Over inheemse, vreemde en bastaardwoorden in het Nederlands

Heuven, V.J.J.P. van

Over inheemse, vreemde en bastaardwoorden [1]

Vincent J. van Heuven

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, enz. 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 naarmate ze langer geleden in onze taal zijn binnengedrongen.

Bij discussies over aanpassing van uitheemse woorden aan het Nederlands kijken taalkundigen uitsluitend naar de hoorbare eigenschappen van die woorden. Het doet er absoluut niet toe hoe we die woorden schrijven. Niettemin is er een zwakke overeenkomst tussen taalkundige structuur en spelling van uitheemse woorden. Naarmate de structuurkenmerken van uitheemse woorden sterker afwijken van de inheemse norm, neemt ook de kans toe dat we in die woorden ongewone letters en lettercombinaties (grafieën) zullen aantreffen.

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 (bijv. *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 (bijv. *consequent, extract, apathie*). Ze bevatten alleen nog maar inheemse

1 Dit onderzoek is uitgevoerd in samenwerking met Anneke Neijt (KUN). De noodzakelijke computerprogramma's zijn ontwikkeld door Maarten Hijzelendoorn (RUL).

klanken, maar kunnen uitheems aandoen door afwijkende klankcombinaties, bijv. *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, bijv. *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 dikwijls onnederlandse klanken (bijv. *garage, douche, goal*). Vreemde woorden behouden meestal de spelling van de taal waaruit ze afkomstig zijn.

Volgens zijn instellingsbeschikking had de spellingcommissie de opdracht om een vergaand consequente spellingsregeling te ontwerpen voor alleen de bastaardwoorden. 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 wezenlijk belang geweest om te komen tot een scherpe afbakening van de drie typen woorden binnen de Nederlandse woordenschat. Daarbij heeft de commissie niet willen volstaan met de gebruikelijke intuïtieve indeling, maar heeft geprobeerd te komen tot een automatisch toepasbaar stelsel van criteria waarmee deze indeling uitgevoerd kan worden. Over deze pogingen gaat dit stuk.

Als het mogelijk is 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. Voor 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 spellingsregime 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, kunnen echter zonder bezwaar omgespeld worden volgens de inheemse spellingsystematiek.

Vraagstelling en plan van aanpak

De eerste vraag die we met ons onderzoek willen beantwoorden is: *kunnen we, zonder naar de spelling te kijken, 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 systeem. 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 bij mijn weten 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 taalkundige 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 slechts van toepassing op niet-samengestelde woorden. Samenstellingen moeten eerst worden opgesplitst in kleinste betekenisdragende woorddelen (morfemen) omdat anders geen eenduidige status bepaald kan worden. In bijv. het samengestelde woord *spraak+synthese* is het eerste lid, *spraak* van inheemse oorsprong, terwijl het twee lid, *synthese*, afkomstig uit het Grieks, de bastaardstatus heeft. 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 en Van Heuven, 1993).

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 uitgefilterd worden. In onze aanpak onderscheiden we in feite vijf van zulke filters, waarbij ieder filter eigenschappen toetst op een bepaald niveau in de taalkundige structuur van een woord. Ik bespreek deze niveaus eerst globaal; daarna wordt per niveau in een aparte paragraaf een toelichting gegeven.

- per klank (klankfilter): bevat het woord uitsluitend inheemse klanken (klanken)?
- per lettergreep (syllabefilter): is de opeenvolging van klanken binnen iedere syllabe legaal?
- per woord (syllabe-opeenvolgingfilter): is de opeenvolging van syllaben legaal?
- per woord (klemtoonfilter): ligt de klemtoon op de juiste syllabe?
- per woord (buigingsfilter): is de verbuiging 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 licht genoeg zijn om een woord de bastaardstatus 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).

Klankfilter

Om vast te stellen of een woord alleen maar inheemse klanken bevat, stellen we ons de uitspraak voor van dat woord door een Algemeen Beschaafd spreker van het Nederlands. Deze uitspraak wordt op papier vastgelegd, getranscribeerd, in de vorm van een globale klankrepresentatie. Daarbij maken we gebruik van uitsluitend de klanken die voorkomen in de inventaris van het *Centre for Lexical Information*, de Celex-databank (Max Planck Instituut, Nijmegen). Woorden die niet kunnen worden weergegeven met uitsluitend de Celex-klanken, zijn per definitie vreemd. Binnen de Celex-klankinventaris bevindt zich echter ook een aantal uitheemse klanken, zoals aangegeven in tabel 1 (waarin de Celex-notatie is aangepast; zie bijlage E voor de overige klanksymbolen):

