http://lexikos.journals.ac.za

Moroccorp: tien miljoen woorden uit twee Marokkaans-Nederlandse chatkanalen

Tom Ruette, *KU Leuven, Leuven, België en Humboldt-Universität zu Berlin, Berlijn, Duitsland (tom.ruette@hu-berlin.de)*
en
Freek Van de Velde, *KU Leuven, Leuven, België en FWO-Vlaanderen, België (freek.vandevelde@arts.kuleuven.be)*

Samenvatting: In dit artikel stellen we een nieuw corpus voor van computer-gemedieerde communicatie in het Nederlands door Marokkaans-Nederlandse taalgebruikers, dat bestaat uit tien miljoen woorden chat-materiaal. We behandelen de achtergrond, de compilatiemethode en de interne structuur van het corpus, en we leggen het verband tussen ons eigen werk en eerdere pogingen om een corpus van Nederlandse chattaal te bouwen. We hebben ook een *Stable Lexical Marker analyse* uitgevoerd en een gevalstudie over een welbekende morfosyntactische eigenschap van het Marokkaans Nederlands om op die manier de representativiteit van het corpus te beoordelen.

Trefwoorden: NEDERLANDS, MAROKKAANS NEDERLANDS, CORPUS, CHAT, STABLE LEXICAL MARKER ANALYSIS, ETNISCH NEDERLANDS, REPRESENTATIVITEIT

Abstract: Moroccorp: Ten Million Words from two Moroccan Dutch Chat Channels. In this article we introduce a new corpus of computer-mediated communication in Dutch by Maroccan-Dutch language users, consisting of ten million words of chat material. We treat the background, the compilation method and the inner structure of the corpus, and we relate our efforts to previous attempts to build a corpus of Dutch chat language. We also conducted a *Stable Lexical Marker analysis* and a case study on a well-known morphosyntactic feature of Moroccan Dutch to assess in this manner the representativity of the corpus.

Keywords: DUTCH, MOROCCAN DUTCH, CORPUS, CHAT, STABLE LEXICAL MARKER ANALYSIS, ETHNIC DUTCH, REPRESENTATIVITY

1. Inleiding[1]

De studie van 'taalcontact' mag zich al een tijdje verheugen in een groeiende interesse onder taalkundigen, zoals blijkt uit het bestaan van een eigen tijdschrift (*Journal of Language Contact*, Brill) en tal van inleidingen bij gereputeerde uitgevers zoals die van Thomason (2001) bij Edinburgh University Press, die van Matras (2009) bij Cambridge University Press of die van Hickey (2010) bij

Blackwell. De toename is voor een deel toe te schrijven aan de beschikbaarheid van dataverzamelingen, zoals bijvoorbeeld transcripties van taalproductie van twee- of meertalige kinderen in de Childes database (MacWhinney 2000), of multilinguale corpora (zie Xiao 2008 voor een overzicht). De studie van taalcontact bevindt zich op het snijpunt van de sociolinguïstiek, taalverwerving, twee-/meertaligheid en de historische taalkunde, en er wordt vrij breed aangenomen dat taalcontact een uiterst belangrijke aanstoker of determinant is in taalverandering (zie Weinreich 1953; Thomason en Kaufman 1988; Harris en Campbell 1995; Croft 2000; Heine en Kuteva 2003, 2005; Drinka 2010).

Ook voor het Nederlands is er intensief onderzoek verricht naar het effect van taalcontact. De Lage Landen zijn al eeuwenlang een gebied waarin sprekers van verschillende talen met elkaar in contact komen. Meer zelfs: het uitzicht van het Nederlands is in grote mate een resultante van taalcontact (Buccini 1995, 2010), en dat geldt eigenlijk voor de Germaanse taalfamilie in het algemeen (Hawkins 1990: 60-61; Roberge 2010). Die impact van vreemde elementen is er in recente jaren niet minder op geworden. De demografische samenstelling in sommige Nederlandse en Vlaamse steden zoals Rotterdam en Genk laat diepe sporen na in het Nederlands dat daar gesproken wordt (Cornips en de Rooij 2003).

Ook vanuit de sociolinguïstiek is er veel aandacht voor variëteiten of 'lecten' die zich niet of niet alleen op een geografische dimensie laten definiëren, maar die in belangrijke mate gedefinieerd worden volgens de sociaal-culturele dimensie. Voor het Nederlands valt onder meer te denken aan 'polderneder-lands', 'tussentaal', 'straattaal' en 'cités'. In die variëteiten zit vaak wel een geografische component: poldernederlands (Stroop 1998) wordt gesproken in Nederland, tussentaal (o.a. Geeraerts 2002) wordt gesproken in Vlaanderen en heeft een Brabantse invloed (pace Taeldeman 2008), cités (Ramaekers 1998) is beperkt tot Genk en ook straattaal (Appel 1999) is regiogebonden. Maar een goed begrip van deze variëteiten kan niet om de sociale dimensie heen: poldernederlands wordt of werd in eerste instantie geproken door jonge, hoogopgeleide vrouwen (Stroop 1998), en de sociale stratificatie van de tussentaal is uitvoerig onderzocht in Plevoets (2008). Voor cités en straattaal geldt de (groot)stedelijke context als bepalend, en een van de factoren die een grote impact hebben op die context is de aanwezigheid van sprekers die het Nederlands niet als moedertaal hebben. Op die manier ondergaat het Nederlands invloeden van buitenaf en de studie van deze hedendaagse lecten (cf. Ruette et al. 2014) kan dan ook een bijdrage leveren aan de historische taalkunde, net zoals omgekeerd de historische taalkunde inzichten kan bijbrengen over de langetermijneffecten van zulke vormen van beïnvloeding.

De toenemende belangstelling voor taalcontact en lectale variatie en de groeiende technologische mogelijkheden binnen de Digital Humanities, hebben ertoe geleid dat er corpora aangelegd zijn van wat met een enigszins beladen term wel eens 'etnisch Nederlands' genoemd wordt. In dit artikel willen we het fonds beschikbaar onderzoeksmateriaal uitbreiden door een nieuw corpus voor

te stellen en publiek beschikbaar te maken voor onderzoek binnen de historische taalkunde, de taalverwerving, de sociolinguïstiek en de studie van taalcontact. Het gaat om een Nederlands chattaal-corpus van 10.000.000 woorden geschreven door bezoekers van chatkanalen die zich richten op Nederlanders met een Marokkaanse achtergrond, wier taalgebruik zich laat omschrijven als een etnolectische variëteit die in de literatuur vaak met de term 'Marokkaans Nederlands' of 'Moroccan flavored Dutch' aangeduid wordt (zie Nortier en Dorleijn 2008). Dit corpus hebben we de naam 'Moroccorp' gegeven.

Op de taalkundige en ideologische problemen bij het afzonderen van zo'n etnolect komen we terug in Sectie 2.3. We willen er hier wel al op wijzen dat we de term louter descriptief of 'fenomenologisch' opvatten, zonder meteen ook ontologische uitspraken te doen over de status van het etnolect.

Het aanleggen van een corpus van Marokkaans Nederlands is geen nieuw idee, maar het bestaande arsenaal aan ruwe onderzoeksdata is aan uitbreiding toe. Sommige van de 'etnische Nederlandse' corpora zijn particuliere initiatieven en zijn niet vrij beschikbaar (b.v. het corpus van Jaspers 2004). Andere kunnen wel vrij geconsulteerd worden. Zo is er de Dutch Bilingualism Database (DBD), die vrij beschikbaar is via de website van het Max Planck Institut (MPI) in Nijmegen[2] en een sectie Marokkaans Nederlands bevat. De opnames in de DBD betreffen gesproken Nederlands, wat aan de ene kant fijnmazige analyse toelaat, ondermeer op het klankniveau, maar wat aan de andere kant praktische beperkingen heeft opgelegd aan de omvang van het verzamelde materiaal. De precieze omvang in aantal tokens van het Marokkaans-Nederlandse DBD-corpus is niet bekend, maar beslaat ongeveer 30 uur aan opnames. Het Moroccorp is in vergelijking veel groter, aangezien het chattaal van meerdere maanden omvat.

