Fryske Akademy nr. 1075

De tienduizend dingen

Foarmjouwing: Roelof Koster
Opmaak: Jan Tiemersma
Omslach: skilderij fân Michael Berkhemer (detail)

Afûk, Postbus 53, 8900 AB Ljouwert

NUR 610
ISBN 978 90 62739 67 7

Neat út dizze útjefte mei op hokker wize dan ek fermannichfâldige wurde sûnder dat dêr skriftlike tastimming fan de útjouwer oan foarôf giet.

www.afuk.nl
www.fryske-akademy.nl

Subsidiënt: Boersma-Adema Stichting te Leeuwarden

De tienduizend dingen

Feestbundel voor Reinier Salverda

Onder redactie van

Hanno Brand
Ben Groen
Eric Hoekstra
Cor van der Meer

Ljouwert 2013

geen rekening te houden met de uitwisseling van grammaticale informatie. Er is een directe relatie tussen de vorm en de betekenis zodat de informanten het woord makkelijk verwerken. Met andere woorden, er is sprake van *mapping* tussen de vorm en de betekenis.
De aanpassing van de grammaticale informatie tussen de woorden en tussen de woordgoepen blijft bij de informanten een struikelblok. Bij de tussentaalzinnen wordt de grammaticale informatie niet altijd uitgewisseld. Om goede Nederlandse zinnen te vormen moet men rekening houden met de juiste uitwisseling van de grammaticale informatie. In de beginfase van de tussentaal letten de taalleerders meer op de betekenis dan op de grammatica. De betekenis moet uitgedrukt worden door middel van talige elementen en de grammatica maakt de ordening van de betekenis doelmatig en doeltreffend. Riyanto heeft onderzocht dat Indonesiërs die Nederlands beheersen meer op de betekenis letten dan op de grammatica, terwijl de moedertaalsprekers van het Nederlands het omgekeerde doen.[17]
De *Processability Theory* blijft teveel aan de veilige kant. De theorie bepaalt het percentage van 70% als minimale beheersing bij Sep, Inv, en V-einde. Met zo'n percentage is het moelijk om de theorie te testen. De theorie zou ruimte moeten hebben voor hogere percentages, bijvoorbeeld 80%. Met een hoger percentage zal het resultaat van het onderzoek er anders uitzien. De theorie past dus eerder bij beginnende taalleerders. Voor de gevorderden of *near native* heeft men een veel hoger percentage, bijvoorbeeld 90%. Dat is een uitdaging voor verder onderzoek.
Onderzoeken naar tussentaal inspireren de mensen om niet negatief te oordelen over taalproducten van tweede-taalleerders. Negatieve reacties op de tussentaal passen hierbij ook niet. De docenten en leraren mogen de leerders zeker niet ontmoedigen. Ze moeten de leerders juist aanmoedigen. De zinnen die ze produceren hebben een ingewikkeld proces doorlopen in hun hoofden. Ze hebben een beperkte tijd om alles te verwerken. Hun woordenschat is nog beperkt en ze kennen ook nog weinig grammaticale regels. De tussentaal is het resultaat van de noodzaak bij de leerders om in snel tempo ideeën en concepten uit te drukken in talige elementen terwijl ze nog een beperkte woordenschat en een beperkte grammatica hebben en ze de eerste taal of een vreemde taal al beheersen. Dat ze toch iets kunnen zeggen in de tweede taal in welke vorm dan ook moet men accepteren als een geweldige prestatie. Ze hebben in hun hoofden veel moeite gedaan.

17 S. Riyanto, *Syntactische en semantische middelen bij de interpretatie van Nederlandse zinnen* (MA-thesis Universiteit Leiden 1990).

Sluitdagen *en* afpakschuur: *de Indisch-Nederlandse woordenschat*[1]

Nicoline van der Sijs
Taalwetenschap, Radboud Universiteit Nijmegen
en Meertens Instituut (KNAW) te Amsterdam

Slamat

Sinds zich, vanaf circa 1600, groepen Nederlanders in het huidige Indonesië hebben gevestigd, hebben het Nederlands en het Maleis invloed uitgeoefend op elkaar. Zo ontstonden er enkele creooltalen of mengtalen: het Petjoh met Maleis en het Javindo met Javaans als grammaticale basis. De woordenschat van beide creooltalen is voor een groot deel Nederlands, met veel leenwoorden uit Indonesische talen. Deze creooltalen ontstonden vooral na 1800, doordat pas toen grote groepen Nederlanders zich voor langere tijd in Indonesië vestigden.[2]

Naast de Nederlandse creooltalen was en bleef het Nederlands in Indonesië in gebruik, en het werd telkens ververst door nieuw uitgezonden Nederlanders. Dit Nederlands ontwikkelde zich onder invloed van het Maleis tot het Indisch-Nederlands: een variëteit van het Nederlands waarin allerlei Maleise eigenaardigheden binnenslopen.[3] Er was sprake van een glijdende schaal die liep van Europees-Nederlands, via Indisch-Nederlands, naar Petjoh en Maleis: de creooltaal Petjoh lag in het continuüm dus het dichtst bij het Maleis en het verst van het Nederlands.[4]

1 Ik dank Hans Beelen, Gosse Bouma, Joep Kruijsen, Maarten Marx, Martin Reynaert en René van Stipriaan voor hun nuttige opmerkingen.

