

Taalvouten op online datingprofielen: een corpusonderzoek

Het verband tussen het opleidingsniveau van de profieigenaar en de aanwezigheid van
taalfouten en syntactische complexiteit in profielteksten van OkCupid

The relationship between the educational level of the profile owner and the presence of
language errors and syntactic complexity in profile texts on OkCupid



Radboud Universiteit Nijmegen

MA scriptie

Naam: Djuna Kanters (s1025515)

Begeleider: Dr. Nathan Vandeweerd

Tweede beoordelaar: Dr. Jet Hoek

Datum: 06-07-2023

Aantal woorden: 8947

Samenvatting

Dit onderzoek biedt een gedetailleerd inzicht in het taalgebruik in profielteksten op de online datingsite OkCupid. De studie diende als replicatiestudie van Van der Zanden et al. (2018) en trachtte een verband te vinden tussen het opleidingsniveau van profieleigenaren en het taalgebruik in hun profieltekst, specifiek gezien de aanwezigheid van taalfouten en syntactische complexiteit. 300 profielteksten uit een aselechte steekproef werden geanalyseerd op 5 subcategorieën taalfouten en op 2 subcategorieën van syntactische complexiteit. Daarbij werd onderscheid gemaakt tussen 3 groepen opleidingsniveaus: laag-, middelhoog- en hoogopgeleide profieleigenaren. Ten eerste werd geconstateerd dat over het algemeen de meeste taalfouten (apostrof- en spelfouten) door laagopgeleiden werden gemaakt. Voor grammaticale fouten werd er geen verschil gevonden tussen de profieleigenaren met verschillende opleidingsniveaus. Bovendien werd gevonden dat syntactische complexiteit vaker voorkwam in profielteksten van hoger opgeleide profieleigenaren dan in die van laagopgeleide profieleigenaren. Zoals verwacht produceerden hoogopgeleide profieleigenaren de langste T-units. Echter produceerden middelhoogopgeleiden langere *noun phrases* (NP) dan hoog- en laagopgeleiden. De bevindingen suggereerden dat over het algemeen een hoger opleidingsniveau van profieleigenaren samenhangt met meer syntactische complexiteit in een tekst. Op basis van deze resultaten kon voorzichtig geconcludeerd worden dat het opleidingsniveau van profieleigenaren in bepaalde mate invloed heeft op het taalgebruik op hun online dating profiel.

Online dating, taalfouten, syntactische complexiteit, opleidingsniveau, profielteksten

Introductie

In de hedendaagse digitale wereld is de manier waarop interpersoonlijke relaties ontstaan aanzienlijk veranderd ten opzichte van voor het digitale tijdperk. Tegenwoordig komt er vaak online communicatie aan bod voordat *face-to-face* ontmoetingen plaatsvinden (McKenna et al., 2002). Conventioneel daten heeft plaats gemaakt voor online daten (Rosenfeld et al., 2019; Smith & Duggan, 2013). Datingplatform Tinder, dat voornamelijk tijdens de coronacrisis groeide, kende tijdens het meest recente meetmoment in begin 2023 meer dan 75 miljoen maandelijkse gebruikers (Tinder, 2023). Online dating is onder andere gegroeid omdat meer mensen toegang tot het internet hebben, waardoor communicatie nu steeds meer online kan plaatsvinden (Fox et al., 2005; Morales, 2009).

Online datingsites zijn zodanig ingericht dat men profielen van anderen bekijkt en beoordeelt, en contact zoekt met andere profieleigenaren (Hancock et al., 2007). Profieleigenaren kiezen zelf een profielfoto, beantwoorden een aantal vragen over hun opleidingsniveau, leeftijd en lichaamstype en schrijven een profieltekst waarin ze meer informatie over zichzelf geven (Rosen et al., 2008). Online daten is een proces waarin profielen worden bekeken, reacties worden achtergelaten en chatgesprekken worden begonnen. Vaak ontstaan relaties pas nadat er ook *face-to-face* communicatie heeft plaatsgevonden. Echter gaat het online proces vaak pas over op fysieke ontmoetingen wanneer twee potentiële partners veel online hebben gecommuniceerd (McKenna et al., 2002). Hierbij is de impressie die iemand vormt van een datingsite gebruiker cruciaal. Als deze impressie positief is, zal er namelijk eerder worden overgegaan van online gesprekken naar een fysieke ontmoeting (Finkel et al., 2012).

Impressievorming bij online daten

Impressies worden gevormd op basis van beschikbare *cues* over een profieleigenaar. Met *cues* worden signalen die iets zeggen over persoonlijke kenmerken bedoeld (Ellison et al., 2006). Bij computer-gemedieerde communicatie (CMC) zijn er minder *cues* beschikbaar dan bij *face-to-face* communicatie, omdat er geen signalen zoals gezichtsuitdrukking, lichaamsbouw en non-verbale signalen bekend worden (Antheunis et al., 2020). Doordat er minder *cues* aanwezig zijn in CMC, wegen de *cues* die wel beschikbaar zijn zwaarder in impressievorming, volgens de *Social Information Processing* (SIP) theorie van Walther & Burgoon (1992). De SIP-theorie beweert dat gesprekspartners zich aanpassen aan de capaciteit van het communicatiemedium, waardoor relatievorming via CMC even succesvol kan zijn als relatievorming in *face-to-face* communicatie mits er genoeg tijd beschikbaar is. Daarnaast proberen profieleigenaren door het gebrek aan beschikbare *cues* met andere strategieën hun onzekerheid tegenover andere profieleigenaren te verkleinen, bijvoorbeeld door veel vragen te stellen aan hun gesprekspartner om elkaar zo goed mogelijk te leren kennen (Antheunis, 2009; Eek, 2009).

De *cues* die aanwezig zijn op een datingprofiel zijn bijvoorbeeld de profielfoto en de profieltekst (Fiore et al., 2008; Ellison et al., 2006). In het proces van zelfpresentatie zijn *cues* de aspecten die bewust door mensen worden gekozen om zichzelf te presenteren (Goffman, 1956) en in online communicatie worden deze signalen afgegeven om een specifieke indruk op anderen te maken (Donath, 2002). Vooral de profieltekst op een datingprofiel heeft veel invloed op impressievorming, omdat er bepaalde *cues* in taalgebruik zitten die onbewust worden afgegeven (Ellison et al., 2006; Van der Zanden et al., 2020). Deze onbedoelde *cues* zijn volgens eerder onderzoek erg waardevol bij het vormen van impressies (Walther & Parks, 2002; Wotipka & High, 2016). Wotipka & High (2016) stellen namelijk dat de onbedoelde *cues* een accurater beeld geven van een persoon omdat ze een verband laten zien tussen iemands gepresenteerde identiteit en daadwerkelijke identiteit. Onbedoelde *cues* die invloed hebben op

impresievorming in online dating zitten onder andere verstopt in taalgebruik (Lea & Spears, 1992; Walther & D'Addario, 2001; Van der Zanden et al., 2020). Een voorbeeld van een onbedoelde *cue* in taalgebruik is de aanwezigheid van taalfouten.

Taalfouten

Een taalfout is het foutief gebruik van een woord, leesteken of grammaticale constructie en wordt gemaakt wanneer bij iemand de taal nog niet goed genoeg verworven is (Richards & Schmidt, 2002; Norrish, 1983). Taalfouten worden gezien als signalen die aangeven dat een persoon een taal nog niet onder de knie heeft (Hendrickson, 1978). Echter is het maken van taalfouten niet volledig gerelateerd aan taalverwerving, zo kunnen taalfouten namelijk ook het gevolg zijn van onoplettendheid van de spreker of schrijver (Corder, 1974; Kreiner et al., 2002). Onder taalfouten vallen onder andere spelfouten (*theur house* in plaats van *their house*), grammaticafouten (*looking for a girl* in plaats van *I am looking for a girl*) en interpunctiefouten zoals apostroffouten (*youll* in plaats van *you'll*).

Taalfouten zijn voorbeelden van kleine *cues* in taalgebruik die de trigger kunnen zijn voor een negatieve impressie van de tekst of auteur, omdat ze onbewust gemaakt worden en daarom dienen als signalen die informatie geven over hoe de auteur daadwerkelijk is (Van der Zanden et al., 2020). Eerder onderzoek heeft dan ook bewezen dat het maken van taalfouten een negatieve invloed heeft op de impressie die lezers vormen van de auteur (Ellison et al., 2006; Figueredo & Varnhagen, 2005; Kreiner et al., 2002; Liu & Ginther, 2002).

Zo bleek ten eerste uit een experiment in een onderwijssetting, uitgevoerd door Figueredo & Varnhagen (2005), dat studenten de auteur van een tekst die non-homofone fouten bevatte (een woord dat niet bestaat maar wel hetzelfde klinkt) als minder schrijfvaardig, intelligent en van een minder academisch niveau beschouwden dan wanneer de tekst van de auteur geen taalfouten bevatte. Een voorbeeld van een non-homofone fout is de spelling

'favorite' in plaats van *'favourite'*. Ten tweede bleek uit drie kleine experimenten van Kreiner et al. (2002) dat auteurs die spelfouten maken als minder schrijfvaardig werden beoordeeld. Hierbij hadden de non-homofone spelfouten een negatievere invloed op de geschatte schrijfvaardigheid dan typefouten, fouten die onbewust worden gemaakt tijdens het typen (*thubgs* in plaats van *things*). Fouten zoals typefouten worden namelijk meer als slordig gezien dan als een effect van een lage schrijfvaardigheid (Kreiner et al., 2002). Echter worden typfouten voornamelijk gerelateerd aan de onoplettendheid en gebrek aan interesse van de auteur, waar grammaticale fouten worden geassocieerd met de bekwaamheid en intelligentie van de auteur (Kreiner et al., 2002; Queen & Boland, 2016; Van der Zanden et al., 2020).

Taalfouten in online dating

Een type taalfout die vaak voorkomt in online dating is de ellips (Van der Zanden et al., 2018). Ellipsen zijn zinnen waarin het onderwerp en/of de persoonsvorm wordt weggelaten, wat resulteert in een spreektaalige zinsconstructie zoals 'looking for a nice girl', waar 'I am' is weggelaten. In formele taal worden ellipsen niet als grammaticaal correct beschouwd, maar in gesproken taal en korte online teksten zijn ellipsen gebruikelijker (Van der Zanden et al., 2018; Mulder & Hulstijn, 2011). Andere studies beschrijven dat ellipsen bewust worden ingezet om teksten spreekwoordelijk over te laten komen (Klooster, 2001; Tesak et al., 1995; Verheijen, 2016). Van der Zanden et al. (2018) vonden in hun studie dat maar liefst 38% van de gecodeerde taalfouten ellipsen bleken te zijn. Hoewel in de studie van Van der Zanden et al. (2018) niet werd onderzocht of de aanwezigheid van taalfouten in een profieltekst invloed had op de impressie die over de profieleigenaar werd gevormd, toonden Ellison en collega's (2006) dit wel aan. Zij vonden bijvoorbeeld dat de aanwezigheid van spelfouten in een datingprofiel de indruk bij participanten wekte dat de profieleigenaar geen goede educatie had gehad (Ellison et

al., 2006). Bovendien wekten de spelfouten de indruk dat naast het taalgebruik van de auteur, ook andere zaken uit zijn leven zouden lijden onder een gebrek aan oplettendheid.