Het ontwikkelen van het Moroccorp is een inspanning om *Kommunikationserreignisse* (Zeige Te versch.) te verzamelen die de gevolgen van taalcontact tussen Nederlands en Marokkaans of Berbers weerspiegelen. Het Moroccorp doet dit door een substantiële verzameling aan te leggen van Marokkaans Nederlands — dat wil zeggen: taalproductie door Nederlandstaligen met een Marokkaanse achtergrond — waarin redactionele inmenging zo klein mogelijk gehouden is. Zoals in Sectie 2 in detail beschreven wordt, hebben we online materiaal vergaard uit chatberichten. Het idee om chattaal op internet te gebruiken als 'proxy' voor spontane en informele taalproductie met een laag redactioneel gehalte, is niet nieuw. In Androutsopoulos (2006) wordt een overzicht gegeven van onderzoek naar zogenoemde computer-gemedieerde communicatie, waaronder ook chattaal valt. Het idee om chattaal te verzamelen is voor het Nederlands eerder al gevolgd in het ConDiv-corpus (zie Grondelaers et al. 2000) en in het chatcorpus van Vandekerckhove (zie Vandekerckhove en Nobels 2010). We willen hierbij aanmerken dat een chatcorpus uiteraard slechts een partiële relevantie voor het taalkundige onderzoek heeft. Een belangrijke beperking volgt uit het feit dat enkel het taalgebruik van een kleine groep mensen in één zeer specifiek register gedocumenteerd wordt, waardoor de

gevonden resultaten niet zonder meer veralgemeend kunnen worden. Hierdoor is er vooral aandacht gegaan naar specifieke kenmerken van zogenoemde *internet-taal* (Crystal 2011), vanuit het perspectief van de conversatie-analyse (Herring 2010) of van taalverloedering (Tagliamonte en Denis 2008). Het zou anderzijds te ver gaan om chattaal zondermeer als onbruikbaar ter zijde te schuiven. Voor het Marokkaans Nederlands is echter de bruikbaarheid van internetdata onder de aandacht gebracht door Boumans (2002), zij het met een veel beperkter corpus (ca. 40.000 woorden).

2. Samenstelling van het corpus

Het samenstellen van een corpus waarin de taal van één of meerdere chatkanalen wordt verzameld vereist een aantal weloverwogen methodologische keuzes. We bespreken de genomen stappen kort in Sectie 2.1. In die sectie beargumenteren we ook de stelling dat het door ons verzamelde taalmateriaal representatief is voor het Nederlands van taalgebruikers met een Marokkaanse achtergrond in openbare chatkanalen. Daarna, in Sectie 2.2, geven we een korte kwantitatieve inkijk in het corpus door een overzicht te geven van het aantal woorden, het aantal gebruikers, etc. Tot slot is het noodzakelijk om in Sectie 2.3 enkele deontologische afwegingen te maken.

2.1 Methodologie

Technisch gezien is het niet zo moeilijk om een corpus te maken op basis van de conversaties in een chatkanaal. De meeste chatprogramma's voorzien de functionaliteit om een log bij te houden van het kanaal waarop men aangemeld is. Met behulp van die functionaliteit is het mogelijk om snel en eenvoudig relatief grote hoeveelheden chatconversaties te verzamelen. Voor het Moroccorp maakten we gebruik van het open-source programma *irssi*.[3] Deze methode wordt algemeen gevolgd bij het maken van chatcorpora, zoals bijvoorbeeld in het *Dortmunder Chat Corpus*.

De chatconversaties in het Moroccorp zijn afkomstig uit twee chatkanalen. Het eerste chatkanaal heet *#maroc* en is bereikbaar via de server *irc.marocchat.net*. Het tweede chatkanaal heet *#maroc.nl* en is bereikbaar via de server *irc.scarynet.org*. Met behulp van de automatische logfunctionaliteit in het gebruikte chatprogramma konden we de conversaties in deze chatkanalen gemakkelijk opslaan. Het Moroccorp is gebaseerd op de logs van beide chatkanalen uit de zomer van 2012 (van mei 2012 tot en met september 2012). In mei en in juni werd het eerste chatkanaal gelogd, en van juli tot en met september werd het tweede chatkanaal gelogd. Onze logactiviteiten werden geregeld onderbroken uit technische overwegingen en om ervoor te zorgen dat onze passieve aanwezigheid op het chatkanaal niet zou opvallen — meer hierover in Sectie 2.3.

Beide chatkanalen hebben het Nederlands als voertaal, en richten zich

tegelijkertijd tot een publiek met Marokkaanse achtergrond. Hoewel de chatkanalen in principe ook toegankelijk zijn voor niet-Marokkaanse Nederlanders kunnen we er door de specificiteit van de chatkanalen van uitgaan dat het leeuwendeel van de chatters van Marokkaanse origine is. Voor die assumptie zijn er enkele kwalitatieve aanwijzingen. Zo verwijzen vele gebruikers in hun gebruikersnaam naar een Marokkaanse identiteit, bv. "vrouwuitmarokka", "ZeTLa-Maroc", "Femmedumaroc", "safouan_maroc", "maroc-meid". Daarnaast bevatten bijna 22.000 lijnen uit het chatcorpus (ongeveer 1% van alle lijnen in het corpus) een referentie naar Marokko, wat aantoont dat dit een belangrijk onderwerp is voor de chatters. Principieel sluit dit niet uit dat er hier gechat wordt over, zeg maar, Marokko als vakantieland, maar zelfs een eenvoudige vluchtige kijk in het Moroccorp doet anders vermoeden. Verdere aanwijzingen dat het hier gaat over het Nederlands van taalgebruikers met een Marokkaanse achtergrond halen we uit de lexicale analyse en de morfosyntactische analyse in Sectie 3. Uiteraard is het niet uit te sluiten dat er ook enkele niet-Marokkaanse taalgebruikers op het (openbare) chatkanaal actief zijn. Een aanwijzing hiervoor vinden we in Fragment 1, waarin chatter "TurksDraakje31" in zijn pseudoniem een expliciet niet-Marokkaanse achtergrond uitdrukt. Dat zien we niet als een probleem, aangezien een zekere hoeveelheid ruis in elk corpus onvermijdbaar is. Zo bevatten bijvoorbeeld ook klassieke krantencorpora vaak niet-journalistieke teksten: reclame, lezersbrieven, aankondigingen enzovoort. In feite is een belangrijk doel van de taalkundige analyse in Sectie 3 het aantonen dat deze hoeveelheid ruis relatief beperkt is.

Daarnaast blijken beide chatkanalen opgebouwd te zijn rond een overzichtelijke groep van kernleden. Dit zorgt ervoor dat de conversatie bijna nooit chaotisch wordt — dus zonder verschillende gesprekken die door elkaar gevoerd worden. Dat is namelijk vaak een probleem in openbare chatkanalen, waar soms heel veel mensen tegelijkertijd chatten (Herring 2010). Echter, de vrij gebalanceerde tussenvorm in de door ons verzamelde chatkanalen, met nooit al te veel chatters tegelijkertijd en een goed te volgen conversatie, lijkt vanuit linguïstisch oogpunt bijzonder attractief: er is voldoende animositeit, waardoor de snelheid van communiceren (zoals in een druk chatkanaal) zijn invloed kan hebben op het taalgebruik, maar tegelijkertijd is de groep chatters niet zo groot dat er geen coherent 'gesprek' gevoerd kan worden. Dat maakt het Moroccorp, los van zijn mogelijke etnolinguïstische waarde, ook interessant voor conversatie-analyse.

