Big Pimpin'. Een big data-benadering van de verspreiding van het leenwoord pimpen in het Nederlands

Pijpops, Dirk; De Pascale, Stefano; Van de Velde, Freek; Zenner, Eline

Published in:
Taal en Tongval

DOI:
10.5117/TET2023.1.005.PIJP

Publication date:
2023

Amsterdam University Press

TAAL & TONGVAL 75.1 (2023) 73-113
HTTPS://DOI.ORG/10.5117/TET2023.1.005.PIJP

Big Pimpin'. Een big data-benadering van de verspreiding van het leenwoord *pimpen* in het Nederlands

Dirk Pijpops
Lilith, Faculté de Philosophie et Lettres, Université de Liège
dirk.pijpops@uliege.be

Stefano De Pascale
FWO Vlaanderen, Vrije Universiteit Brussel, Brussel en Quantitative Lexicology and Variational Linguistics (*QLVL*), *KU Leuven*
stefano.depascale@kuleuven.be

Freek Van de Velde
Quantitative Lexicology and Variational Linguistics (*QLVL*), *KU Leuven*
freek.vandevelde@kuleuven.be

Eline Zenner
Quantitative Lexicology and Variational Linguistics (*QLVL*), *KU Leuven*
eline.zenner@kuleuven.be

Abstract

This article illustrates some of the opportunities and challenges of pursuing a big data approach in linguistic research. To do so, we investigate the diffusion of the loan verb *pimpen* 'to fancify' in Dutch based on Twitter data. First, we focus on the derivations of the verb (e.g.: *terugpimpen* 'to pimp back', *herpimpen* 'to repimp', etc.) and plot the diversity of these forms through time, using the Chao-Wang-Jost estimation of Shannon entropy. We follow this up with an alternation study that compares *pimpen* not only to its 'native' alternative *opleuken*, but also its most frequent derivation *oppimpen*, using multinomial regression. It is found that, while *pimpen*'s early expansion in Dutch has proceeded at breakneck speed, resulting e.g. in a plethora of derivations that has so far gone undetected, its current momentum seems to be waning.

Keywords: big data, Twitter, entropie, multinomiale regressie, alternantie

1. Inleiding

Er is al veel geschreven over de rol van corpora en kwantitatieve methodes in de ontwikkeling van de taalwetenschappen in de laatste decennia.[1] Anno 2023 is de impact die een doorgedreven empirische aanpak heeft voortgebracht voor voornamelijk het variationele en gebruiksgebaseerde taalkundige onderzoek glashelder. Zo gaat er meer dan ooit aandacht naar de minutieuze uitwerking van protocols voor dataverzameling en -annotatie, wordt het palet aan kwantitatieve analysetechnieken (eerst ontleend aan de fundamentele statistiek, maar recenter ook aan de meer toegepaste *machine learning*) steeds gevarieerder en wordt de theorievorming sterk bepaald door wat kwantitatief kan worden geoperationaliseerd en getoetst.

Na de eerste pleidooien voor een sterke verbintenis tussen een kwantitatieve *forma mentis* en de variationele en gebruiksgebaseerde taalkunde (Geeraerts 2006), en de daarop volgende programmatische consolidatie van corpusgebaseerd onderzoek (zie Janda 2013; Divjak et al. 2016; maar ook de vele statistiekhandboeken voor taalwetenschappers die het laatste decennium zijn uitgegeven, waaronder Gries 2009/2013/2021; Levshina 2015; Desagulier 2018; Brezina 2018; Winter 2020), lijkt recent het moment aangebroken voor een kritische reflectie en opmaak van de balans. In het lokale forum is dit te zien in Kestemont en Van Hulle (2019) of Van de Velde en van der Horst (2021) en replieken, internationaal is er bijvoorbeeld het themanummer van Sönning en Werner (2021).

Op het moment dat er openlijk wordt gediscussieerd over de impact van de kwantificering van het taalkundige onderzoek, zet de schaalvergroting die zowel de diversiteit als kwantiteit van bronnen voor dataverzameling aangaat, zich onverminderd voort, en wel op drie vlakken. Ten eerste, dankzij de alomtegenwoordigheid van smartphones en de constante connectiviteit die daarvan het gevolg is, zijn sprekers (of beter: smartphonegebruikers) steeds en snel bereikbaar als dataleveranciers, rekening houdend met de uitdagingen van de privacywetgeving. Getuige daarvan zijn grote crowdsourcingsprojecten, vaak onder de noemer van *Citizen Science*, waarmee via apps grote hoeveelheden gesproken taaldata werden verzameld (van der Sijs 2020; zie het themanummer van Hilton en Leemann 2021). Ten tweede ontstaan er megaconsortia van onderzoeksgroepen die niet, in tegenstelling tot vroeger, enkel gremia zijn waarop resultaten worden besproken en

gedeeld, maar die al in een eerdere fase actief samenwerken om omvangrijke databanken samen te stellen. De tendens is zichtbaar in de psychologische wetenschappen, met projecten zoals ManyLabs (Klein et al. 2014; Many Labs 2 et al. 2018) en ManyBabies (The ManyBabies Consortium 2020), maar ook in taalkundige ondernemingen neemt dit soort strategische krachtenbundeling toe, naar het voorbeeld van de CHILDES- (MacWhinney 2000) en WALS-projecten (Haspelmath et al. 2005). Ten slotte komen we bij de nieuwe mijlpaal in het proces van schaalvergroting in de corpuslinguïstiek: de opkomst van Big Data.

De term Big Data had oorspronkelijk een meer gelimiteerde en precieze toepassing dan het huidige veralgemeende gebruik. De term omvatte immers niet enkel verandering in de grootteordes van data die worden verzameld, maar ook in het type data: volledig ongestructureerd en in een continue stroom aangeleverd en accumulerend, en bovendien zeer heterogeen, zoals dat het geval is voor bijvoorbeeld data uit sociale media, meteorologische data en fintech data (Mayer-Schönberger en Cukier 2013). Het betreft collecties die aan de zogenaamde 5 V's beantwoorden: *volume, velocity, variety, veracity* en *value*. *Big Data*-datasets worden in terabytes (=1000 gigabytes) en petabytes (=1000 terabytes) uitgedrukt. Zulke datasets kunnen bijgevolg ook niet in statische bestanden opgeslagen worden. Het gevolg is dat zulke verzamelingen niet meer via traditionele software en statistische technieken kunnen worden geanalyseerd, maar eerder aan de hand van bijvoorbeeld *machine learning*-technieken, zoals *deep learning*- en *ensemble learning*-algoritmes (L'Heureux et al. 2017). Zolang je een dataset in bijvoorbeeld Microsoft Excel kunt openen (limiet: 2 gigabyte), en in R kunt analyseren (R Core Team 2014), heb je dus nog niet te maken met Big Data. Zelfs wanneer we binnen de taalkunde komen aanzetten met datasets van honderdduizenden meetpunten, die zeker vergeleken met talige datasets uit het verleden een enorme toename betekenen, kunnen we op basis van de V-parameters eigenlijk nog niet spreken van Big Data. Een mogelijkheid is om te spreken over 'big data', zonder hoofdletters, om de nieuwe invulling van de term te onderscheiden van het technische gebruik ervan.

Het gros van het huidige taalkundige onderzoek kan dus eerlijkheidshalve niet als 'Big Data' in de originele betekenis worden beschouwd. De opkomst van gigantische webgebaseerde corpora (van verschillende miljarden tokens, zoals de TenTen-corpora aangeboden door SketchEngine (Jakubiček et al. 2013) heeft desondanks wel al een enorme impact gehad op de gebruiksgebaseerde en variationele taalkunde. Hoewel we zulke corpora nog niet als 'Big Data' kunnen beschouwen, kunnen ze zeker onder de noemer 'big data' (met kleine letters) ondergebracht worden. De opportuniteiten die

dat soort megacollecties aanbieden, zijn nog niet voldoende ontgonnen, ten dele door de methodologische uitdagingen die daarmee gepaard gaan.

In deze bijdrage willen we daarom het potentieel verkennen van zulke grote dataverzamelingen om eerder *small data*-onderzoek aan te vullen. We doen dat concreet door ons eerdere onderzoek naar de ontwikkeling van het werkwoord *pimpen* (Van de Velde en Zenner 2010) te completeren op twee vlakken die vooralsnog, precies door een gebrek aan de nodige hoeveelheid data, onbelicht zijn gebleven. Ten eerste laat een big data-benadering toe om bijzondere aandacht te besteden aan laagfrequente vormen. Dat leidt ons tot een nieuwe operationalisering van derivationeel-morfologische productiviteit, die rekening kan houden met het ontstaan van laagfrequente nieuwvormingen, in ons geval de ontwikkeling van nieuwe samengestelde werkwoorden met *pimpen* als grondwoord, zoals *ontpimpen, afpimpen* en *doorpimpen*. Vorige corpusgebaseerde methodes voor het meten van productiviteit schieten immers tekort wat de integratie van laagfrequente attestaties betreft, grotendeels omdat kleine dataverzamelingen geen betrouwbare schatting gaven van zulke attestaties. Ten tweede zullen we een alternantiestudie uitvoeren waarbij we de tekstuele context strikt onder controle houden door die te beperken tot één woordvorm. De grote hoeveelheid data stelt ons immers in staat zulke strikte beperkingen op te leggen. In dit tweede deel van het onderzoek duidt het *big data*-etiket dan ook op de initiële dataverzameling, niet op de uiteindelijke dataset waarbij de tekstuele context strikt beperkt wordt – die is weliswaar nog steeds groot, maar niet enorm. In dit deel van het onderzoek verkennen we ook het gebruik van multinomiale regressieanalyse voor grammaticale alternantiestudies. Zulke studies beperken zich typisch tot twee alternanten, geanalyseerd aan de hand van binaire logistische regressie, wat in vele gevallen een simplistische assumptie en operationalisering is. In onze studie laten we die beperking vallen en verkennen we het gebruik van deze onderbenutte multinomiale techniek bij een alternantie waarbij meer dan twee varianten in competitie onderzocht kunnen worden.

De rest van het artikel is als volgt opgebouwd. We introduceren eerst het geval *pimpen* en vatten de relevante resultaten van de vorige studies erover samen (sectie 2). Daarna bestuderen we aan de hand van een grote Twitterdataset de morfologische familie van *pimpen*, die bestaat uit alle lemma's die de stam *pimp* delen (over alle woordklassen heen) en die dus verwant zijn op de syntagmatische as (sectie 3). In ons geval bestaat de morfologische familie van *pimpen* voornamelijk uit samengestelde werkwoorden zoals *doorpimpen, uitpimpen* of *ontpimpen*. Vervolgens gaan we in op een multinomiale alternantiestudie van de conceptuele familie

van bijna-synoniemen op de paradigmatische as, die samen de betekenis 'pimpen' lexicaliseren, namelijk *pimpen*, *opleuken* en *oppimpen* (sectie 4). We eindigen deze paper met een discussie over onze resultaten in het licht van recente big data-ontwikkelingen in de taalkunde (sectie 5).

2. Van Pimp My Ride naar pimpen

Het werkwoord *pimpen* vindt in het Nederlands snel ingang na de uitzending van de populaire Engelstalige MTV-serie *Pimp My Ride* in 2004 (Van de Velde en Zenner 2010; De Pascale et al. 2022). In het Engels is het lemma dan al eeuwenlang in gebruik als substantief verwijzend naar 'een persoon die een deel van de opbrengsten van een prostituee opneemt, gewoonlijk in ruil voor het aanbrengen van klanten, het bieden van bescherming etc.' (Oxford English Dictionary, lemma *pimp*, eigen vertaling) en als werkwoord om de handelingen en gedragingen van zo'n individu te benoemen ('optreden als pooier' of 'iemand prostitueren', OED, lemma *pimp*), dit in zowel letterlijke als figuurlijke betekenis (bv. *He's pimping himself here, knowing he needs the publicity but hating himself for playing the game* 'Hij is zichzelf hier aan het pimpen in de wetenschap dat hij die publiciteit nodig heeft, al verwijt hij zichzelf dat hij het spel meespeelt', OED, lemma *pimp*). Het televisieprogramma uit 2004 introduceert een nieuw gebruik van het werkwoord, namelijk de betekenis 'opleuken, versieren, met name op een opzichtige manier'. In de show zien we immers de opzichtige make-over van sjofele auto's die jonge bestuurders aan de programmamakers ter beschikking stellen. De auto wordt dan bewust verfraaid op zo'n manier dat die er uiteindelijk uitziet als een voertuig dat stereotiep gelinkt kan worden aan onder meer pooiers (cf. *pimpmobile,* Gold 1985). Het programma, dat zich afspeelde in de regio van Los Angeles, steunde daarbij sterk op de lokale hiphopcultuur, maar kende niettemin ook een groot succes in het buitenland, getuige daarvan de uitzending in Nederland en België, maar ook de vele nationale bewerkingen (in Brazilië, Indonesië etc.).[2]

Van de Velde en Zenner (2010) sommen een aantal beperkingen van *pimpen* op, die het onwaarschijnlijk maakten dat het woord zou doorbreken in het Nederlands. Ze beschrijven vervolgens waarom het parcours van *pimp* van vast onderdeel van de Engelse eigennaam *Pimp My Ride* naar een volledig regelmatig Nederlands werkwoord *pimpen* het beste kan worden beschreven vanuit de constructiegrammatica. Deze eerste corpusgebaseerde studie, met een beperkte selectie krantenartikelen tussen 1998 en 2009, kon zo al een aantal patronen vaststellen. Ten eerste bevestigen de auteurs

de al vermoede causale link tussen de eerste uitzending van de show in 2004 en de opkomst van de nieuwe 'opleuken'-betekenis van *pimpen*, door aan te tonen dat *pimpen* in die toepassing volledig ontbreekt in de data voor 2004. Na de introductie van *Pimp My Ride* blijft de nieuwe betekenis ook duidelijk de meest frequente. Ten tweede observeren Van de Velde en Zenner (2010) hoe de numerieke groei van de 'opleuken'-betekenis snel gepaard gaat met toegenomen variatie zowel op formeel als op semantisch vlak. Zo begint de vaste woordcombinatie *Pimp My Ride*, die in zijn geheel verwijst naar de show, al in het eerste jaar na uitzending interne syntactische compositionaliteit en productiviteit te vertonen, met name in de variatie van het tweede en derde slot, die respectievelijk een ander voornaamwoord (i.e. POSS) en een ander zelfstandig naamwoord (i.e. N) bevatten dan *my* en *ride*. In wat volgt, zullen we dit de '*pimp* POSS N'-constructie noemen. Ook op semantisch vlak vertoont het sjabloon variatie, door in eerste instantie andere voertuigen (2) toe te laten in het N-slot dan auto's (1), daarna niet-gemotoriseerde objecten (3) en ten slotte zelfs bezielde objecten (4).