2 H. van Rheeden, 'The mixed language of the Indos in Batavia', in P. Bakker & M. Mous (red.), *Mixed languages* (Amsterdam 1994) pp. 223-237. H. van Rheeden, *Het Petjo van Batavia. Ontstaan en structuur van de taal van de Indo's* (Amsterdam 1995). J.W. de Vries, 'The language of the Indo-Dutch', in: M.A. Bakker & B.H. Morrison (eds.), *Studies in Netherlandic Culture and Society* (Lanham 1994) pp. 213-226. J.W. de Vries, 'Mengtalen in de archipel: Nederlands in vreemde mond', *Thema's en trends in de sociolinguïstiek 2. Toegepaste taalkunde in artikelen* 52 (1995) 2 pp. 71-78. Over de woordenschat: R. Cress, *Petjoh. Woorden en wetenswaardigheden uit het Indisch verleden* (Amsterdam 1998). V.E. de Gruiter, *Het Javindo* (Den Haag 1990). F.S. Loen, *Petjoh Indisch woordenboek* (Rotterdam 1994).

3 R. Salverda, 'Between Dutch and Indonesian: colonial Dutch in time and space', te verschijnen in: F. Hinskens & J. Taeldeman (eds.), *Language and Space: Dutch* (Berlin 2013). M.C. van den Toorn, 'De taal van de Indische Nederlanders', *De Nieuwe Taalgids* 50 (1957) pp. 218-226. J.W. de Vries, 'Indisch-Nederlands: verleden, heden en toekomst', in: W. Willems (red.), *Sporen van een Indisch verleden* 1600-1942 (Leiden 1992) pp. 125-140.

4 J. de Vries in: Van der Sijs (red.)(2005) pp. 59-78. Salverda, 'Between Dutch and Indonesian'.

Het Indisch-Nederlands wordt momenteel alleen nog – in zeer beperkte mate – gehoord in Nederland, want na de onafhankelijkheid van Indonesië in 1949 zijn de meeste Indisch-Nederlandse sprekers naar Nederland getrokken.

Door het taalcontact zijn er leenwoorden uitgewisseld: Nederlandse leenwoorden drongen door tot Indonesische talen, met name het Maleis (tegenwoordig Bahasa Indonesia genoemd)[5] en andersom.[6] Bovendien werden er in het Nederlands van Indonesië nieuwe woorden gevormd, of kregen bestaande woorden nieuwe betekenissen. Die woorden waren nodig voor het benoemen van de voor Nederlanders nieuwe Aziatische wereld. De nieuw ontstane woordenschat en de aanpassingen en vernieuwingen die het Nederlands doormaakte in Indonesië zijn tot nu toe slechts zeer cursorisch beschreven.[7] Dat bracht Reinier Salverda ertoe om een recent artikel over de geschiedenis van het Nederlands in Indonesië te besluiten met de wens dat er een uitgebreid historisch-kritisch etymologisch woordenboek van het Indisch-Nederlands moge worden vervaardigd.[8] Een dergelijk woordenboek is in zijn ogen essentieel voor een beter inzicht in de leenwoorduitwisseling die heeft plaatsgevonden in de context van het Indonesisch-Nederlandse taalcontact. Minstens zo interessant als de uitgewisselde leenwoorden zijn de nieuwe Nederlandse woorden en betekenissen die in de Indonesische context zijn ontstaan. Daarover weten we vrijwel niets. Elders heb ik al eens de hypothese geopperd dat allerlei veranderingen in het Nederlands – nieuwe Nederlandse woorden of constructies – in het tweetalige Indië zijn ontstaan of geaccepteerd geraakt.[9] Lang niet al die veranderingen zijn doorgedrongen tot het Europees-Nederlands, maar een deel wel. Zo vond Jan Stroop onlangs dat de uitdrukking *dat klopt (niet)* in Indonesië is gevormd.[10]

5 R. Jones (ed.), *Loan-words in Indonesian and Malay* (Leiden 2007). N. van der Sijs, *Nederlandse woorden wereldwijd* (Den Haag 2010).

6 J. van den Berg, *Soebatten, sarongs en sinjo's. Indische woorden in het Nederlands* (Den Haag 1990). P. Mingaars e.a., *Indisch lexicon. Indische woorden in de Nederlandse literatuur* (Utrecht 2005). N. van der Sijs (bezorger), *Uit Oost en West. Verklaring van 1000 woorden uit Nederlands-Indië* van P.J. Veth, met aanvullingen van H. Kern en F.P.H. Prick van Wely (Amsterdam 2003). N. van der Sijs, *Groot Van Dale Leenwoordenboek. De invloed van andere talen op het Nederlands* (Utrecht, Antwerpen: 2005).

7 F.P.H. Prick van Wely, *Neerlands Taal in 't verre Oosten, eene bijdrage tot de kennis en de historie van het Hollandsch in Indië* (Semarang/Soerabaia 1906). F.P.H. Prick van Wely, *Viertalig Aanvullend Hulpwoordenboek voor Groot-Nederland. Vermeerderd met door Prof. H. Kern herzien etymologisch aanhangsel* (Weltevreden 1910).

8 Salverda, 'Between Dutch and Indonesian'.

9 N. van der Sijs, 'Het ongezochte vinden', *NRC Handelsblad*, Wetenschapsbijlage 13 oktober (2012) p. 2.

10 http://nederl.blogspot.nl/2013/05/klopt.html#more.