Tabel 1.

fonetische tekens voor vreemde klanken		
notatie	als in	transcriptie
íé	analyse	aa - n aa - l **íé** - z e
óé	rouge	r **óé** - zj e
úú	centrifuge	s e n - t r ie - f **úú** - zj e
àà	pass	p **àà** s
èè	serre	s **èè** - r e
èù	freule	f r **èù** - l e
òò	roze	r **òò** - z e
ã	restaurant	r e s - t oo - r **ã**
ẽ	mannequin	m a - n e - k **ẽ**
õ	plafond	p l aa - f **õ**
ũ	parfum	p a r - f **ũ**
gh	spaghetti, goal	**gh** oo l
zj	journaal	**zj** oe r n aa l

Woorden die in hun transcriptie één of meer klanksymbolen bevatten uit deze tabel zijn uitheems (en zelfs vreemd).

Syllabefilter

Klanken worden in de taal samengenomen tot syllaben (lettergrepen). Niet iedere willekeurige combinatie van klanken levert een correcte syllabe op. De mogelijke klankopeenvolgingen in een syllabe worden verantwoord door regels die van taal tot taal kunnen verschillen. 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 bijv. Neijt, 1991 en verwijzingen aldaar) hebben opgesteld, bevat een opsomming van in beginsel alle klankopeenvolgingsregels, voor zover van toepassing op het strikt inheemse deel van de Nederlandse woordenschat. Daarbij is het gemakkelijker om de beperkingen op toegestane klankopeenvolgingen apart te specificeren voor de medeklinkers die aan de klinker voorafgaan (de onset) en voor de toelaatbare combinaties van klinkers en daaropvolgende slotmedeklinkers (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 medeklinkers na de klinker) van een eindsyllabe vertonen grotere variëteit dan die van woordmediale lettergrepen. In dit verband tellen voor- en achtervoegsels als marginale syllaben: de coda van een voorvoegsel (bijv. *ont-* als in *ontzien*) en de onset van een achtervoegsel (bijv. *-ster* als in *bedriegster*) vertonen dezelfde ruimere mogelijkheden als de syllaben van woordeinde en -begin. Dezelfde woordmarginale status hebben we gegeven aan onsets die in een woord voorkomen na de (kwasi-)voorvoegsels *ge-, be-, ver-, te-, je-, me-, de-* en aan rijmen voor de (kwasi-)achtervoegsels *-de* en *-te*.

We hebben er vanaf gezien om bij de formulering van onze klankopeenvolgingsbeperkingen 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 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.

Woordmediale onset en rijm
Inheemse onsets zijn:

(1) - p, b, t, d, k, f, v, s, z, ch, g, m, n, l, r, j, w (d.w.z. alle Nederlandse medeklinkers behalve [h] en [ng])
- st
- NIL (uitsluitend in achtervoegsels)

Inheemse rijmen zijn

(2) - Korte klinker binnen dezelfde lettergreep gevolgd door p, b, t, d, k, f, s, ch, g, m, n, ng, r, l
- Sjwa, lange klinker of tweeklank
- Sjwa gevolgd door r
- Lange klinker gevolgd door p, b, t, d, k, f, s, ch, g, m, n, r, l

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

(3) - NIL, p, b, t, d, k, f, v, s, z, g, m, n, l, r, j, w, h (d.w.z. niets plus de klanken die genoemd zijn onder 1, zonder [st] en met [h] in de plaats van [ch])
- pl, pr, bl, br, tr, tw, dr, dw, kn, kl, kr, kw, fn, fl, fr, vl, vr, sp, st, sch, sm, sn, sl, sj, zw, gn, gl, gr
- spr, str, schr, spl

Een rijm aan het woordeinde mag bevatten:

(4) a Korte klinker gevolgd door
- p, b, t, d, k, f, s, ch, g, m, n, ng, l, r (d.w.z. alle medeklinkers behalve [v, z, j, w, h])
- ps, pt, bs, bt, ts, ds, ks, kt, fs, ft, st, chs, cht, ms, mt, ns, nt, ngs, ngt, ls, lt, rs, rt (d.w.z. alle medeklinkers behalve [v, z, g, j, w, h], gevolgd door [s, t])
- mp, lp, rp, md, nd, ld, rd, ngk, lk, rk, lv, rv, lg, rg, mz, nz, lz, rz, lch, rch, lm, rm, ln, rn
- rts, chts, lts, nts, ngks, lps, rft, rkt, mpt, rps, rst, lst, ngst, lft, rfst

b Lange klinker 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
- [ie] gevolgd door [lp, rv, rp]

c Tweeklanken 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, ch, g, m, n, l, r
- nd, rd, ld, nt, rt, lt, ns, rs, ls

e Lange klinker plus heterorgane halfklinker: [aaj, eew, iew, ooj, oej, uuw] optioneel gevolgd door [s, t, st, ts].