Voorts zullen we tegen het einde van het artikel durven te beweren dat het Moroccorp representatief is voor Nederlandse chattaal van taalgebruikers met een Marokkaanse achtergrond (zoals die gesproken werd in de zomer van 2012), omdat dit — zover ons bekend is — de enige twee chatkanalen zijn waar de deelnemers overwegend van Marokkaanse origine zijn en in het Nederlands chatten. Dit neemt uiteraard niet weg dat er een kans bestaat dat er toch nog andere, misschien private chatkanalen zijn waarop Marokkaanse Nederlanders converseren. Desalniettemin lijkt ons de kwantiteit van tien miljoen woorden

chattaal gedurende een beperkte periode een aanwijzing voor de grote populariteit van de door ons opgenomen chatkanalen. Het is echter wel zo dat we enkel toegang hebben tot de conversaties in het publieke gedeelte van het chatkanaal. Beide chatkanalen geven de gebruikers immers ook de mogelijkheid om private gesprekken te voeren in een ad-hoc sub-chatkanaal. Uiteraard zijn die private gesprekken niet vrij toegankelijk en ze zijn ook niet opgenomen in het Moroccorp. Daarom is het goed om toe te voegen dat het Moroccorp enkel chattaal bevat van taalgebruikers die zich ervan bewust zijn dat ze zich op een publiek forum begeven. Tot slot moeten we ook nog opmerken dat het bij chatcorpora doorgaans onmogelijk is om precieze socio-demografische gegevens van de chatters te achterhalen. Zo kunnen we niet met zekerheid weten wat het sociale profiel is dat het Moroccorp beschrijft, inclusief de etniciteit van de chatters. Daarom wijdt dit artikel hieronder veel plaats aan het aannemelijk maken van de Marokkaanse etniciteit van de meerderheid van de chatters. Ondanks deze beperkingen, zijn we er van overtuigd dat het Moroccorp een nuttige bron voor het taalonderzoek kan zijn.

2.2 Enkele cijfers

In deze paragraaf belichten we het corpus van een technische en kwantitatieve kant.

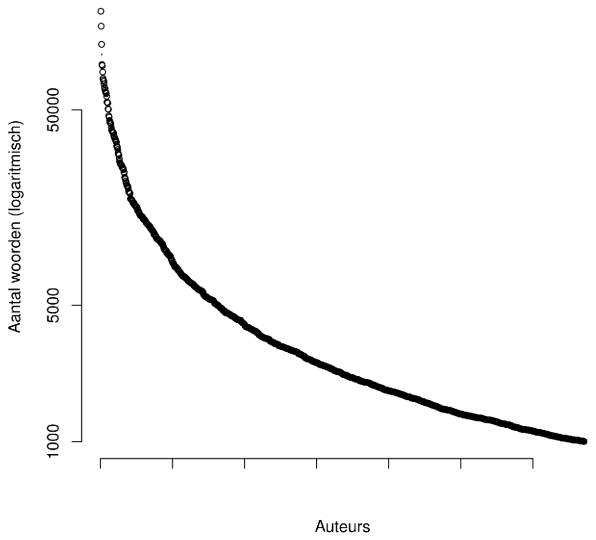
Tabel 1: Hoeveelheid woorden in het Moroccorp

chatkanaal	**log**	**opgeschoond**
marocchat.net:#maroc	5.542.949	2.739.330
scarynet.org:#maroc.nl	19.310.048	7.533.347
Totaal	24.852.997	10.272.677

Het volledige corpus zoals het werd gelogd in het chatprogramma telt bijna vijfentwintig miljoen woorden.[4] Van het eerste chatkanaal hebben we vijf miljoen woorden in de logs, en van het tweede chatkanaal hebben we negentien miljoen woorden gelogd. Echter, de ruwe logs van het chatprogramma vereisen nog enkele opruimactiviteiten om de chatlogs in een bruikbaar corpus te veranderen. Ten eerste willen we boodschappen verwijderen die niet door mensen, maar door de computer gegenereerd zijn. Het is namelijk zo dat de software die nodig is om een chatkanaal uit te baten een heleboel automatische aankondigingen genereert. Daarnaast hebben de gebruikers van een chatkanaal de mogelijkheid om met behulp van zogenaamd *bots*[5] de functionaliteit van het kanaal te vergroten. Deze teksten zijn formeel gemakkelijk te herkennen en kunnen dan ook eenvoudig en automatisch weggefilterd worden. Ten tweede bevat de ruwe chatlog informatie over de server waarop het chatkanaal zich

baseert, of over het chatprogramma waarmee een gebruiker zich heeft aangemeld, of de tijd waarop een berichtje de wereld werd ingestuurd. Deze metainformatie wordt verwijderd zodat het corpus uiteindelijk regels bevat volgens het formaat *auteur {tab} boodschap*. Na deze opruimactiviteiten is het aantal woorden in het volledige corpus nog ruim tien miljoen (10.272.677). Met deze hoeveelheid woorden is het Moroccorp naar huidige standaarden niet zo groot, maar te vergelijken qua grootte met het Corpus Gesproken Nederlands (Nederlandse Taalunie 2004).

Figuur 1: Aantal woorden per auteur (enkel auteurs met meer dan 1000 woorden zijn hier afgebeeld)



Niet alle gebruikers dragen evenveel woorden bij aan het corpus, zoals we zien in Figuur 1. De grafiek uit Figuur 1 wordt iets tastbaarder als we de twintig *nicknames* (een soort van schuilnaam) die het grootste aantal woorden bij elkaar gechat hebben presenteren in Tabel 2. Het blijkt inderdaad dat er een beperkt aantal gebruikers zijn die een belangrijk deel van het chatcorpus vullen. De meest actieve gebruiker produceert maar liefst drie keer zoveel woorden als het nummer twintig van de productiefste chatters. Dat is in principe geen probleem voor de validiteit van het Moroccorp, maar wel voor de representativiteit: de gebruikers van het corpus kunnen niet zonder meer de bevindingen veralgemenen, en bij een kwantitatieve studie met inferentiële statistiek wordt het aangeraden om gebruik te maken van een *random effect* dat de specificiteit van iedere chatter ondervangt.

Tabel 2: De twintig meest productieve chatters

Nickname	**Aantal woorden**
Manal	160.258
koekje	134.301
Aketoef	108.481
lady_mamita	85.482
kipsate	84.435
groentebol	78.249
KYRA	72.585
nourdine	71.264
meisjepraatveel	70.069
awayagher	67.956
BijzondereMan	64.982
Zorro	64.620
Kanjer_	62.979
nabil_1984	61.912
Okegoed	61.422
Moslima21	60.688
TheMo	58.834
fessiaa	54.792
Fir3auwn	54.650
Dora	54.263

2.3 Deontologische afwegingen

Het compileren van een corpus van chatconversaties vereist dat de onderzoeker zich aanmeldt op het chatkanaal en zonder inmenging in de conversatie taalmateriaal verzamelt, teneinde de *observer's paradox* (Labov 1966) te ondervangen. Hoewel deze houding sociolinguïstisch methodologisch gebruikelijk is, kan men zich hierbij deontologische vragen stellen. In hoeverre is het ongemerkt 'opnemen' van spontaan taalgebruik zonder dat de taalgebruiker hiervan op de hoogte is aanvaardbaar? Daarnaast wordt het 'idlen' op een chatkanaal — aangemeld zijn zonder participatie — niet geapprecieerd.

Ter illustratie kunnen we een stukje chatconversatie weergeven van het moment waarop ontdekt wordt dat we al enkele dagen 'idlen'. Dit stukje is uiteraard ook opgenomen in het Moroccorp.