Er is nooit systematisch onderzoek gedaan naar nieuwvormingen in Indonesië, zodat we niet weten om welke woorden en constructies het gaat, welke daarvan geabsorbeerd zijn door het Europees-Nederlands en welke zijn verdwenen. Evenmin weten we of nieuwvormingen in Indonesië morfologische bijzonderheden vertonen die te herleiden zijn tot invloed van het Maleis of het Indisch-Nederlands. Ter ere van Reinier Salverda zal ik in dit stuk bekijken of het mogelijk is met behulp van corpusonderzoek nieuwe Indisch-Nederlandse woorden en betekenissen op te sporen die een plekje verdienen in het door hem gewenste historisch-kritische woordenboek.

Een eerste verkennend corpusonderzoek naar het Indisch-Nederlands

Tot voor kort gebeurde onderzoek naar nieuwe woorden, betekenissen of constructies noodgedwongen met de hand: een onderzoeker excerpeerde zoveel mogelijk teksten en noteerde de woorden die in zijn ogen potentieel 'nieuw' waren. Die methode is in de praktijk in hoge mate subjectief en foutgevoelig. In de huidige digitale wereld is het mogelijk automatisch, aan de hand van een corpus, vast te stellen wanneer en waar een woord voor het eerst is gebruikt, op voorwaarde dat alle relevante bronnen digitaal beschikbaar zijn. Om Indisch-Nederlandse nieuwvormingen op te sporen, kunnen we de computer een corpus Nederlandstalige teksten uit Nederland laten vergelijken met een vergelijkbaar corpus uit Indonesië. De computer kan woorden en woordcombinaties tellen en aangeven welke alleen of eerder voorkomen in Indonesië, en daar wellicht zijn ontstaan. Op die manier krijgt het onderzoek naar nieuwvormingen een statistische basis: in plaats van toevallig, handmatig bij elkaar gezochte voorbeelden worden de gegevens automatisch uit een enorm tekstbestand geëxtraheerd. Uiteraard moeten onderzoekers die gegevens vervolgens wel op hun validiteit beoordelen.

Voor systematisch onderzoek zijn dus twee dingen nodig: onderzoekscorpora en een onderzoeksinstrumentarium waarmee de corpora kunnen worden geanalyseerd. Geen van beide ligt momenteel gemakkelijk binnen het bereik van een onderzoeker, maar in de toekomst zal dat verbeteren. Binnen het NWO-groot project Nederlab – een laboratorium voor onderzoek naar de veranderingspatronen in de Nederlandse taal en cultuur – zijn we bezig een onderzoeksomgeving te bouwen waar data en tools worden aangeboden.[11] De onderzoeksomgeving is momenteel in opbouw, maar we kunnen er al wel testjes mee doen, zoals ik hieronder zal laten zien.

11 Zie http://www.nederlab.nl/. Ik dank de technici Matthijs Brouwer, Matthijs Droës, Hennie Brugman, Marc Kemps-Snijders, Maarten Marx, Martin Reynaert, Erik Tjong Kim Sang, Rob Zeeman en Junte Zhang voor hun hulp.

Achter de schermen wordt binnen Nederlab gewerkt aan de Historische Kranten die zijn gedigitaliseerd door de Koninklijke Bibliotheek.[12] De kranten uit het jaar 1900 zijn inmiddels geconverteerd naar het FoLiA-formaat.[13] Dit houdt onder meer in dat de woorden zijn gelemmatiseerd en zijn voorzien van woordsoortinformatie. Daarna zijn de woordvormen geïndexeerd. Bij die indexering zijn hoofdletters automatisch omgezet naar kleine letters, om het aantal woordtypes te beperken: anders zouden woorden die kapitaal of klein kapitaal staan – iets wat veel voorkomt in kranten –, en woorden aan het begin van een zin die per definitie met een hoofdletter zijn gespeld, als aparte woordtypes gelden. Het totale corpus bevat 130.152.587 woorden (tokens). Zonder omzetting van de hoofdletters in kleine letters zijn er 5.289.418 verschillende woordvormen (types), na omzetting van de hoofdletters is dit aantal gereduceerd tot 4.797.631. Hiervan komen er 3.851.986 slechts eenmaal voor, dus tachtig procent. Dat is veel hoger dan de gemiddeld vijftig procent hapaxen die (Engelse) teksten bevatten.[14] Hieronder zullen bijzondere eigenschappen uit de kranten ter sprake komen die waarschijnlijk verantwoordelijk zijn voor het afwijkende percentage hapaxen in de kranten.

De kranten zijn oorspronkelijk gepubliceerd in Nederland, Indonesië, Suriname en de Antillen, en ze kunnen per regio worden gesorteerd. Zo kunnen we corpora voor iedere regio samenstellen. Aan de hand van die corpora kunnen we bekijken welke statistisch significante verschillen er bestaan in de frequenties van woorden of woordcombinaties in krantencorpora uit Nederland en uit Indië. Woorden of uitdrukkingen die statistisch significant vaker of eerder in Indische kranten voorkomen, zijn wellicht nieuwvormingen in Indonesië: woorden dus die waarschijnlijk opgenomen dienen te worden in een uitvoerig historisch-kritisch woordenboek van het Indisch-Nederlands.