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

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 gewichtsverdeling tussen de lettergrepen. Inheemse woorden bevatten in beginsel slechts één syllabe met een volle klinker, d.w.z. een klinker anders dan sjwa. Deze syllabe noemen we de kernlettergreep. De kernlettergreep mag vooraf worden gegaan door één ultralichte lettergreep (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 (in het programma is dit deelfilter voor klankopeenvolgingen gedefinieerd):

Gewichtsverdeling binnen woord

(5) a Niet meer dan één lettergreep bevat een volle klinker of tweeklank.
b De volle lettergreep mag worden voorafgegaan door
ge-, be-, ver-, te-, je-, me-, de-.
c De volle lettergreep mag worden gevolgd door maximaal één medeklinker plus
- *-e, -er, -en, -el, -em, -ig, -ik, -eld, -end, -erd, -ens, -erd, -ers, -end, -eren, -ige, -elig, -elijk, -erik.*
- *-ij, -uw, -uwe, -ing, -aar, -ond, -and, -og, -ik, -erik.*

De tweede groep beperkingen op syllabe-opeenvolgingen betreft de mogelijke onsets van een ultralichte lettergreep na bepaalde rijmen van de kernlettergreep (Kager en Zonneveld, 1986, 208 en verder). Het rijm van de kernlettergreep bevat twee segmenten (een lange klinker of tweeklank, dan wel een korte klinker plus medeklinker), optioneel gevolgd door een willekeurige medeklinker, die op zijn beurt weer gevolgd mag worden door [s, t] of door een sjwa plus [m, n, ng, l, r]. De aaneengesloten reeks medeklinkers van kernlettergreep en de volgende ultralichte lettergreep 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 wrijfklank [v, z] onmiddellijk na een korte klinker: **puzzel* met een korte [u] is duidelijk een uitheems woord; het zou inheems kunnen worden als de uitspraak met een lange [uu] (*puzel*) gangbaar wordt. Schema (6) vat mogelijkheden samen:

Kernrijm plus volgende onset

(6) Alleen de volgende kernrijm-onsetcombinaties zijn toegestaan:
a Lange klinker gevolgd door
- iedere medeklinker behalve [f, s]
- st, nt, rt, gd, nd, ld, rd, jk, rz, rn, lj, rst, nst
b Tweeklank gevolgd door
- iedere medeklinker behalve [s]
- st, nd
c Korte klinker gevolgd door
- sp, ps, sk, ks, st, ts, ft, cht, nt, lt, rt, mp, rp, mb, rb, nd, ld, rd, ngk, rk, lv, rv, ng, nk, ns, ls, rs, nz, lz, rz, lg, rg, rm, ln, nj, rw, kst, mst, nst, lst, rst, ngst
d Sjwa plus nul of meer medeklinkers

Klemtoonfilter
Inheemse woorden dragen de klemtoon op de meest linkse volle klinker of tweeklank. Ligt de klemtoon op een andere syllabe dan is het woord uitheems.

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

(7) Inheemse zelfstandige naamwoorden vormen hun meervoud als volgt:
   a geen meervoud (bijv. verzamelnamen),
   b *-n* en/of *-s* na een ultralichte syllabe,
   c *-en* in alle andere gevallen.

   Inheemse bijvoeglijke naamwoorden:
   a eindigen op *-en* (*houten, stalen*)
   b nemen een buigings-*e* (*grote, kleine*)

Wanneer een zelfstandig of bijvoeglijk naamwoord zich niet gedraagt volgens (7) of daarnaast nog een andere meervoudsvorm kent (bijv. *musea* naast *museums*), dan is het betreffende woord uitheems.

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 Quintus-Prologprogramma en vervolgens getest op het RUL-morfeemlexicon, een bestand met klankrepresentaties van morfemen. Dit lexicon is in de jaren 1986-1990 ontwikkeld als onderdeel van een voorleesmachine (Van Heuven en Pols, 1993). Het bevat ongeveer 12.500 ongelede Nederlandse morfemen en 4.000 onregelmatige gelede vormen (onregelmatig omdat de betekenis niet bepaald is door de som van de betekenissen van de samenstellende delen). Iedere vorm is voorzien van een 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 verbuiging van naamwoorden 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 (Van der Veen en Van der Sijs, 1990). De etymologiecode in het morfeemlexicon is eenvoudig; alleen inheems en uitheems zijn aangegeven, bij uitheems niet nader uitgewerkt naar de taal van herkomst.