Fragment 1: Corpusmakers betrapt op chatloggen.

```
_adil_relax  hij is de mol
...
_adil_relax  undercover onderzoeker
...
_adil_relax  kulnet.kuleuven.be
...
TurksDraakje31 denk dat ie zijn laptop vergete is dicht te klappe
TurksDraakje31 heb ik soms ook wel;)
...
_adil_relax  4 dagen vriend
_adil_relax  bijna 5
TurksDraakje31 dalijk is ie dood ofso
_adil_relax  zware chatlogger is dat
...
_adil_relax  datamining
_adil_relax  waar hebben we het over
_adil_relax  ze zijn bezig met onderzoek
TurksDraakje31 weet ik veel
...
TurksDraakje31 zwaai dan ff
TurksDraakje31 misgien kom je in het nieuws:D
...
_adil_relax  onderzoeker
_adil_relax  undercover
...
_adil_relax  research
...
_adil_relax  ga een tagcloud maken
TurksDraakje31 ja maar wrm kicke hem dan niet
...
_adil_relax  "meest getype woorden lijst"
...
_adil_relax  hahaha allemaal scheldwoorden
```

Uit dit fragment uit het eerste chatkanaal blijkt dat de chatters bekend zijn met het fenomeen chatloggen. Er ontstaat geen agitatie noch verontwaardiging, maar een zekere gelatenheid. De enige consequentie is dat onze aanwezigheid op het chatkanaal niet langer gewenst en toegelaten is. In Grondelaers et al. (2000) wordt daarom geadviseerd om op voorhand contact op te nemen met een verantwoordelijke van het chatkanaal. Wij hebben deze stap niet ondernomen omdat een chatkanaal per definitie openbaar is en het principieel niet nodig is om toestemming te vragen. Bovendien bestaat er duidelijkheid bij de chatters over de vrije toegankelijkheid van de chatkanalen. Het tweede chatkanaal plaatst zelfs een permanente boodschap bovenaan in het chatprogramma waarin wordt gewaarschuwd voor het ontbreken van privacy. Daarnaast hebben we geen enkele poging ondernomen om de chatconversaties te registreren die in een expliciet private context gevoerd werden. Aangezien de chatters hun eigen anonimiteit waarborgen door het gebruik van een pseudoniem, hebben we ook geen stappen ondernomen om het Moroccorp verder te anonimiseren.

Een andere deontologische overweging is dat etnolectisch corpusonderzoek een aantal risico's inhoudt, die gemakkelijk uit het oog verloren kunnen worden bij een (in het algemeen behartenswaardige) onbevangen aanpak. Dat heeft hiermee te maken dat het afgrenzen van etnolecten culturele implicaties heeft. Clyne (2000: 86) definieert etnolecten als "varieties of a language that mark speakers as members of ethnic groups who originally used another language or distinctive variety", en Androutsopoulos (2001: 2) hanteert de volgende definitie: "a variety of the majority language (or 'host language') which is used by and regarded as a vernacular for speakers of a particular ethnic descent and is marked by certain contact phenomena". Die definities zien er op het eerste gezicht vrij onschuldig uit, maar o.a. Jaspers (2008) wijst erop dat het afzonderen van een etnolect ideologische consequenties kan hebben door er al te vanzelfsprekend van uit te gaan dat taalgebruik in de eerste plaats aangestuurd wordt door etnie, en daardoor een onverantwoord homogeniserend, essentialistisch beeld op kan hangen van de sprekers, die daar als culturele minderheid hinder van kunnen ondervinden. Dat is natuurlijk niet onze bedoeling met het aanleggen van een corpus Marokkaans Nederlands, en we moeten bedacht zijn op dergelijke effecten. Toch kan geredelijk betwijfeld worden of we ons met het aanleggen van een corpus van een Marokkaans Nederlands etnolect schuldig maken aan cultureel essentialisme. Uit eerder onderzoek blijkt immers dat er zich in de taalfeiten wel een variëteit Marokkaans Nederlands laat onderscheiden (zie Nortier en Dorleijn 2008), en het lijkt ons interessant om collegae-onderzoekers de mogelijkheid te bieden die variëteit ook verder te onderzoeken in een wat groter tekstbestand (iets waar die collega's ook expliciet om vragen, zie Nortier en Dorleijn 2008: 140 en Hinskens 2011: 126). In principe biedt ons corpus ook de mogelijkheid om met taaldata in de hand de constructie van zo'n etnolect te falsifiëren. Als de eerder vastgestelde patronen in het taalgedrag van de sprekers van het etnolect niet stabiel zijn, of aanzienlijke variatie vertonen, of niet uniek zijn voor het etnolect, dan kan dat leiden tot een gemotiveerde revisie van de bestaande afbakening, maar niemand is gebaat bij een gebrek aan data. Ons corpus heeft alvast het voordeel dat het omvangrijker is dan wat er totnogtoe publiekelijk beschikbaar is. Met het gebruik van de term 'Marokkaans Nederlands' (cf. infra) en de naam van het corpus ('Moroccorp') hebben we geen ontologische pretenties, en we willen er in dit artikel verder ook uitdrukkelijk op wijzen dat het er ons niet om te doen is de variëteit in kwestie te bestempelen als 'onvolmaakt' of 'slecht' Nederlands — als die termen überhaupt al bruikbaar zouden zijn. Heel vaak zijn sprekers van een etnolect trouwens goed in staat de standaardtaal te spreken, maar zien ze daar in sommige contexten bewust van af (Hinskens 2011: 104). Trouwens, ook het standaardnederlands kan als een selectieve variëteit beschouwd worden, die geen usurperende claim kan leggen op het predicaat 'hét Nederlands'. Bovendien vestigen we nog even de nadruk op het feit dat wat standaardnederlands genoemd wordt, zelf de resultante is van verschillende periodes van diepgaand taalcontact in het verleden (zie Sectie 1, Inleiding), zodat je zou kunnen zeggen

dat elke variëteit van het Nederlands historisch gezien een etnolect is.

3. Hoe 'Marokkaans' is het Nederlands in het Moroccorp?

In deze sectie gaan we na hoe 'Marokkaans' het Nederlands is dat verzameld werd in het Moroccorp. Hiervoor zullen we twee taalkundige indicatoren gebruiken, namelijk een lexicale analyse en een morfologische analyse. Tegelijkertijd kunnen deze indicatoren aantonen hoe bruikbaar het Moroccorp is voor het taalkundige onderzoek.

Alvorens we tot de twee kleine gevalstudies kunnen overgaan moeten we bondig het Nederlands van taalgebruikers met een Marokkaanse achtergrond bespreken. Hoewel leden van de Marokkaanse gemeenschap in Nederland een verschillende moedertaalachtergrond hebben (Berber of Arabisch), en er uiteraard grote linguïstische verschillen zijn tussen eerste-, tweede- en ondertussen ook derde-generatie-Marokkanen, is het toch mogelijk een etnische variant van het Nederlands af te zonderen (zie El-Aissati et al. 2005). Die variant zal intern nog wel individuele verschillen vertonen, maar dat is bij de standaardtaal en bij sociolecten niet anders. Verschillende detailstudies en overzichtsartikelen hebben fonologische, lexicale en grammaticale kenmerken van dat Marokkaans Nederlands beschreven (El-Aissati et al. 2005; Nortier en Dorleijn 2008; Hinskens 2011). Een geschreven corpus zoals het Moroccorp, ook al ligt het zeer waarschijnlijk dichter bij de impromptu gesproken taal dan overvloedig geredigeerde journalistieke, wetenschappelijke of politieke teksten, is minder geschikt voor fonologisch onderzoek, maar op lexicaal en grammaticaal vlak is het corpus zeer wel bruikbaar. Hier moeten we uiteraard nog toevoegen dat het Moroccorp slechts een bepaald aspect van Marokkaans Nederlands belicht, namelijk het taalgebruik in de context van computer-gemedieerde communicatie.