12 http://kranten.kb.nl/.

13 Voor uitleg zie: http://proycon.github.io/folia/index.nl.html.

14 F. Fengxiang, 'An Asymptotic Model for the English Hapax/Vocabulary Ratio', *Association for Computational Linguistics* 2010: http://aclweb.org/anthology/J/J10/J10-4003.pdf.

Vijftig woorden die specifiek zijn voor Indonesische kranten
Hieronder staan de vijftig meest voorkomende woorden die alleen in Indonesische kranten voorkomen en niet, of hoogstens een enkele maal, in kranten gepubliceerd in de andere regio's. Tussen haakjes staat het aantal voorkomens in de kranten van 1900.

scheepsberichten (758)	ontvang-avonden (300)	prauwenveer (162)
dexe (711)	spreekdagen (297)	vendulocaal (151)
xng (609)	dragonbrand (279)	pandgr (138)
nederlandsch-indie (553)	emmahaven (278)	fort-de-kock (136)
totok (505)	batavia-nieuws (276)	soesman's (136)
vendutie (501)	pagoejaman (263)	sluitdagen (134)
teker (477)	passar (255)	getien (132)
xyn (443)	giang (249)	assistent-wedono (132)
postsluiting (420)	soei (249)	neera (132)
tjjn (391)	batavia-bladen (209)	shanghay (131)
avagyee (352)	xeker (194)	k.m.p.u. (123)
wedono (343)	hokwei (191)	pryce (123)
postsluitingen (338)	volks-bibliotheek (189)	bindjey (121)
vervaldagen (300)	pandgoederen (184)	tulke (114)
telfs (320)	xonder (182)	pandhouder (112)
residentie-huize (312)	opwekkertjes (180)	krakal (109)
deten (311)	vendutien (180)	

Op figuur 1 is van twee willekeurige woorden, *prauwenveer* en *sluitdagen*, de verspreiding in de tijd (binnen het jaar 1900) en over de gebieden gevisualiseerd.
Sommige woorden uit het rijtje zijn onherkenbaar als Nederlands woord, bijvoorbeeld *xng* of *soei*. Om erachter te komen wat deze woorden betekenen, heb ik ze nagezocht in de originele kranten en de context ervan bekeken. Daarbij bleek dat het om verschillende typen woorden gaat.[15]

15 In een statistisch onderzoek gebaseerd op het Engelstalige corpus van Google Books merken de auteurs op dat niet alle woordvormen (1-grammen) Engelse woorden zijn: zij onderscheiden drie categorieën niet-woorden: "(i) 1-grams with nonalphabetic characters (*l8r*, 3.14159), (ii) misspellings (*becuase*, *abberation*), and (iii) foreign words (*sensitivo*)." Zij berekenden dat het percentage van dergelijke niet-woorden in hun corpus liep van 51% in 1900 tot 31% in 2000. Zie: J.-B. Michel e.a., 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Science* 14 January 2011: 331 (6014), 176-182. Published online 16 December 2010.

Figuur 1: Het voorkomen van prauwenveer *en* sluitdagen *in kranten uit 1900*

Leesfouten

Een deel van de woorden blijkt terug te gaan op een leesfout: de door de KB gehanteerde optische tekenherkenning (OCR) waarmee de kranten zijn 'gelezen', heeft tekens verkeerd geïnterpreteerd. Zo moeten *dexe, xng, teker, xyn, tjjn, telfs, deten, xeker, xonder, getien* en *tulke* worden gelezen als *deze, ???, zeker, zijn, zijn, zelfs, dezen, zeker, zonder, gezien, zulke.* Het 'woord' *xng* is niet te herleiden tot een enkele leesfout, het staat op allerlei plaatsen waar de computer er eenvoudigweg niet uitkwam. Het is bepaald verontrustend voor de betrouwbaarheid van de ge-OCRde kranten als onderzoeksmateriaal dat deze leesfouten zo vaak voorkomen dat er hoogfrequente woorden door worden gecreëerd: maar liefst elf van de vijftig meest frequente woorden, bijna een kwart, blijken een leesfout. Het hoge percentage leesfouten is onder andere te verklaren door de slechte leesbaarheid van het krantenzetsel. Van de ene kant leidt dit tot 'spookwoorden' als *xeker* en *xyn*, maar we mogen ook vermoeden dat wel bestaande frequente woorden in de statistiek ten onrechte niet opduiken omdat de betreffende vindplaatsen niet goed zijn gelezen. Opmerkelijk is dat leesfouten als *dexe* komen bovendrijven als kenmerkend voor kranten uit Indonesië uit het jaar 1900 – deze leesfouten komen immers, zo blijkt uit een kleine steekproef in de Historische Kranten van de KB, in andere jaren minstens zo vaak voor in kranten afkomstig uit Nederland of Suriname. Om dergelijke toevallige uitkomsten te vermijden is het dus nodig het onderzoek te herhalen op een veel groter krantencorpus.

Het scheelt veel werk als dergelijke leesfouten uit toekomstig onderzoek worden uitgesloten door ofwel de gelezen tekst door een correctieprogramma te halen, ofwel slimme zoekprogramma's te ontwerpen die leesfouten omzeilen.[16] Dit is ook binnen Nederlab een punt van aandacht.