Niet alle vormen kwamen zowel in het morfeemlexicon als in het etymologische woordenboek voor; de doorsnee van beide verzamelingen leverde een kleine 5.000 morfemen op. Van ieder morfeem is vervolgens door het computerprogramma vastgesteld of deze volgens onze filters inheems of uitheems is. Deze uitkomst is vergeleken met de etymologiecode uit het woordenboek. Er zijn dan vier verschillende resultaten mogelijk, die in tabel 2 zijn aangegeven tezamen met de vastgestelde aantallen.

Tabel 2. Inheems en uitheems volgens het computerprogramma en volgens het woordenboek, eerst in absolute getallen, daarna in percentages van het aantal woorden in het gegevensbestand.

	Inheems volgens computer-programma	Uitheems volgens computer-programma	Totaal
Inheems volgens woorden-boek	Terechte acceptatie 2243 91%	Valse verwerping 220 9%	2463 100%
Uitheems volgens woorden-boek	Valse acceptatie 308 12%	Terechte acceptatie 2182 88%	2490 100%

De resultaten laten zien dat onze filters over de hele linie ongeveer 90% van de woorden correct indelen. De twee mogelijke soorten fouten 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%. Ik bespreek 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 geleed zijn

(bijv. *veertien, alledaags, aardappel*). Daarnaast zijn er een aantal vormen in die naar de stand vandaag weliswaar als ongeleed moeten worden aangemerkt, maar die dat vroeger niet waren (zgn. historisch gelede woorden, bijv. *oorlog, middag, twaalf*). Zulke gelede woorden bevatten illegale klankopeenvolgingen op syllabegrenzen of bevatten meer dan één volle klinker, 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 klanktranscriptie, 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.

Valse acceptaties. Twaalf procent van de etymologisch uitheemse woorden werd door het filter niettemin 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 een-lettergrepige klassieke woorden (bijv. *straat, vorm, som*) en een aantal meer-lettergrepige woorden met één volle klinker (bijv. *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 spellingsconventies wordt geschreven. Van de exotisch gespelde 40 zijn er opvallend veel homofoon met een inheems woord, bijv.: *ether - eter, lynx - links, pact - pakt.*

Vreemde woorden

Het algoritme is in beginsel ook in staat binnen de categorie uitheemse woorden de bastaardwoorden van de vreemde woorden te scheiden. Bastaardwoorden passeren altijd het klankfilter; woorden met een vreemde klank worden door het klankfilter onmiddellijk ontmaskerd als vreemd. Voorts is buiging een krachtig middel om vreemde woorden te ontmaskeren. Waar ons buigingsfilter een uitgang *-en* voorschrijft, worden vreemde woorden die *-s* hebben als vreemde woorden herkend (bijv. *club* is vreemd wegens *clubs* in plaats van **clubben*).

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 getalsmatig 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: vreemde woorden (verworpen door klank- of buigingsfilter) worden nooit omgespeld: zij behouden hun buitenlandse spellinguiterlijk. Van de bastaardwoorden is de spellingsystematiek apart onderzocht en in regels gevangen (zie verder de bijdrage van Neijt aan deze bundel).

Conclusies

De vragen die we 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 is hierbij dat de formele karakteristiek op geen enkele manier gebruik maakt van het spellingsbeeld 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 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 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. Overwogen kan worden om deze groep woorden volgens de inheemse spellingsconventies te schrijven (zie verder de bijdrage van Neijt aan deze bundel).

Literatuur

Heemskerk, J.S.M. en V. J. van Heuven 1993, 'MORPA, a MORphological PArser for a Dutch text-to-speech system,' in: V.J. van Heuven en L.C.W. Pols (eds.) *Analysis and synthesis of speech, towards high-quality text-to-speech generation.* Mouton de Gruyter, Berlijn, p. 67-85.

Kager, R. en W. Zonneveld 1986, 'Schwa, syllables and extrametricality in Dutch,' *The Linguistic Review* 5, p. 197-221.

Neijt, A.H. 1991, *Universele fonologie.* Foris Publications, Dordrecht.

Veen, P.A.F. van der, 1990, *Etymologisch woordenboek, de herkomst van onze woorden.* Van Dale Lexicografie, Utrecht/Antwerpen.