve taalgebruikers, voornamelijk omdat door de bewerkelijkheid van dit soort onderzoek maar een heel kleine steekproef van beslissingen getoetst zou kunnen worden. In plaats daarvan hebben we gemeend er beter aan te doen de indeling volgens onze regels te vergelijken met de werkelijke herkomst van woorden, zoals we die kunnen vinden in een etymologisch woordenboek. We gaan er dan van uit dat alle woorden van klassiek Griekse of Latijnse herkomst bastaardwoorden zijn, en alle recentere ontleningen (uit Frans of Engels) vreemd. Zo komen we op onze tweede vraag:

> *Welke overeenkomst is er tussen de indeling volgens onze regels en de werkelijke etymologische herkomst van de woorden? Hoeveel woorden zijn er bij voorbeeld die volgens onze regels volkomen inheems zijn maar in werkelijkheid uitheems?*

Omdat we de deugdelijkheid van onze criteria willen kunnen toetsen aan een zo groot mogelijk deel van het lexicon, en om daarbij objectieve toepasbaarheid te

waarborgen, is besloten om het stelsel van criteria te formaliseren en te implementeren in de vorm van een computerprogramma.

Bij het vaststellen van criteria om woorden te ordenen op een schaal van inheems naar vreemd, verwijzen we alleen naar eigenschappen van de talige structuur van een woord zoals dat in het huidige Nederlands wordt uitgesproken, en niet naar de spelling. Meer in het bijzonder letten we alleen op de klankvorm van woorden en op hun buigingsvormen. Onze criteria zijn bovendien slechts van toepassing op niet-samengestelde woorden. Samenstellingen moeten eerst worden opgesplitst in hun kleinste betekenisdragende woorddelen (morfemen), omdat anders geen eenduidige status bepaald kan worden. In b.v. het compositum *spraaksynthese* is het eerste morfeem *spraak* van inheemse oorsprong terwijl het tweede lid *synthese*, afkomstig uit het Grieks, de bastaardstatus heeft. Het is dan niet zinvol om een tussenliggende herkomstcategorie toe te kennen (halverwege inheems en bastaard), of om de herkomststatus volledig te laten afhangen van hetzij het eerste lid (prosodisch hoofd) hetzij het tweede lid (morfosyntactisch hoofd) volgens een of andere left of right hand head rule, zoals die in andere delen van de grammatica wel gebruikt worden (resp. bij klemtoontoekenning en woordsoortbenoeming, cf. Trommelen & Zonneveld 1986: 149).

Overigens is het taaltechnologisch mogelijk om in de grote meerderheid (ca. 90%) van de voorkomende gevallen woorden automatisch op te splitsen in hun samenstellende morfemen (Heemskerk & van Heuven, 1993), zodat ook voor de praktijk de analyse van samenstellingen geen onoverkomelijk probleem hoeft te zijn.

3. De criteria

De toelatingscriteria waaraan een woord moet voldoen om erkend te worden als inheems, vatten we op als een filter: inheemse woorden passeren het filter ongehinderd, terwijl woorden die op een of andere grond uitheems zijn door het filter uitgezeefd worden. In onze aanpak onderscheiden we in feite vijf van zulke filters, waarbij ieder filter eigenschappen toetst op een specifiek niveau in de talige structuur van een woord. We bespreken deze niveaus eerst globaal; daarna wordt per niveau in een aparte paragraaf een toelichting gegeven.

1. per foneem (klankfilter): bevat het woord uitsluitend inheemse fonemen?
2. per lettergreep (syllabefilter): is de opeenvolging van fonemen binnen iedere syllabe legaal?
3. per woord (syllabe-opeenvolgingfilter): is de opeenvolging van syllaben legaal?
4. per woord (klemtoonfilter): ligt de klemtoon op de juiste syllabe?
5. per woord (flectiefilter): is de flectie inheems?

De algemene gedachte is dan dat een woord uitheems is zodra het ook maar aan één filter niet voldoet. Daarna moet worden vastgesteld welke schendingen van

welke criteria kunnen worden toegestaan om een woord toch nog de bastaard-status te geven. Woorden die een (of meer) van de ernstigere criteria schenden, worden afgewezen als vreemd. Dit stuk zal vooral gaan over het vaststellen van de grens tussen inheems en uitheems (d.w.z. bastaard en vreemd tezamen). Aan het einde van het artikel zullen we gelegenheid hebben kort in te gaan op mogelijkheden ook een verdere tweedeling aan te brengen in de groep van uitheemse woorden in termen van bastaard en vreemd.