Namen

Een tweede categorie woorden die niet direct herkenbaar zijn, bestaat uit woorden die in het origineel zijn gespeld met een hoofdletter. Zoals hierboven uitgelegd zijn bij de indexering alle hoofdletters vervangen door kleine letters. Dit heeft voordelen (het aantal verschillende woordtypes

16 Voor automatische correctie van OCR-fouten heeft Martin Reynaert TICCL ontwikkeld. Voor uitleg zie: M. Reynaert, 'Character confusion versus focus word-based correction of spelling and OCR variants in corpora', *International Journal on Document Analysis and Recognition* vol. 14 (2010) nr. 2, pp.173-187 (http://repository.uvt.nl/id/ir-uvt-nl:oai:wo.uvt.nl:444358z). Op het INL is een historisch computationeel lexicon opgebouwd dat ingezet kan worden. Uit het onderzoek van Michel e.a. uit 2010 blijkt echter dat de woordenschat in de 20e eeuw enorm toeneemt en dat daarvan slechts een deel is opgenomen in woordenboeken en lexica.

wordt erdoor verminderd), maar, zo blijkt nu, ook duidelijke nadelen. Het is namelijk vaak niet meer direct duidelijk in welke gevallen er sprake is van namen. Bij Batavia-nieuws, Batavia-bladen, Nederlandsch-Indië en Shanghay kom je daar wel op. Maar je moet de originele kranten raadplegen om erachter te komen dat Avaygee, Giang, Hokwei en Neera namen van schepen zijn. Soei blijkt een achternaam. Bindjey, Emmahaven en Fort-de-Kock zijn geografische namen. Dragonbrand is een merknaam voor een bepaald soort petroleum. Pagoejaman is de naam van een mijnbouwmaatschappij. Soesman's en Pryce zijn onderdelen van firmanamen die voluit luiden: Soesman's Kantoor en John Pryce & Co.

Uit deze lijst blijkt dat het aantal namen zeer hoog ligt, een derde van het totaal aantal woorden: zestien van de vijftig hoogst frequente woorden die (vrijwel) alleen in Indische kranten voorkomen. Dat namen zo hoog scoren, kan verklaard worden uit het feit dat nieuws regionaal is, en veel namen die in Indonesische kranten worden vermeld, dus niet in Nederlandse kranten voorkomen. Daarnaast ligt het percentage namen sowieso hoger dan in andere tekstsoorten zoals literaire werken of wetenschappelijke publicaties: kranten berichten per definitie over nieuws dat gaat over mensen, bedrijven, gebeurtenissen die zich afspelen op een bepaalde plaats. Uit een steekproef in enkele complete jaargangen kranten blijkt dat ongeveer 6,8 procent van het totale aantal tokens een naam is, terwijl het percentage namen in de gemiddelde roman tussen de 2 à 3 ligt.

In woordenboeken is het tot nu toe vanwege omvangbeperkingen niet gebruikelijk om namen op te nemen, behalve in specifieke gevallen, bijvoorbeeld als ze spreekwoordelijk zijn geworden (*praten als Brugman*) of als ze soortnamen zijn geworden (*beaujolais, luxaflex, amsterdammertje*). Encyclopedieën nemen wel namen op, maar lang niet alle.[17]

De digitale tijd levert nieuwe mogelijkheden en stelt ook nieuwe eisen. Het argument dat een woordenboek te dik wordt als er ook namen in worden opgenomen, vervalt nu woordenboeken niet meer worden gedrukt maar op internet beschikbaar komen. Het verschil tussen woorden en namen is arbitrair: taalgebruikers hebben uitleg nodig over beide. In mijn ogen zou een belangrijk onderdeel van het samen te stellen historisch-kritische woordenboek van het Indisch-Nederlands eruit bestaan een lijst van namen op te stellen die in het verleden in Indonesië werden gebruikt. De beschrijving zou per naam minimaal moeten bestaan uit een jaartal of periode waarin de naam is aangetroffen en wat voor soort naam het is: de naam van een persoon, plaats, bedrijf, schip, vliegtuig, enzovoorts.

17 Voor Indonesië zijn van belang: J. Paulus e.a. (red.), *Encyclopaedie van Nederlandsch-Indië*, 9 delen (Den Haag, Leiden 1917-1939). G.F.E. Gonggryp, *Geïllustreerde encyclopaedie van Nederlandsch-Indië* (Leiden 1934).

Een dergelijke namenlijst zou een belangrijk hulpmiddel vormen bij het lezen en begrijpen van teksten die in het verleden in Nederlands-Indië zijn gedrukt. Namen zijn voor veel geesteswetenschappelijk onderzoek relevant: zo doen letterkundigen en historici veel onderzoek naar het beeld van (fictieve en reële) personen door de tijd heen en de relaties die in het verleden hebben bestaan tussen verschillende personen. De namenlijst kan ook gebruikt worden voor de ontwikkeling van computerprogrammaatjes, zogenaamde tools, die namen in teksten als zodanig automatisch herkennen. Dergelijke Named Entity Recognition-programma's werken op basis van een set welomschreven kenmerken (waaronder spelling: woorden die beginnen met een hoofdletter en niet aan het begin van een zin staan, worden als naam geïnterpreteerd) en de herkenning wordt verbeterd doordat ze leren van trainingsvoorbeelden. Voor het Nederlands is het bekendste programma NERD (Named Entity Recognition for Dutch), ontwikkeld door Bart Desmet in Gent.[18] Een andere standaard, die ook voor Nederlandstalige teksten gebruikt wordt, is de Stanford Named Entity Recognizer (NER).[19]
Er zijn wel initiatieven om een lijst van namen samen te stellen, die de computer kan gebruiken als hulpmiddel bij de naamherkenning. Binnen het project LINKS wordt onder leiding van Kees Mandemakers bijvoorbeeld gewerkt aan een computationeel namenlexicon van 19de- en begin 20ste-eeuwse achternamen in Nederland. Om de programma's voor Named Entity Recognition – waarvan de werking, zeker voor oudere teksten, momenteel meestal matig is – te verbeteren zou een zo omvangrijk mogelijk computationeel namenlexicon moeten worden samengesteld waarin ook namen uit Indonesië een plaatsje krijgen.