Als eerste indicator van de Marokkaanse kleuring van het Nederlands in het Moroccorp voeren we een lexicale analyse uit. We proberen een gevoel te krijgen voor de inhoud en de kenmerken van de twee chatkanalen door op zoek te gaan naar woorden die typisch zijn voor de verzamelde chatkanalen. Een eenvoudige frequentielijst zoals in Tabel 3 is hiervoor niet bijzonder inzichtelijk, want we vinden enkel woorden terug die sowieso hoogfrequent zijn in het Nederlands. Dit versterkt uiteraard wel onze claim dat het Nederlands de voertaal is in de beide chatkanalen, en dat het taalgebruik in het Moroccorp een reële variëteit is van het Nederlands.

Om woorden te vinden die echt typisch zijn voor ons corpus maken we gebruik van de zogenaamde *Stable Lexical Marker analyse* uit Speelman et al. (2006, 2008) en De Hertog et al. (2013). Ruwweg vergelijkt deze techniek de woordfrequenties uit een doelcorpus met de woordfrequenties uit een referentiecorpus. Als de frequentie van een woord uit het doelcorpus statistisch gezien opvallend hoger is dan in het referentiecorpus, dan wordt dit woord een *Lexical Marker*. Als dat woord dan ook nog eens consequent frequenter is doorheen

willekeurige opdelingen van het doelcorpus, dan noemen we dit woord een *Stable Lexical Marker*. Als referentiecorpus nemen we hier de Nederlandse chatcomponent uit het ConDiv corpus (Grondelaers et al. 2000).[6] De restrictie tot de Nederlandse chatcomponent (en dus de exclusie van de Vlaamse chatcomponent) is gemotiveerd door het bijzonder kleine aantal Vlamingen in de beide chatkanalen. De relatieve afwezigheid van Vlamingen hebben we vastgesteld met behulp van een telling van het aantal *ge*-vormen, een persoonlijk voornaamwoord voor de tweede persoon enkelvoud. Die *ge*-vorm is namelijk niet verspreid in Nederland. Het blijkt dat van de meerdere duizend chatters slechts ongeveer veertig chatters in het totaal ongeveer 100 *ge*-vormen produceren.

Tabel 3: Lijst van de twintig meest frequente tokens in het Moroccorp

Positie	**Woord**	**Frequentie**	**Positie**	**Woord**	**Frequentie**
1	je	290.435	11	van	79.986
2	ik	240.415	12	in	71.296
3	is	160.412	13	met	70.156
4	niet	128.978	14	ben	68.340
5	een	128.817	15	wat	64.696
6	de	118.019	16	jij	63.760
7	en	113.334	17	op	62.810
8	dat	101.814	18	maar	59.607
9	het	100.868	19	ze	56.332
10	die	84.419	20	heb	55.272

In Tabel 4 worden de resultaten van de Stable Lexical Marker analyse gepresenteerd. Technisch gesproken krijgt elk woord in een Stable Lexical Marker analyse een score toegekend, waarbij een sterk positieve score impliceert dat het woord opvallend frequent gebruikt wordt in het geteste corpus. Het zou echter niet inzichtelijk zijn om elk woord hier apart te representeren, en daarom hebben we de twintig hoogst scorende woorden hier in groepen gesorteerd.[7] We vinden enkele woorden die duidelijk refereren aan Marokko en Marokkaans (groepen 1, 2 en 4), en ook een groep die verwijst naar de belangrijkste religie in Marokko (groep 6). Dat zijn lexicale domeinen die prominent zijn in het taalgebruik van Nederlanders met een Marokkaanse achtergrond, zoals El-Aissati et al. (2005: 174-175) opmerken — al kunnen die natuurlijk net zo goed voorkomen in het taalgebruik van andere Nederlanders wanneer ze het toevallig over deze onderwerpen hebben. Daarnaast zien we ook dat er enkele gebruikelijke Nederlandse chatafkortingen als markant worden aangeduid (groep 3), en enkele woorden die kunnen duiden op typische onder-

werpen in de chatkanalen (groep 5).

Tabel 4: Stable Lexical Marker analyse: groepering van de twintig meest typische woorden voor de Marokkaans-Nederlandse chatkanalen, in vergelijking met algemene Nederlandse chatkanalen

	Woorden	**Uitleg**
1	salaam, salam, wslm (wasalaam), ewa, beslama	Marokkaanse groetwoorden
2	marokkaanse, marokko, marokkanen	zelfreferenties
3	gwn (gewoon), wrm (waarom)	internet afkortingen
4	wollah (vloek), hmdl (hamdoullah)	Marokkaanse uitroepen
5	broeder, trouwen, dame, vader	topicwoorden
6	islam, moslim, ramadan, allah	religieuze termen

Deze lexicale analyse toont aan dat het Moroccorp een vorm van het Nederlands die aanleunt bij computer-gemedieerde communicatie (Tabel 4, groep 3), maar ook lexicale verwijzingen bevat die de veronderstelling dat de chatters een Marokkaanse achtergrond hebben bevestigt (Tabel 4, groep 1 en 4). Daarnaast wijzen ook de andere woordgroepen uit Tabel 4 op de in Marokko gangbare religie (groep 6). We beklemtonen dat deze lexicale analyse natuurlijk geen definitief uitsluitsel geeft over de status van Marokkaans Nederlands als etnolect, maar wel onze overtuiging sterkt dat het Moroccorp het taalgebruik van Nederlanders met een Marokkaanse achtergrond beschrijft.

3.2 Morfologische analyse

Omdat tweede-generatie-Marokkanen het Nederlands meestal niet van huis uit hebben meegekregen, maar er doorgaans pas mee in aanraking zijn gekomen op de schoolbanken, vertoont hun taalgebruik hier en daar aspecten van wat met een technische term *fossilisatie* heet (Matras 2009: 75). Die fossilisatie neemt natuurlijk minder spectaculaire vormen aan dan die in het taalgebruik van eerste-generatie-sprekers, en het Nederlands op de chatkanalen is voor een ongetrainde blik vaak moeilijk te onderscheiden van dat van moedertaalsprekers, maar de omvang van het Moroccorp laat toe statistisch significante patronen aan te wijzen die onttrokken blijven aan het blote oog.

Een wel typisch te noemen verschijnsel in Marokkaans Nederlands, ook van die sprekers die vroeg met het Nederlands in aanraking gekomen zijn op school, betreft de adjectiefinflectie in het Nederlands. Het gaat om de inflectie in nominale constituenten van het type *een mooi verhaal*, waarin Marokkaans-Nederlandssprekenden de neiging vertonen 'onterecht' een flexie-e (sjwa) toe te

voegen. Als het Moroccorp dit patroon vertoont, zou hier opnieuw uit blijken dat het een waardevol instrument kan zijn voor de studie van het Nederlands door taalgebruikers met een Marokkaanse achtergrond.

Om te begrijpen wat hier precies aan de hand is, moet eerst iets gezegd worden over het eigenaardige systeem van adjectiefinflectie in het hedendaags Nederlands. Dat systeem kan in zijn meest eenvoudige vorm weergegeven wordt als in (1), de talrijke kleine uitzonderingen en de regionale variatie niet te na gesproken in het Nederlands (zie daarvoor onder andere Blom 1994; Weerman 2003; Tummers et al. 2004, 2005; Plevoets et al. 2009).