3.1. *Klankfilter*

Om vast te stellen of een woord alleen maar inheemse fonemen bevat, stellen we ons de uitspraak voor van dat woord door een Algemeen Beschaafd spreker van het Nederlands. Deze uitspraak wordt op papier vastgelegd in de vorm van een globale fonematische transcriptie. Daarbij maken we gebruik van uitsluitend de fonemen die voorkomen in de inventaris van de Celex databank. Woorden die niet kunnen worden weergegeven met uitsluitend de Celex-fonemen, zijn per definitie vreemd. Binnen de Celex-foneeminventaris bevindt zich echter ook een aantal uitheemse klanken (de z.g. leenfonemen of marginale fonemen volgens Cohen, Ebeling, Fokkema & van Holk 1978: 25-27), zoals aangegeven in tabel 1:

Tabel 1: Celex inventaris van consonanten en vocalen (foneemsymbolen plus voorbeeldwoorden) benodigd voor een klankcodering van Nederlandse woorden, uitgesplitst naar inheemse en uitheemse fonemen.

consonanten				vocalen			
	inheems		uitheems		inheems		uitheems
p	*p*ak			A	p*a*k	A~	restaur*ant*
b	bak			E	p*e*t	E~	enf*in*
t	*t*ak			I	p*i*t		
d	*d*ak			O	tr*o*m	O~	bonb*on*
k	kat	g	*g*oal	U	d*u*n	U~	parf*um*
f	*f*ee			@		de	
v	*v*ee			a:	pr*aa*t	A:	h*a*lf-time
s	*s*op	S	dou*ch*e	e:	m*ee*t	E:	s*e*rre
z	*z*ou	Z	gara*g*e	i.	r*ie*p	i:	analyse
x	la*ch*			o:	r*oo*m	O:	z*o*ne
G	*g*eel			u.	r*oe*m	u:	r*ou*ge
m	*m*at			y.	f*uu*t	y:	centrif*u*ge
n	*n*at			&:	r*eu*s	U:	fre*u*le
N	*zing*			EI	r*ij*p		
l	*l*aat			AU	k*ou*		
r	*r*at			UI	l*ui*s		
j	*j*as						
w	*w*ie						
h	*h*ard						

Woorden die in hun transcriptie één of meer klanksymbolen bevatten uit de kolommen onder "uitheems" in de symbolenlijst zijn uitheems (en zelfs vreemd).

3.2. *Syllabefilter*

Fonemen worden in de taal samengenomen tot syllaben. Niet iedere willekeurige combinatie van klanken levert een correcte syllabe op. De mogelijke klankopeenvolgingen in een syllabe worden verantwoord door z.g.n. fonotaktische regels. Die regels zijn van taal tot taal verschillend. Een inheems Nederlands woord mag bij voorbeeld niet beginnen met de klankopeenvolging /skr/ terwijl dat in het Engels wel mag (*scream, scrabble*). Het syllabefilter dat we, grotendeels aan de hand van beschikbare overzichten op dit gebied (zie b.v. Neijt, 1991 en verwijzingen aldaar) hebben opgesteld, bevat een opsomming van in beginsel alle fonotaktische regels, voor zover van toepassing op de Nederlandse erfwoordenschat. Daarbij is het gemakkelijker om de beperkingen op toegestane foneemopeenvolgingen apart te specificeren voor de consonanten die aan de vocaal voorafgaan (de onset) en voor de toelaatbare combinaties van vocalen en daaropvolgende slotconsonanten (het rijmdeel van de syllabe). Bovendien maken we onderscheid tussen syllaben die in het midden van een woord kunnen voorkomen (mediale syllaben), tegenover syllaben die alleen aan het begin of het eind van woorden kunnen staan (marginale syllaben). De onset van een beginsyllabe en de coda (alle consonanten na de vocaal) van een eindsyllabe vertonen grotere variëteit dan die van woordmediale syllaben. In dit verband tellen voor- en achtervoegsels als marginale syllaben: de coda van een prefix (b.v. *ont-* als in *ontzien*) en de onset van een suffix (b.v. *-ster* als in *bedriegster*) vertonen dezelfde ruimere mogelijkheden als syllaben aan resp. het woordeinde en -begin. Dezelfde woordmarginale status hebben we gegeven aan onsets die in een woord voorkomen na de (kwasi-)prefixen *ge-, be-, ver-, te-, je-, me-, de-* en aan rijmen voor de (kwasi-)suffixen *-de* en *-te*.