(1) Deze Twingo is **gepimpt** en nu helemaal klaar voor de verkoop! De ramen zijn getint en de grille zwart gewrapt. http://t.co/qdjentod6v

(2) De Morgen: Levering gepimpte C-130 voor Belgisch leger op til: Het Belgisch leger maakt zich klaar voo.. http://tinyurl.com/degokr

(3) @USERNAME Zo werkt de italiaanse keuken niet. Geen italiaan die er nog maar aan denkt om z'n bolognese recept te **pimpen** tegen de saaiheid.

(4) @USERNAME Hoe zou ge een chihuahua dan plastisch willen **pimpen**?

Het laatste stadium in de ontwikkeling van *pimpen,* volgens de metingen in het 2010-artikel, was het verschijnen van finiete vormen van het werkwoord en daarmee dus de voltooiing van het werkwoordsparadigma. Kortom, het neologisme *pimpen* heeft een aantal syntagmatische en semantische 'schematiserende' veranderingen ondergaan, van eigennaam naar semi-vast constructiesjabloon tot volwaardig Nederlands werkwoord.

Ondanks de duidelijke patronen had deze eerste studie ook enkele beperkingen, wat de aanzet heeft gegeven voor het vervolgonderzoek in De Pascale et al. (2022). Hoewel het krantencorpus uit 1998-2009 al enkele systematische inzichten heeft verschaft, is een formeel register wellicht niet de meest geschikte bron om de veranderingen in het gebruik van het

eerder informele *pimpen* ten volle te bestuderen. Ten tweede was de dataset van *pimp*-vormen in die studie te klein (N=246), en de tijdsspanne te kort, om de sterkte en het bereik van de evoluties aangehaald in de vorige alinea grondig en met zekerheid te beschrijven. In de vervolgstudie van De Pascale et al. (2022) werd daarom een aanzienlijk grotere dataset aangelegd, deze keer bestaande uit publiek beschikbare tweets, berichten op Twitter dus. Hoewel onderzoek aantoont dat Twitter niet als een monolithisch register met een in de tijd stabiele gebruikerspopulatie kan worden beschouwd (Grondelaers et al. 2021), geeft het platform wel toegang tot geschreven taal die informeler is dan de bij Van de Velde en Zenner (2010) gebruikte krantendata. Bovendien telt deze Twitterdataset 163.046 attestaties uit de periode 2007–2020, waarin de letterreeks **pimp** voorkomt (uitgezonderd de samenkoppelingen en afleidingen, cfr. infra), wat ongeveer 800 keer meer is dan de dataset uit de studie van 2010. Van deze Twitterdataset annoteerden De Pascale et al. (2022) 4596 voorkomens manueel met semantische en syntactische parameters, die onder andere betrekking hebben op de keuze van voornaamwoord, de semantiek van het object en de werkwoordsvorm.

De replicatie van De Pascale et al. (2022) maakt een onderscheid tussen het gebruik van *pimpen* binnen het constructiesjabloon [*pimp* POSS N] en buiten dat sjabloon, als 'vrije' werkwoordsvorm. Op basis van de verschillende semantische en syntactische parameters wordt een 'deconstructionaliseringsscore' berekend. Een regressie met deze score als afhankelijke variabele bevestigt grotendeels het schematiseringstraject van *pimpen* dat al in de eerste studie geobserveerd werd, al lijkt de trend van ontwikkeling minder uitgesproken te zijn.[3] Wat het gebruik van *pimpen* buiten het sjabloon betreft, lijkt er volgens diezelfde deconstructionaliseringsscore geen verdere formele en semantische differentiatie plaats te vinden van 2007 tot en met 2020. Hoewel de diachrone trendlijn stabiel blijft, verbergt ze het feit dat in de latere jaren de *span* van deze scores groter wordt. Dat betekent dat doorheen de jaren zowel minder als meer geschematiseerde constructies met *pimpen* frequenter worden. Dit effect werd verklaard vanuit het 'a rising tide lifts all the boats'-fenomeen. Daarbij geldt dat een nieuw gebruik (c.q. het 'vrije' *pimpen* buiten het sjabloon) het oude gebruik (c.q. het sjabloon [*pimp* POSS N]) niet verdringt of vervangt, maar er net voor zorgt dat de frequentie van de oude toepassing ook een toename ondervindt onder invloed van primingmechanismen (zie voor een ander geval van dit fenomeen de *ever/ooit*-alternantie beschreven in Zenner, Heylen en Van de Velde 2018).

In dit artikel breiden we het onderzoek naar *pimpen* verder uit, en kijken we in het bijzonder naar de verhouding van het werkwoord ten opzichte van (paradigmatische en syntagmatische) verwante vormen in het lexicale

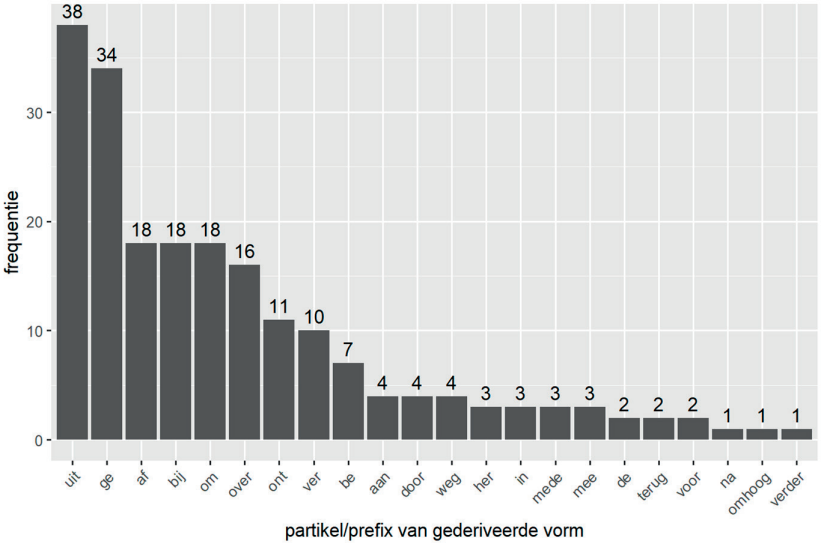
veld. Het doel van de studie is niet enkel om de gebruikspatronen van deze ontlening beter in kaart te brengen vanuit twee perspectieven die tot nu toe onderbelicht bleven (i.e. morfologische productiviteit en onomasiologische competitie). *Pimpen* vormt ditmaal ook een ideale gevalstudie om de methodologische uitdagingen van 'big data' voor taalkundig-theoretisch onderzoek te testen.

3. Morfologische productiviteit berekenen met gecorrigeerde entropiematen

3.1. Kwalitatieve verkenning

De ontwikkeling van het werkwoord *pimpen* geschetst in de vorige paragraaf is niet enkel zichtbaar in de verankering van het inflectionele paradigma, maar ook in de mate waarin neologismen opduiken, in het bijzonder dan samenstellingen met *pimpen* als grondwoord en een partikel en in minder mate afleidingen met *pimpen* en een prefix. Figuur 1 geeft de frequentie weer van de verschillende geattesteerde derivaties van *pimpen* in Twitter tussen 2007 en 2020 (waarbij enkel rekening wordt gehouden met niet-gescheiden inflectionele vormen).[4]

In totaal konden we 203 tokens (i.e. ongeveer 0.001% van alle *pimp*-voorkomens) vinden die terug te leiden zijn tot een afgeleide vorm van **pimpen,* gespreid over 21 lemma's (met *medepimpen* en *meepimpen* als één lemma gerekend). Figuur 1 toont bovendien de verwachte Zipfiaanse frequentieverdeling van zulke afleidingen, waarbij tegenover enkele frequente lemma's een langere staart aan infrequente vormen staat. De omvang van deze verzameling kan het best beoordeeld worden met de volgende informatie indachtig. Ten eerste is deze verzameling met 203 nevenvormen bijna even groot als de dataset van 246 *pimpen*-grondwoordvormen die werd gebruikt in de studie van Van de Velde en Zenner (2010). Ten tweede is deze verzameling ook vele malen groter dan wat men in een referentiecorpus als SoNaR (Oostdijk et al. 2013) of het Corpus Hedendaags Nederlands (CHN 2021) zou vinden, die respectievelijk 500 miljoen en 1 miljard woorden tellen. *Oppimpen* buiten beschouwing gelaten, vindt men in SoNaR slechts 3 attestaties van respectievelijk *uitpimpen*, *verpimpen* en *overpimpen* en in het CHN 2 voorkomens van *volpimpen* en *overpimpen*.[5] Kortom, de evidente schaalvoordelen van ad-hoc gecompileerde enorme datasets stellen taalkundigen in staat om verschijnselen te bestuderen die omwille van hun intrinsieke eigenschappen (i.e. lexicaal-morfologisch domein, recent karakter, in volle ontwikkeling, veel variatie, niet frequent) voordien nauwelijks empirisch te toetsen waren met zelfs grote, maar statische corpora.



Figuur 1: Frequentieverdeling van elk partikel of prefix in de morfologische familie van *pimpen*. In alle gevallen gaat het om samengestelde werkwoorden, buiten bij *ge-*, waar het gaat om de deverbale substantivering *gepimp*

De selectie van 203 gevallen is tevens al ruim genoeg om beginnende vormen van zowel formele als semantische differentiatie te observeren in de derivaties. Deze selectie toont een complexe configuratie binnen de morfologische familie van *pimpen*. Enerzijds valt op dat door de groei van de familie semantische specificatie en diversiteit optreedt, aangezien elk nieuw werkwoord minstens een nuance of facet van de activiteit van het 'opleuken' lexicaliseert, afhankelijk van de betekenis van het specifieke partikel of voorvoegsel. Dat geldt voor alle werkwoorden opgenomen in Tabel 1. Anderzijds merken we dat de groei op semantisch vlak niet enkel diversificatie inhoudt, maar tegelijkertijd ook unificatie. Zo drukken sommige nieuwvormingen elk een betekenis uit die zo wederzijds verwant zijn dat ze een nieuw synoniemenveld, en dus als het ware één conceptuele categorie, gaan vormen. De voorbeelden van betekenis A van *uitpimpen* en *verpimpen* tonen aan dat de twee neologismen grotendeels hetzelfde concept lexicaliseren ('overdragen van een persoon om bepaalde diensten te laten leveren'). Die betekenis gaat op zijn beurt terug op het oorspronkelijke Engelse gebruik van *to pimp* – een gebruik dat dan weer niet werd overgenomen in het Nederlandse *pimpen*. *Uitpimpen* kan trouwens ook een leenvertaling zijn van het Engelse *to pimp out* dat zelf synoniem is aan *to pimp*. Bij *verpimpen* ligt zo'n

Tabel 1: Polysemie in afgeleide vormen van *pimpen*

Werk-woord	Betekenis	Voorbeeld
uitpimpen	A. uitleveren/uitlenen/overdragen van een levend wezen (vaak een vrouw) met de bedoeling om dat wezen seksuele diensten te laten leveren	(1) *Poor girl werd gewoon gecaffished en daarna **uitgepimpd***
		(2) *Omdat ik ook geld wil verdienen met mijn huisdier ga ik vanaf nu mijn kat **uitpimpen**. Je vindt hem op IG Talentless_cat. Want hij kan niets.*
	B. het beëindigen van het pimpen	(3) *Als je **uitgepimpt** bent is er in Gouda nog een ongepimpte versie waar je je mag uitleven ;-) RT PogingTwee Gaat balkon verder oppimpen*
		(4) *Mijn nieuwe Nokia N95 8 gB **uitpimpen** to the max en dus hoort daar ook Twitter bij :P*
afpimpen	A. het pimpen tot een einde brengen	(5) *En dan nu nog even snel een hele mooie naam **afpimpen** en de werkdag is weer voorbij… Welterusten allemaal!*
		(6) *even bezig met schrijfdossier **afpimpen** dan filmpje kieken like always :) #geeflobi #zegmaarchallas*
	B. (volkomen) uitgeput zijn door het pimpen	(7) *@wolzak Geen idee wie t is maar hij ziet er zelf ietwat **afgepimpt** uit. Blik vooruit vanaf nu of minder denken ☺*
verpimpen	A. aangeven dat het object door het pimpen overgaat in andere handen	(8) *@lamazone Is dat geen idee; een social network waar je je vrijgezel vrienden kan **verpimpen** :)?*
	B. aangeven dat het pimpen een verandering van het object veroorzaakt	(9) *@kimmetjekoop Ehu…. ja lekker! We hebben alleen geen kids morgen, omdat we hun kamers gaan **verpimpen**…*

rechtstreekse invloed uit het Engels minder voor de hand. Vervolgens kunnen we voor de werkwoorden *uitpimpen, afpimpen* en *verpimpen* al polysemie ontwaren, wat aantoont dat sommige neologismen mogelijk al in een verder stadium zitten, waarin isomorfisme wordt opgegeven en semantische variatie mogelijk wordt. Op het formele vlak, ten slotte, treffen we bij *overpimpen* instabiliteit aan wat betreft de scheidbare of onscheidbare status van het werkwoord (cf. Tabel 2). Met deze verzameling (klein in absolute waarde, maar groot in vergelijking met traditionelere datasets) blijft de kwalitatieve analyse van de morfologische familie van *pimpen* beperkt, maar de geobserveerde variatie aan vormen en betekenissen, hun oorsprong en onderlinge verband, verdient een diepere reflectie in toekomstig onderzoek.