Leenwoorden

In het lijstje van vijftig woorden zit een aantal waarvan in één oogopslag duidelijk is dat het een leenwoord uit een Indonesische taal betreft. Zonder nader onderzoek weten we van deze woorden dat ze specifiek zijn voor het Indisch-Nederlands en een plaats verdienen in een historisch-kritisch lexicon van het Indisch-Nederlands. Het gaat om: (*assistent-*)*wedono, krakal, passar, prauwenveer* en *totok*. Dergelijke leenwoorden zijn veelal opgenomen in de bestaande lexica met Indische woorden in het Neder-

18 B. Desmet & V. Hoste, 'Dutch Named Entity Recognition using Classifier Ensembles', in: T. Markus, P. Monachesi & E. Westerhout (eds.), *Computational Linguistics in the Netherlands* (Utrecht 2010): http://lt3.hogent.be/media/uploads/publications/2010/Desmet2010a.pdf.

19 Zie http://nlp.stanford.edu/software/CRF-NER.shtml.

lands.[20] Die lexica vermelden ook de betekenissen: *wedono* is een 'inlands districtshoofd op Java', *krakal* staat voor 'strafarbeid aan de wegenbouw', *passar* is hetzelfde als *pasar* 'markt', *prauwenveer* is een 'veerdienst met prauwen', en *totok* tot slot staat voor een 'volbloed Hollander'.
Ongetwijfeld zijn nog niet alle Indonesische leenwoorden in de bestaande lexica opgenomen, maar op dit terrein is al fors voorwerk gedaan voor de toekomstige samensteller van een historisch-kritisch lexicon van het Indisch-Nederlands.

Potentiële nieuwvormingen binnen het Indisch-Nederlands

De woorden die nu nog overblijven uit het lijstje van vijftig, zijn Nederlandse samenstellingen en afleidingen die, althans in potentie, in het Indisch-Nederlands zijn ontstaan of in die taal hun betekenis hebben gekregen. Het gaat om de volgende woorden: *k.m.p.u., ontvang-avonden, opwekkertjes, pandgoederen, pandgr, pandhouder, postsluiting(en), residentie-huize, scheepsberichten, sluitdagen, spreekdagen, vendulocaal, vendutie(n), vervaldagen, volks-bibliotheek.*
Veel van deze woorden lijken op het eerste oog heel gewone Nederlandse doorzichtige samenstellingen, zonder specifiek Indische connectie. Om erachter te komen of ze daadwerkelijk in het Indisch-Nederlands zijn gevormd of een specifieke betekenis in het Indisch-Nederlands hebben gekregen, moeten we nagaan of ze ook, en al eerder, in het Europees-Nederlands voorkwamen. Omdat we momenteel slechts gegevens van één jaargang kranten hebben, kunnen we hierover geen informatie uit de kranten halen: daarvoor zouden we de woordenschat van meerdere jaargangen moeten vergelijken. Ik heb deze woorden daarom nagezocht in het *Woordenboek der Nederlandse Taal* (WNT), in de Grote Van Dale uit 2005, in het *Viertalig Aanvullend Hulpwoordenboek voor Groot-Nederland* van Prick van Wely uit 1910, en in de zoekmachine Kronos (http://www.zoekkronos.nl/), die de mogelijkheid biedt woorden te zoeken in een groot aantal gedateerde teksten, waaronder de complete Digitale Bibliotheek voor de Nederlandse Letteren (DBNL). Hieronder beschrijf ik de resultaten per woord.

- *K.m.p.u.* is een afkorting van *kilometer per uur*, en daar is niets Indisch-Nederlands aan. Het lijkt toeval dat deze afkorting veel in Indische kranten rond 1900 voorkomt. In andere jaren komt hij ook voor in kranten die in Nederland zijn gedrukt, zo blijkt uit de Historische Kranten van de KB.