We hebben ervan afgezien om bij de formulering van onze fonotaktische beperkingen op inheemse syllaben gebruik te maken van de formele middelen van de generatieve fonologie, zoals het specificeren van natuurlijke klassen van klanken met behulp van distinctieve kenmerken. In plaats daarvan hebben we onze regels in normale taal uitgeschreven in de vorm van min of meer systematisch geordende lijsten, zoals hieronder weergegeven. Volledigheidshalve verstaan we onder korte vocalen de bovenste zes klinkers uit de derde kolom van tabel 1, de sjwa is de zesde klinker, lange vocalen zijn de laatste tien en diftongen de laatste drie.

Woordmediale onset en rijm

Inheemse onsets zijn:

(1) a p, b, t, d, k, f, v, s, z, x, G, m, n, l, r, j, w (d w z alle Nederlandse consonanten behalve h en N)
b st
c NIL (uitsluitend in suffixen)

Inheemse rijmen zijn:

(2) a Korte vocaal binnen dezelfde syllabe gevolgd door p, b, t, d, k, f, s, x, G, m, n, N, r, l
b Sjwa, lange vocaal of diftong
c Sjwa gevolgd door r
d Lange vocaal gevolgd door p, b, t, d, k, f, s, x, G, m, n, r, l

Woordmarginale onsets en rijmen

Aan het begin van een inheems woord zijn de volgende onsets toegestaan:

(3) a NIL, p, b, t, d, k, f, v, s, z, G, m, n, l, r, j, w, h (d w z niets plus de combinaties die genoemd zijn onder 1a, met h in de plaats van x)
b pl, pr, bl, br, tr, tw, dr, dw, kn, kl, kr, kw, fn, fl, fr, vl, vr, sp, st, sx, sm, sn, sl, sj, zw, Gn, Gl, Gr
c spr, str, sxr, spl

Een rijm aan het woordeinde mag bevatten:

(4) a Korte vocaal gevolgd door
- p, b, t, d, k, f, s, x, G, m, n, N, l, r (d w z alle consonanten behalve /v, z, j, w, h/)
- ps, pt, bs, bt, ts, ds, ks, kt, fs, ft, st, xs, xt, ms, mt, ns, nt, Ns, Nt, ls, lt, rs, rt (d w z alle consonanten behalve /v, z, G, j, w, h/, gevolgd door /s, t/)
- mp, lp, rp, md, nd, ld, rd, Nk, lk, rk, lv, rv, lG, rG, mz, nz, lz, rz, lx, rx, lm, rm, ln, rn
- rts, xts, lts, nts, Nks, lps, rft, rft, rkt, mpt, rps, rst, lst, Nst, lft, rfst

b Lange vocaal gevolgd door
- NIL, p, t, d, k, v, s, z, G, m, n, l, r, j, w
- ps, pt, ts, ks, kt, st, ms, mt, ns, nt, ls, lt, rs, rt, js, jt, ws, wt
- Gd, md, nd, ld, rd, rz, rn, rts, tst
- /ɪ / gevolgd door lp, rv, rp

c Diftongen gevolgd door
- NIL, p, b, t, d, k, f, v, s, z, G, m, n, l
- st

d Sjwa gevolgd door
- NIL, p, k, t, s, x, G, m, n, l, r
- nd, rd, ld, nt, rt, lt, ns, rs, ls

e Lange vocaal plus heterorgane semivocaal /a j, e w, ɪ w, o j, u j, y w/ optioneel gevolgd door /s, t, st, ts/

Een syllabe is uitheems zodra de opeenvolging van fonemen die erin voorkomt niet gedekt wordt door een van de hierboven genoemde mogelijkheden.