Tabel 2: Formele variatie in afgeleide vormen van *pimpen*

Werkwoord	Betekenis	Voorbeelden
overpimpen	te intens, te veel of in te hoge mate een object pimpen	scheidbaar samengesteld: *Van die boeren met hun **overgepimpte** auto's en loeiharde, marginalemuziek*
		onscheidbaar samengesteld: *Ga die **overpimpte** dopingwedstrijd geen 3 weken mn TL laten verpleuren.*

3.2. Kwantitatieve analyse

De verzameling van afgeleide neologismen en hun frequentie in Figuur 1 stelt een statische weergave voor van de morfologische familie van *pimpen*. De vraag rijst hoe deze categorie is geëvolueerd in een relatief korte periode van 17 jaar, met andere woorden welk diachroon patroon we aantreffen in de morfologische productiviteit van *pimpen*. Veelgebruikte operationaliseringen in het corpusgebaseerd onderzoek naar morfologische productiviteit en rijkdom, zoals *expanding* en *potential productivity* (zie Baayen 2009 voor een overzicht) zijn voornamelijk gebaseerd op de combinatie van tellingen op type-niveau enerzijds, en meer bepaald het aantal *hapax legomena* (of singletons), en de grootte van het corpus anderzijds. Als morfologische productiviteit wordt gedefinieerd als de mate waarin een categorie in staat is om nieuwe woorden te genereren, dan zijn zulke singletons de beste corpusgebaseerde benaderingen voor die neologismen. Ook vanuit probabiliteitstheorie zijn er goede argumenten om te focussen op *hapax legomena*, vanwege hun nut bij berekeningen voor frequentieverdelingen met Good-Turing correctie (Pierrehumbert en Granell 2018).[6]

Er zijn echter twee nadelen verbonden aan voorgaande operationaliseringen. Ten eerste houden ze geen rekening met de frequentieverdeling van de types die geen singletons zijn. Ten tweede veronderstellen ze dat de grootte van het onderliggende corpus is gekend. In een benadering waarbij een studie-onafhankelijk corpus wordt gebruikt waarin de tellingen van het fenomeen ter zake worden uitgevoerd, zijn deze productiviteitsmaten zeker geschikt. Wanneer de grootte van het corpus echter niet berekend kan worden, zoals het geval is bij het gebruik van Twitter-data, schieten deze maten tekort.

In de laatste jaren zijn voor de bovenstaande uitdagingen nieuwe impulsen gekomen uit de biostatistiek, waarin al langer maten in omgang zijn voor de berekening van variatie in soorten in het planten- en dierenrijk. De evolutie om zowel concepten als methodes uit de exacte wetenschappen over

te nemen in de taalkunde is al langer aan de gang, en steunt op potentiële gelijkenissen tussen talige en biologische fenomenen (Van de Velde en van der Horst 2021). Het beschrijvingskader dat bij uitstek de consensus vormt tussen het biologische en taalkundige perspectief, is dat van de informatietheorie van Claude Shannon (Shannon 1948), en de centrale maat voor op informatietheorie-gestoelde studies is de entropiemaat (e.g.: Piantadosi et al. 2012; Stoll et al. 2012 etc.). Zo ontleent Moscoso del Prado Martin (2014; 2015) gecorrigeerde entropiematen uit de ecologische statistiek, waar ze al langer in zwang zijn als indicatoren van biodiversiteit (Gotelli en Chao 2013) om morfologische en syntactische variatie te onderzoeken, en recent publiceerden Kestemont et al. (2022) onderzoek naar de reconstructie van verloren manuscripten aan de hand van indices uit diezelfde biologische discipline.

In zijn biostatistische interpretatie kwantificeert de Shannon-entropie de mate van diversiteit in een verzameling met geobserveerde soorten waarin elke soort met een probabiliteit voorkomt, zoals in Formule 1. Entropie heeft een vaste ondergrens van 0, waarin volledige uniformiteit regeert en slechts een soort geobserveerd werd. Hoe groter de entropie, hoe groter de diversiteit (in het aantal soorten en in hun verdeling) in de verzameling. De taalkundige interpretatie ligt in lijn met de biologische: hoe gevarieerder de verzameling van morfologische derivaties van een woord, hoe hoger de entropie.[7]

$$H_{Sh} = -\sum_{i=1}^{S} p_i log\, p_i$$

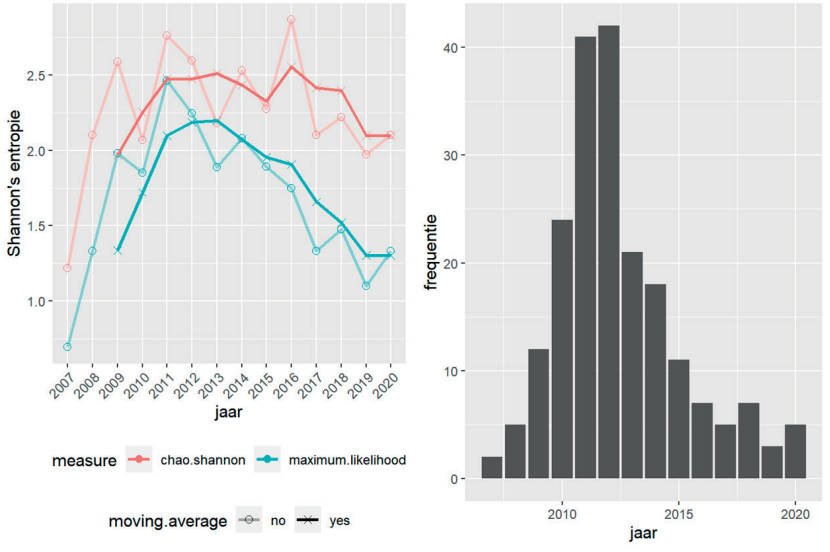
Formule 1: De berekening van Shannon-entropie H_{sh} op basis van de probabiliteit p_i van elke soort i in de verzameling van S geobserveerde soorten

Hoewel de Shannon-entropie belangrijke voordelen heeft vergeleken met traditionelere diversiteitsmaten, is de maat als dusdanig niet vrij van schadelijke correlaties met steekproefgrootte. Als de gerapporteerde diversiteit te afhankelijk is van de grootte van de steekproef dan wordt het onmogelijk om de effectieve diversiteit van steekproeven van verschillende omvang met elkaar te vergelijken. Die correlatie sluipt al in de manier waarop de probabiliteiten van de soorten worden berekend. Een populaire schatting van de probabiliteiten is Maximum Likelihood Estimation (MLE), die gebaseerd is op de relatieve frequentie van de soorten. MLE-gebaseerde schattingen houden echter systematisch een onderschatting in van de diversiteit, aangezien ze geen rekening houden met *ongeziene* soorten. Daarom hebben Chao, Wang en Jost (2013) een nieuwe schatter van de Shannon-entropie

ontworpen, die een aantal cruciale vormen van informatie in de formule integreert. Ten eerste maakt hun formule gebruik van de zogenaamde 'soortenaccumulatiecurves', die de verhouding tussen de (toenemende) grootte van de steekproef en het (toenemende) aantal soorten weergeeft, gelijkaardig aan de *vocabulary growth curves* uit Baayen (2001). De intuïtie achter het gebruik van deze curve is dat de richtingscoëfficiënt (*slope*) van de curve de waarschijnlijkheid voorstelt dat het $(K + 1)$-ste individu in de steekproef een soort vertegenwoordigt die in de vorige steekproef van grootte K werd gemist. Opeenvolgende richtingscoëfficiënten van de curve geven dus de snelheid aan waarmee tijdens grotere steekproeven nieuwe soorten worden ontdekt (Chao Wang en Jost, 2013, 1093). Ten tweede incorporeert deze verbeterde schatter niet enkel de telling van singletonsoorten (i.e. *hapax legomena* in een taalkundige context), maar ook van doubletonsoorten (i.e. *dis legomena* in een taalkundige context). De exacte berekening van de schatter is te vinden in Bijlage 1.

Wanneer de Chao-Wang-Jost schatter wordt toegepast op onze verzameling *pimp*-derivaties, valt uit Figuur 2a op te maken dat de schatter (in rood) inderdaad een stuk minder afhankelijk is van de onderliggende steekproefgrootte dan een diversiteitswaarde die slechts gebaseerd is op niet-gecorrigeerde relatieve frequenties (in blauw). Zoals te zien is in Figuur 2b zijn de grootste steekproeven te vinden in de jaren 2011 en 2012, en in de jaren erna verminderen zowel het aantal verschillende morfologische derivaties als hun frequentie drastisch (onder de 10 voorkomens in totaal). Het zou onvoorzichtig zijn om de daling weerspiegeld in Figuur 2b rechtstreeks te interpreteren als een afnemende productiviteit van *pimpen*. Het is namelijk zo dat Twitter een toename aan nieuwe gebruikers heeft gekend precies in de jaren 2012-2014 en dat het microbloggingkanaal daarna met een stagnerende populariteit werd geconfronteerd (Turpijn, Kneefel, en van der Veer 2015; Grondelaers et al. 2021). De gecorrigeerde entropiemaat probeert exact dat soort effecten uit te filteren, en toont dat de diversiteit/productiviteit van *pimpen* nog lang stabiel blijft (tot 2018) en daarna een knik ervaart, echter lang niet zo ingrijpend en vooral niet zo systematisch als de niet-gecorrigeerde maat zou aangeven. Een eenvoudige Spearmancorrelatie tussen de grootte van de verzameling en de gemeten entropie geeft aan dat de MLE-gebaseerde entropie veel hoger correleert (ρ = 0.92) met de onderliggende verzameling dan de gecorrigeerde maat (ρ = 0.63).

Hoewel we in deze korte illustratie aan de hand van de morfologische familie van *pimpen* hebben kunnen aantonen dat de berekening van morfologische productiviteit gebaat kan zijn bij zorgvuldig gecorrigeerde maten afgeleid uit de biostatistiek, blijken niet alle problematische kwesties van de baan. De belangrijkste is dat in voorgaande berekeningen, en ook in meer



Figuur 2: (a) Verloop van Shannon's entropie voor de morfologische familie van *pimpen*; (b) Totale tokenfrequentie van morfologische familie per jaar

recent werk in de biostatistiek (Chao et al. 2021), de temporele afhankelijkheid van steekproeven nog niet voldoende uitgewerkt is in het beschrijvingskader. Zo lijkt de informatie van geobserveerde, maar infrequente soorten in een vroeger tijdstadium niet opgenomen in de diversiteitsberekening van een later tijdstip. De frequentieverdeling van een later tijdstip zou met andere woorden moeten kunnen worden gecorrigeerd door een frequentieverdeling van een vroeger tijdstip. In Figuur 2a werd deze afhankelijkheidsrelatie eenvoudigweg opgelost door een lopend gemiddelde met een *span* van 3 jaar te nemen over de entropiewaarden, maar de correctie zou dus idealiter plaatsvinden op de onderliggende verdeling. Met een lopend gemiddelde zitten we op het terrein van de *time series analysis*. *Time series analysis* is een familie van technieken, met toepassingen in de psycholinguïstiek (Baayen et al. 2018) en – mondjesmaat – in de historische taalkunde (Koplenig en Müller-Spitzer 2016; Koplenig 2017; Rosemeyer en Van de Velde 2021; Van de Velde en Petré 2020).[8]

4. Multinomiale analyse: *pimpen, opleuken* en *oppimpen*

De vorige sectie richtte zich op de formele uitwaaiering van het werkwoord *pimpen*; deze sectie gaat in op zijn functionele ontwikkeling. *Pimpen* in de

betekenis 'opleuken' kwam de Nederlandse taal niet binnen als lexicalisering van een voorheen totaal onbekend concept, zoals dat wel het geval was voor pakweg *computer*.[9] Het inheemse alternatief *opleuken* was immers al beschikbaar, sinds ten laatste 1988 (etymologiebank, s.v. *opleuken*). In deze subsectie onderzoeken we hoe *pimpen* zich verhoudt ten opzichte van dit alternatief. We kiezen voor *opleuken*, en niet voor bijvoorbeeld *verfraaien* of *optuigen*, om aan te sluiten bij Van de Velde en Zenner (2010). Uit het onderzoek voorgesteld in de vorige sectie bleek echter dat naast het simplex *pimpen* ook het splitsbaar samengestelde werkwoord *oppimpen* voorkomt. Dit lijkt formeel een mengvorm van *pimpen* en *opleuken*. Het onderscheidt zich van de andere morfologische afleidingen zoals *uitpimpen* en *ontpimpen* doordat er geen duidelijk betekenisverschil is met *pimpen*, en het veel frequenter is dan de andere afleidingen (zie voetnoot 3).

Pimpen, opleuken en *oppimpen* hebben dus dezelfde betekenis, wat impliceert dat ze inwisselbaar zijn volgens de sociolinguïstische onderzoekstraditie en dus als een sociolinguïstische alternantie bestudeerd kunnen worden (Labov 1972, Tagliamonte 2012, zie Pijpops 2020 voor een overzicht van de verschillende definities van het begrip *alternantie*). We beperken ons tot drie varianten, viz. *pimpen, opleuken* en *oppimpen*, aangezien we in deze sectie een aantal a priori hypotheses willen testen met inferentiële statistiek. Het gebruik van vier of meer varianten zou volgens ons leiden tot een modelcomplexiteit die de interpretatie te sterk zou bemoeilijken (cf. Van de Velde, Franco, en Geeraerts 2019, 335-336; Fahy et al. 2022). Deze studie is met de keuze voor drie varianten daarmee slechts een eerste stap in de richting van realistischere maar noodzakelijkerwijze complexere statistische modellering. Indien het doel van de studie zou zijn hypotheses te genereren in plaats van te testen, zou het wel gemakkelijker zijn meer varianten op te nemen, dan weliswaar met exploratieve technieken (cf. Pijpops, Speelman, en van den Bosch te versch.).