In de context van online dating zijn profielteksten zonder taalfouten de standaard waar men zich doorgaans aan houdt, ondanks dat online datingsites een informele sfeer hebben (Ellison et al., 2006). In informele settingen wordt niet altijd waarde gehecht aan foutloze teksten, zo bewijzen ook Surkyn et al. (2022). In hun studie werd aangetoond dat slechts 57% van de scholieren tussen de 16 en 20 jaar oud correcte spelling belangrijk vindt in berichten op Whatsapp en Facebook Messenger, terwijl 93% van de scholieren dat belangrijk vindt in schoolopdrachten en berichten naar docenten. Dit verschijnsel zou zich mogelijk ook voor kunnen doen op online datingsites, aangezien het niet perfect willen overkomen op een profiel door het tonen van kwetsbaarheid kan resulteren in meer positieve reacties van andere profieleigenaren (Park et al., 2011). Men voelt namelijk een minder hoge drempel om te reageren op anderen wanneer iemand zich op een menselijke manier presenteert, ten opzichte van wanneer iemand perfect over wil komen. Echter is het bewezen dat het schenden van taalnormen tijdens het online daten een negatief effect zal hebben op de impressies die over de tekst en auteur worden gevormd (Van der Zanden et al., 2020). Dit is ook in lijn met de *Language Expectancy Theory* van Burgoon & Miller (1985), die inhoudt dat men bepaalde normen en verwachtingen opstelt over welk taalgebruik gepast is in welke context. In de context van online dating zijn foutloze teksten de norm (Ellison et al., 2006).

Taalfouten door opleidingsniveau

Bovenstaande studies toonden aan dat het maken van taalfouten een negatieve invloed heeft op impressievorming tijdens het online daten (Ellison et al., 2006; Hargittai, 2006). Het opleidingsniveau van een persoon blijkt een belangrijke voorspeller te zijn voor het maken van taalfouten. Zo hebben verschillende studies al ondervonden dat laagopgeleiden doorgaans meer

taalfouten maken dan hoogopgeleiden in een onderwijssetting (Bonset, 2010; Van Hout, 1999; Schijf et al., 2010), en blijken deze resultaten ook terug te keren in een online context (Van der Zanden et al., 2018; Hargittai, 2006; Van der Zanden et al., 2020). Een belangrijk onderdeel van impressievorming tijdens het online daten, is het geschatte opleidingsniveau van de profieleigenaar (Schoot, z.d.; Eek, 2009). Zo toonde een studie van Eek (2009) aan dat voor mensen op online datingsites het opleidingsniveau van potentiële partners een aspect is waar mensen hun impressies op baseren. Het opleidingsniveau van profieleigenaren wordt op bijna alle datingprofielen vermeld, waardoor het automatisch behoort tot de beschikbare informatie over een profieleigenaar.

Replicatiestudie

De huidige studie dient als replicatie van het onderzoek van Van der Zanden et al. (2018), waarbij wordt getracht dezelfde resultaten omtrent taalfouten te vinden, namelijk dat laagopgeleiden meer taalfouten maken in de profieltekst op hun online datingprofiel dan hoogopgeleiden. Daarvoor ligt de focus op een selectie van grammaticale, spel- en interpunctiefouten waarop in de studie van Van der Zanden et al. (2018) werd gecodeerd. De taalfouten die rechtstreeks zijn overgenomen uit de studie van Van der Zanden et al. (2018) zijn ellipsen, lexicaal homofone spelfouten en non-homofone spelfouten. Hier zijn nog twee subcategorieën taalfouten aan toegevoegd die voornamelijk interessant zijn om in een Engelstalige context te onderzoeken. Dat waren onbepaald lidwoordfouten, omdat het onbepaald lidwoord in het Engels aangepast zou moeten worden op een klinkerklank ('*an apple*' en niet '*a apple*') en apostroffouten, vanwege de samentrekkingen tussen het onderwerp en persoonsvorm waar een apostrof tussen geplaatst hoort te worden ('*I am*' wordt '*I'm*').

De studie van Van der Zanden et al. (2018) was toegespitst op taalfouten, maar ook lexicale verfijndheid en lexicale diversiteit werden onderzocht. Lexicale verfijndheid houdt de

complexiteit van een woord in en wordt gemeten door de lengte van een woord en hoe frequent het gebruikte woord voorkomt in een taal (Vermeer, 2016). Hoogopgeleiden gebruiken doorgaans lexicaal complexere, en vaak langere woorden die minder frequent voorkomen, terwijl laagopgeleiden lexicaal simpelere woorden gebruiken die minder divers zijn. De verwachting was dan ook dat het meten van de lengte en frequentie van woorden een verschil zou aantonen tussen laag- en hoogopgeleiden (Van der Zanden et al., 2018; Le Dorze & Bédard, 1998; Mulder & Hulstijn, 2011).

Lexicale diversiteit betreft hoeveel unieke woorden een tekst bevat. Teksten die geschreven zijn door hoogopgeleiden zijn vaak lexicaal meer divers dan teksten die laagopgeleiden schrijven, omdat hoogopgeleiden vaak beschikken over een grotere woordenschat (Bernstein, 1964; Tummers & Deveneyns, 2016). In de studie van Van der Zanden et al. (2018) werden Nederlandse profielteksten van online datingsites ($n = 1570$) geanalyseerd door automatische tools. Er werd ook getest of hoogopgeleiden hun taalgebruik aanpassen wanneer ze op een datingsite zitten voor zowel hoog- als laagopgeleiden. Lexicale verfijndheid werd geoperationaliseerd door het aantal letters per woord en de woordfrequentie uit het grootste corpus van geschreven Nederlands te meten. Lexicale diversiteit werd gemeten door naar de originele vorm van een woord te kijken (lemma) en vervolgens het aantal unieke woorden te meten aan de hand van de lemma type-token-ratio (TTR). De TTR geeft een indruk van de lexicale diversiteit door de verhouding te geven van hoeveel verschillende woorden iemand gebruikt ten opzichte van het totaal aantal woorden. Het is de meest gebruikte meting van lexicale diversiteit in taalkundig onderzoek (oa. Daller et al., 2003; Lu, 2012; Jelsma, 2015). De resultaten van de studie van Van der Zanden et al. (2018) toonden aan dat, in de context van online daten, het taalgebruik op lexicaal niveau verschilt tussen laag- en hoogopgeleiden. Zo bleek dat profielteksten van hoogopgeleiden meer lexicale verfijndheid en diversiteit bevatten dan profielteksten van laagopgeleiden.

Syntactische complexiteit

De bevindingen van Van der Zanden et al. (2018) laten dus zien dat het maken van taalfouten en lexicale verfijndheid en diversiteit in een tekst in verband staan met het opleidingsniveau van de auteur. Naast lexicale kenmerken van taal zijn ook syntactische aspecten van teksten al vaak onderzocht in onderwijssettingen (oa. Hunt, 1970; Staples et al., 2016; Biber et al., 2011).

In onderzoek naar syntactische complexiteit, de moeilijkheidsgraad van de gebruikte grammatica in een zin, worden het meest frequent de gemiddelde lengte van een T-unit en de gemiddelde lengte van een naamwoordgroep, in het Engels *noun phrase* (NP), gebruikt als meting (Bulté & Housen, 2012; Beers & Nagy, 2009; Staples et al., 2016). Hunt (1970) beschrijft de T-unit als ‘de kleinste eenheid in discourse waar in geknipt kan worden zonder dat er zinsfragmenten buitenvallen’. Hunt (1970) illustreert de T-unit in het volgende voorbeeld, waarin elk cijfer een nieuwe T-unit aanduidt:

1. I like the movie we saw about Moby Dick the white whale
2. the captain said if you can kill the white whale Moby Dick I will give this gold to the one that can do it
3. and it is worth sixteen dollars
4. they tried and tried
5. but while they were trying they killed a whale and used the oil for the lamps
6. they almost caught the white whale. (p. 4-5)

De gemiddelde lengte van een T-unit is volgens Hunt (1970) de sterkste manier om syntactische complexiteit te meten, omdat dit de beste indicatie kan geven van het niveau van de schrijver.

Omdat de T-unit zich richt op zinsdelen, meet hij de complexiteit op zinsniveau en daarbij geldt dat hoe langer een T-unit is hoe grammaticaal complexer de zin is (Klaassen, 2021). Hoewel de gemiddelde lengte van een T-unit regelmatig als meting gebruikt wordt in onderzoek naar syntactische complexiteit, is er echter ook kritiek geleverd op de accuraatheid en omvattendheid

van de meting en de geschiktheid van de meting (Gaies, 1980; Biber et al., 2011; Beers & Nagy, 2009; Staples et al., 2016; Klaassen, 2021). Toch is er, ongeacht deze uitingen van kritiek, genoeg bewijs dat een T-unitmeting een helder beeld van syntactische complexiteit in een tekst kan geven (oa. Bulté & Housen, 2012; Cragg & Nation, 2006; Crowhurst & Piche, 1979). De onbetrouwbaarheid van de T-unit als meting voor syntactische complexiteit zou zich vooral in academische teksten voordoen, omdat de teksten complex zijn op een manier die de T-unit niet goed weer kan geven. Biber et al. (2011) constateerden voorheen dat de gemiddelde lengte van T-units dan ook een beter passende meting is voor het meten van syntactische complexiteit binnen informele alledaagse communicatie. Academische teksten zijn namelijk gecompliceerd op een manier die een T-unitmeting niet goed vast kan stellen (Klaassen, 2021). Omdat de huidige studie zich richt op een informele online context, wordt verwacht dat het analyseren van de gemiddelde lengte van een T-unit een geschikte meting zal zijn voor het analyseren van de syntactische complexiteit in profielteksten op datingsites.

Naast T-units zijn ook naamwoordelijke kenmerken van teksten belangrijke metingen in onderzoek op het gebied van syntactische complexiteit. Naamwoordelijke kenmerken zijn te meten door de lengte van de *noun phrase* (NP), een meting die voornamelijk wordt gebruikt in academische teksten vanwege hun complexe karakter (Staples et al., 2016). De gemiddelde lengte van een NP is een geschikte meting om de syntactische complexiteit in academische teksten te analyseren omdat deze in zulke typen teksten vaak zeer complex zijn (Klaassen, 2021). Hoewel de gemiddelde lengte van een NP voornamelijk als meting wordt ingezet in een academische context, zal de meting ook in de huidige studie over profielteksten op online datingsites gebruikt worden. Omdat de profielteksten afkomstig zijn van profieleigenaren met een opleidingsniveau variërend van laag tot hoog, wordt verwacht een verschil te vinden in de naamwoordelijke complexiteit tussen laag- en hoogopgeleiden, ondanks dat de profielteksten geen academisch doeleinde hebben. In het voorbeeld hieronder, afkomstig uit de studie van

Staples et al. (2016), wordt de complexiteit van NP's geïllustreerd. In deze passage staan slechts twee werkwoorden (onderstreept) en worden de NP's (**dikgedrukt**) uitgebouwd met complementen (*cursief*), waardoor de tekst zeer complex wordt:

Selectivity [of the harvest [on Putauhinu Island]] translates into *large* **differences** [in *harvest* rates [among *weight* **classes**]. [. . .] There is **evidence** [for such **links** [between **characteristics** [of *young* **individuals**] and *life history* **traits** [of adults]] [in many taxa]]. (p. 150)

Ook voor de NP's geldt dat hoe meer woorden de NP bevat, hoe complexer hij is (Staples et al., 2016).