3.3. *Syllabe-opeenvolgingfilter*

In het syllabe-opeenvolgingfilter wordt gecontroleerd of twee groepen van eigenschappen voldoen aan de inheemse norm. De eerste groep heeft te maken met de gewichtverdeling tussen de lettergrepen. Inheemse woorden bevatten in beginsel slechts één syllabe met een volle vocaal (d.w.z. een klinker anders dan sjwa), die we de nucleaire syllabe noemen. De nucleaire syllabe mag vooraf worden gegaan door maximaal één ultralichte syllabe (d.w.z. een lettergreep die een sjwa bevat) en worden gevolgd door maximaal twee van zulke ultralichte lettergrepen. De syllaben *-ing* (*haring*) en *-uw(e)* (*zenuw, weduwe*) tellen in dit verband als ultralicht. Het volgende deelfilter somt de besproken mogelijkheden op:

Gewichtverdeling binnen woord

(5) a. Niet meer dan één syllabe bevat een volle vocaal of diftong
b. De volle syllabe mag worden voorafgegaan door
- G@, b@, v@r, t@, j@, m@, d@

c. De volle syllabe mag worden gevolgd door maximaal één consonant plus
- @, @r, @n, @l, @m, @x, @k, @ld, @nd, @rd, @ns, @rt, @rs, @nt, @r@n, @G@, @l@x, @l@k, @r@k
- EI, y.w, y.w@, IN, a:r, Ond, And, OG, Ik, @rIk

De tweede groep beperkingen op syllabe-opeenvolgingen betreft de mogelijke onsets van een ultralichte lettergreep na specifieke rijmen van de nucleaire syllabe (Kager & Zonneveld 1986: 208ff.). Het rijm van de nucleaire syllabe bevat twee segmenten (een lange vocaal of diftong, dan wel een korte vocaal plus consonant), optioneel gevolgd door een willekeurige consonant, die op zijn beurt weer gevolgd mag worden door een coronale obstruent (/s, t/) of door een sjwa plus een optionele liquida (/m, n, N, l, r/). De aaneengesloten reeks consonanten van de nucleaire syllabe en de volgende ultralichte syllabe vertoont een aflopende sonorantiegraad. Hiermee wordt onder meer verantwoord dat woorden als *schamper* en *dorpel* legaal zijn terwijl dat niet het geval is met **schapmer* en **doprel*. Ten slotte verbieden we een stemhebbende fricatief (/v, z/) onmiddellijk na een korte vocaal: **puzzel* met een korte /U/ is duidelijk een uitheems woord; het zou inheems kunnen worden als de uitspraak met een lange /y./ (*puzel*) gangbaar wordt. Schema (6) vat de mogelijkheden samen:

Nucleaire rijm plus volgende onset

(6) Alleen de volgende nucleaire rijm-onsetcombinaties zijn toegestaan:
a. Lange vocaal gevolgd door
- iedere consonant behalve /f, s/
- st, nt, rt, Gd, nd, ld, rd, jk, rz, rn, lj, rst, nst

b. Diftong gevolgd door
- iedere consonant behalve /s/
- st, nd

c. Korte vocaal gevolgd door
   - sp, ps, sk, ks, st, ts, ft, xt, nt, lt, rt, mp, rp, mb, rb, nd, ld, rd, Nk, rk, lv, rv, nG, nk, ns, ls, rs, nz, lz, rz, lG, rG, rm, ln, nj, rw, kst, mst, nst, lst, rst, Nst

d. Sjwa plus nul of meer consonanten

3.4. *Klemtoonfilter*

Inheemse woorden dragen de klemtoon op de meest linkse volle vocaal of diftong. Ligt de klemtoon op een andere syllabe, dan is het woord uitheems.