Bij de multinomiale analyse testen we zes hypotheses. Een eerste hypothese stelt dat *pimpen* vaker zal voorkomen in latere tweets, aangezien dit de nieuwe variant is, en dat deze opkomst ten koste gaat van *opleuken* (Van de Velde en Zenner 2010). *Oppimpen* zou dan wat later ten tonele verschijnen, en de opkomst van die afleiding zou op zijn beurt vooral ten koste gaan van *pimpen.*

Een tweede hypothese voorspelt een verschil in register, waarbij we vermoeden dat het inheemse *opleuken* de voorkeur draagt in een formeel register, en het leenwoord *pimpen* vaker voorkomt in een informeel register. De reden is dat het hier om een leenwoord gaat, en *pimpen* vanwege zijn oorspronkelijke betekenis in het Engels bovendien als aanstootgevend

aangevoeld kan worden, terwijl *opleuken* 'braver' zou klinken. *Oppimpen* zou zich als mengvorm dan weer tussen beide in bevinden: het partikel *op* is duidelijk inheems, wat het werkwoord minder herkenbaar maakt als uitheemse variant. Al onze data bestaan uit tweets, dus deze hypothese kunnen we niet rechtstreeks testen. We kunnen de hypothese echter wel testen door de aanname te maken dat langere tweets doorgaans in een formeler register geschreven zijn dan kortere tweets. Langere tweets vereisen immers een langere tijd om op te stellen, zodat de auteur er waarschijnlijk ook enige redactionele controle overheen laat gaat voor die de tweet verzendt. Hieruit volgt de voorspelling dat *opleuken* vaker voorkomt in langere tweets, gevolgd door *oppimpen*, en ten slotte *pimpen*. Deze hypothese wordt geoperationaliseerd door de lengte van de tweet te meten in het aantal tekens. Hierbij wordt het woord van de variant zelf, bv. *opgeleukt* of *gepimpt*, als slechts één teken geteld. De reden is dat *opgeleukt* bijvoorbeeld meer tekens bevat dan *gepimpt*. Indien we deze tekens zouden meetellen, zou de voorspelling van de hypothese circulair zijn.

Een derde hypothese betreft de sociale invloed van de twitteraar. Een succesvolle opkomst van een nieuwe talige variant zoals *pimpen* houdt vaak in dat de eerste gebruikers een zeker sociaal aanzien genieten dat uitstraalt op de talige variant (Blythe en Croft 2012). We vermoeden dat dit ook geldt voor *pimpen*: de plotse doorbraak van *pimpen* is wellicht deels te verklaren doordat zijn eerste gebruikers een breed sociaal bereik hadden. We beschikken over drie maatstaven om dit sociale bereik te meten, namelijk het aantal volgers van een twitteraar, het aantal mensen dat gevolgd wordt door de twitteraar, en het aantal tweets dat hij of zij al uitgestuurd heeft. We verwachten hierbij dat *oppimpen*, als nieuwere variant dan *pimpen*, nog vaker dan *pimpen* gebruikt wordt door meer sociaal verbonden en actievere twitteraars.

Een vierde hypothese luidt dat er een verschil zal zijn tussen mannelijke en vrouwelijke twitteraars. Engelse leenwoorden worden doorgaans vaker gebruikt door mannen dan door vrouwen (Sharp 2001; Zenner, Speelman, en Geeraerts 2014). We verwachten dan ook dat mannen de sterkste voorkeur hebben voor *pimpen*, terwijl vrouwen *opleuken* verkiezen, en het meer 'vernederlandste' *oppimpen* zich tussen beide in bevindt.

Een vijfde hypothese stelt dat *opleuken* vaker zal voorkomen in Nederlandse tweets dan in Belgische. De reden is dat het adjectief *leuk* vaker voorkomt in Nederland dan in België, en we daarom vermoeden dat dat ook voor *opleuken* zal gelden. Voor *oppimpen* verwachten we dat het zich als mengvorm tussen *opleuken* en *pimpen* in bevindt.

Tot slot verwachten we als de zesde hypothese een aantal interactie-effecten tussen het jaartal en de overige variabelen. *Pimpen* heeft als innovatie *from below* wellicht zijn eerste successen geboekt in informeel taalgebruik, maar we weten uit eerder onderzoek dat het ook al in formeel taalgebruik is doorgebroken (Van de Velde en Zenner 2010). Bovendien stammen onze vroegste data uit 2007, 3 jaar na de introductie van *pimpen* in het Nederlands door het programma *Pimp my Ride*. Daarom vermoeden we dat de sterkste groei van *pimpen* tussen 2007 en 2020 eerder in de formelere registers zal plaatsvinden. In de informelere registers zou het zich immers al stevig gevestigd hebben. Voor *oppimpen* geldt dat echter niet: het gaat hier om een jongere variant, zodat we de sterkere groei wel nog in de informelere registers verwachten. Voor de sociale invloed van de twitteraar geldt mutatis mutandis hetzelfde. De groei van *pimpen* is wellicht het sterkste bij twitteraars met een beperkter sociaal bereik, aangezien twitteraars met een breed sociaal bereik het al vroeger zijn gaan gebruiken. Voor het jongere *oppimpen* geldt dan weer dat de groei waarschijnlijk wel eerder zit bij de twitteraars met een breed bereik. Voor geslacht verwachten we dat mannen het Engelse leenwoord *pimpen* vaker zouden gebruiken dan vrouwen. Naarmate de band met het Engels door de tijd echter zwakker wordt en *pimpen* salonfähiger wordt, zou het echter steeds populairder kunnen worden bij vrouwen. Daarom verwachten we een sterkere stijging van *pimpen* bij vrouwen dan bij mannen, en voor het nog meer Nederlands klinkende *oppimpen* geldt hetzelfde. Tot slot verwachten we dat de groei van *pimpen* en *oppimpen* in België minder gehinderd wordt dan in Nederland. Het Nederlands-Nederlandse *opleuken* zou in het noorden immers beter in staat zijn weerwerk te bieden dan in het zuiden.

Om de laatste drie hypotheses te testen, moet de dataset nog verrijkt worden met de nodige informatie over het geslacht en de nationaliteit van de twitteraar. We hebben geen rechtstreekse toegang tot informatie over geslacht, maar kunnen wel een aanwijzing vinden in de gebruikersnamen van de twitteraars. Deze gebruikersnamen kunnen we vergelijken met een lijst voornamen die traditioneel aan jongens worden gegeven, en een lijst namen die traditioneel aan meisjes worden gegeven. Op de website *naamkunde.net* zijn zulke voornaamlijsten vrij beschikbaar (http://www.naamkunde.net/?page_id=293, geraadpleegd op 20 januari 2022). Deze lijsten bevatten alle meisjes- en jongensnamen die tussen 1983 en 2006 vaker dan 26 maal zijn aangemeld bij de Sociale Verzekeringsbank, de instantie die in Nederland kinderbijslag uitbetaalt.

De gebruikersnamen zijn via het volgende algoritme vergeleken met de namenlijsten. Eerst werd de gebruikersnaam opgeschoond. Vervolgens werd

de waarde 'onbekend' toegekend aan alle gebruikersnamen korter dan twee tekens, of waarbij er een exacte overeenkomst was in de lijst jongensnamen én meisjesnamen. Indien er slechts in één lijst een exacte overeenkomst was, werd het geslacht van die lijst toegekend. Voor de overige gevallen werden de jongensnamen en meisjesnamen geselecteerd die qua Levenshteinafstand het dichtst aanleunden bij de gebruikersnaam. Indien het verschil in afstand tussen de dichtstbijzijnde jongensnaam en de dichtstbijzijnde meisjesnaam kleiner was dan drie, of indien meer dan de helft van de gebruikersnaam aangepast moest worden om tot de dichtstbijzijnde opgelijste naam te komen, werd aangenomen dat het geslacht onbekend was. In de overige gevallen werd het geslacht van de dichtstbijzijnde naam genomen. De reden om Levenshteinafstand te gebruiken is dat vele gebruikersnamen die zich niet in de namenlijsten bevonden, spellings- of uitspraakvarianten waren van namen die wel opgenomen waren. Zo bevond *Abdillah* zich niet in de lijsten, maar *Abdellah*, *Abdallah* en *Abdullah* wel. Ook *Artheur* bevond zich niet in de lijsten, maar *Arthur* wel.

Om een indicatie van de nationaliteit van de twitteraar te krijgen, maakten we gebruik van de variabelen GEBRUIKERSLOCATIE, PLAATSLAND en PLAATSNAAM. De eerste variabele geeft aan wat de twitteraar zelf heeft aangegeven als woonplaats, terwijl de twee laatste variabelen automatisch zijn toegevoegd door Twitter en het land en de gemeente betreffen van waaruit de tweet verstuurd is. Geen van alle drie was beschikbaar voor alle tweets in de dataset. Deze variabelen hebben we vergeleken met twee lijsten van plaatsnamen in België en Nederland die als volgt zijn samengesteld. Eerst is de lijst genomen van Belgische gemeenten en deelgemeenten tot 2019 van de Belgische federale overheidsdienst Statbel, en de lijst van Nederlandse gemeenten uit 2006 en 2020 van de Nederlandse overheidsdienst Centraal Bureau voor Statistiek. Aan deze lijsten zijn handmatig de provincienamen toegevoegd, de namen van de landen zelf, alsook enkele varianten van die namen, zoals *Belgique* en *Netherlands*, en de namen van enkele regio's, zoals *Achterhoek* of *Vlaanderen*. De lijsten zijn in de eerste plaats vergeleken met de variabele GEBRUIKERSLOCATIE, en slechts in de tweede plaats met PLAATSLAND en PLAATSNAAM. Wanneer iemand vanuit Nederland tweet, maar een Belgische woonplaats heeft opgegeven, schatten we de kans immers groter dat het gaat om een Belgische bezoeker in Nederland, dan dat een Nederlander betreft die woonachtig in België, maar op bezoek is zijn of haar land van herkomst.

De vergelijking met de lijsten van bekende plaatsnamen is via het volgende algoritme gebeurd. De variabele GEBRUIKERSLOCATIE is eerst wat opgeschoond, en vervolgens werd per waarde van GEBRUIKERSLOCATIE gezocht

naar een exacte overeenkomst in de lijsten. Indien er in beide lijsten een exacte overeenkomst was, werd het land beschouwd als onbekend. Indien er in slechts één lijst een exacte overeenkomst werd gevonden, werd het land van die lijst toegevoegd. Als er geen exacte overeenkomst werd gevonden, werd de langste Belgische en Nederlandse plaatsnaam geselecteerd die een onderdeel oftewel 'substring' vormde van de gebruikerslocatie. Indien er zo geen plaatsnaam gevonden werd, of de Belgische en Nederlandse plaatsnaam even lang waren, of de plaatsnaam korter was dan zes tekens, werd het land als onbekend beschouwd. Deze laatste regel werd toegevoegd om te voorkomen dat bijvoorbeeld de gebruikerslocatie *best city in the world* gematcht werd met de Nederlandse gemeente Best. In de andere gevallen werd het land van de langste plaatsnaam geselecteerd. Indien het land op basis van de voorgaande procedure onbekend was, maar de variabele PLAATSLAND aanduidde dat de tweet uit België of Nederland kwam, werd het land op basis van die variabele toegevoegd. Als de tweet geen waarde had voor PLAATSLAND, maar wel voor PLAATSNAAM, en het land was nog steeds onbekend, werd dezelfde procedure als voor GEBRUIKERSLOCATIE toegepast op PLAATSNAAM. Als het land nog steeds onbekend was, werd ten slotte de waarde *onbekend* toegekend. Waar de toekenning van geslacht dus gebeurde op basis van Levenshtein-afstand, werd voor de nationaliteit gekozen voor een toekenning op basis van substrings. De reden daarvoor is dat gebruikerslocatie vaak waarden had zoals *aalbeke kortrijk* of *aalsmeer netherlandsholland*.

We beperkten de data nu tot alle voorkomens van twitteraars waaraan zowel een geslacht als een nationaliteit was toegekend. Zo hielden we 36.969 van de 229.328 voorkomens over – veel minder, maar nog steeds een comfortabel hoog aantal. Dit is opnieuw een belangrijk voordeel van big data: ze stellen ons in staat kieskeurig te zijn in de selectie van data. De nog steeds grote hoeveelheid overblijvende data stelt ons bovendien in staat de talige context van de voorkomens strikt te beperken, door de woordvorm van de voorkomens constant te houden. Zo worden de overgebleven data in hoge mate vergelijkbaar, zodat we een scherp beeld krijgen op de overblijvende factoren die de keuze tussen de varianten bepalen (Grondelaers et al. 2008, 158–160).

Vervolgens is gecontroleerd hoe betrouwbaar deze toegekende waarden voor geslacht en land waren. Dit gebeurde door willekeurig 500 voorkomens uit de dataset te selecteren waarbij de hierboven beschreven algoritmes een land en het geslacht hadden toegekend, en die handmatig te controleren. We konden natuurlijk niet zeker zijn van het geslacht of de nationaliteit van de twitteraar in kwestie, maar we konden de keuzes van de algoritmen wel

vergelijken met onze eigen intuïtieve keuzes. Zo werd nationaliteit Nederland toegekend voor een voorkomen van *pimpen* waarbij GEBRUIKERSLOCATIE de waarde *grens van Nederland & Belgie* had. Het algoritme had immers de substring *Nederland* gevonden. De substring *Belgie* werd ook gevonden, maar deze was korter dan *Nederland*, zodat het algoritme koos voor Nederland. Hier hadden we zelf echter de waarde *onbekend* toegekend. Op die manier bleek dat er in 490 van de 500 manueel gecontroleerde voorkomens, oftewel 98%, een waarde voor nationaliteit was toegekend die overeenkwam met onze eigen keuze. Voor geslacht was er een overeenkomende waarde toegekend in 480 van de 500 gevallen, oftewel 96%. Dit leek voldoende hoog om de toegekende geslachten en nationaliteiten als betrouwbaar te beschouwen.

We kozen ervoor om de data verder te beperken tot alle voorkomens van een onverbogen voltooid deelwoord en zijn spellingsvarianten (zoals *gepimpt, gepimped* en *opgeleukd*). Deze vorm heeft de volgende voordelen. Ten eerste bevinden er zich bij de scheidbaar samengestelde werkwoorden *opleuken* en *oppimpen* zelden andere woorden tussen het scheidbare werkwoordspartikel en de rest van het werkwoord, in tegenstelling tot de finiete vormen. Dit maakt dat we meer vertrouwen hebben dat de voorkomens van *oppimpen* correct zijn geclassificeerd. De enige gevallen waarbij er zich wel andere woorden kunnen bevinden tussen het werkwoordspartikel en de rest van het werkwoord, zijn voorkomens waarbij een hulpwerkwoord zich tussen beide in bevindt in de werkwoordelijke eindgroep, zoals in (5). Ten tweede is het voltooid deelwoord van deze werkwoorden steeds formeel ondubbelzinnig te herkennen als voltooid deelwoord. Dat is niet het geval voor de infinitief, die dezelfde vorm heeft als de meervoudsvorm van de onvoltooid tegenwoordige tijd. De enige uitzondering hierop is de vorm *pimped*, die rechtstreeks uit het Engels ontleend is en zowel dienst kan doen als deelwoord en als onvoltooid verleden tijd. Daarom is deze vorm ook uitgesloten. Tot slot is geen gebruik gemaakt van het onvoltooide deelwoord of van het voltooide deelwoord met *e*-uitgang, omdat we ons zoveel mogelijk wilden beperken tot werkwoordelijke voorkomens.