Syntactische complexiteit door opleidingsniveau

Eerder onderzoek heeft al aangetoond dat de syntactische complexiteit van teksten en het opleidingsniveau van de auteur met elkaar in verband staan. Ten eerste werd in een studie van Hunt (1970) onderzocht of het aantal opleidingsjaren een effect had op de lengte van de T-units die werden geproduceerd. De participanten in de studie van Hunt (1970) bestonden uit kinderen van 9 tot 18 jaar oud, mensen met dezelfde baan die de middelbare school hadden afgerond en volwassenen die werkten als redacteur. Zijn resultaten toonden aan dat de participanten die minder opleidingsjaren hadden genoten teksten schreven die minder woorden per bepaling en minder woorden per T-unit bevatten, twee metingen die de moeilijkheid van grammatica aanduiden. Dit gegeven impliceert een verband tussen grammatica en opleidingsniveau. Daarnaast bleek uit onderzoeken van Loban (1976) en Richardson et al. (1976) dat studenten met een hoge taalvaardigheid meer woorden per bepaling en per T-unit produceerden in de teksten die zij schreven dan studenten met een lagere taalvaardigheid.

Ten tweede werd door Staples et al. (2016) aangetoond dat in academische teksten de grammaticale complexiteit vaak tot uiting komt in het schrijven van uitgebreide

naamwoordgroepen (NP). Staples et al. (2016) toonden aan dat studenten aan het eind van hun universitaire opleiding ingewikkeldere NP's gebruikten dan universitaire studenten aan het begin van hun opleiding, terwijl kenmerken op zinsniveau afnamen. Deze studie bevestigt dat voornamelijk naamwoordelijke kenmerken belangrijk zijn in de grammaticale complexiteit van Engelstalige teksten (Staples et al., 2016). Biber et al. (2011) stellen namelijk dat mensen naarmate zij schrijfvriendiger worden, meer gebruik zullen maken van het uitbreiden van NP's. Teksten van ontwikkelde schrijvers zullen daarom complexere NP's bevatten dan teksten van minder ontwikkelde schrijvers. Dit maakt dat het produceren van NP's in profielteksten op online datingprofielen mogelijk gestuurd wordt door opleidingsniveau.

Syntactische complexiteit in impressievorming

Zoals hierboven beschreven werd, zijn het vooral de onbedoelde *cues* die in bijvoorbeeld taal worden afgegeven waardoor anderen hun impressies over die persoon vormen. Deze aspecten geven namelijk een eerlijker beeld van een persoon (Walther & Parks, 2002; Wotipka & High, 2016). Een studie van Van Bree (2000) toonde aan dat men zich niet van alle onderdelen van taal even bewust is, en het gebruik van de syntactische constructies vaak onbewust plaatsvindt (Van Bree, 2000). Syntactische complexiteit heeft daarom mogelijk invloed op de impressies die van auteurs worden gevormd. Aangezien grammaticale aspecten bijdragen aan vloeiende zinnen in een tekst en een indruk geven van iemands schrijfstijl (Beers & Nagy, 2009), zijn deze onderdelen belangrijk om een goede indruk op lezers achter te laten. Het gebruik van zinnen die uit eenvoudige structuren bestaan zal namelijk ten koste gaan van de impressie die van de schrijver wordt gevormd (National Council of Teachers of English, 2004). Dit onderdeel van impressievorming hangt samen met het geschatte opleidingsniveau van de auteur (Ellison et al., 2006). Omdat syntactische complexiteit invloed kan hebben op impressievorming is het relevant om het te onderzoeken in de context van online dating.

Huidige studie

Als replicatiestudie van Van der Zanden en collega's (2018) zal de huidige studie de verschillen in taalgebruik in profielteksten tussen laag- en middelhoog- en hoogopgeleiden in termen van taalfouten onderzoeken. In de studie van Van der Zanden et al. (2018) lag de focus op het onderzoeken van taalaccommodatie door hoogopgeleide profieleigenaren op twee typen datingsites. Ze onderzochten in welke mate hoogopgeleiden hun taalgebruik aanpasten aan hun doelgroep wanneer zij communiceerden op een datingsite specifiek voor hoogopgeleiden. In de huidige studie blijft dat onderdeel uit, aangezien slechts de datingprofielen van één datingsite worden geanalyseerd (OkCupid).

Ook zullen de lexicale verfijndheid en lexicale diversiteit, zoals onderzocht door Van der Zanden et al. (2018), in de huidige studie niet aan bod komen. In plaats daarvan wordt syntactische complexiteit als variabele toegevoegd, omdat er verwacht wordt een positief verband te vinden tussen het opleidingsniveau van profieleigenaren en de aanwezigheid van syntactische complexiteit in hun profielteksten. Deze bevinding zou in lijn zijn met wat eerder onderzoek heeft aangetoond, namelijk dat men syntactisch complexere teksten schrijft naarmate hij hoger opgeleid is (Hunt, 1970; Loban, 1976; Richardson et al., 1976; Staples et al., 2016; Biber et al., 2011). De huidige analyse is een combinatie van een handmatige taalfoutenanalyse en een automatische analyse van de syntactische complexiteit in profielteksten. Met deze replicatie wordt getest of de onderzoeksresultaten over taalfouten in profielteksten van laag- en hoogopgeleiden uit het onderzoek van Van der Zanden et al. (2018) gewaarborgd kunnen worden. Wanneer corresponderende resultaten gevonden worden, zal dit de betrouwbaarheid van de bevindingen van Van der Zanden et al. (2018) vergroten. Aan de andere kant, wanneer resultaten worden gevonden die niet overeenkomen met de bevindingen van Van der Zanden et al. (2018), zal dit nieuwe redenen geven voor vervolgonderzoek naar taalfouten in online dating. Op basis van bovenstaande gegevens wordt de volgende onderzoeksvraag geformuleerd:

‘Wat is het verband tussen het opleidingsniveau van profieleigenaren en de aanwezigheid van taalfouten en syntactische complexiteit in profielteksten op hun online datingprofiel?’

Om de onderzoeksvraag te beantwoorden worden eerst vijf hypothesen met deelhypothesen gevormd over taalfouten:

H1: Profielteksten geschreven door profieleigenaren op OkCupid zullen minder grammaticale fouten zoals onbepaald lidwoordfouten (H1a) en ellipsen (H1b) bevatten naarmate het opleidingsniveau van de profieleigenaar hoger is.

H2: Profielteksten geschreven door profieleigenaren op OkCupid zullen minder apostroffouten bevatten naarmate het opleidingsniveau van de profieleigenaar hoger is.

H3: Profielteksten geschreven door profieleigenaren op OkCupid zullen minder lexicaal homofone spelfouten (H3a) en non-homofone spelfouten (H3b) bevatten naarmate het opleidingsniveau van de profieleigenaar hoger is.

De laatste twee hypothesen richten zich op de syntactische complexiteit:

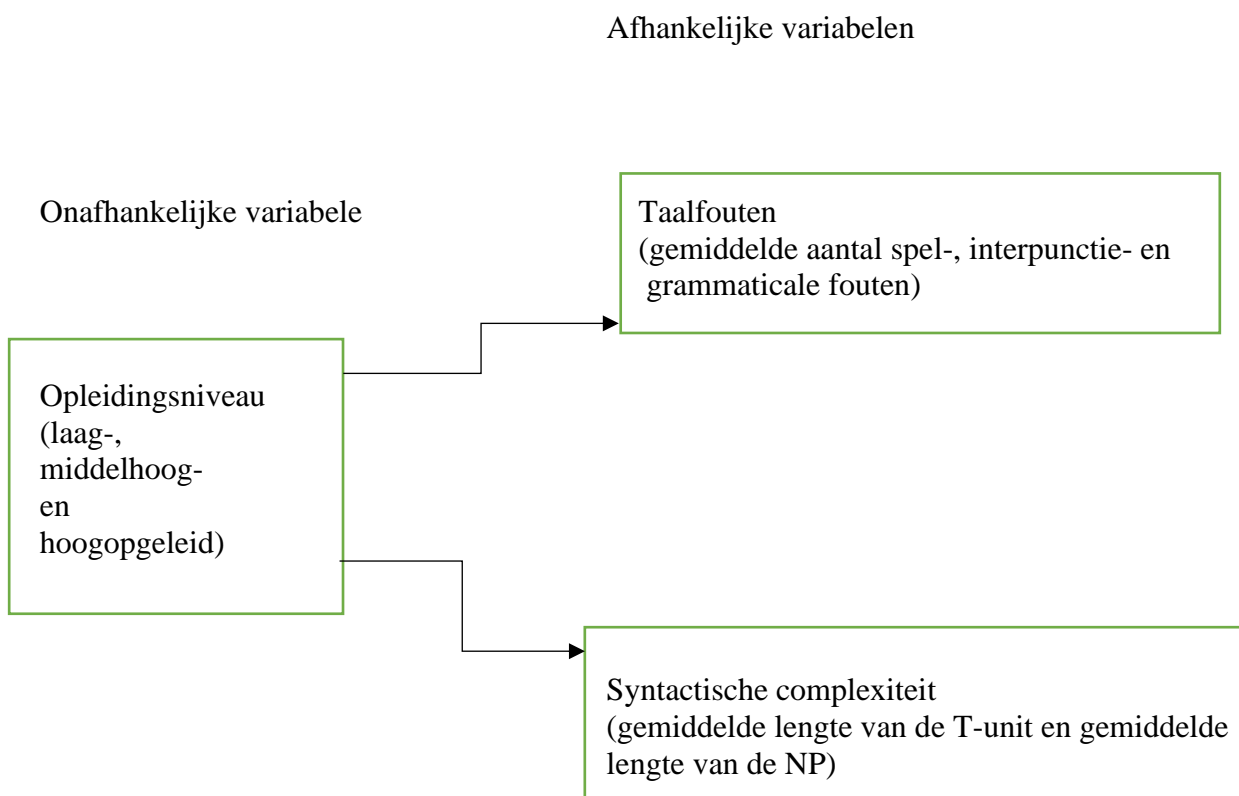
H4: Profielteksten geschreven door profieleigenaren op OkCupid zullen meer woorden per T-unit bevatten naarmate het opleidingsniveau van de profieleigenaar hoger is.

H5: Profielteksten geschreven door profieleigenaren op OkCupid zullen meer woorden per NP bevatten naarmate het opleidingsniveau van de profieleigenaar hoger is.

Het huidige onderzoek ligt ten grondslag aan gedetailleerder onderzoek naar impressievorming op online datingprofielen. Inzicht in taalfouten en syntactische complexiteit op datingprofielen, dat verkregen wordt in de huidige studie, kan leiden tot specifiek onderzoek naar de aspecten die meewegen in impressievorming tussen twee potentiële partners. Indien in de huidige studie verschillen worden gevonden in de aantallen taalfouten en de syntactische complexiteit tussen

profiel eigenaren met verschillende opleidingsniveaus, zou het interessant zijn om vervolgens meer inzicht te verkrijgen in het gevolg van de aanwezigheid van taalfouten in profielteksten op de impressies die van profiel eigenaren worden gevormd. Bovendien is het relevant om meer informatie te krijgen over het verband tussen syntactische complexiteit in profielteksten en de waargenomen aantrekkelijkheid van deze profiel eigenaren, voornamelijk omdat syntactische complexiteit nog niet eerder in de context van online dating is onderzocht.