20 Opgesomd in noot 6.

- *Ontvang-avonden* wordt door Van Dale omschreven als 'avond waarop men bezoek ontvangt'. Van Dale legt geen connectie met Indonesië, maar die is er wel: het woord is opgenomen door Prick van Wely, die verwijst naar *receptie*, en de oudste citaten die Kronos vermeldt, zijn afkomstig uit overzeese gebieden, vooral Indië maar ook Zuid-Afrika: het gaat om avonden waarop een bestuurder, zoals een resident, bezoek ontvangt.
- *Opwekkertjes* blijkt bij nazoeken in de Historische Kranten de rubrieksnaam voor 'korte grappige anekdotes' in Indische kranten te zijn.
- *Pandgoederen* en *pandgr* (dat laatste blijkt een verkorting te zijn) staat in Van Dale: dit woordenboek omschrijft *pandgoed* als 'goed dat in pand is gegeven'. Het WNT vermeldt *pandgoed* een enkele maal in de 16de eeuw, maar uit Kronos blijkt dat het woord *pandgoederen* omstreeks 1900 voornamelijk voorkomt in bronnen uit Indonesië. Het woord lijkt overzee een nieuw leven te zijn begonnen, nadat het na de 16de eeuw uit het Nederlands was verdwenen.
- *Pandhouder* wordt door Van Dale omschreven als 'pandnemer, pandbezitter'. WNT en Kronos geven bronnen uit Nederland en elders. Dit woord lijkt geen Indische connectie te hebben.
- *Postsluiting(en)* wordt door Van Dale omschreven als 'uiterste tijdstip waarop nog post voor een bep. mail aangenomen wordt'. Dat *mail* suggereert een overzeese connectie, want in het verleden betekende *mail* 'brievenpost van en naar overzeese gebieden'. In het WNT komt het woord niet voor en de paar citaten uit Kronos zijn uit Indië en Suriname. Het woord is dus typisch voor overzeese gebieden, al staat dat niet expliciet in Van Dale.
- *Residentie-huize* (de verbogen naamval komt doordat het in de kranten telkens voorkomt in de vaste verbinding *ten residentie-huize*) is uiteraard een woord dat in Indonesië is ontstaan, want *resident* is de titel van de gewestelijke bestuurshoofden in Indonesië.
- *Scheepsberichten* wordt door Van Dale omschreven als '(rubriek met) berichten over varende schepen'. Uit Kronos blijkt dat het woord omstreeks 1900 overzees, met name in Indonesië, gebruikt werd en daar dus zal zijn ontstaan.
- *Sluitdagen* staat niet in Van Dale of het WNT, Kronos vermeldt twee citaten waarin het iets anders betekent. Uit de Historische Kranten blijkt dat het gaat om dagen waarop men geen post per schip kon versturen. Dit woord is kennelijk in Indonesië ontstaan en het interessante ervan is dat de woordvorming afwijkt van die van het Europees-Nederlands: daarin is het immers *sluitingsdagen*.
- *Spreekdagen* staat niet in Van Dale of WNT. Kronos helpt ook niet. Uit de Historische Kranten blijkt dat het gaat om de dagen waarop men een onderhoud kon hebben met een bestuurder (resident, gouverneur, secretaris) in Indonesië.

- *Vendulocaal* of *vendulokaal* blijkt (WNT, Prick van Wely) een Indisch-Nederlands woord voor 'veilingzaal': veilingen werden vaak gehouden in Indonesië omdat veel tijdelijk gevestigde ambtenaren hun huisraad kwijt moesten als ze teruggingen naar Nederland. Zie ook het volgende woord.
- *Vendutie(n)* beschrijft Van Dale als: 'vendu (m.n. in Indië wegens vertrek naar elders)'. En het WNT meldt: 'Blijkbaar een in de 17de e. in O. -Indië ontstane vervorming van ouder *venditie* (zie ald.), onder invloed van *vendu*. De uitspraak [ven'dy(t)si] is ontstaan onder invloed van de spelling. In Z. Afrika zijn beide vormen, ouder *vandisie* en jonger *vendusie*, bewaard.'
- *Vervaldagen* zijn volgens Van Dale data waarop een wissel vervalt. In de Historische Kranten uit 1900 wordt *vervaldagen* echter uitsluitend gebruikt in de vaste combinatie *vervaldagen van vendutiën*, dus in de betekenis 'dagen dat er geen openbare veilingen worden gehouden, dat de openbare veiling vervalt'. Die combinatie is kennelijk in het Indisch-Nederlands ontstaan.
- *Volks-bibliotheek* als 'voor het grote publiek bestemde bibliotheek' heeft volgens de verschillende bronnen geen specifiek Indische connectie.

Van de veertien woorden blijken er maar liefst elf te zijn ontstaan in Indonesië of daar een specifieke betekenis te hebben aangenomen: *ontvangavonden, opwekkertjes, pandgoederen, postsluiting(en), residentie-huize, scheepsberichten, sluitdagen, spreekdagen, vendulocaal, vendutie(n)* en *vervaldagen*. Morfologisch onderscheiden deze woorden zich niet van Europees-Nederlandse woorden, behalve *sluitdagen*. Het zijn allemaal samenstellingen of afleidingen, behalve *vendutie*, een leenwoord waarvan in Indonesië de klank is veranderd. De meeste staan in de woordenboeken niet als Indisch-Nederlandse woorden te boek, terwijl ze dat wel degelijk blijken te zijn. De betekenisomschrijvingen zijn in de woordenboeken onvolledig. Uit deze kleine terreinverkenning blijkt dat door corpusonderzoek nieuwgevormde Nederlandse woorden en betekenissen in Indonesië aan het licht kunnen komen. De woorden die we hebben gevonden, zijn qua vorming of betekenis niet heel verrassend of veelzeggend voor het Indisch-Nederlands. Ik heb daarom ook gekeken naar minder frequente woorden uit de kranten, om te zien of daaronder interessante woordvormingen zitten. Dat wordt wel steeds lastiger: naarmate de frequentie van woordvormen afneemt, neemt het percentage fout gelezen woorden toe. Toch vond ik onder de woorden die meer dan 20 keer in de kranten voorkomen, nog enkele interessante gevallen die lexicografen tot nu toe zijn ontgaan, zoals *afpakschuur* 'schuur waarin producten definitief worden ingepakt', *damespraatjes* (blijkt een rubriek in de krant), *getelefoond* ('voortaan zou ons het belangrijkste nieuws worden getelefoond'), *misdrijfzaak*