(5) *Ik zie dat je je boot wat op hebt gepimpt.*

Nu hielden we een dataset met 6759 observaties over, afkomstig van 5570 twitteraars. Veruit de meeste twitteraars, met name 4818, hadden slechts één tweet in onze dataset, maar er zijn er toch een aantal met meerdere tweets. De analysetechniek die we willen gebruiken, multinomiale logistische regressie (zie hieronder), neemt aan dat alle observaties onafhankelijk van elkaar zijn, maar dat is natuurlijk niet het geval voor observaties met

dezelfde auteur. Daarom kiezen we ervoor om per twitteraar willekeurig één observatie te kiezen, en de overige te verwijderen uit de dataset (vgl. Van de Velde en Pijpops 2021, 173). Een alternatief zou zijn om een *random effect* TWITTERAAR te integreren in het model, maar daarvoor hebben we niet gekozen vanwege het hoge aantal waarden dat zo'n *random effect* zou hebben, en omdat het leeuwendeel van die waarden slechts één observatie zou hebben. Door de data op die manier te beperken, bevatte onze dataset nog 3736 voorkomens van *pimpen*, 1032 van *opleuken* en 802 van *oppimpen*. Dit is nog steeds ruim voldoende voor een alternantieonderzoek. Tabellen 5 en 6 in de Bijlage 2 laten zien hoe deze voorkomens verdeeld zijn over de variabelen GESLACHT en LAND.

De volgende keuze is welke statistische techniek we gebruiken om de hypotheses te testen op de data. Aangezien onze afhankelijke variabele drie waarden heeft, namelijk *pimpen, opleuken* en *oppimpen*, en we de invloed van meerdere onafhankelijke variabelen tegelijk willen nagaan, kiezen we voor multinomiale logistische regressie (Gries 2021, 344–353). Multinomiale logistische regressie is een veralgemening van binomiale logistische regressie. In tegenstelling tot bij binomiale logistische regressie kan bij multinomiale logistische regressie de afhankelijke categoriale variabele meer dan twee waardes hebben. Bovendien kan multinomiale regressie net als binomiale regressie interactie-effecten integreren zoals die voorspeld worden door de zesde hypothese. Een belangrijke aanname die de techniek echter maakt, is de onafhankelijkheid van irrelevante alternatieven. Die houdt in dat de kansverhouding tussen twee varianten niet geaffecteerd wordt door het bestaan van een derde variant.

Aangezien multinomiale regressie geen wijd gebruikte techniek is in de taalkunde (maar zie Levshina 2015), verdient deze aanname een woordje uitleg. Stel dat je een café uitbaat waarbij er drie dranken op het menu staan, namelijk limonade, koffie en whisky. Je wil nu de keuze van je klanten tussen deze drie opties modelleren met een multinomiaal logistisch regressiemodel. De aanname van onafhankelijke irrelevante alternatieven houdt dan in dat de mogelijkheid om limonade te bestellen, geen effect heeft op de onderlinge kansverhouding tussen koffie en whisky. Stel dat een klant in een bepaalde situatie 20% kans heeft om limonade te bestellen, 40% kans om koffie te bestellen, en 40% kans om whisky te bestellen. De kansverhouding tussen koffie en whisky is dus 1. Indien limonade van de kaart wordt geschrapt, houdt de aanname in dat die kansverhouding bewaard blijft, en de kans op koffie en whisky dus gelijkmatig stijgt tot elk 50%. In dit voorbeeld is dat aannemelijk, maar dat is niet noodzakelijk het geval. Stel dat de drie dranken whisky, limonade en cola zijn, waarbij een klant in een gegeven situatie 80%

kans heeft om limonade te bestellen, 10% op cola, en 10% kans op whisky. Nu is het waarschijnlijk dat het verdwijnen van limonade in grotere mate ten goede komt aan de cola dan aan de whisky. Wie zin heeft in limonade zal immers waarschijnlijker uitwijken naar cola dan naar whisky, aangezien zowel limonade als cola frisdranken zijn. Voor dit laatste voorbeeld is het dan ook niet aangeraden om multinomiale regressie te gebruiken. Hier ligt het meer voor de hand eerst een binomiaal regressiemodel te bouwen dat de keuze tussen whisky en frisdrank voorspelt, en vervolgens een tweede binomiaal model te bouwen dat binnen de frisdranken de keuze tussen limonade en cola voorspelt.

Dezelfde redenering is op zich ook mogelijk bij ons eerste voorbeeld. Zowel limonade als whisky zijn dranken die doorgaans koud of alleszins niet warm geserveerd worden, terwijl koffie een warme drank is. Is het dan niet aannemelijk dat klanten eerst een keuze maken tussen warme en koude dranken, en vervolgens een keuze tussen de verschillende koude dranken? Dat is mogelijk, maar een andere redenering lijkt even aannemelijk: zowel koffie als whisky zijn dranken die de klant een warm gevoel bezorgen, terwijl limonade voor verfrissing zorgt. Hetzelfde geldt voor *pimpen, opleuken* en *oppimpen*. Enerzijds is het aannemelijk dat *opleuken* en *oppimpen* in dezelfde categorie geplaatst kunnen worden aangezien het beide scheidbaar samengestelde werkwoorden zijn met hetzelfde partikel *op*. Anderzijds kunnen we argumenteren dat *pimpen* en *oppimpen* beide vormen van *pimpen* zijn, en dus samen horen. Aangezien beide redeneringen aannemelijk zijn, is multinomiale regressie een geschikte analysetechniek.

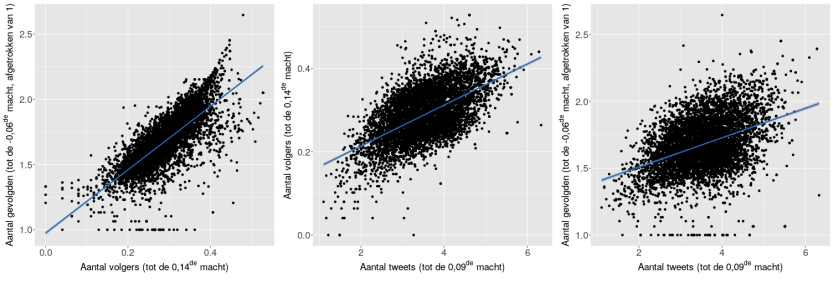
Een andere reden voor de keuze voor multinomiale regressie is dat onze hypotheses verschillende verhoudingen tussen de drie varianten voorspellen. Het is met andere woorden niet het geval dat *oppimpen* zich naar verwachting voor elke onafhankelijke variabele tussen *opleuken* en *pimpen* in bevindt, en we kunnen dus niet zomaar ordinale regressie gebruiken (Gries 2021).

Voor de analyse en visualisatie is gebruikt gemaakt van R (R Core Team 2014), de R-packages *effects* (Fox et al. 2016), *ggplot2* (Wickham 2016), *nnet* (Venables en Ripley 2002), *HandTill2001* (Hand en Till 2001) en *MASS* (Venables en Ripley 2002), alsook enkele functies uit Gries (2021).

De categorische variabelen GESLACHT en LAND zijn via dummycodering geïmplementeerd in het model. De numerieke variabelen zijn eerst getransformeerd. De variabele JAAR, die varieert tussen 2007 en 2020, is eenvoudigweg gecentreerd, terwijl we voor de variabelen TWEETLENGTE, AANTAL VOLGERS, AANTAL GEVOLGDEN en AANTAL TWEETS een Box-Cox-procedure gevolgd hebben om de transformatie te bepalen die de variabelen het best normaliseert (Box en Cox 1964). De niet-getransformeerde verdelingen van

deze variabelen zijn te vinden in Bijlage 3. Aangezien een Box-Cox-procedure niet kan omgaan met nul-waarden, en de variabelen AANTAL VOLGERS en AANTAL GEVOLGDEN over nul-waarden beschikken, is 1 opgeteld bij deze variabelen.[10] De procedure gaf aan de exponenten 0,56, 0,14, -0,06 en 0,09 te gebruiken om respectievelijk de variabelen TWEETLENGTE, AANTAL VOLGERS, AANTAL GEVOLGDEN en AANTAL TWEETS te normaliseren. Omdat de negatieve exponent -0,06 de variabele AANTAL VOLGERS 'omdraait' – de laagste waarde wordt de hoogste waarde en omgekeerd – en dit tegenintuitief is bij het interpreteren van de resultaten, wordt de variabele opnieuw 'omgedraaid' door de resulterende waarde af te trekken van 1.

Vervolgens is gecontroleerd op (multi)collineariteit tussen de onafhankelijke variabelen. De variabelen AANTAL VOLGERS, AANTAL GEVOLGDEN en AANTAL TWEETS operationaliseren hetzelfde concept, namelijk de sociale invloed van de twitteraar. We verwachtten dan ook dat deze variabelen sterk zouden correleren met elkaar, en dat is ook het geval, zoals de spreidingsdiagrammen in Figuur 3 laten zien. We hebben er daarom voor gekozen enkel de variabele AANTAL VOLGERS te behouden in de analyse, aangezien die volgens ons het nauwst aansluit bij het te meten concept, namelijk het sociale bereik van de twitteraar.



Figuur 3: Spreidingsdiagrammen van de correlaties tussen de variabelen AANTAL VOLGERS, AANTAL GEVOLGDEN en aantal TWEETS

Voorts bestaat er ook een kleine tot middelgrote correlatie tussen JAAR en TWEETLENGTE (Pearson's correlatie: 0,32, Cohen 1988). De reden hiervoor is dat Twitter sinds 2006 het maximaal aantal tekens per tweet voor een deel van de gebruikers heeft opgetrokken van 140 tot 280, en sinds 2007 voor alle gebruikers. Daarom is besloten TWEETLENGTE te centreren per jaar. De achterliggende redenering is dat wie een lange tweet verstuurt in 2005, dat ook zou doen in 2008, ook al botst hij of zij op verschillende bovenlimieten. De percentielen van de variabele zouden dan in elk jaar

Tabel 3: Schattingen van het multinomiaal logistisch regressiemodel

Voorkomens *pimpen*:	3736	AUC:	0,621
Voorkomens *opleuken*:	1032	AIC:	9030,10
Voorkomens *oppimpen*:	802	R2:	0,121

Onafhankelijke variabele	Waarde afhankelijke variabele	Schatting coëfficient	Standaard-fout	Z-waarde	P-waarde
INTERCEPT	*opleuken*	-2,68	0,19	-14,28	< 0,0001
	oppimpen	-1,68	0,19	-8,78	< 0,0001
JAAR	*opleuken*	0,26	0,07	3,81	0,0001
	oppimpen	0,05	0,08	0,60	0,5467
TWEETLENGTE	*opleuken*	0,18	0,01	12,95	< 0,0001
	oppimpen	0,05	0,01	3,93	0,0001
AANTAL VOLGERS	*opleuken*	3,98	0,62	6,40	< 0,0001
	oppimpen	0,62	0,66	0,94	0,3453
GESLACHT (referentiewaarde: vrouw)	*opleuken*	0,28	0,08	3,64	0,0003
	oppimpen	0,06	0,08	0,73	0,4642
LAND (referentiewaarde: Nederland)	*opleuken*	-1,01	0,21	-4,75	< 0,0001
	oppimpen	-0,88	0,22	-3,94	0,0001
INTERACTIE JAAR – TWEETLENGTE	*opleuken*	-0,01	0,00	-3,00	0,0027
	oppimpen	-0,01	0,00	-1,19	0,2340
INTERACTIE JAAR – AANTAL VOLGERS	*opleuken*	-0,27	0,21	-1,26	0,2092
	oppimpen	0,09	0,24	0,38	0,7014
INTERACTIE JAAR – GESLACHT	*opleuken*	0,06	0,03	1,92	0,0549
	oppimpen	0,07	0,03	2,14	0,0322
INTERACTIE JAAR – LAND	*opleuken*	-0,14	0,06	-2,17	0,0302
	oppimpen	-0,17	0,08	-2,16	0,0308

ongeveer hetzelfde type tweet bevatten qua relatieve lengte. Daarnaast is het ook zinvol deze variabele te centreren in het licht van de interpretatie: tweets kunnen immers niet korter zijn dan 1 teken. Alle andere correlaties tussen de onafhankelijke variabelen waren zwakker dan 0,2.