Figuur 1. Analysemodel.



Methode

Materiaal

Om de hypothesen te toetsen werden de profielteksten van 300 datingprofielen geanalyseerd. De 300 profielen waren een aselekt gegenereerde deelsteekproef van de 33.256 profielteksten die aan de criteria voor het opleidingsniveau van de profieleigenaar en het aantal woorden voldeden (zie sectie *Procedure*). Het corpus was afkomstig van de Amerikaanse datingsite OkCupid en werd gegenereerd door Kim en Escobedo-Land (2015). In 2022 heeft de Ethics Assessment Committee Humanities (EACH) van de Faculteit der Letteren en de Faculteit der Filosofie, Theologie en Religiewetenschappen van de Radboud Universiteit toestemming verleend voor het gebruiken van deze database in taalkundig onderzoek. Hiervoor is door de auteur van het huidige onderzoek een geheimhoudingsverklaring getekend.

De datingprofielen in het corpus beschikten over persoonlijke informatie waarvan de volgende gegevens het meest belangrijk waren voor de huidige studie: leeftijd, opleidingsniveau, etniciteit, baan en geslacht. In Figuren 3, 4 en 5 worden de beschrijvende statistieken van de leeftijd, etniciteit en banen van de profieleigenaren uit de deelsteekproef weergegeven. De profielteksten op de datingsite OkCupid zijn opgedeeld in negen onderdelen waarin profieleigenaren een open antwoord gaven op een vraag. In de huidige studie werd slechts één component hiervan meegenomen, namelijk: “een samenvatting over mijzelf”, net zoals in het artikel van Van der Zanden (2018). Een voorbeeld van een antwoord in het onderdeel van de profieltekst, vertaald naar het Nederlands, is als volgt:

‘i would like to start by describing myself as a fun, outgoing, laid back guy. i'm a really easy going person once you get to know me and give me a chance. a lot of my friends would consider me as the person who they can always rely on. i am honest and loyal, and i expect the same amount from my future significant other. i like to be goofy and laugh and live life to the fullest. at the same time i know when to be serious. i enjoy the

outdoors and taking pictures of the wonderful sceneries. i have just recently adopted an interest in photography. i am even considering taking up classes to learn more about photography and expand my knowledge. i enjoy different kinds of music and i enjoy playing the guitar. i own 18 guitars :) it's become a hobby of mine to collect them.'

Nederlandse vertaling:

'ik zou willen beginnen met mezelf te omschrijven als een leuke, uitgaande, relaxte kerel. ik ben een heel gemakkelijk persoon als je me eenmaal leert kennen en me een kans geeft. veel van mijn vrienden zouden me als beschouwen de persoon op wie ze altijd kunnen rekenen. ik ben eerlijk en loyaal en ik verwacht hetzelfde van mijn toekomstige wederhelft. ik hou van gek zijn, lachen en het leven met volle teugen te leven. tegelijkertijd weet ik wanneer ik serieus moet zijn. ik geniet van buiten zijn en neem graag foto's van de prachtige landschappen. ik ben onlangs geïnteresseerd geraakt in fotografie. ik overweeg zelfs om lessen te volgen om meer over fotografie te leren en mijn kennis uit te breiden. ik luister naar verschillende soorten muziek en ik speel graag gitaar. ik bezit 18 gitaren :) het is een hobby geworden om ze te verzamelen.'

Er werd een aantal stappen uitgevoerd om de deelsteekproef te genereren. Ten eerste werden alle datingprofielen waar de samenvatting niet is ingevuld, geëxcludeerd van het corpus. Ten tweede werd er een selectie gemaakt van de opleidingsniveaus die werden geïnccludeerd in deze studie. Er werd onderscheid gemaakt tussen drie groepen opleidingsniveaus om een accurate representatie van de niveaus te kunnen weergeven. Gebaseerd op eerder onderzoek worden in de huidige studie alleen opleidingen die zijn afgerond geïnccludeerd (Dijk, 2020; Hubers & de Hoop, 2013). De derde groep is een uitzondering, daar zijn ook de niveaus geïnccludeerd wanneer profieleigenaren nog bezig waren met de opleiding. Omdat de derde groep opleidingen bevat die bij uitstek hoger zijn dan de opleidingsniveaus, werd besloten dat ook profielteksten

van profieigenaren die nog bezig waren met die opleiding te includeren. De categorisatie van de opleidingsniveaus wordt weergegeven in Tabel 2.

Tabel 2. Categorisatie en frequentie van opleidingsniveaus in de deelsteekproef van OkCupid

<i>Opleidingsniveau</i>	<i>Opleiding</i>	<i>Frequentie</i>
Laagopgeleid	high school	3 (1%)
	graduated from high school	97 (32%)
Middelhoogopgeleid	college/university	1 (0%)
	graduated from college/university	98 (33%)
	graduated from two-year college	1 (0%)
Hoogopgeleid	graduated from masters program	51 (17%)
	working on masters program	17 (6%)
	graduated from ph.d program	9 (3%)
	working on ph.d program	6 (2%)
	graduated from med school	3 (1%)
	working on med school	5 (2%)
	graduated from law school	3 (1%)
	working on law school	6 (2%)
Totaal		300 (100%)

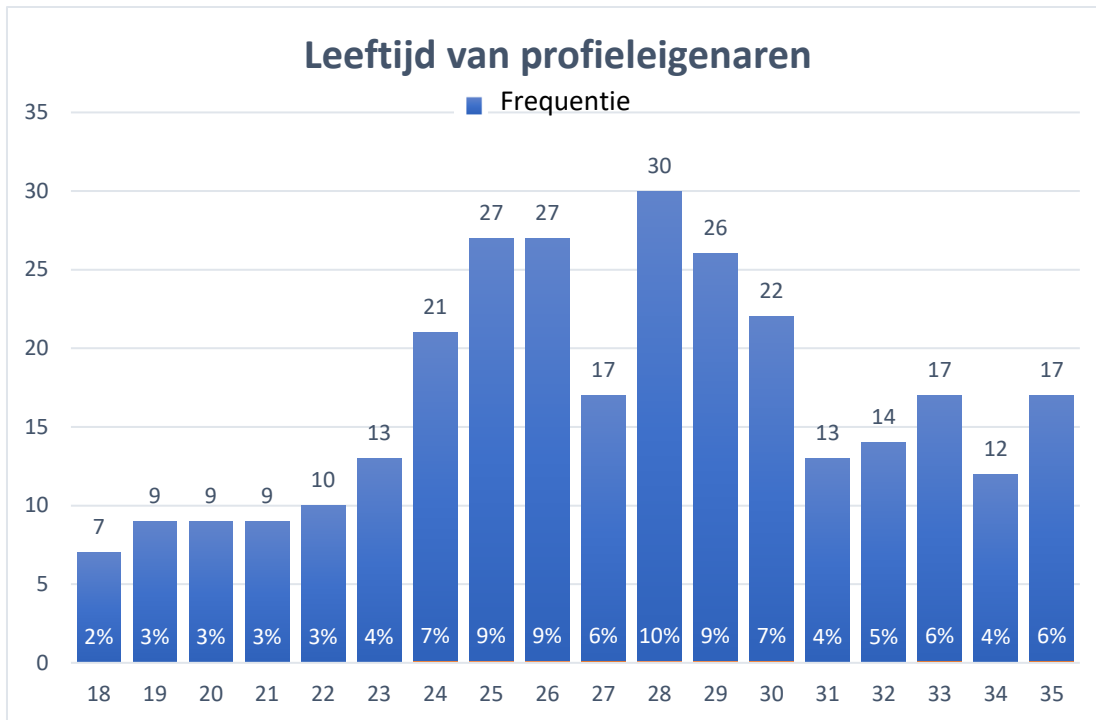
Ten derde werd op leeftijd geselecteerd. Alleen profielen van mensen tussen de 18 en 35 jaar oud werden geïncludeerd in deze studie. Deze beslissing is gemaakt omdat de leeftijdsgroep van 18 tot 35 jaar wordt beschouwd als de doelgroep voor datingplatforms. Omdat de meesten in deze leeftijdscategorie actief zijn op dergelijke platformen wordt specifiek deze groep

afgebakend. Ten slotte werd een minimum en maximum aantal woorden ingesteld voor de teksten in de deelsteekproef. De profielteksten varieerden in woordaantal ($M = 102,36$, $SD = 75,34$), en er werd gekozen voor een minimum van 100 woorden en een maximum van 250 (2 keer SD) woorden als selectiecriteria zodat de profielteksten niet te veel verschilden van lengte. Profielteksten met een woordaantal daarbuiten werden dus geëxcludeerd.

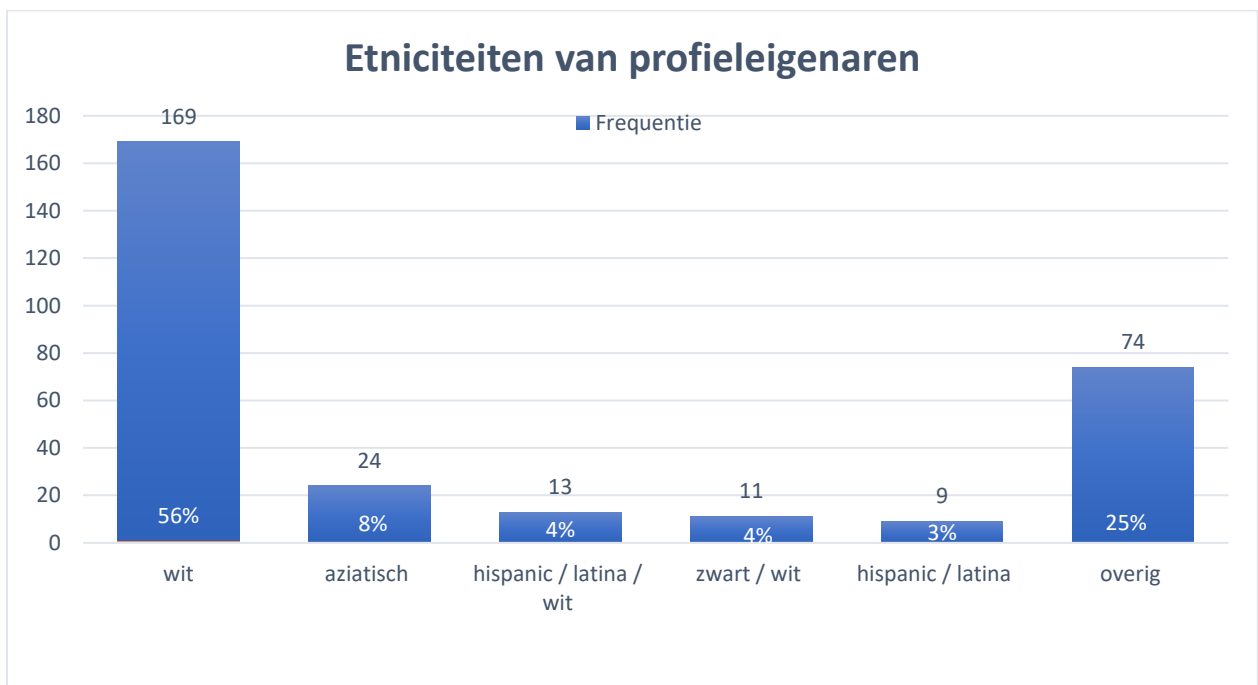
Van de overgebleven profielteksten werd een deelsteekproef van 300 datingprofielen gegenereerd. In de deelsteekproef werd een gelijke representatie van geslacht en opleidingsniveau gepresenteerd. Zo werden er 100 datingprofielen uit elke van de 3 groepen opleidingsniveaus te selecteren, en binnen deze 100 datingprofielen een gelijke verdeling tussen mannen en vrouwen te waarborgen. In Bijlage 1 wordt een uitgebreide beschrijving gegeven van hoe de aselecte deelsteekproef werd gegenereerd.