'misdrijf', *ondercollecteur* 'bepaalde rang bij belastinginners', *vaarbeurten* 'invalreis van een schip voor een ander', en *verponding* 'grondbelasting', met de samenstellingen *verpondingsbelasting* en *verpondingswaarde* (*verponding* bestond eerder in het Europees-Nederlands maar was daaruit inmiddels verdwenen; het woord kreeg een nieuwe betekenis in Indonesië).
Ook komen onder deze minder frequente gevallen de woorden *breidel* 'censuur' en *breidelen* 'censureren' voor, waarvan bekend is dat ze in Indonesië zijn gevormd. Maar in kranten uit 1900 vinden we geen vermelding van bijvoorbeeld *voorkinderen* 'kinderen uit een (eerdere) relatie met een inheemse concubine' of *haatzaai(-artikelen), haatzaaien*, waarvan bekend is dat ze Indisch zijn. *Haatzaai-artikelen* zijn in 1914 opgenomen in het Wetboek van Strafrecht voor Nederlandsch-Indië.[21] Vanaf 1893 komt de combinatie *haat zaaien* in verschillende contexten in kranten voor.
Ook ontbreken in de kranten van 1900 woorden als *knoerthard, knotsgek, reuzeleuk, piepklein* en *tjokvol*, waarvoor ik tot nu toe als oudste datering een bron in Indonesië heb gevonden, wat suggereert dat het neologismen zijn die daar zijn ontstaan.[22] Maar dat moet nog maar bewezen worden. Duidelijk is in ieder geval dat voor corpusonderzoek één jaargang van een krant onvoldoende is. De krantencorpora zouden bovendien moeten worden aangevuld met corpora van in Indonesië gepubliceerde romans, non-fictie, brieven, dagboeken etc., zodat er zoveel mogelijk gevarieerd taalgebruik beschikbaar komt.

Tabee

Wil men systematisch nieuw materiaal verzamelen voor een historisch-kritisch woordenboek van het Indisch-Nederlands, dan lijkt corpusonderzoek de aangewezen methode. De kranten die we momenteel tot onze beschikking hebben, blijken geen ideale bron, vanwege de vele OCR-leesfouten en doordat namen niet direct herkenbaar zijn. Dat laatste kunnen we binnen Nederlab in een volgend stadium verbeteren. Ook is het de moeite waard na te gaan in hoeverre het mogelijk is automatisch leesfouten in de OCR te ondervangen. Want zelfs uit deze kleine test bleek dat de kranten interessante nieuwe gegevens opleveren. Door correctie van OCR-fouten zal het aantal hapaxen (eenmaal voorkomende woorden), dat momenteel in de kranten op tachtig procent ligt, dichter bij het gemiddelde van vijftig procent komen.

21 G.C. Haverkate, 'Haatzaaien', *Strafblad* mei 2013, 163-164. N. van der Sijs, *Klein uitleenwoordenboek* (Den Haag 2006).

22 N. van der Sijs, 'Het ongezochte vinden', *NRC Handelsblad*, Wetenschapsbijlage 13 oktober (2012) p. 2.

Hoe kan de samenstelling van een historisch-kritisch lexicon het best worden aangepakt? Uiteraard dient te worden voortgebouwd op bestaande kennis. Dat betekent dat de bestaande woordenboeken over het Indisch-Nederlands[23] een beginpunt kunnen vormen. Daaraan kunnen namen en eventueel andere relevante informatie uit encyclopedieën worden toegevoegd.[24] Op basis van omvangrijke corpora kan de computer woorden en betekenissen verzamelen die typerend lijken voor het Nederlands in Indonesië. Ook woordcombinaties die specifiek zijn voor Indonesië kunnen zo worden herkend. Een voorbeeld daarvan uit de kranten uit 1900 is de constructie *aangeslagen vendutiën*, die kennelijk werd gebruikt voor 'bekend gemaakte, geafficheerde veilingen'.
Alle door de computer geleverde gegevens moeten door een competente onderzoeker op hun waarde worden geschat. Ik hoop natuurlijk dat Reinier Salverda, die in de gelukkige positie is dat hij na zijn pensionering meer tijd zal hebben, dit *handschoentje* (pun intended) oppakt. Het is veel werk, maar het resultaat zal er ook naar zijn. Een complete beschrijving van het Indisch-Nederlandse lexicon vormt niet alleen een welkome en noodzakelijke aanvulling op de lexicografie van het Nederlands, we kunnen er ook uit leren welke semantische, morfologische, fonologische en syntactische kenmerken typerend zijn voor het Indisch-Nederlands.
Tjoba, Reinier.

23 Zie noot 6 en 7.
24 Zie noot 17.