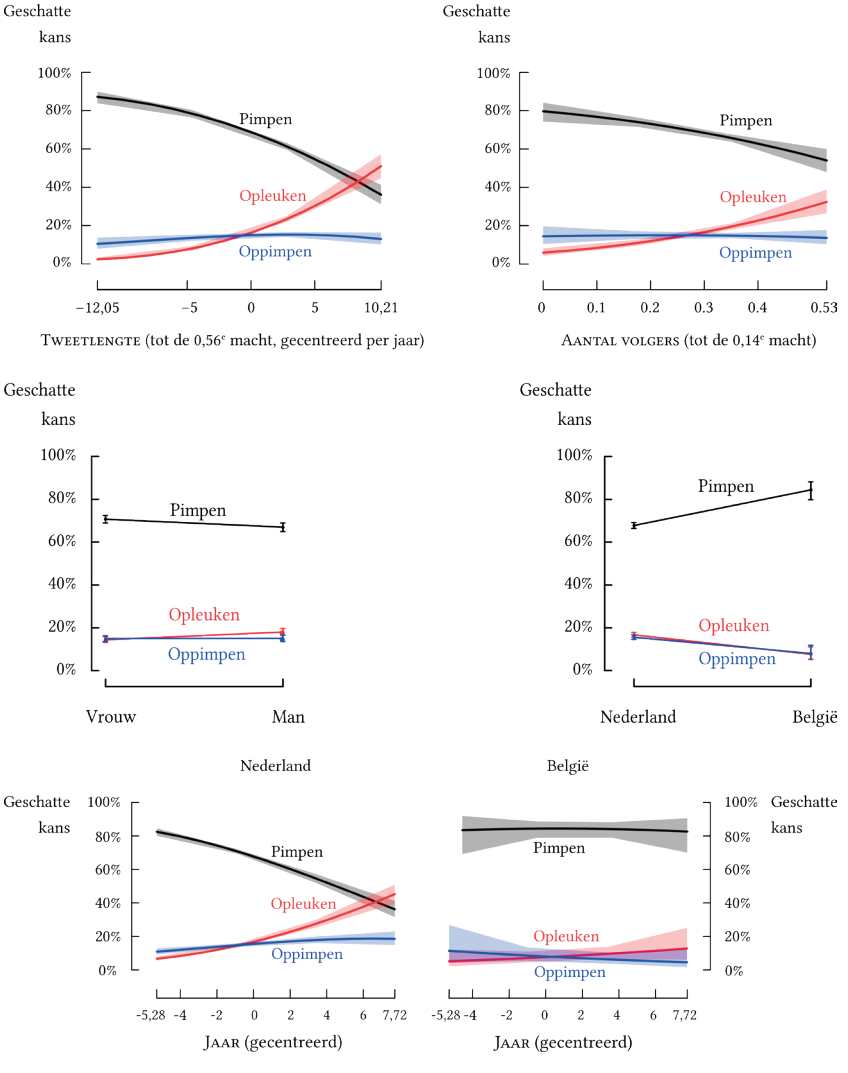
Tot slot is het model gefit op de data met *pimpen* als referentiewaarde. Hiervoor is gebruik gemaakt van de functie *multinom* uit het R-package *nnet* (Venables en Ripley 2002). De specificaties van het model zijn te vinden in

Tabel 4: ANOVA-tabel, vergelijkingen van het gefitte model met telkens een model waarbij de onafhankelijke variabele in kwestie verwijderd is, als main effect en als interactie

Onafhankelijke variabele	Daling AIC	P-waarde
Jaar	194,10	< 0,0001
Tweetlengte	188,20	< 0,0001
Aantal volgers	33,73	< 0,0001
Geslacht	13,55	0,0002
Land	79,06	< 0,0001

Tabel 3 en Tabel 4. In Figuur 4 staan de effectgrafieken van de *main effects* tweetlengte, aantal volgers, geslacht en land, alsook de interactie tussen jaar en land. De effectgrafieken van de overige interacties zijn te vinden in Figuur 6 in Bijlage 4. De AUC (*Area Under the ROC-Curve*, ook bekend als C-index) kan voor een multinomiaal model niet op dezelfde manier berekend worden als voor een binomiaal model. We hebben daarom gebruik gemaakt van de veralgemening beschreven door Hand en Till (2001). De resulterende AUC-waarde in Tabel 3 is laag. Dat is op zich niet verwonderlijk voor een model waar de responsvariabele drie waarden heeft. Hoe meer waarden die heeft, hoe moeilijker het immers is voor het model om de juiste waarde te voorspellen. Toch is het ook niet uit te sluiten dat er nog factoren zijn die de keuze tussen *pimpen*, *oppimpen* en *opleuken* bepalen en die ontbreken in het model, wat ook zou resulteren in een lage AUC-waarde. Zo speelt een semantische specialisatie van de werkwoorden bijvoorbeeld wellicht ook een rol.

Onze eerste hypothese blijkt niet bevestigd. *Pimpen* neemt af doorheen de tijd, althans in Nederland, ten opzichte van een toename van *oppimpen* en vooral van *opleuken*. In België blijven de verhoudingen tussen de drie varianten dan weer stabiel. Misschien is de glorieperiode van *pimpen* voorbij. Labov (2001, 228) stelde al dat media-gedreven taalverandering doorgaans een korte levensduur heeft. Dat lijkt ons wat te sterk gesteld – *pimpen* is intussen opgenomen in woordenboeken, zoals het Van Dale Groot Woordenboek en het Algemeen Nederlands Woordenboek, en er wordt geen bezwaar tegen gemaakt op normatieve websites zoals de taaladvieswebsite van de Vlaamse Overheid en Onze Taal (https://www.vlaanderen.be/taaladvies/pimpen, https://onzetaal.nl/taaladvies/pimpen, geraad-pleegd op 20 januari 2022).[11] Het lijkt dus weinig waarschijnlijk dat het woord binnenkort geheel zal verdwijnen. Het is echter wel mogelijk dat *pimpen* net door zijn initiële en plotse succes een belangrijk deel van



Figuur 4: Effectgrafieken van het model in Tabel 3

zijn sociale aantrekkingskracht al verloren heeft, en zijn gebruik daardoor sinds 2007 achteruitgaat. Om het cru te stellen: wat opgenomen is in Van Dale, is niet meer hip. De hypothese voor *oppimpen* is wel bevestigd in Nederland: dit werkwoord maakt er nog wel een kleine stijging door, al lijkt die ook in kracht af te nemen in de latere jaren. Mogelijk is de sterkere weerbarstigheid van *oppimpen* ten opzichte van *pimpen* te verklaren doordat

dat werkwoord zich meer aan het Nederlands heeft aangepast door het partikel *op* op te nemen.

Onze tweede hypothese is wel bevestigd. Het gebruik van *pimpen* neemt af in langere tweets, waar het gebruik van *opleuken* toeneemt. *Oppimpen* bevindt zich, ook zoals verwacht, tussen beide in, en vertoont geen duidelijke opgang of neergang in langere tweets – al is er wel een significant verschil met *pimpen*. Deze hypothese was gebaseerd op een verschil in register, en we interpreteren de resultaten dan ook zo. *Pimpen* blijkt de meest informele variant, gevolgd door *oppimpen*, terwijl *opleuken* formeler is.

De derde hypothese is niet bevestigd. Naarmate de twitteraar meer volgers heeft, is hij kennelijk minder geneigd om *pimpen* en in mindere mate *oppimpen* te gebruiken. Mogelijk speelt hier opnieuw een effect van register. In onze dataset bevinden zich namelijk een aantal tweets van bekende Nederlanders en Vlamingen. Zulke mensen met veel volgers lopen natuurlijk een grotere kans dat iemand aanstoot neemt aan hun taalgebruik, en zijn daarom misschien meer geneigd mogelijk aanstootgevende woorden als *pimpen* of *oppimpen* te vermijden.

De vierde hypothese is ook niet bevestigd. Vrouwen hebben een voorkeur voor *pimpen*, al is het verschil tussen beide geslachten klein, en lijkt er bij *oppimpen* geen sprake te zijn van een verschil tussen vrouwen en mannen. De vijfde hypothese is dan weer wel bevestigd. *Opleuken* is duidelijk populairder in Nederland. Opvallend is wel dat *oppimpen* zich hierbij lijkt te gedragen zoals *opleuken*, niet zoals *pimpen*, en Nederlanders er ook een voorkeur voor vertonen. Dit ondersteunt de stelling dat *oppimpen* een mengvorm is van *pimpen* en *opleuken*, en niet zomaar een onafhankelijke morfologische uitbreiding van *pimpen*.

Wat de zesde hypothese betreft, zien we enkel een duidelijke interactie tussen JAAR en LAND. De effectgrafieken van de overige interacties zijn te vinden in Bijlage 4. In Nederland neemt het gebruik van *pimpen* doorheen de tijd af ten opzichte van *opleuken* en in mindere mate *oppimpen*, terwijl JAAR in België nauwelijks een effect lijkt te hebben. Mogelijk speelt hier een *rising-tide-lifts-all-the-boats*-effect (Zenner, Heylen, en Van de Velde 2018). Het initiële succes van *pimpen* in de jaren na 2004 zorgt er mogelijk voor dat er meer gepraat wordt over allerlei vormen van opleuken, wat in de jaren daarna het gebruik van het werkwoord *opleuken* opdrijft. Die opdrijving zou enkel of voornamelijk plaatsvinden in Nederland, aangezien *opleuken* vooral daar sterk staat als alternatief voor *pimpen*. Toch moeten we hoe dan ook voorzichtig zijn met deze interpretatie, gezien het lage aantal observaties uit België (zie Bijlage 2).

5. Conclusies

Onze *big data*-benadering van *pimpen* heeft een aantal nieuwe inzichten opgeleverd in de gevalstudie. Enerzijds is de uitbreiding van het werkwoord *pimpen* in het Nederlands een stuk verder gegaan dan eerst werd aangenomen. We wisten al dat het werkwoord zich had verspreid vanuit de eigennaam *Pimp my Ride* over de constructie [pimp POSS N] naar de finiete vormen *pimpt, pimpte* en *pimpten*, maar de uitwaaiering heeft zich kennelijk doorgezet naar een hele resem productieve afleidingen, zoals *aanpimpen, bepimpen* en *herpimpen*. Deze geprefigeerde en scheidbaar samengestelde werkwoorden zijn talrijk, en hebben uiteenlopende, vaak meervoudige betekenissen. Bovendien zijn de meeste infrequent, zodat ze zonder een grote hoeveelheid data verborgen zouden zijn gebleven. Prefigering en de vorming van scheidbaar samengestelde werkwoorden blijken dus erg productieve processen, en ze kunnen nieuwe werkwoorden in het Nederlands binnen enkele jaren of misschien zelfs sneller al met derivationele ornamenten uitrusten. Anderzijds lijkt *pimpen* aan momentum in te boeten. De diversiteit aan nieuwvormingen zoals *doorpimpen, wegpimpen,* enzovoort lijkt de laatste jaren af te nemen, en het ziet ernaar uit dat de alternatieven *opleuken* en *oppimpen* aan een relatieve opkomst bezig zijn. *Pimpen* blijkt ook niet meer zo populair bij invloedrijke taalgebruikers. Twitteraars met veel volgers kiezen liever voor *opleuken*.

Onze voornaamste doelen met dit artikel waren te illustreren dat big data nieuwe kansen bieden voor de taalkunde, maar ook dat we aangepaste methodes nodig hebben om die kansen ten volle te benutten. De kansen zijn menigvuldig. Ten eerste stellen big data ons in staat extreem infrequente, maar theoretisch interessante vormen in beeld te brengen, zoals *napimpen, terugpimpen* en *herpimpen.* Ten tweede geven ze ons de mogelijkheid om kieskeurig te zijn in de selectie van corpusdata, zodat zowel de situationele als talige context strak onder controle gehouden kan worden. Dit hebben we gedaan door ons in de alternantiestudie te beperken tot de data waarvoor we een goede inschatting konden maken van de nationaliteit en het geslacht van de twitteraar, en waarbij het alternerende werkwoord als een onverbogen voltooid deelwoord verscheen. Ten derde laten ze ons toe gemakkelijk meer dan twee alternerende vormen met elkaar te vergelijken, wat de 'accountability' van de studie ten goede komt (Labov 1969; Szmrecsanyi et al. 2016). Een *big data*-benadering heeft echter ook zijn eigen beperkingen. Ook hiervan zijn enkele voorbeelden ter sprake gekomen in de studie. Zo is de grootte van het corpus niet noodzakelijk gekend, is meta-informatie doorgaans niet eenvoudig beschikbaar, en laten traditionele statistische

technieken het soms afweten. Deze beperkingen kunnen we gelukkig deels ondervangen door aangepaste methoden. Zo hoeft de grootte van het corpus niet gekend te zijn om gebruik te maken van de Shannon-entropiemaat voor diversiteit, bijvoorbeeld op basis van de Chao-Wang-Jost schatter. Het gebrek aan meta-informatie hebben we kunnen opvangen door gebruik te maken van publiek beschikbare namen- en plaatslijsten, die ons in staat stelden met algoritmes een inschatting te maken van de nationaliteit en het geslacht van de twitteraar. Het wordt wel sterk aangeraden om een deelverzameling van deze inschattingen manueel te controleren, zodat je een beoordeling kan maken van hun betrouwbaarheid. Ten slotte kunnen nieuwe statistische methodes zoals multinomiale regressie de beperkingen van traditionele methodes opvangen.

De mogelijkheden van big data zijn nog niet ten volle verkend, en we zijn er zeker van dat deze trend zich de komende jaren nog sterk voort zal ontwikkelen. De taalkunde staat gelukkig niet alleen op dit onbekende en onontgonnen terrein. Ook de andere wetenschappen, zoals de biologie en de statistiek, zijn volop bezig de kansen te verkennen die big data bieden. Meer dan ooit reiken big data dan ook de gelegenheid aan om van elkaar te leren.

Referenties

Baayen, R. Harald. 1996. "The effects of lexical specialization on the growth curve of the vocabulary." *Computational Linguistics* 22 (4): 455–480.

Baayen, R. Harald. 2001. *Word frequency distributions*. Dordrecht: Springer Science en Business Media.

Baayen, R. Harald. 2009. "Corpus linguistics in morphology: Morphological productivity." In *Corpus Linguistics. An International Handbook. Volume 2*, geredigeerd door Anke Lüdeling en Merja Kytö, 899–919. Berlijn/New York: De Gruyter Mouton.

Baayen, R.Harald, Jacolien van Rij, Cecile de Cat, en Simon Wood. 2018. "Autocorrelated Errors in Experimental Data in the Language Sciences: Some Solutions Offered by Generalized Additive Mixed Models" In *Mixed-Effects Regression Models in Linguistics*, geredigeerd door Dirk Speelman, Kris Heylen, en Dirk Geeraerts, 49–69. Cham: Springer International Publishing.

Blythe, Richard A., en William Croft. 2012. "S-curves and the mechanisms of propagation in language change." *Language* 88 (2): 269–304.

Box, George, en David Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society. Series B* (*Methodological*) 26 (2): 211–252.

Brezina, Vaclav. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.

Chao, Anne, Peter A. Henderson, Chun-Huo Chiu, Faye Moyes, Kai-Hsiang Hu, Maria Dornelas, en Anne E. Magurran. 2021. "Measuring temporal change in alpha diversity: A framework integrating taxonomic, phylogenetic and functional diversity and the iNEXT. 3D standardization." *Methods in Ecology and Evolution* 12 (10): 1926–1940.

Chao, Anne, Y.T. Wang, en Lou Jost. 2013. "Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species." *Methods in Ecology and Evolution* 4 (11): 1091–1100.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral science*2e editie. Hillsdale: Lawrence Erlbaum Associates.