Vervolgens werden in deze deelsteekproef alle profielteksten ontdaan van foutcodes, en werden de teksten gecorrigeerd op interpunctiefouten en hoofdletterfouten. Deze stappen werden ondernomen zodat de teksten automatisch geanalyseerd kunnen worden op syntactische complexiteit door de tool TAASSC (zie sectie *Procedure*). Het was daarvoor van belang dat de online tool alle woorden zou herkennen, en de tekst dus geen fouten meer bevatte. Foutcodes die vaak terug werden gevonden waren bijvoorbeeld '
' voor 'enter' en '&' voor '&'.

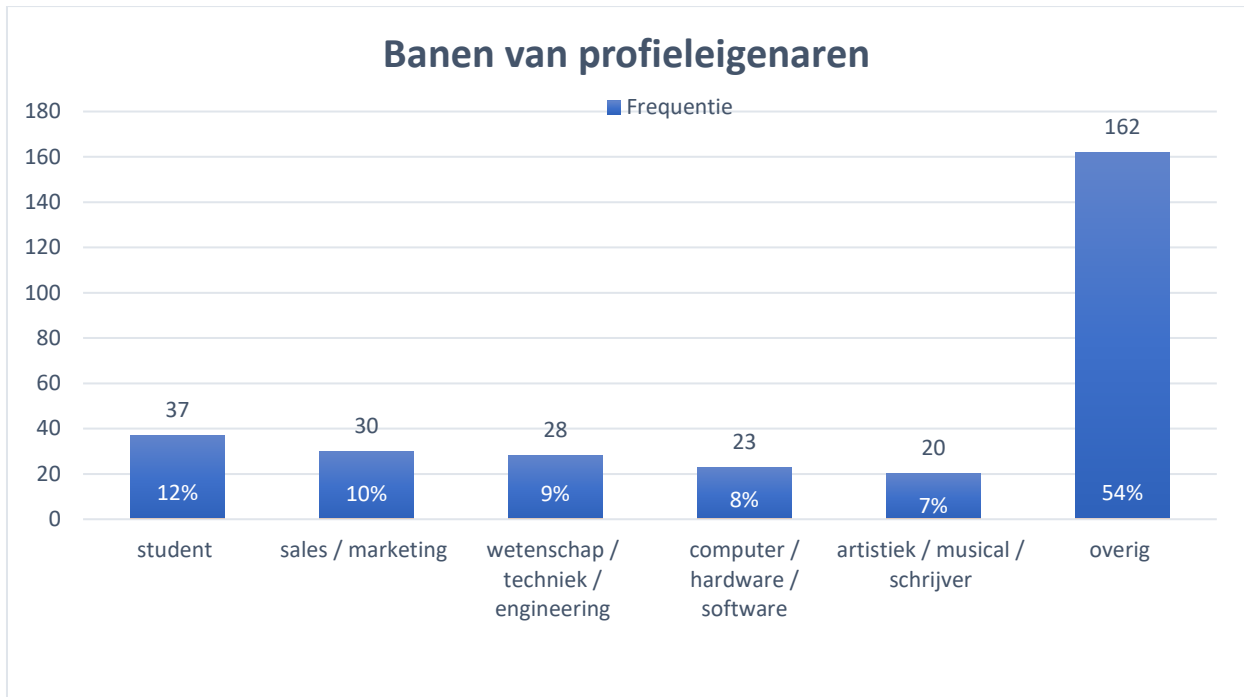
Figuur 3. Verdeling van leeftijd in de deelsteekproef van OkCupid



Figuur 4. Verdeling van etniciteit in de deelsteekproef van OkCupid



Figuur 5. Verdeling van banen in de deelsteekproef van OkCupid



Procedure

Taalfouten

In de huidige studie werd met een selectie van taalfouten gewerkt die gebaseerd was op de studie van Van der Zanden et al. (2018). Binnen deze selectie vallen grammaticale fouten (onbepaald lidwoordfout en ellipsen), interpunctiefouten (apostroffout) en spelfouten (lexicaal homofone fouten en non-homofone fouten). In Tabel 6 wordt een voorbeeld gegeven van elke hoofdcategorie, subcategorie en een voorbeeld van een fout binnen elke subcategorie. In onderstaande tabellen worden de subcategorieën van de taalfouten weergegeven. In Bijlage 2 van dit onderzoek wordt het uitgebreide codeboek weergegeven.

Tabel 6. Categorieën, subcategorieën en voorbeelden van de geanalyseerde taalfouten

Categorie	Subcategorie	Voorbeeld
Grammaticale fouten	1. Onbepaald lidwoordfout	A honest guy i.p.v An honest guy
	2. Ellips van onderwerp en/of gezegde	Working on a show i.p.v I am on a show
working		
Interpunctiefouten	1. Apostroffout	Im i.p.v I'm
Spelfouten	1. Lexicaal homofone spelfout	Their vs. There Hole vs. Whole
	2. Non-homofone spelfout	Favorate i.p.v. Favourite

Alle 300 profielteksten uit de deelsteekproef werden in aparte kolommen gecodeerd op de categorieën taalfouten in een Excel-bestand. Er werd gecodeerd via een turfsysteem, waarin per kolom werd gecodeerd hoeveel fouten per subcategorie voorkomen. Om de betrouwbaarheid van de coderingen aan te tonen werd een subset van 10%, overeenkomend met 30 profielteksten, door een tweede codeur gecodeerd. Deze subset werd willekeurig gegenereerd uit de deelsteekproef en het dubbelcoderen vond afzonderlijk van de eerste codeur plaats. De interbeoordelaarsbetrouwbaarheid van de variabele taalfout was zeer goed voor de subcategorieën apostroffouten ($r(30) = 1.00, p < .001$), lexicaal homofone spelfouten ($r(30) =$

1.00, $p < .001$) en non-homofone spelfouten ($r(30) = .99, p < .001$), goed voor de subcategorie ellipsen ($r(30) = .86, p < .001$) en adequaat voor de subcategorie onbepaald lidwoordfouten ($r(30) = .70, p < .001$). Uiteindelijk werden de gecodeerde fouten per subcategorie genormaliseerd over 100 woorden, zodat de spreiding in het aantal woorden tussen de profielteksten gelijk werd getrokken.

Syntactische complexiteit

De syntactische complexiteit wordt aan de hand van twee metingen door de automatische tool TAASSC gecodeerd. TAASSC is een tool die wordt ingezet voor de syntactische analyse van een tekst. De tool meet grammaticale aspecten die zijn opgedeeld in vijf categorieën: *'clause complexity'*, *'noun phrase complexity'*, *'syntactic sophistication'*, *'components'* en *'syntactic complexity analyzer'* (Kyle, 2006; Lu, 2010). De tool werd gedownload via de website linguisticanalysistools.org. Om de syntactische complexiteit te meten in de profielteksten uit de huidige studie werden twee metingen van TAASSC gebruikt: de gemiddelde lengte van een T-unit en de gemiddelde lengte van een NP. De gemiddelde lengte van een T-unit werd gemeten door de 'Syntactic Complexity Analyzer' van Lu (2010) en gaf het gemiddelde aantal woorden per T-unit als uitkomst van de analyse. De gemiddelde lengte van de NP werd berekend aan de hand van de informatie in het 'Output text' bestand in de analyse van TAASSC. In dit bestand werden per zin alle NP's weergegeven met een corresponderende code voor de bijbehorende syntactische categorie (bijvoorbeeld: *amod_long-nn_car-dobj_rides*). Tussen elk woord stond een koppelteken om ze te onderscheiden, en aan de hand van een formule kon de gemiddelde lengte per NP van elke profieltekst worden berekend. In de berekening van beide metingen van de syntactische complexiteit bleek dat profieltekst 135 en profieltekst 227 niet aan de juiste criteria voldeden om geanalyseerd te kunnen worden door TAASSC. Profielteksten 135 en 227 zijn gedichten en opsommingen waarin alleen komma's stonden en geen punten. Beide teksten

werden konden daarom niet door TAASSC geanalyseerd worden, omdat de tool geen zinnen kan herkennen die niet worden afgesloten door een punt. Beide analyses voor syntactische complexiteit zijn voltooid zonder deze twee profielteksten wat erin resulteerde dat opleidingsgroep 2 en 3 (beiden $n = 99$) een profieltekst minder bevatten dan opleidingsgroep 1 ($n = 100$).

Statistische toetsing

Om te toetsen of er verschillen waren in de aantallen taalfouten tussen de drie opleidingsniveaus werden statistische analyses uitgevoerd. Bij zes van de zeven metingen werd de assumptie van normaliteit geschonden, waardoor non-parametrische toetsen uitgevoerd moesten worden. Voor alle subcategorieën van de taalfouten en voor de gemiddelde lengte van de T-unit werden zes Kruskal-Wallistoetsen uitgevoerd en Mann-Whitneytoetsen als post-hoc toets om inzicht te krijgen in welke opleidingsniveaus significant van elkaar verschilden. Tot slot werd voor de gemiddelde lengte van de NP een eenweg variantie-analyse uitgevoerd, aangezien aan alle assumpties werd voldaan.

Resultaten

Taalfouten

In deze studie werd met een deelsteekproef van 300 profielteksten gewerkt. Hiervan werden 100 profielteksten geschreven door laagopgeleide profieleigenaren, 100 door middelhoogopgeleide profieleigenaren en 100 door hoogopgeleide profieleigenaren. In deze 300 profielteksten werden in totaal 582 taalfouten gevonden. Van de gevonden fouten waren de meesten ellipsen (49%), apostroffouten (30%) of non-homofone spelfouten (13%) en

bestonden de overige taalfouten uit lexicaal homofone spelfouten (7%) en onbepaald lidwoordfouten (1%). In Tabel 7 worden de aantallen per subcategorie taalfouten getoond.

Tabel 7. Beschrijvende statistieken van de gevonden taalfouten in 300 profielteksten

<i>Taalfouten</i>	<i>n (%)</i>
Onbepaald lidwoordfout	6 (1%)
Ellips	283 (49%)
Apostroffout	173 (30%)
Lexicaal homofone spelfout	38 (7%)
Non-homofone spelfout	82 (13%)
Totaal	582 (100%)

De statistische analyses voor de subcategorieën van de taalfouten en de gemiddelde lengte van de T-unit voldeden niet aan de assumptie van normale verdeeldheid en om die reden werden er non-parametrische toetsen uitgevoerd voor deze variabelen. Omdat er meermaals werd getest op dezelfde dataset zijn Dunn-Bonferroni-correcties doorgevoerd voor de p-waarden. Daarnaast werden voor de significante resultaten uit de analyses Mann-Whitneytoetsen uitgevoerd om inzichten te krijgen in welke groepen van elkaar verschilden en om de effectgroottes te meten. Ten eerste bleek uit een Kruskal-Wallis toets geen significant verschil in het aantal gemaakte onbepaald lidwoordfouten ($H(2) = 1.62, p = .445$) en het aantal ellipsen ($H(2) < 1$) tussen profieleigenaren met verschillende opleidingsniveaus.