(1) (a) predicatief gebruik: adj-Ø (*dat boek is moeilijk*)
(b) attributief gebruik: adj-ə (*het moeilijk-e boek*)
(b') behalve: [+sg -def +neutr] NPs: adj-Ø (*een moeilijk boek*)

Het systeem is syntactisch-semantisch slecht gemotiveerd. Dat komt door de 'behalve'-clausule in (b'). Door die weg te halen ontstaat een relatief transparant, gemotiveerd systeem waarbij afwezigheid van inflectie predicatief gebruik markeert — daarmee aansluiting zoekend bij de adverbia (zie Diepeveen en Van de Velde 2010) — terwijl de inflectionele sjwa attributief gebruik markeert. Door (b') te schrappen krijgt de inflectionele sjwa bij adjectieven met andere woorden een signifié. Het is precies deze vereenvoudiging die massaal toegepast wordt door (jonge) L2-leerders van het Nederlands (zie Ziemann et al. 2011), en met name ook Marokkaans-Nederlandssprekenden (zie Blom et al. 2008). Het resultaat is het systeem in (2) (zie ook Van de Velde en Weerman Te verschijnen 2014).

(2) (a) predicatief gebruik: adj-Ø (*dat boek is moeilijk*)
(b) attributief gebruik: adj-ə (*het moeilijk-e boek, een moeilijk-e boek*)

De representativiteit van het Moroccorp kan nu afgelezen worden aan de mate waarin het dit soort sjibboletten bevat. Om nu te kijken of het corpus deze trend laat zien moeten we twee dingen doen. Allereerst moeten we op zoek gaan naar het patroon. Dat is moeilijker dan het lijkt: het aantal adjectieven en nomina dat in deze constructie kan voorkomen is groot, en zoeken in een niet-syntactisch geannoteerd corpus leidt dan bijna onvermijdelijk tot een hoop valse treffers, die manueel gefilterd moeten worden. Dit klemt te meer omdat onzijdige woorden veruit in de minderheid zijn in het Nederlands (Ziemann et al. 2011: 185), terwijl we daar net naar op zoek zijn. Uit praktische overwegingen is daarom gekozen een steekproef te doen bij vijf frequente onzijdige woorden, te weten: *verhaal, boek, land, onderscheid* en *verschil*.[8] Concreet is er gezocht naar patronen van een onbepaald lidwoord (*een*) gevolgd door een willekeurig woord, gevolgd door een van deze vijf woorden. Dat levert, na het uitwieden van de irrelevante hits (b.v. *een harrypotter boek* (samenstelling), *een eigen land* (principieel onverbuigbaar adjectief) etc.) een verdeling op van geflecteerde versus ongeflecteerde vormen. De tweede stap is om deze verde-

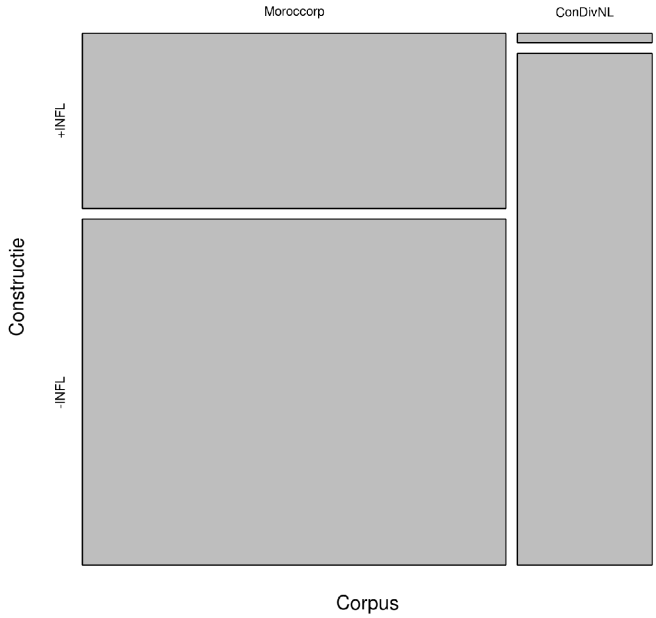
ling te vergelijken met een controlegroep. In principe is het immers best mogelijk dat het Moroccorp veel correct geflecteerde patronen bevat: het is beslist niet zo dat alle sprekers van het Marokkaans Nederlands altijd de 'onterecht' geflecteerde variant gebruiken, en principieel is het ook mogelijk dat niet-Marokkaanse Nederlandstaligen actief zijn op het chatkanaal. De vraag is daarom of de informanten in het Moroccorp de gemarkeerde vorm significant vaker gebruiken dan eentalig-Nederlandse chatters. Voor die controlegroep moeten we ons dus wenden tot een vergelijkbaar corpus. Daarvoor zijn we te rade gegaan bij het Nederlandse IRC-materiaal in het Condiv-corpus — ongeveer 7 miljoen tokens in totaal (Grondelaers et al. 2000). Qua genre is dit een bijna perfecte match met het Moroccorp, en het lijkt aannemelijk dat er ook een grote gelijkenis is qua sociaal profiel. Een mogelijke storende factor, die we echter niet verder onderzoeken, zou kunnen worden gevonden in de toch wel tien jaar die ligt tussen het verzamelen van het ConDiv corpus en het Moroccorp. Uit de resultaten, weergegeven in Tabel 5, blijkt dat er een merkbaar verschil is tussen Marokkaans Nederlands en eentalig Nederlands (zie Figuur 2 voor een visualisatie in een mozaïekgrafiek). In het eentalige Nederlands komt de constructie eigenlijk niet voor, terwijl die in het Moroccorp in een derde van de gevallen gebruikt wordt. De associatie is statistisch significant.

Tabel 5: Verdeling van geflecteerde en ongeflecteerde adjectieven bij onzijdige nomina na een onbepaald lidwoord. (Chi-kwadraat, met Yates' continuity correction = 20,6927, vg = 1, p-waarde < 0,001)

	+INFL	**-INFL**
	(*een mooie verhaal*)	(*een mooi verhaal*)
bron:moroccorp	59	117
bron:condivnl	1	55

We merken bij deze kwantitatieve resultaten op dat de gemarkeerde +INFL vorm geen noodzakelijke en voldoende voorwaarde is om een bepaalde vorm van het Nederlands als 'Marokkaans' te bestempelen. Aan de ene kant komt de +INFL vorm ook (zeer beperkt) voor in het algemene corpus (ConDivNL), en aan de andere kant komt ook de ongemarkeerde -INFL vorm voor in het Moroccorp. Meer nog, de ongemarkeerde vorm blijkt duidelijk in de meerderheid. Dit toont aan dat zelfs een wijdverbreid stereotiep kenmerk van Marokkaans Nederlands op zichzelf wellicht niet voldoende is om deze vorm van Nederlands eenduidig af te zonderen als een etnolect. Een corpus van voldoende grote omvang, dat op een verantwoorde manier onderzocht wordt met behulp van statistische technieken, kan niettemin op een genuanceerde manier patronen blootleggen. En die patronen moeten verklaard worden. Het Moroccorp biedt die mogelijkheid.

Figuur 2: Mozaïekgrafiek van geflecteerde en ongeflecteerde adjectieven bij onzijdige nomina na een onbepaald lidwoord



4. Conclusie

Tot slot van dit artikel vatten we bondig de kenmerken van het Moroccorp samen. Het corpus heeft het taalgebruik opgenomen van een gemeenschap van (waarschijnlijk overwegend) Nederlanders met Marokkaanse afkomst. Het taalgebruik is beperkt tot conversaties in openbare chatkanalen waarin de voertaal het Nederlands is. Gedurende enkele maanden hebben we de conversaties in twee chatkanalen opgenomen totdat het corpus ongeveer tien miljoen woorden bevatte. De twee gevolgde chatkanalen zijn specifiek gericht op Marokkaanse Nederlanders, en hoogstwaarschijnlijk de enige van zulke aard die openbaar toegankelijk zijn. We durven daarom beweren dat het Moroccorp representatief is voor het Nederlands dat gebruikt wordt door Marokkaanse Nederlanders in openbare chatgroepen. Het Moroccorp bevat geen private berichten, en biedt daarom enkel inzicht in het taalgebruik van een groep mensen die zich bewust zijn van hun openbaarheid. Aangezien de chatters hun eigen anonimiteit waarborgen door het gebruik van een pseudoniem, hebben we geen stappen ondernomen om het Moroccorp verder te anonimiseren.