Corpus Hedendaags Nederlands – CHN (Versie 3.0). Beschikbaar op het Instituut voor de Nederlandse Taal: http://hdl.handle.net/10032/tm-a2-s8. Geraadpleegd op 8/08/2022.

De Pascale, Stefano, Dirk Pijpops, Freek Van de Velde, en Eline Zenner. "Reassembling the Pimped Ride: A Quantitative Look at the Integration of a Borrowed Expression." *Frontiers in Communication* 7: 777312.

Desagulier, Guillaume. 2018. *Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics*. Dordrecht: Springer International Publishing.

Divjak, Dagmar, Natalia Levshina, en Jane Klavan. 2016. "Cognitive Linguistics: Looking back, looking forward." *Cognitive Linguistics* 27 (4): 447–463.

Fahy, Matthew, Jesse Egbert, Benedikt Szmrecsanyi, en Douglas Biber. 2022. "Comparing Logistic Regression, Multinomial Regression, Classification Trees and Random Forests Applied to Ternary Variables: Three-Way Genitive Variation in English." *Data and Methods in Corpus Linguistics*: *comparative approaches,* 194–223. Cambridge/New York: Cambridge University Press.

Fox, John, Sanford Weisberg, Michael Friendly, Jangman Hong, Robert Andersen, David Firth en Steve Taylor. 2016. "Effect Displays for Linear, Generalized Linear, and Other Models." R package version 3.2.

Gale, William A., en Geoffrey Sampson. 1995. "Good-turing frequency estimation without tears." *Journal of quantitative linguistics* 2 (3): 217–237.

Geeraerts, Dirk. 2006. "Methodology in Cognitive Linguistics." In *Cognitive Linguistics: Current Applications and Future Perspectives*, geredigeerd door Gitte Kristiansen, Michel Achard, René Dirven, and Francisco Ruiz de Mendoza Ibañez, 21–49. Berlijn/New York: Mouton de Gruyter.

Good, Irving J. 1953. "The population frequencies of species and the estimation of population parameters." *Biometrika* 40 (3-4): 237–264.

Gold, David L. 1985. "Nouns Ending in -Mobile." *American Speech* 60 (4): 362–366.

Gotelli, Nicholas J., en Anne Chao. 2013. "Measuring and estimating species richness, species diversity, and biotic similarity from sampling data." In *Encyclopedia of Biodiversity*, geredigeerd door Simon A. Levin, 195–211. 2e editie. Waltham, MA: Academic Press.

Gries, Stefan Th. 2009. *Statistics for linguistics with R. A practical introduction*. 1e editie. Berlijn: De Gruyter.

Gries, Stefan Th. 2013. *Statistics for linguistics with R. A practical introduction*. 2e editie. Berlijn: De Gruyter.

Gries, Stefan Th. 2021. *Statistics for linguistics with R. A practical introduction*. 3e editie. Berlijn: De Gruyter Mouton.

Grondelaers, Stefan, Dirk Speelman, en Dirk Geeraerts. 2008. "National variation in the use of er 'there'. Regional and diachronic constraints on cognitive explanations." In *Cognitive Sociolinguistics. Language Variation, Cultural Models, Social Systems*, geredigeerd door Gitte Kristiansen, en René Dirven, 153–203. Berlijn/New York: Mouton de Gruyter.

Grondelaers, Stefan, Esther Veerbeek, Roeland Van Hout, en Astraea Blonk. 2021. "What happened to Twitter when adolescents left? The "Great Exodus" and its consequences for social media-based research on syntactic diffusion." (Paper gepresenteerd tijdens *New Ways of Analyzing Variation* 49 (NWAV49), University of Texas in Austin).

Hand, David J., en Robert J. Till. 2001. "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems." *Machine learning* 45 (2): 171–186.

Haspelmath, Martin, Matthew S. Dryer, David Gil, en Bernard Comrie. 2005. *The world atlas of language structures*. Oxford: Oxford Univerisity Press.

Heidbuchel, Hendrik J.P. 1962. *ABN Woordenboek*. Hasselt: Heideland.

Hilton, Nanna H., en Adrian Leemann. 2021. "Editorial: using smartphones to collect linguistic data." *Linguistics Vanguard* 7, s1: 20200132.

Jakubiček, Miloš, Adam Kilgarriff, Vojtěch Kovár, Pavel Rychly, en Vít Suchomel. 2013. "The TenTen corpus family." In *Proceedings of the 7th International Corpus Linguistics Conference CL*, 125–127. Lancaster: Lancaster University

Janda, Laura A., ed. 2013. *Cognitive Linguistics – The Quantitative Turn. The Essential Reader*. Berlijn, Boston: De Gruyter Mouton.

Kestemont, Mike, en Dirk Van Hulle, eds. 2019. "Theorie en de digitale geesteswetenschappen: Ten geleide." *Tijdschrift voor Nederlandse Taal-en Letterkunde*. 135 (4)

Kestemont, Mike, Folgert Karsdorp, Elisabeth de Bruijn, Matthew Driscoll, Katarzyna A. Kapitan, Pádraig Ó Macháin, Daniel Sawyer, et al. 2022. "Forgotten books: The application of unseen species models to the survival of culture." *Science*. 375 (6582): 765–769.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams Jr, Štěpán Bahnik, Michael J. Bernstein, Konrad Bocian, et al. 2014. "Investigating variation in replicability." *Social psychology* 45 (3): 142-152.

Klein, Richard A., Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginal B. Adams Jr., Sinan Alper, Mark Aveyard, et al. 2018. "Many Labs 2: Investigating variation in replicability across samples and settings." *Advances in Methods and Practices in Psychological Science* 1 (4): 443–490.

Koplenig, Alexander. 2017. "Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions." *Digital Scholarship in the Humanities*. 32 (1): 159–168.

Koplenig, Alexander, en Carolin Müller-Spitzer. 2016. "Population size predicts lexical diversity, but so does the mean sea level--why it is important to correctly account for the structure of temporal data." *PloS One* 11 (3): e0150771.

L'Heureux, Alexandra, Katarina Grolinger, Hany F. Elyamany, en Miriam A.M. Capretz. 2017. "Machine Learning with Big Data: Challenges and Approaches." *IEEE Access*. 5: 7776–7797.

Labov, William. 1969. Contraction, Deletion, and Inherent Variability of the English Copula. *Language: Journal of the Linguistic Society of America*. 45.4: 715–762.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 2001. *Principles of linguistic change, vol. 2: Social factors.* Oxford: Blackwell.

Levshina, Natalia. 2015. *How to do linguistics with R*. Amsterdam: John Benjamins.

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1,* 3e editie. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Manybabies Consortium. 2020. "Quantifying sources of variability in infancy research using the infant-directed-speech preference." *Advances in Methods and Practices in Psychological Science* 3 (1): 24–52.

Moscoso del Prado Martin, Fermín. 2014. "Grammatical change begins within the word: Causal modeling of the co-evolution of Icelandic morphology and syntax." In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 36.

Moscoso del Prado Martin, Fermín. 2015. "Vocabulary, grammar, sex, and aging." *Cognitive Science* 41 (4): 950–975.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, en Ineke Schuurman. 2013. "The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch." In *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, geredigeerd door Peter Spyns, en Jan Odijk, 219–247. Heidelberg: Springer.

Pijpops, Dirk. 2020. "What is an alternation? Six answers." *Belgian Journal of Linguistics* 34: 283–294.

Pijpops, Dirk, Dirk Speelman, en Antal van den Bosch. Te verschijnen. "Generating hypotheses for alternations at low and intermediate levels of schematicity. The use of Memory-Based Learning." *Linguistics Vanguard.*

Mayer-Schönberger, Viktor, en Kenneth Cukier. 2013. *Big data: A revolution that will transform how we live, work, and think*. Boston/New York: Houghton Mifflin Harcourt.

Piantadosi, Steven T., Harry Tily, en Edward Gibson. 2012. "The communicative function of ambiguity in language." *Cognition* 122 (3): 280–291.

Pierrehumbert, Janet, en Ramon Granell. 2018. "On Hapax Legomena and Morphological Productivity." In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 125–130. Brussel: Association for Computational Linguistics.

R Core Team. 2014. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Wenen.

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423.

Rosemeyer, Malte, en Freek Van de Velde. 2021. "On cause and correlation in language change. Word order and clefting in Brazilian Portuguese." *Language Dynamics and Change* 11 (1): 130–166.

Sharp, Harriet. 2001. *English in spoken Swedish: a corpus study of two discourse domains*. Stockholm: Almqvist en Wiksell International.

Sönning, Lulas, en Valentin Werner. 2021. "The replication crisis, scientific revolutions, and linguistics." *Linguistics* 59 (5): 1179–1206.

Speelman, Dirk. 2014. "Logistic regression: A confirmatory technique for comparisons in corpus linguistics." In *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy,* geredigeerd door Dylan Glynn, en Justyna A. Robinson, 487–533. Amsterdam: John Benjamins.

Stoll, Sabine, Balthazar Bickel, Elena Lieven, Netra Paudyal, Goma Banjade, Toya N. Bhatta, Martin Gaenszle, et al. 2012. "Nouns and verbs in Chintang: children's usage and surrounding adult speech." *Journal of Child Language* 39 (2): 284–321.

Szmrecsanyi, Benedikt, Douglas Biber, Jesse Egbert, en Karlien Franco. 2016. "Toward more accountability: Modeling ternary genitive variation in Late Modern English." *Language Variation and Change* 28 (1): 1–29.

Tagliamonte, Sali A. 2012. *Variationist sociolinguistics: change, observation, interpretation*. Chichester: Wiley-Blackwell.

Turpijn, Loes, Samantha Kneefel en Neil der Veer. 2015. *Nationale social media onderzoek 2015*. Amsterdam: Newcom Research en Consultancy.

van der Sijs, Nicoline. 2020. In hoeverre houden geëmigreerde Nederlanders en Vlamingen in de eenentwintigste eeuw vast aan de Nederlandse taal en cultuur? *Internationale Neerlandistiek*. 58.1: 5–21.

Van de Velde, Freek en Eline Zenner. 2010. "Pimp my Lexis: het nut van corpusonderzoek in normatief taaladvies." In *Liever meer of juist minder? Over normen en variatie in taal* geredigeerd door Els Hendrickx, Karl Hendrickx, Willy Martin, Hans Smessaert, William Van Belle, en Joop van der Horst, 51–68. Gent: Academia Press.

Van de Velde, Freek, Karlien Franco en Dirk Geeraerts. 2019. "Reality check voor de kwantitatieve Nederlandse taalkunde: laveren tussen de Scylla van het conservatisme en de Charybdis van de zelfgenoegzaamheid. " *Tijdschrift voor Nederlandse Taal- en Letterkunde* 135 (4): 329–343.

Van de Velde, Freek en Peter Petré. 2020. "Historical linguistics." In *The Routledge handbook of English language and digital humanities,* geredigeerd door Svenja Adolphs en Dawn Knight, 328–359. Londen: Routledge.

Van de Velde, Freek en Joop van der Horst. 2021. "De taalwetenschap: een plaatsbepaling." *Verslagen en mededelingen van de KANTL* 130 (1): 5–23.

Van de Velde, Freek en Dirk Pijpops. 2021. "Investigating Lexical Effects in Syntax with Regularized Regression (Lasso)." *Journal of Research Design and Statistics in Linguistics and Communication Science* 6 (2): 166–199.

Van Hout, Roeland, en Anne Vermeer. 2007. "Comparing measures of lexical richness." In *Modelling and Assessing Vocabulary Knowledge*, geredigeerd door Helmut Daller, James Milton, en Jeanine Treffers-Daller, 93-115. Cambridge: Cambridge University Press

Venables, William, en Brian Ripley. 2002. *Modern applied statistics with S*. 4te editie. New York: Springer.

Wickham, Hadley. 2016. *ggplot2: Elegent graphics for data analysis.* New York: Springer.

Winter, Bodo. 2020. *Statistics for linguists: An introduction using R*. London: Routledge.

Zenner, Eline, Dirk Speelman, en Dirk Geeraerts. 2015. "A sociolinguistic analysis of borrowing in weak contact situations: English loanwords and phrases in expressive utterances in a Dutch reality TV show." *International Journal of Bilingualism* 19 (3): 333–346.

Zenner, Eline, Kris Heylen, en Freek Van de Velde. 2018. "Most borrowable construction ever! A large-scale approach to contact-induced pragmatic change." *Journal of Pragmatics* 133: 134–149.

Notes

1. Dit artikel werd uitgegeven met steun van de Universitaire Stichting van België.
2. Hoewel de integratie van *pimpen* in het Nederlands een duidelijke boost heeft gekregen dankzij de show, en het plausibel is dat het een verbastering vormt van het Engelse werkwoord *to pimp* (zoals het gebruik in de titel), kunnen we nog niet helemaal uitsluiten dat *pimpen* ook als denominatieve (comparatieve) verbalisering van *pimp* 'pooier' deel is gaan uitmaken van het Nederlandse lexicon.
3. Hierbij moeten enkele valkuilen vermeld worden. De eerste tweets uit deze nieuwe datasets dateren van 2007, dus vier jaar nadat de show werd geïntroduceerd en in een periode waarin de schematisatie van het sjabloon al in volle ontwikkeling is, volgens de resultaten uit Van de Velde en Zenner (2010). Dat de nieuwe data toch eenzelfde ontwikkeling toont, is deels ook te wijten aan de mogelijke heruitzending van *Pimp My Ride* in de jaren 2013 en 2014, wat de piek in voorkomens van de vaste eigennaam kan verklaren.
4. In feite bevat de morfologische familie van *pimpen* naast de grondvorm zelf ook een morfologische derivatie die we in de alternantiestudie in sectie 4 bestuderen, nl. *oppimpen*. We hebben evenwel beslist om *oppimpen* uit de diversiteitsberekening te houden omdat dit type onzes inziens enerzijds een aparte semantische status heeft als bijna-synoniem van *pimpen*, in vergelijking met de andere derivaties, en anderzijds omdat de frequentie van *oppimpen* aanzienlijk hoger ligt dan het tweede meest frequente type *uitpimpen* (wat dan weer mogelijk kan duiden op een aparte status binnen de familie).
5. Bij SoNaR stamt een deel van het materiaal weliswaar van voor 2004, maar bij het CHN dateerde, op het moment van raadpleging, het meest recente materiaal van 2019.
6. Good-Turingcorrectie steunt op het inzicht van Good (1953, dat hij toeschrijft aan Alan Turing) dat de probabiliteitsmassa van ongeziene woorden van de woordenschat, geschat op basis van de relatieve frequenties in steekproeven, groot genoeg is om de probabiliteit van de geziene woorden aanzienlijk te vertekenen. Een betrouwbare schatting van die probabiliteit van ongeziene woorden wordt dan berekend als het aantal unieke hapax legomena in de steekproef gedeeld door de grootte van de steekproef (Gale and Sampson 1995; Baayen 1996).
7. In de (toegepaste) taalkunde werden al lang voor de introductie van de Shannon-entropie eigen diversiteitsmaten uitgewerkt (bv. type-token ratio, Guirauds index, Ubers index, Herdans index, zie Van Hout en Vermeer 2007), maar geen van die maten biedt de mathematische voordelen die entropie biedt. Een vergelijking valt buiten het bestek van dit artikel, maar zie Moscoso del Prado Martin (manuscript).