Vervolgens toonde een Kruskal-Wallis H toets een significant verschil aan in het aantal gemaakte apostroffouten tussen de verschillende opleidingsniveaus ($H(2) = 26.13, p < .001$).

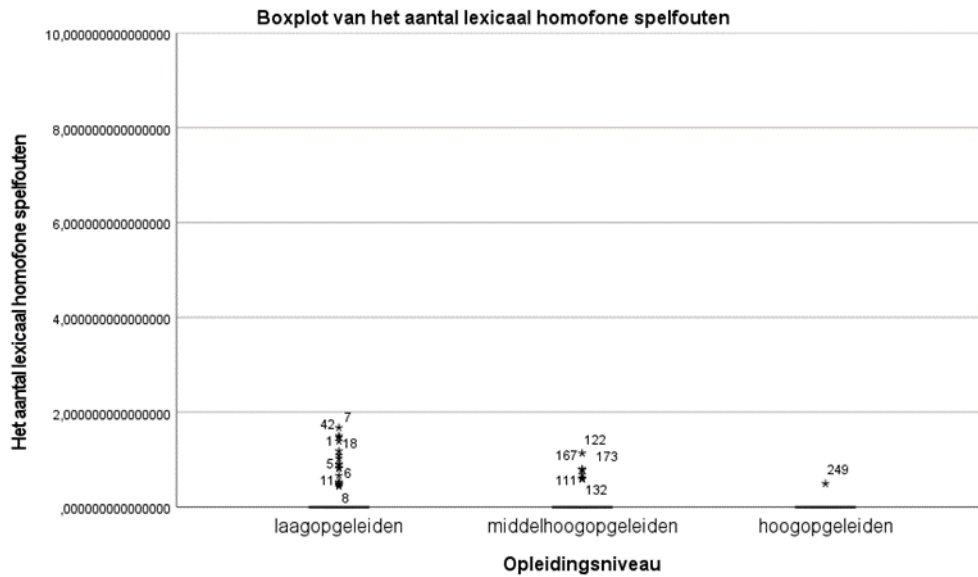
De resultaten demonstreerden dat het aantal gemaakte apostroffouten per 100 woorden door laagopgeleide profieleigenaren ($M = 0.82$, $SD = 1.68$) hoger was dan voor zowel middelhoogopgeleide profieleigenaren ($z = 3.68$, $p < .001$, Bonferroni-correctie; $M = 0.23$, $SD = 0.91$) als hoogopgeleide profieleigenaren ($z = -4.38$, $p < .001$, Bonferroni-correctie; $M = 0.06$, $SD = 0.25$). Het aantal gemaakte apostroffouten door middelhoog- en hoogopgeleide profieleigenaren verschilde niet van elkaar ($p = 1.000$, Bonferroni-correctie). In Figuur 8 wordt de spreiding van de data weergegeven.

Figuur 8. Boxplot van het aantal apostroffouten in profielteksten op OkCupid



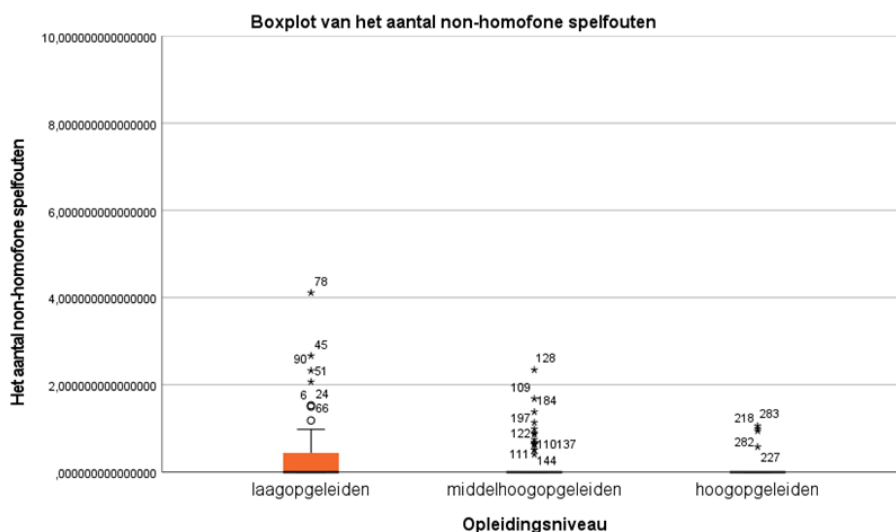
Bovendien bleek uit een Kruskal-Wallis H toets een significant verschil in het aantal gemaakte lexicaal homofone spelfouten tussen opleidingsniveaus ($H(2) = 20.84$, $p < .001$). Laagopgeleide profieleigenaren maakten meer lexicaal homofone spelfouten ($M = 0.17$, $SD = 0.39$) dan middelhoogopgeleide profieleigenaren ($z = -2.57$, $p = .008$, Bonferroni-correctie; $M = 0.05$, $SD = 0.20$) en hoogopgeleide profieleigenaren ($z = -4.25$, $p < .001$, Bonferroni-correctie; $M = 0.00$, $SD = 0.05$). Tussen het aantal gemaakte lexicaal homofone spelfouten door middelhoogopgeleiden en hoogopgeleiden zat geen significant verschil ($p = .420$, Bonferroni-correctie). Zie Figuur 9 voor een visualisatie van de data.

Figuur 9. Boxplot van het aantal lexicaal homofone spelfouten in profielteksten op OkCupid



Ten slotte toonde een Kruskal-Wallistoets een significant verschil aan in het aantal gemaakte non-homofone spelfouten tussen opleidingsniveaus ($H(2) = 18.84, p < .001$). Laagopgeleide profieleigenaren maakten meer non-homofone spelfouten ($M = 0.40, SD = 1.31$) dan hoogopgeleide profieleigenaren ($z = -4.30, p < .001$, Bonferroni-correctie; $M = 0.04, SD = 0.18$). Het aantal gemaakte non-homofone spelfouten door middelhoogopgeleide profieleigenaren verschilde niet significant van laagopgeleide ($p = .078$, Bonferroni-correctie) en hoogopgeleide profieleigenaren ($p = .104$, Bonferroni-correctie). In Figuur 10 wordt de spreiding van het aantal non-homofone spelfouten weergegeven. De gemiddelden en standaardafwijkingen van alle subcategorieën taalfouten worden in Tabel 11 weergegeven.

Figuur 10. Boxplot van het aantal non-homofone spelfouten in profielteksten op OkCupid



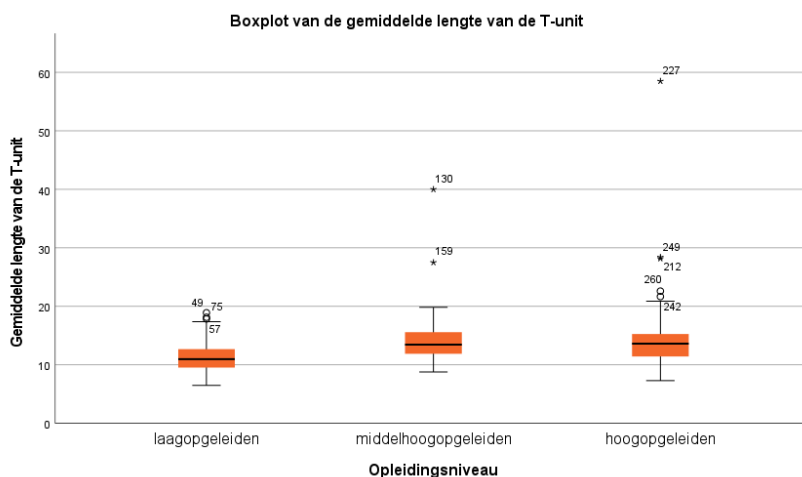
Tabel 11. Gemiddelden en standaardafwijkingen (tussen haakjes) van het aantal taalfouten in profielteksten op OkCupid

	Opleidingsniveau		
	Laagopgeleiden	Middelhoogopgeleiden	Hoogopgeleiden
	$n = 100$	$n = 100$	$n = 100$
<i>Taalfouten</i>			
Onbepaald lidwoordfout	0.02 (0.12)	0.01 (0.08)	0.00 (0.05)
Ellips	0.89 (1.44)	0.61 (1.14)	0.54 (0.73)
Apostroffout	0.82 (1.68)	0.23 (0.91)	0.06 (0.25)
Lexicaal homofone spelfout	0.17 (0.39)	0.05 (0.20)	0.00 (0.05)
Lexicaal non-homofone spelfout	0.40 (1.31)	0.14 (0.39)	0.04 (0.18)

Syntactische complexiteit

Ook in de statistische analyse van de gemiddelde lengte van de T-unit werd de assumptie van normaliteit geschonden, waardoor het uitvoeren van een non-parametrische toets van toepassing was. Een Kruskal-Wallistoets toonde een significant verschil aan in de gemiddelde lengte van de T-unit tussen de verschillende opleidingsniveaus ($H(2) = 39.56, p < .001$). De gemiddelde lengte van de T-unit geproduceerd door laagopgeleide profieeigenaren was korter ($M = 11.36, SD = 2.70$) dan die geproduceerd door middelhoogopgeleide profieeigenaren ($z = -5.72, p < .001$, Bonferroni-correctie; $M = 13.97, SD = 3.99$) en hoogopgeleide profieeigenaren ($z = -5.13, p < .001$, Bonferroni-correctie; $M = 14.19, SD = 5.92$). De gemiddelde lengte van de T-unit verschilde niet significant tussen middelhoog- en hoogopgeleiden ($p = 1.000$, Bonferroni-correctie). Zie Figuur 12 voor de spreiding van de gemiddelde lengte van de T-unit.