3.5. *Flectiefilter*

Onze filters leggen slechts twee beperkingen op aan de verbuiging van inheemse woorden. In het eerste geval wordt geëist dat nomina een regelmatig meervoud vormen; in het tweede geval wordt verlangd dat adjectieven een buigingsvorm op *-e* bezitten. De beperkingen zijn hieronder geformuleerd:

(7) Inheemse nomina vormen hun meervoud als volgt:
   a. geen meervoud (bij collectiva)
   b. /-n/ en/of /-s/ na ultralichte syllabe
   c. /-@n/ in alle andere gevallen

   Inheemse adjectieven (attributief bij meervoudig nomen):
   a. eindigen op /@n/ (*houten, stalen*)
   b. nemen een buigings /-@/ (*grote, kleine*)

Wanneer een nomen (of adjectief) zich niet gedraagt volgens (7) en/of daarnaast nog een andere meervoudsvorm kent (b.v. *musea* naast *museums*), dan is het betreffende woord uitheems.

We merken hierbij op dat vele aanvullende morfologische berperkingen geformuleerd kunnen worden die alleen gelden voor etymologisch gemotiveerde delen van de Nederlandse woordenschat. Zo kunnen inheemse stammorfemen in de regel alleen gecombineerd worden met inheemse derivationele suffixen, terwijl woorden van Grieks-Latijnse herkomst (bij uitstek de bastaardwoorden) een sterke neiging vertonen alleen klassieke derivationele suffixen toe te staan. Van inheemse adjectiva leiden we nomina af met *-heid*, van klassieke adjectiva met *-iteit*. Voorshands hebben we ervan afgezien deze complicaties in het filtersysteem te betrekken.

4. Kwantitatieve evaluatie en foutanalyse

Om te kunnen nagaan hoe goed deze filters in staat zijn inheemse en uitheemse woorden uit elkaar te houden, zijn alle genoemde beperkingen geïmplementeerd in een Prolog programma en vervolgens getest op het RUL-morfeemlexicon. Dit lexicon is in de jaren 1986-1990 ontwikkeld als onderdeel van een voorleesmachine (van Heuven & Pols, 1993). Het bevat ongeveer 12.500 ongelede Nederlandse morfemen en 4.000 complexe vormen waarvan de beteke-

nis niet bepaald is door de som van de betekenissen van de samenstellende delen. Iedere vorm is voorzien van een abstracte uitspraakcodering, met inbegrip van syllabegrenzen en klemtoonpositie. Ook zijn de vormen voorzien van hun morfologische valenties, codes die aangeven met welke andere morfemen zij verbindingen kunnen aangaan. In deze valentie-informatie ligt ook de flectie van nomina en adjectieven gecodeerd. Wat niet in het RUL-morfeemlexicon was aangegeven, is de status van de vormen in termen van inheems/bastaard/vreemd zoals die wordt aangevoeld door de Nederlandse taalgemeenschap. Zoals boven uiteengezet is deze informatie vooralsnog niet in te brengen. Wel hebben we informatie over de werkelijke herkomst van de vormen ingevoerd, door deze af te leiden uit een computerleesbare versie van Van Dale's Etymologisch Woordenboek[1] (van der Veen & van der Sijs, 1990). De etymologiecode in het morfeemlexicon is vooralsnog binair, en specificeert de herkomst van een vorm als inheems of uitheems.

De doorsnee van het morfeemlexicon en het etymologisch woordenboek leverde een verzameling van een kleine 5.000 morfemen op. Van iedere vorm is vervolgens automatisch vastgesteld of deze volgens onze filters inheems of uitheems is. Deze uitkomst is vergeleken met de opgeslagen etymologiecode van dezelfde vorm. Er zijn dan vier verschillende resultaten mogelijk, die in tabel 2 zijn aangegeven tezamen met de vastgestelde aantallen.

Tabel 2: Aantallen en percentages inheems/uitheems-beslissingen als functie van de werkelijke etymologie van de testmorfeemverzameling.

	beslissing algoritme		
werkelijke herkomst	inheems	uitheems	totaal
inheems	terechte acceptatie	onterechte verwerping	
	2243	220	2463
	91%	9%	100%
uitheems	onterechte acceptatie	terechte acceptatie	
	308	2182	2490
	12%	88%	100%

De resultaten laten zien dat onze filters over de hele linie ongeveer 90% van de woorden in de correcte etymologische categorie indelen. De twee mogelijke soorten fout komen in ruwweg gelijke mate voor: werkelijk inheemse vormen worden ten onrechte als uitheems afgewezen in 9% van de gevallen; werkelijk uitheemse vormen worden ten onrechte als inheems geaccepteerd in 12%. We bespreken de twee fouttypen achtereenvolgens.