8. Er bestaan geavanceerdere technieken dan een eenvoudig lopend gemiddelde, die ook de autocorrelatie verdisconteren en via een conservatievere schatting van de standaardfout de kans op Type 1-fouten verkleint, maar voor dit artikel zien we af van verdere analyse.
9. Puristische alternatieven zoals *rekentuig* werden pas voorgesteld na de intrede van *computer* in het Nederlands (Heidbuchel 1962).
10. Dit kan conceptueel gerechtvaardigd worden door de twitteraar zelf zowel als volger van zichzelf en als gevolgd door zichzelf te beschouwen – wie tweet, leest immers steeds zijn of haar eigen tweets.
11. In het Algemeen Nederlands Woordenboek wordt *pimpen* wel gelabeld als informeel, en de taaladvieswebsite van de Vlaamse Overheid vermeldt dat het woord vooral voorkomt in jongerentaal.

Bijlage 1

Chao-Wang-Jost entropieschatter

f_1 is het aantal singletons/hapax legomena types
f_2 is het aantal doubletons/dis legomena types
n is de grootte van de steekproef, i.e. de totale frequentie van de tokens van de types
X_i is de frequentie van een woordtype in de steekproef
A is de geschatte gemiddelde relatieve frequentie van singletons/hapax legomena (onder 3 mogelijke scenario's in de steekproef)
H_{cwj} is de eigenlijke Chao-Wang-Jost entropieschatter

$$A = \begin{cases} \dfrac{2f_2}{(n-1)f_1 + 2f_2} \; als f_2 > 0 \\ \dfrac{2}{(n-1)(f_1 - 1) + 2} \; als f_2 = 0, f_1 \neq 0 \\ 1 \; als f_2 = f_1 = 0 \end{cases}$$

$$H_{cwj} = \sum_{1 \leq X_i \leq n-1} \frac{X_i}{n} \left(\sum_{k=X_i}^{n-1} \frac{1}{k} \right) + \frac{f_1}{n} (1 - A)^{-n+1} \left\{ -log(A) - \sum_{r=1}^{n-1} \frac{1}{r} (1 - A)^r \right\}$$

Bijlage 2

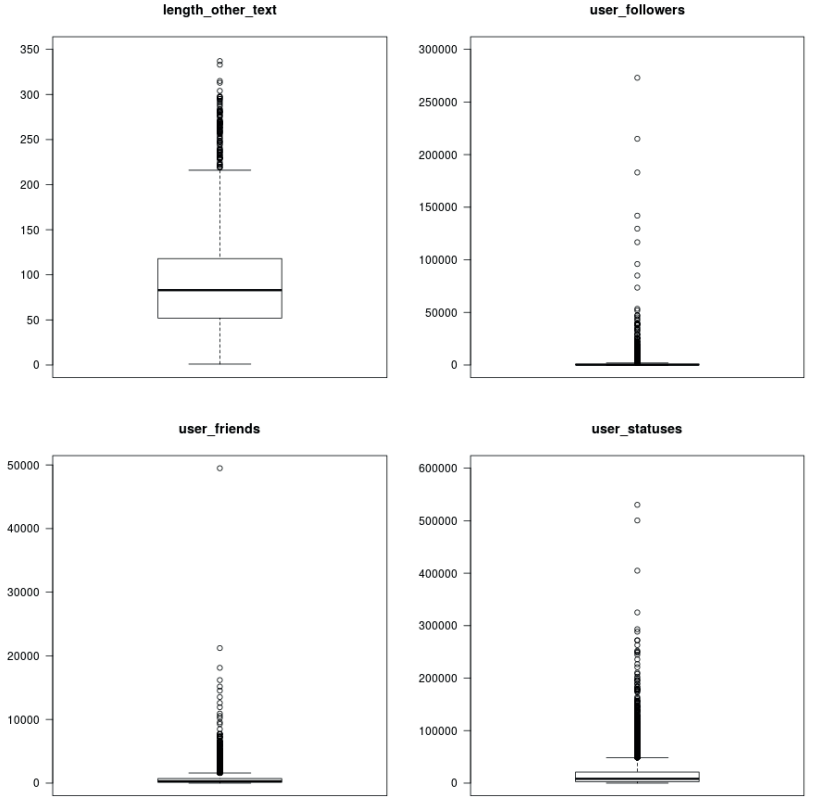
Tabel 5: Verdeling van de data waarop het multinomiaal model is gebaseerd over de responsvariabele en de variabele *GESLACHT*.

	pimpen	*opleuken*	*oppimpen*
Vrouw	2127	465	439
Man	1609	567	363

Tabel 6: Verdeling van de data waarop het multinomiaal model is gebaseerd over de responsvariabele en de variabele land. Het leeuwendeel van onze tweets komt uit Nederland. Er zijn natuurlijk meer Nederlandstalige Nederlanders dan Nederlandstalige Belgen, maar wellicht is Twitter ook populairder in Nederland dan in België.

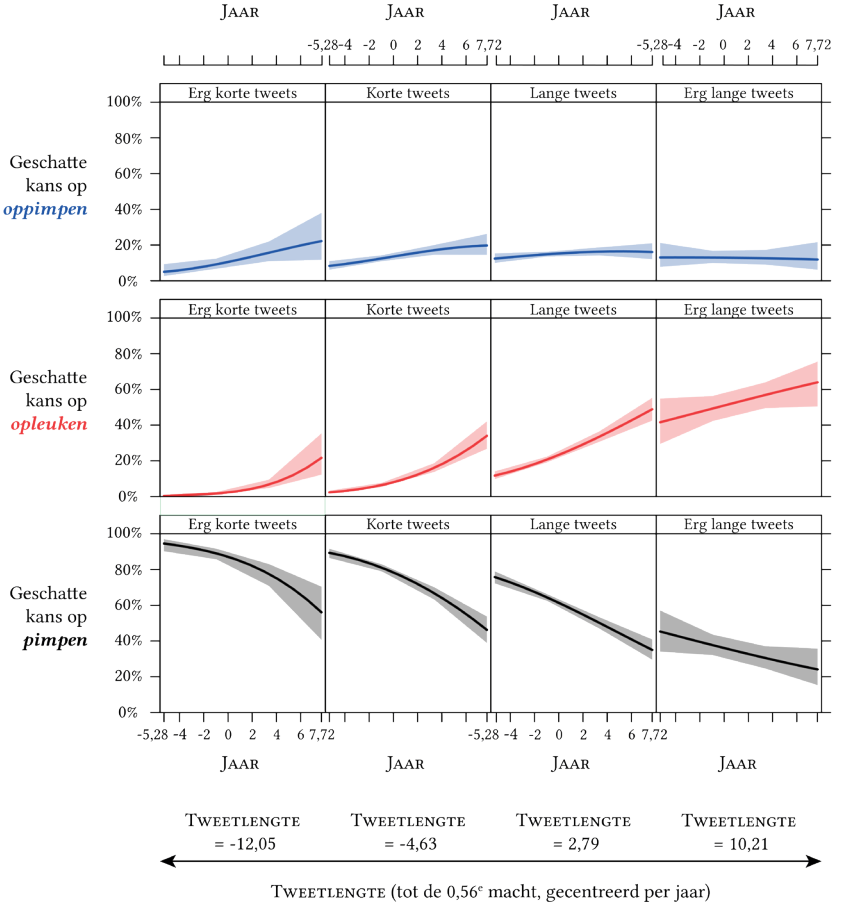
	pimpen	*opleuken*	*oppimpen*
Nederland	3454	994	776
België	282	38	26

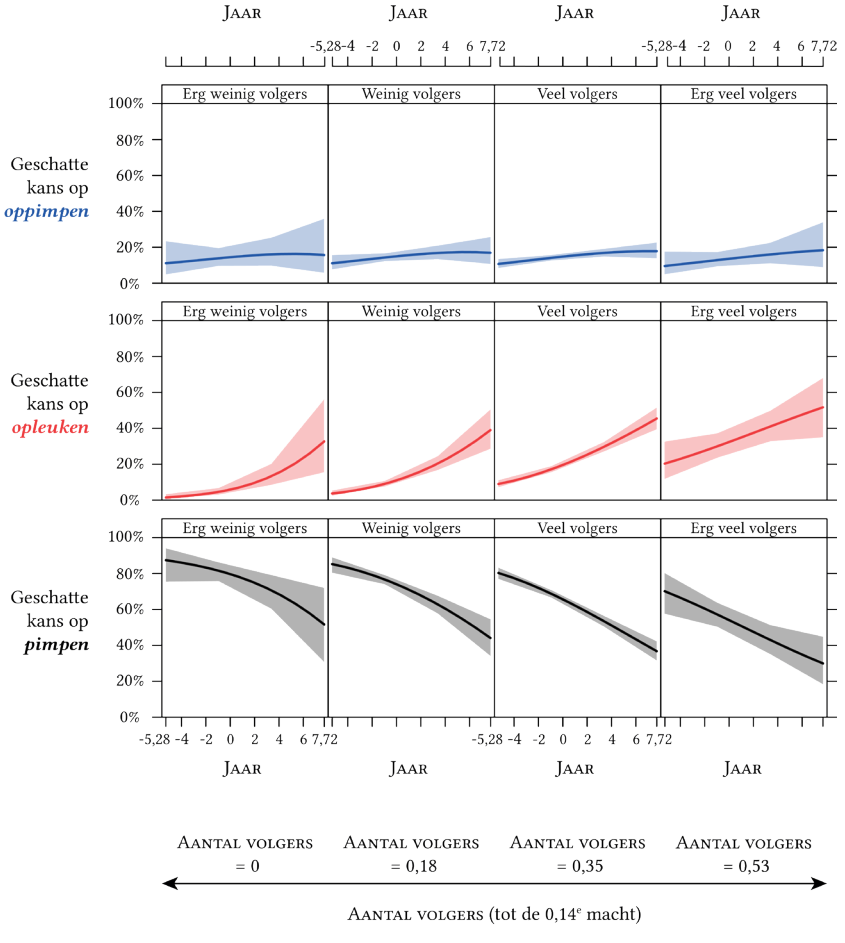
Bijlage 3

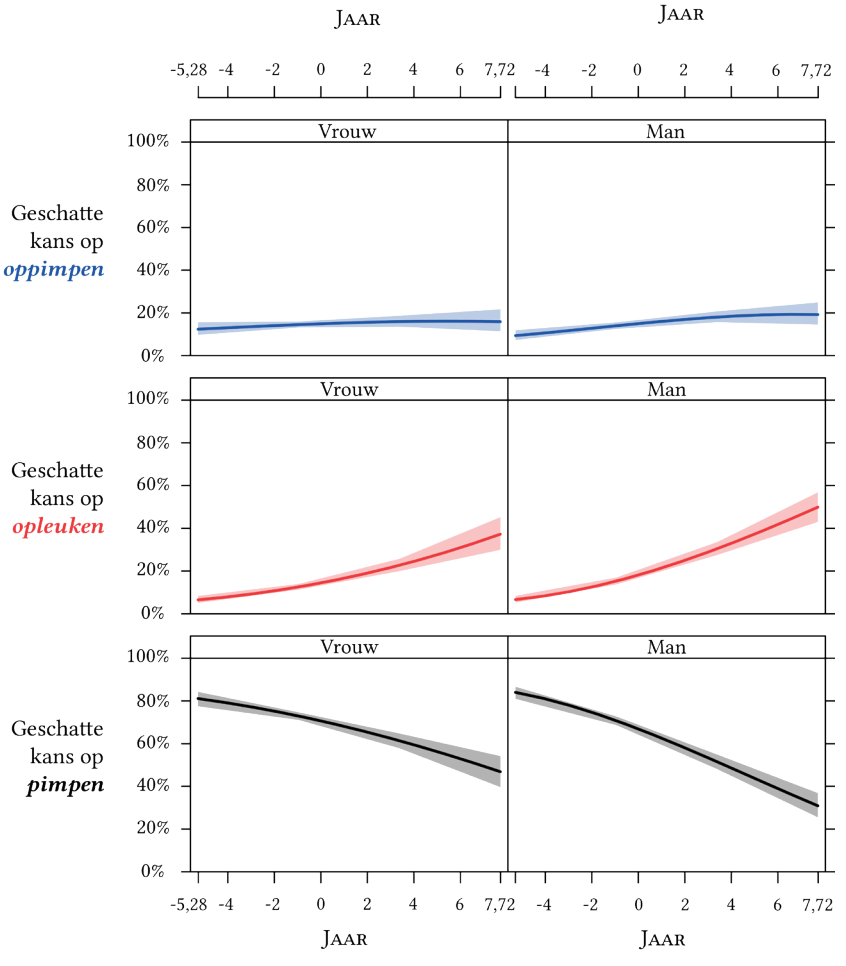


Figuur 5: Boxplots van de niet-getransformeerde variabelen TWEETLENGTE, AANTAL VOLGERS, AANTAL GEVOLGDEN en AANTAL TWEETS. Bij TWEETLENGTE hebben een aantal voorkomens een waarde die hoger ligt dan 280 tekens, de maximale lengte van een tweet. Dit komt omdat er één of een aantal speciale tekens zijn opgenomen, die achterliggend als meerdere tekens gecodeerd zijn.

Bijlage 4







Figuur 6: Effectgrafieken van de interacties tussen JAAR en respectievelijk TWEET-LENGTE, AANTAL VOLGERS en GESLACHT.