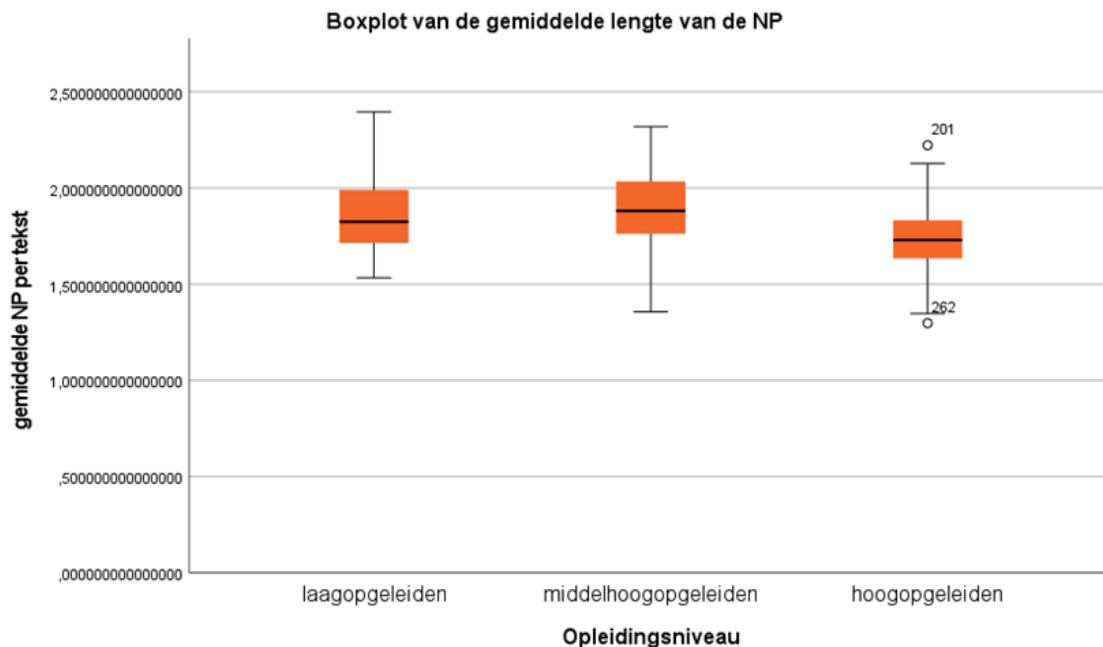
Figuur 12. Boxplot van de gemiddelde lengte van de T-unit in profielteksten op OkCupid



Tot slot werd een statistische analyse uitgevoerd voor de gemiddelde lengte van de NP. De data waren normaal verdeeld en homogeen, waardoor een eenweg variantieanalyse uitgevoerd kon worden. Een eenweg variantie-analyse van opleidingsniveau op de gemiddelde lengte van de NP toonde een significant hoofdeffect van opleidingsniveau aan ($F(2, 295) = 15.87, p < .001, \eta^2 = .10$). Met andere woorden, 10% van de variantie in de scores van de gemiddelde lengte van

de NP wordt verklaard door het opleidingsniveau van de profieleigenaar. De gemiddelde lengte van de NP in profielteksten van middelhoogopgeleiden ($M = 1.89$, $SD = 0.21$) bleek langer dan die van hoogopgeleiden ($p < .001$, Bonferroni-correctie; $M = 1.74$, $SD = 0.18$). Daarnaast was de gemiddelde lengte van NP's langer in profielteksten van laagopgeleiden ($M = 1.85$, $SD = 0.19$) dan in profielteksten van hoogopgeleiden ($p < .001$, Bonferroni-correctie; $M = 1.74$, $SD = 0.18$). De gemiddelde lengte van NP's in profielteksten van laagopgeleiden en middelhoogopgeleiden verschilde niet van elkaar ($p = .667$, Bonferroni-correctie). In Figuur 13 wordt de spreiding van de gemiddelde lengte van de NP weergegeven. De gemiddelden en standaardafwijkingen van de gemiddelde lengte van de T-unit en NP worden in Tabel 14 weergegeven.

Figuur 13. Boxplot van de gemiddelde lengte van de NP in profielteksten op OkCupid



Tabel 14. Gemiddelden en standaardafwijkingen (tussen haakjes) van de syntactische complexiteit in profielteksten op OkCupid

	Opleidingsniveau		
	Laagopgeleiden	Middelhoogopgeleiden	Hoogopgeleiden
	<i>n</i> = 100	<i>n</i> = 99	<i>n</i> = 99
<i>Syntactische complexiteit</i>			
Gemiddelde lengte van de T-unit	11.36 (2.70)	13.97 (3.99)	14.19 (5.92)
Gemiddelde lengte van de NP	1.85 (0.19)	1.89 (0.21)	1.74 (0.18)

Conclusie en discussie

Dit onderzoek geeft een dieper inzicht in het taalgebruik van profieleigenaren op online datingprofielen. Aan de hand van een selectie van vijf subcategorieën taalfouten en twee syntactische complexiteitsmetingen werd onderzocht of er een verband is tussen het opleidingsniveau van profieleigenaren en het aantal gemaakte taalfouten en de syntactische complexiteit in hun profieltekst. In overeenstemming met eerder onderzoek verschilde het taalgebruik van profieleigenaren met verschillende opleidingsniveaus van elkaar op een aantal subcategorieën van taalfouten en syntactische complexiteit.

Taalfouten

Ten eerste toonden de resultaten aan dat het gemiddelde aantal gemaakte onbepaald lidwoordfouten (H1a) niet verschilde tussen laag-, middelhoog- en hoogopgeleiden. Fouten als ‘*a honest person*’ (*an honest person*) werden dus niet minder vaker gemaakt naarmate het opleidingsniveau van profieleigenaren toenam, terwijl dit wel in lijn der verwachting lag.

Hypothese 1a werd dus verworpen. Zoals eerdere studies aantoonde, maken mensen die onderwijs hebben gevolgd op een laag niveau meer grammaticale fouten dan zij die onderwijs hebben gevolgd op een middelhoog of hoog niveau (Bonset, 2010; Van Hout, 1999; Schijf et al., 2010). Een mogelijke reden voor het feit dat deze resultaten niet werden gereproduceerd kan zijn dat de huidige studie andere soort participanten had dan de bovengenoemde studies. De studies van Bonset (2010) en Schijf et al. (2010) werden uitgevoerd met scholieren van het basis- en middelbaaronderwijs als participanten. In de huidige studie werden ‘laagopgeleide profieeigenaren’ gedefinieerd als profieeigenaren die hun middelbareschooldiploma al hadden behaald. Aangezien taalregels over het onbepaald lidwoord in middelbaar onderwijs worden aangeleerd (Bonset, 2010; Schijf et al., 2010), zou het onwaarschijnlijk dat de profieeigenaren in de huidige studie nog onbepaald lidwoordfouten maken. Hier werd ook bewijs voor gevonden, aangezien onbepaald lidwoordfouten zeer zeldzaam bleken te zijn in het corpus van OkCupid. Slechts 1% van de gevonden taalfouten in de deelsteekproef bestond uit onbepaald lidwoordfouten, wat erin kan resulteren dat het onmogelijk was om significante verschillen te detecteren tussen de drie groepen opleidingsniveaus van profieeigenaren. Mogelijk zijn de enkele onbepaald lidwoordfouten in de deelsteekproef afkomstig van profieeigenaren van wie Engels hun tweede taal is (L2). Een studie van Zdorenko & Paradis (2012) illustreerde dat L2-sprekers meer moeite hebben met het aanleren van taalregels over lidwoorden in hun tweede taal, omdat ze dit moeten combineren met de kennis van lidwoorden in hun moedertaal. Dit zou serieuze complicaties kunnen opleveren in het toepassen van taalregels in de L2, en daarom zou het kunnen verklaren dat L2-sprekers mogelijk meer onbepaald lidwoordfouten maken dan L1-sprekers (Lardiere, 2009). Alhoewel onbepaald lidwoordfouten niet werden geanalyseerd in het gereproduceerde onderzoek van Van der Zanden en collega's (2018), toonden hun resultaten wel aan dat er minder grammaticale taalfouten werden gemaakt door profieeigenaren met een hoger opleidingsniveau. Echter werden in hun studie alle subcategorieën van taalfouten

samengenomen, waardoor er geen inzicht is in de resultaten per specifieke taalfout. Toch duidt deze bevinding erop dat er geen overeenstemming is met eerder onderzoek, waar werd aangetoond dat laagopgeleiden meer grammaticale fouten maken dan hoogopgeleiden in een onderwijssetting (Bonset, 2010; Van Hout, 1999; Schijf et al., 2010) en in een online datingcontext (Van der Zanden et al., 2018). Er kan voorzichtig worden geconcludeerd dat het analyseren van onbepaald lidwoordfouten als subcategorie van taalfouten in de huidige studie niet relevant was. Zoals de resultaten aantoonde kwamen onbepaald lidwoordfouten zelden voor in profielteksten geschreven door profieleigenaren die tussen de 18 en 35 jaar oud en in bezit waren van minstens een middelbareschooldiploma.

Ten tweede demonstreerden de resultaten dat er geen verschil zat tussen het aantal ellipsen dat voorkwam in de profielteksten van laag-, middelhoog- en hoogopgeleiden. Met andere woorden, fouten als ‘[I am] looking for a nice guy’ werden niet minder gemaakt naarmate het opleidingsniveau van profieleigenaren toenam. Hypothese 1 werd dus in zijn geheel verworpen. Dit resultaat correspondeerde met de resultaten van Van der Zanden et al. (2018) en zou mogelijk uitgelegd kunnen worden aan de hand van het spreektaalige karakter van ellipsen. Verschillende studies beschreven al dat het bestempelen van ellipsen als taalfouten discutabel is (Van der Zanden et al., 2018; Mulder & Hulstijn, 2011). Van der Zanden et al. (2018) lichtten in hun Nederlandstalige studie toe dat ellipsen volgens de regels van het Standaardnederlands in officiële taal incorrect is, maar dat ellipsen in informelere contexten waar spreektaal wordt gebruikt vaak voorkomen en niet als taalfout gekenmerkt worden. Ellipsen worden in informele teksten vaak gebruikt als stijlfiguren om spreektaal uit te drukken (Klooster, 2001; Tesak et al., 1995; Verheijen, 2016). Net als in de studie van Van der Zanden et al. (2018) werd in de huidige studie gevonden dat het grootste deel van de gevonden fouten in het corpus ellipsen waren (49%). Een mogelijke verklaring voor het feit dat er in de huidige studie geen significant verschil tussen de verschillende opleidingsgroepen werd gevonden, net

als in de studie van Van der Zanden et al. (2018), zou kunnen zijn dat het gebruik van ellipsen niet gestuurd wordt door opleidingsniveau, maar door persoonlijke voorkeuren in informele communicatie. Datingsites hebben een informeel karakter (Ellison et al., 2006; Van der Zanden et al., 2020), wat kan resulteren in het feit dat profieleigenaren er de voorkeur aan geven om profieltekst een spreektaalig karakter te geven. Dat verklaart dat het gebruik van ellipsen op dating sites geaccepteerd wordt door zowel laagopgeleiden als middelhoog- en hoogopgeleiden.

Ten derde bleken de gemiddelden van het aantal apostroffouten in profielteksten wel te verschillen tussen laag-, middelhoog- en hoogopgeleide profieleigenaren. De tweede hypothese werd dus aangenomen. Fouten als '*Im* interested in psychology' (*I'm* interested in psychology) werden minder vaak gemaakt naarmate het opleidingsniveau van profieleigenaren toenam, zoals in de lijn der verwachting lag. Deze bevinding is in overeenstemming met een studie van Johnson (1917). In zijn studie werd aangetoond dat eerstejaarscholieren op middelbaar onderwijs maar liefst twee keer zoveel apostroffouten maakten in drie beschrijvingsopdrachten die zij moesten uitvoeren dan eerstejaarsstudenten aan een Amerikaans *college*. Daaruit kunnen we voorzichtig concluderen dat er een verband is tussen iemands opleidingsniveau en het aantal apostroffouten dat diegene maakt, waarbij mensen die lager zijn opgeleid meer apostroffouten maken dan mensen die hoger zijn opgeleid. Bovendien bevestigde een studie van Bonset (2010) dat scholieren op het middelbaar onderwijs beter zijn in het toepassen van taalregels omtrent interpunctie dan basisschoolleerlingen, mogelijk doordat zij al langer op school zitten en dus meer ervaring hebben met het toepassen van deze taalregels. Deze onderbouwingen zouden kunnen verklaren dat in de huidige studie laagopgeleide profieleigenaren meer interpunctiefouten maakten dan zowel middelhoog- als hoogopgeleide profieleigenaren in hun profieltekst. De huidige bevindingen reproduceren de resultaten van Van der Zanden et al. (2018), waar ook werd gevonden dat het aantal gemaakte interpunctiefouten afnam naarmate het opleidingsniveau van profieleigenaren hoger was. Echter zijn in de studie van Van der

Zanden et al. (2018) geen apostroffouten geanalyseerd zoals dat in de huidige studie is gedaan, maar werd wel voor de overkoepelende categorie ‘interpunctiefouten’ gevonden dat deze vaker voorkwamen in profielteksten van laagopgeleiden dan in die van hoogopgeleiden.