Onterechte verwerping. Onze filters waren zo opgezet dat ieder inheems woord erdoor geaccepteerd moest worden. Voorshands is het dus onbegrijpelijk dat zich zoveel onterechte verwerpingen voordoen. Bij nadere inspectie van de

gegevens blijkt een aantal oorzaken aan te wijzen voor de zwakke prestaties van het filter. Er staan vrij veel vormen als ongeleed in het lexicon die in werkelijkheid complex zijn (b.v. *veertien, alledaags, aardappel*). Daarnaast zijn er een aantal vormen die naar de stand van vandaag weliswaar als ongeleed moeten worden aangemerkt, maar die dat vroeger niet waren (zgn. historisch gelede woorden, b.v. *oorlog, middag, twaalf*). Zulke complexe woorden bevatten illegale klankopeenvolgingen op syllabegrenzen en/of bevatten meer dan één volle vocaal, met het gevolg dat ze verworpen worden. In totaal deden zich 69 gevallen van dit soort fout voor in ons materiaal. Wanneer we dit aantal in mindering brengen op de onterechte verwerpingen dan daalt het foutpercentage daar tot 6. Daarnaast blijkt het morfeemlexicon nog steeds codeerfouten te bevatten, met name in de foneemtranscriptie, waardoor illegale klank(opeenvolging)en ontstaan. Ten slotte is een aantal fouten ingeslopen bij het achterhalen van de werkelijke herkomst van de woorden in het etymologisch woordenboek. Wanneer deze ongerechtigheden in het morfeemlexicon verbeterd zijn, zou het aantal onterechte verwerpingen tot 0 teruggebracht moeten kunnen worden.

Onterechte acceptaties. Twaalf procent van de etymologisch uitheemse woorden werd door het filter geaccepteerd als inheems. Dit zijn dus de uitheemse woorden die op formele synchrone kenmerken niet (meer) te onderscheiden zijn van de inheemse woordenschat, vaak – maar niet altijd – als gevolg van ingrijpende aanpassingen. Hieronder valt een aantal monosyllabische klassieke woorden (b.v. *poort, straat, vorm, som*) en een aantal polysyllabische woorden met één volle vocaal (b.v. *simpel, somber, luister*). De taalgemeenschap zal deze woorden hoogst waarschijnlijk als inheems ervaren. In dit verband is het veelzeggend dat 268 van de 308 (87%) gevallen in de huidige voorkeurspelling al volgens de inheemse spellingconventies wordt geschreven. Van de exotisch gespelde 40 zijn er opvallend veel homofoon met een inheems woord, b.v.: *ether – eter, lynx – links, pact – pakt.* Het ligt in de rede om alle woorden in deze categorie te gaan spellen volgens de inheemse conventies.

5. Vreemde woorden

Het algoritme is ongetwijfeld geschikt te maken om binnen de categorie uitheemse woorden de bastaardwoorden van de vreemde woorden te scheiden. Bastaardwoorden passeren altijd het klankfilter; woorden met een vreemd foneem worden door het foneemfilter onmiddellijk ontmaskerd als vreemd. Voorts is een krachtig selectiemiddel voor vreemde woorden het flectiefilter. Vreemde nomina vormen hun meervoud altijd met /-s/. Voor zover ons flectiefilter in die gevallen een uitgang /-@n/ voorschrijft, worden vreemde woorden op flectie ontmaskerd (b.v. *club* is vreemd wegens /klUps/ i.p.v. /klUb@n/).

Op dit ogenblik zijn de ideeën over verdere afbakening van bastaard tegenover vreemd nog tamelijk onuitgewerkt. Evenmin zijn de prestaties van dit deel

van het algoritme kwantitatief geëvalueerd, met als een van de belangrijkste redenen dat het RUL-morfeemlexicon binnen de categorie uitheems geen nadere etymologische uitsplitsingen maakt. Overigens speelt het idee dat bastaard en vreemd van elkaar afgegrensd kunnen worden al wel mee in de voorstellen van de spellingcommissie (Neijt & Zuidema 1994: 97). Vreemde woorden (verworpen door foneem- en/of flectiefilter) worden nooit omgespeld: zij behouden hun buitenlandse spellinguiterlijk, waarbij overigens diacritische tekens vervallen voor zover zij geen effect hebben op de uitspraak van deze woorden. Van de bastaardwoorden is de spellingsystematiek apart onderzocht en in regels gevangen (Neijt 1994).