De omvang van het Moroccorp is ruim tien miljoen woorden, waarmee het een vergelijkbare grootte heeft als het Corpus Gesproken Nederlands. Hoewel het corpus voor een verzameling van geschreven taal relatief klein is, kun-

nen we erop wijzen dat het eenvoudig is om dit corpus verder uit te breiden. Van de opgenomen tien miljoen woorden werden een miljoen woorden geproduceerd door slechts dertien chatters. Dit wijst op een sterke Zipfiaanse curve (Figuur 1, waarin slechts een klein percentage van de taalgebruikers verantwoordelijk is voor de meerderheid van de opgenomen woorden. Aan de ene kant betekent dit dat de hoeveelheid taalgebruikers waarvoor we een voldoende grote hoeveelheid woorden hebben maar klein is, zodat de mogelijkheden tot veralgemening van het corpus niet overschat mogen worden. Aan de andere kant geeft dit ook aan dat de gebruikers van de chatkanalen een hechte gemeenschap vormen, waardoor het mogelijk wordt om ook fijnmazigere conversatie-analyse te doen. Ook voor andere veelgebruikte corpora (b.v. krantenmateriaal) geldt overigens dat ze relatief veel tekst van relatief weinig auteurs of sprekers excerperen.

Uit een kwalitatieve en kwantitatieve analyse van de chatconversaties blijkt dat de voertaal in de kanalen eenduidig Nederlands is, maar dat de uitingen van de gebruikers doorspekt zijn met Berberse of Arabische kenmerken (zie ook Boumans 2002). Op het niveau van het woordgebruik valt op dat typische Arabische en Berberse sjibboletten gebruikt worden in begroetingen en uitroepen. Het woordgebruik, zoals dat geanalyseerd werd in de *Stable Lexical Marker analyse*, verraadt ook dat de conversaties vaak over Marokko, Islam en religieuze tradities gaan. In dit artikel hebben we naast een lexicale analyse ook een stereotiep morfosyntactische kenmerk van zogenaamd 'Marokkaans Nederlands' besproken, namelijk de adjectiefflectie bij onzijdige nomina na een onbepaald lidwoord. Het blijkt dat de gemarkeerde vorm, die als kenmerkend voor Marokkaans Nederlands wordt beschouwd, statistisch significant meer voorkomt in het Moroccorp dan in een algemeen chatcorpus. Hoewel dit geen uitsluitsel geeft over de al dan niet etnolectische status van Marokkaans Nederlands — iets waar dit artikel ook geen definitieve uitspraak over wil doen — sterken deze waarnemingen ons vertrouwen in de bruikbaarheid van het Moroccorp als een waardevolle bron voor onderzoek naar Nederlands van taalgebruikers met een Marokkaanse achtergrond. Het corpus is publiek beschikbaar voor onderzoeksdoeleinden. Wie het hebben wil, kan contact opnemen met de auteurs van dit artikel.

Eindnoten

1. De grafieken en de statistische analyses in dit artikel zijn uitgevoerd met het software pakket R. (R Core Team. 2012. R: A language and environment for statistical computing. Vienna. http://www.R-project.org).
2. http://www.mpi.nl/resources/data
3. http://www.irssi.org
4. De woordtelling in de logs werd uitgevoerd met het UNIX programma *wc*. Het precieze aantal woorden in de logs wijkt waarschijnlijk licht af van de gerapporteerde aantallen omdat de telling gebaseerd is op een bijzonder ruwe tokenisering. De woordtelling in de opgeschoonde versie is uiteraard wel accurater.

5. *Bots* zijn kleine programmaatjes die automatische boodschappen weergeven, zoals bijvoorbeeld quizvragen of, in het geval van deze chatkanalen, de gebedstijden per locatie.
6. De *Stable Lexical Marker analyse* werd hier uitgevoerd op woorden die minstens 30 keer voorkomen in elke opdeling van het doel- en referentiecorpus.
7. We hebben enkele spellingsvarianten, smileys en gebruikersnamen genegeerd.
8. Dit zijn niet de meest frequente onzijdige woorden. De reden daarvoor is dat woorden zoals *weer* en *haar* heel frequent zijn, maar ook homoniem zijn met functiewoorden, en daar hun frequentie aan te danken hebben. We hebben dan ook gekozen voor woorden die redelijk frequent zijn, maar eenduidiger. Volledigheidshalve geven we hier nog even de rang van de gekozen woorden in het corpus aan: *boek* (frequentie 778, rang 1724/333114); *land* (frequentie 1635, rang 943/333114); *onderscheid* (frequentie 96, rang 8202/333114); *verhaal* (frequentie 913, rang 1528/333114); *verschil* (frequentie 796, rang 1699/333114).

Literatuuropgave

Androutsopoulos, J. 2001. *From the Streets to the Screens and Back Again: On the Mediated Diffusion of Variation Patterns in Contemporary German.* Essen: LAUD.

Androutsopoulos, J. 2006. Introduction: Sociolinguistics and Computer-mediated Communication. *Journal of Sociolinguistics* 10(4): 419-438.

Appel, R. 1999. Straattaal. De mengtaal van jongeren in Amsterdam. *Toegepaste taalwetenschap in artikelen* 62(2): 39-56.

Blom, A. 1994. Het ondoorgrondelijk bijvoeglijk naamwoord. *Forum der Letteren* 35(2): 81-94.

Blom, E., D. Polišenská en F. Weerman. 2008. Articles, Adjectives and Age of Onset: The Acquisition of Dutch Grammatical Gender. *Second Language Research* 24: 297-331.

Boumans, L. 2002. Meertaligheid op de Marokkaanse elektronische prikborden. *Levende Talen* 3: 11-21.

Buccini, A. 1995. Ontstaan en vroegste ontwikkeling van het Nederlandse taallandschap. *Taal en Tongval. Themanummer* 8: 8-66.

Buccini, A. 2010. Between Pre-German and Pre-English: The Origin of Dutch. *Journal of Germanic Linguistics* 22(4): 301-314.

Clyne, M. 2000. Lingua Franca and Ethnolects in Europe and Beyond. *Sociolinguistica* 14: 83-89.

Cornips, L. en V. de Rooij. 2003. Kijk, Levi's is een goeie merk: maar toch hadden ze 'm gedist van je schoenen doen 'm niet. Stroop, J. (Red.). 2003. *Waar gaat het Nederlands naartoe? Panorama van een taal*: 131-142. Amsterdam: Bert Bakker.

Croft, W. 2000. *Explaining Language Change: An Evolutionary Approach.* Harlow: Longman.

Crystal, D. 2011. *Internet Linguistics.* London: Routledge.

De Hertog, D., K. Heylen en D. Speelman. 2013. Stable Lexical Marker Analysis: A Corpus-based Identification of Lexical Variation. Soares da Silva, A. (Red.). 2013. *Pluricentricity: Language Variation and Sociocognitive Dimensions*: 117-130. Braga: Aletheia.

Diepeveen, J. en F. Van de Velde. 2010. Adverbial Morphology: How Dutch and German Are Moving Away from English. *Journal of Germanic Linguistics* 22: 381-402.

Drinka, B. 2010. Language Contact. Luraghi, S. en V. Bubenik (Reds.). 2010. *Continuum Companion to Historical Linguistics:* 325-345. Londen: Continuum.

El-Aissati, A., L. Boumans, L. Cornips, M. Dorleijn en J. Nortier. 2005. Turks- en Marokkaans Nederlands. Van der Sijs, N. (Red.). 2005. *Wereldnederlands. Oude en jonge variëteiten van het*

Nederlands: 149-183. Den Haag: Sdu.

Geeraerts, D. 2002. Rationalisme en nationalisme in de Vlaamse taalpolitiek. De Caluwé, J., D. Geeraerts, S. Kroon, V. Mamadouh, R. Soetaert, L. Top en T. Vallen (Reds.). 2002. *Taalvariatie en taalbeleid. Bijdragen aan het taalbeleid in Nederland en Vlaanderen*: 87-104. Antwerpen: Garant.

Grondelaers, S., K. Deygers, H. van Aken, V. van den Heede en D. Speelman. 2000. Het ConDiv-corpus geschreven Nederlands. *NederlandseTaalkunde* 5: 356-363.

Harris, A. en L. Campbell. 1995. *Historical Syntax in Cross-linguistic Perspective*. Cambridge: Cambridge University Press.

Hawkins, J.A. 1990. Germanic Languages. Comrie, B. (Red.). 1990. *The Major Languages of Western Europe*: 58-66. Londen: Routledge.

Heine, B. en T. Kuteva. 2003. On Contact-induced Grammaticalization. *Studies in Language* 27: 529-572.

Heine, B. en T. Kuteva. 2005. *Language Contact and Grammatical Change*. Cambridge: Cambridge University Press.

Herring, S.C. 2010. Who's Got the Floor in Computer-mediated Conversation? Edelsky's Gender Patterns Revisited. *Language@Internet* 7, article 8.

Hickey, R. 2010. *The Handbook of Language Contact*. Oxford: Blackwell.

Hinskens, F. 2011. Emerging Moroccan and Turkish Varieties of Dutch: Ethnolects or Ethnic Styles? Kern, F. and M. Selting (Reds.). 2011. *Ethnic Styles of Speaking in European Metropolitan Areas*: 101-129. Amsterdam/Philadelphia: John Benjamins.

Jaspers, J. 2004. *Tegenwerken, belachelijk doen: talige sabotage van Marokkaanse jongens op een Antwerpse middelbare school. Een sociolinguïstische etnografie.* Doctorale proefschrift. Antwerpen: Universiteit van Antwerpen.

Jaspers, J. 2008. Problematizing Ethnolects: Naming Linguistic Practices in an Antwerp Secondary School. *International Journal of Bilingualism* 12(1–2): 85-103.

Labov, W. 1966. The Social Stratification of (r) in New York City Department Stores. Labov, W. (Red.). 1966. *The Social Stratification of English in New York City*: 63-89. Washington, D.C.: Center for Applied Linguistics.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk. Volume 1: Transcription Format and Programs. Volume 2: The Database*. Mahwah, N.J.: Lawrence Erlbaum.

Matras, Y. 2009. *Language Contact*. Cambridge: Cambridge University Press.

Nederlandse Taalunie. 2004. *Corpus Gesproken Nederlands.* via TST-centrale: http://tst-centrale.org/nl/producten/corpora/corpus-gesproken-nederlands/6-17.

Nortier, J. en M. Dorleijn. 2008. A Moroccan Accent in Dutch: A Sociocultural Style Restricted to the Moroccan Community? *International Journal of Bilingualism* 12(1–2): 125-142.

Plevoets, K. 2008. Tussen spreek- en standaardtaal. Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfosyntactische verschijnselen uit het gesproken Belgisch–Nederlands. Doctorale proefschrift. Leuven: Katholieke Universiteit Leuven.

Plevoets, K., D. Speelman en D. Geeraerts. 2009. De verspreiding van de -e(n)-uitgang in attributieve positie. *Taal en Tongval* 22: 112-143.

Ramaekers, W. 1998. Mi, maak me geen eiers! Het Algemeen Cités. *Onze Taal* 67(4): 94-95.

Roberge, P. 2010. Contact and the History of Germanic Languages. Hickey, R. (Red.). 2010. *The Handbook of Language Contact*: 406-431. Oxford: Wiley-Blackwell.

Ruette, T., D. Speelman en D. Geeraerts. 2014. Semantic Weighting Mechanisms in Scalable Lexi-

cal Sociolectometry. Szmrecsanyi, B. and B. Waelchli (Reds.). 2014. *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*: 178-198. Berlin: De Gruyter.

Speelman, D., S. Grondelaers en D. Geeraerts. 2006. A Profile-based Calculation of Region and Register Variation. Wilson, A., D. Archer and P. Rayson (Reds.). 2006. *Corpus Linguistics around the World*: 181-194. Amsterdam/New York: Rodopi.

Speelman, D., S. Grondelaers en D. Geeraerts. 2008. Variation in the Choice of Adjectives in the Two Main National Varieties of Dutch. Geeraerts, D., G. Kristiansen en Y. Peirsman (Reds.). 2008. *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*: 205-233. Berlijn/New York: Mouton de Gruyter.

Stroop, J. 1998. *Poldernederlands. Waardoor het ABN verdwijnt*. Amsterdam: Bert Bakker.

Taeldeman, J. 2008. Zich stabiliserende grammaticale kenmerken in Vlaamse tussentaal. *Taal en Tongval* 60(1): 26-50.

Tagliamonte, S. en D. Denis. 2008. Linguistic Ruin? LOL! Instant Messaging and Teen Language. *American Speech* 83(1): 3-34.

Thomason, S.G. 2001. *Language Contact: An Introduction*. Edinburgh: Edinburgh University Press.

Thomason, S.G. en T. Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.

Tummers, J., D. Speelman en D. Geeraerts. 2004. Quantifying Semantic Effects. The Impact of Lexical Collocations on the Inflectional Variation of Dutch Attributive Adjectives. Purnelle, G., C. Fairon en A. Dister (Reds.). 2004. *Le poids des mots*: 1079-1088. Louvain-la-Neuve: Presses Universitaires de Louvain.

Tummers, J., D. Speelman en D. Geeraerts. 2005. Inflectional Variation in Belgian and Netherlandic Dutch: A Usage-based Account of the Adjectival Inflection. Delbecque, N., J. van der Auwera en D. Geeraerts (Reds.). 2005. *Perspectives on Variation. Sociolinguistic, Historical, Comparative:* 93-110. Berlijn: Mouton de Gruyter.

Vandekerckhove, R. en J. Nobels. 2010. Destandaardisatie en toe-eigening van schrijftaal. De chatcommunicatie van Vlaamse jongeren. Van der Wal, M. en E. Francken (Reds.). 2010. *Standaardtalen in beweging*: 173-191. Münster: Nodus Publikationen.

Van de Velde, F. en F. Weerman. Te verschijnen 2014. The Resilient Nature of Adjectival Inflection in Dutch. Sleeman P., F. Van de Velde en H. Perridon (Reds.). Te verschijnen 2014. *The Adjective in Germanic and Romance*. Amsterdam: John Benjamins.

Weerman, F. 2003. Een mooie verhaal; veranderingen in uitgangen. Stroop, J. (Red.). 2003. *Waar gaat het Nederlands naartoe? Panorama van een taal*: 249-260. Amsterdam: Bert Bakker.

Weinreich, U. 1953. *Languages in Contact. Findings and Problems*. New York: Publications of the Linguistic Circle of New York.

Xiao, Z. 2008. Well-known and Influential Corpora. Lüdeling, A. en M. Kyto (Reds.). 2008. *Corpus Linguistics: An International Handbook*: 383-457. Berlijn: Mouton de Gruyter.

Zeige, L.E. Te verschijnen. On Cognition and Communication in Usage-based Models of Language Change. Mengden, F. en E. Cousse (Reds.). Te verschijnen. *Usage-based Approaches to Language Change.* Amsterdam/Philadelphia: John Benjamins.

Ziemann, H., F. Weerman en E. Ruigendijk. 2011. Nederlands later geleerd: gebruik van lidwoorden en flexie van bijvoeglijke naamwoorden door duitstalige kinderen en volwassenen. *Internationale Neerlandistiek* 49(3): 183-207.