6. Conclusies

De vragen die we in sectie 2 hebben gesteld, kunnen nu als volgt beantwoord worden. We hebben aangetoond dat een formele karakteristiek van inheemse versus uitheemse woorden goed te geven is. Van cruciaal belang hierbij is dat de formele karakteristiek op geen enkele manier gebruik maakt van het spellingbeeld van de betreffende woorden.

Het is voorshands niet duidelijk of de gemaakte indeling op alle details overeenstemt met de intuïties van de taalgemeenschap; wel kunnen we langs automatische weg 91% van de etymologisch echt inheemse woorden als zodanig aanmerken, terwijl we ervan uit kunnen gaan dat de volle 100% haalbaar is wanneer een aantal codeerfouten in de invoergegevens rechtgezet wordt.

Inheemse woorden zijn voorts verrassend goed af te bakenen van uitheemse woorden. Op basis van de huidige steekproef en de opgestelde synchrone criteria vallen slechts 308 van de 2.490 onderzochte vormen (12%) etymologisch gezien ten onrechte in de inheemse categorie. De overlapping van het werkelijk inheemse en het uitheemse deel van onze woordenschat is dus betrekkelijk gering. Wij gaan ervan uit dat de Nederlandse taalgemeenschap die etymologisch uitheemse woorden die formeel niet meer zijn af te bakenen van de inheemse woorden, als inheems zal willen beschouwen. De teneur van de voorstellen van de spellingcommissie is om deze groep woorden alleen volgens de inheemse spellingconventies te schrijven.

* Een computerleesbare versie van dit woordenboek werd voor onderzoeksdoeleinden aangeleverd door Van Dale Lexicografie (Utrecht) tegen subcommercieel tarief. De Nederlandse Taalunie stelde hiertoe een subsidie beschikbaar. Wij zijn beide instanties erkentelijk voor hun bijdrage.

7. Bibliografie

Cohen, A., C.L. Ebeling, K. Fokkema, & A.G.F. van Holk (1978). *Fonologie van het Nederlands en het Fries; inleiding tot de moderne klankleer.* 's-Gravenhage: Martinus Nijhof.

Heemskerk, J.S.M. & V.J. van Heuven (1993). 'MORPA, a MORphological PArser for a Dutch text-to-speech system', in: V.J. van Heuven & L.C.W. Pols (eds.), p. 67-85.

Heuven, V.J. van & L.C.W. Pols (eds.) (1993). *Analysis and synthesis of speech, towards high-quality text-to-speech generation.* Berlin: Mouton de Gruyter.

Kager, R & W Zonneveld (1986) 'Schwa, syllables and extrametricality in Dutch', in The Linguistic Review, 5, p 197-221
Neijt, A H (1991) *Universele fonologie* Dordrecht Foris
Neijt, A H (1994) 'Van orthografie naar ortografie', in A H Neijt, I Roggema en J J Zuidema (eds) *De spellingcommissie aan het woord*, 's-Gravenhage SDU Uitgeverij Koninginnegracht, p 59-72
Neijt, A H & J J Zuidema (eds) (1994) *Spellingdossier, Deel I Spellingrapport*, Voorzetten van de Nederlandse Taalunie, 44, 's-Gravenhage SDU Uitgeverij Koninginnegracht
Trommelen, M & W Zonneveld (1986) 'Dutch morphology, evidence for the right-hand head rule', in *Linguistic Inquiry* 17, 147-169
Veen, P A F van der & N van der Sijs (1990) *Etymologisch woordenboek, de herkomst van onze woorden* Utrecht Van Dale Lexicografie

Vincent J van Heuven
Vakgroep Algemene Taalwetenschap/Fonetisch Laboratorium RUL

Maarten Hijzelendoorn
Vakgroep Algemene Taalwetenschap/Fonetisch Laboratorium RUL

Anneke H Neijt
Vakgroep Nederlands KUN