Ten slotte verschilde het aantal lexicaal homofone (H3a) en het aantal non-homofone spelfouten (H3b) tussen de drie groepen opleidingsniveaus van profieigenaren. Hypothese 3 werd dus aangenomen. Fouten als ‘I love to go *their*’ (I love to go *there*) kwamen minder vaak voor in profielteksten naarmate het opleidingsniveau van de profieigenaar toenam. Deze bevinding dat laagopgeleide profieigenaren meer lexicaal homofone spelfouten maakten dan middelhoog- en hoogopgeleide profieigenaren komt overeen met de resultaten van eerdere onderzoeken die geciteerd werden in de huidige studie. Schijf et al. (2010) onderzochten de spellingvaardigheid van leerlingen op een middelbare school met niveaus van vmbo-basis tot gymnasium. Uit dit experiment waar zij in een tekst ontbrekende woorden moesten invullen, bleek dat hoe hoger het niveau was dat de scholieren volgden, hoe beter zij waren in spellen. Ook in een online context werden gelijke resultaten gevonden. Hargittai (2006) analyseerde de mogelijke voorspellers van spelfouten in online berichten van internetgebruikers en vond onder andere dat het opleidingsniveau, variërend van ‘lager van middelbare school’ tot ‘PhD’, een belangrijke voorspeller was voor het maken van spelfouten in deze berichten. Net als in de studie van Schijf et al. (2010) constateerde Hargittai (2006) dat zij die een hogere opleiding genoten hadden minder spelfouten maakten dan zij die een lagere opleiding hadden gevolgd, maar dan op online platformen in plaats van in de schoolbanken. Daarop aansluitend werd ook bevestigd dat het aantal gemaakte non-homofone spelfouten verschilde tussen de drie groepen opleidingsniveaus van de profieigenaren. Fouten zoals ‘This is my *favorate* book’ (This is my *favourite* book) kwamen minder vaak voor in profielteksten naarmate het opleidingsniveau van de profieigenaar hoger was. Er werd echter alleen gevonden dat laagopgeleide profieigenaren meer non-homofone spelfouten maakten dan hoogopgeleide profieigenaren.

Verder werd niet gevonden dat laagopgeleiden meer non-homofone spelfouten maakten dan middelhoogopgeleiden of dat middelhoogopgeleide profieeigenaren meer non-homofone spelfouten maakten dan hoogopgeleiden. Deze bevindingen sluiten aan bij de resultaten van de spellingtest van Schijf et al. (2010), die ook aantoonde dat middelhoogopgeleiden en hoogopgeleiden niet van elkaar verschilden voor twee specifieke typen non-homofone spelfouten. Zij illustreerden dat non-homofone spelfouten zoals ‘*handoek*’ (handdoek) en ‘*apoteek*’ (apotheek) voornamelijk worden gemaakt door scholieren met de laagste twee opleidingsniveaus. De foutenaantallen voor de middelhoge en hoge niveaus waren vergelijkbaar. Daarnaast waren in de studie van Schijf et al. (2010) de gevonden grammaticaal-morfologische fouten voor alle niveaus hoog en verschilden de gemiddelden weinig van elkaar. Desalniettemin sluit het gros van de bevindingen aan bij die van Van der Zanden et al. (2018), aangezien de laagopgeleide groep profieeigenaren zowel meer lexicaal homofone als non-homofone spelfouten maakte dan hoogopgeleide profieeigenaren in hun profieltekst.

Syntactische complexiteit

Naast taalfouten werd in deze studie de syntactische complexiteit in profielteksten van OkCupid onderzocht. Er werd verwacht dat profielteksten meer syntactische complexiteit zouden bevatten naarmate het opleidingsniveau van profieeigenaren toenam. Uit de resultaten bleek ten eerste een verschil tussen de gemiddelde lengte van de T-unit en het opleidingsniveau van profieeigenaren. De vierde hypothese werd dus aangenomen. De gemiddelde lengte van de T-unit in profielteksten geschreven door laagopgeleide profieeigenaren bleek korter dan de gemiddelde lengte van de T-unit in profielteksten van middelhoog- en hoogopgeleiden. Dit resultaat komt overeen met eerdere bevindingen uit een studie van Hunt (1970), waarin werd gevonden dat mensen die minder opleidingsjaren hadden gehad teksten schreven waarin de T-units minder woorden bevatten dan in teksten die werden geschreven door mensen die langer onderwijs hadden gevolgd. Ook in studies van Loban (1976) en Richardson et al. (1976) werd

aangetoond dat studenten met een hogere taalvaardigheid meer woorden per T-unit produceerden dan studenten met een lagere taalvaardigheid, waarbij ervan uit wordt gegaan dat taalvaardigheid wordt gestuurd door het opleidingsniveau van de student. Deze bevinding bevestigt dat de lengte van de T-unit die men produceert in verband staat met zijn opleidingsniveau, ook in een online datingcontext.

Tot slot werd een verschil aangetoond in de gemiddelde lengte van de NP tussen de profieleigenaren met verschillende opleidingsniveaus. De verwachting was dat naarmate het opleidingsniveau van profieleigenaren toenam, zij langere NP's zouden produceren in hun profieltekst. Er werd gevonden dat de NP's in de profielteksten van middelhoogopgeleiden langer waren dan de NP's in profielteksten van laag- en hoogopgeleiden. Tussen de profielteksten van laag- en hoogopgeleiden werden geen verschillen in de gemiddelde lengte van NP's gevonden. De vijfde hypothese werd dus verworpen. Deze bevinding ligt deels in lijn met de resultaten uit de studie van Staples et al. (2016), waarin werd aangetoond dat de lengte van NP's stijgt naarmate de opleidingsjaren van mensen stijgen in een academische setting (Staples et al., 2016). In de huidige studie werd ook gevonden dat middelhoogopgeleiden langere NP's produceerden dan laagopgeleiden. Echter waren de NP's in profielteksten van hoogopgeleide profieleigenaren korter dan de NP's in profielteksten van laagopgeleide profieleigenaren, een tegenstrijdige bevinding in vergelijking tot eerder onderzoek (Staples et al., 2016; Biber et al., 2011). Dit verschil is mogelijk te wijten aan het contextverschil waarin de condities zijn onderzocht. Het onderzoek van Staples et al. (2016) onderzocht syntactische complexiteit in een academische context, waarbij geschreven teksten door studenten afkomstig uit een corpus van Britse universiteiten werden geanalyseerd. Mogelijk passen hoogopgeleide profieleigenaren hun taalgebruik aan naar de setting waarin zij communiceren. Met andere woorden maken hoogopgeleide profieleigenaren wellicht minder gebruik van het schrijven van complexe NP's in een informele context zoals online dating dan wanneer zij teksten schrijven

in een academische setting.

Op basis van de resultaten uit de taalfoutenanalyse kan voorzichtig geconcludeerd worden dat het aantal gemaakte taalfouten in profielteksten afnam naarmate het opleidingsniveau van profieleigenaren toenam, alhoewel niet in elke subcategorie taalfouten verschillen tussen laag-, middelhoog- en hoogopgeleiden werden gevonden. Deze inzichten zijn relevant in de context van online dating, omdat ze mogelijk invloed kunnen hebben op de impressie die potentiële partners van elkaar vormen. Taalfouten zijn namelijk voorbeelden van onbedoelde *cues* in taalgebruik (Lea & Spears, 1992; Walther & D’Addario, 2001; Van der Zanden et al., 2020), en eerder onderzoek toonde aan dat juist de *cues* die onbewust worden afgegeven waardevolle factoren zijn in impressievorming (Walther & Parks, 2002; Wotipka & High, 2016). Ook kan aan de hand van de analyse van de syntactische complexiteit de voorzichtige conclusie worden getrokken dat hoger opgeleide profieleigenaren meer gebruik maken van syntactische complexiteit in hun profielteksten dan laagopgeleide profieleigenaren. Met deze conclusie moet echter bedachtzaam worden omgegaan, aangezien de groep met het hoogste opleidingsniveau tegen de verwachting in de minst lange NP’s produceerden. Een suggestie voor toekomstig onderzoek is dan ook om deze bevinding gedetailleerd te analyseren. Al met al komen de resultaten van de huidige replicatiestudie overeen met de studie van Van der Zanden et al. (2018), zo werden vergelijkbare resultaten gevonden voor ellipsen, interpunctiefouten en spelfouten.

Limitaties en aanbevelingen voor vervolgonderzoek

Om de syntactische complexiteit in de 300 profielteksten van OkCupid automatisch te kunnen analyseren werd de linguïstische analysetool TAASSC gebruikt (Kyle, 2006). De profielteksten uit de deelsteekproef werden geüpload in de tool en die berekende de gemiddelde lengte per T-unit. Een limitatie aan de huidige studie is dat de accuraatheid van de tool niet is getoetst voordat de analyse werd uitgevoerd. Hoewel de analysetool TAASSC regelmatig gebruikt voor

syntactische analyses in taalkundig onderzoek, zou het de betrouwbaarheid van de huidige studie hebben versterkt als aangetoond kon worden dat de T-unitmeting van TAASSC in eerste instantie op accuraatheid was gecontroleerd door te oefenen met een kleine dataset.

Een tweede limitatie is het feit dat de huidige replicatiestudie een anderstalig corpus heeft geanalyseerd dan de studie van Van der Zanden et al. (2018) waarop deze replicatie is gebaseerd. Van der Zanden en collega's (2018) werkten met een Nederlands corpus met profielteksten afkomstig van de datingsites Match4Me en Relatieplanet. De subcategorieën taalfouten uit de huidige studie zijn deels gebaseerd op de categorieën van Van der Zanden et al. (2018), maar ook zijn er twee subcategorieën taalfouten toegevoegd waarvan werd verwacht dat deze in het Engelstalig corpus zouden voorkomen. In de huidige studie werd er geen verschil gevonden in grammaticale fouten tussen de groepen profieleigenaren met verschillende opleidingsniveaus. Mogelijk zouden de resultaten van de replicatiestudie een hogere betrouwbaarheid bereiken wanneer ervoor werd gekozen om hetzelfde codeerschema als Van der Zanden en collega's (2018) te hanteren tijdens de taalfoutenanalyse.

Aanbevelingen voor vervolgonderzoek

Zoals eerder is aangegeven, ligt de huidige studie ten grondslag van een diepgaander onderzoek naar impressievorming op online datingsites. De huidige studie was een replicatiestudie van het onderzoek van Van der Zanden et al. (2018), waar het analyseren van syntactische complexiteit in profielteksten aan toe is gevoegd. De resultaten vormden een interessant beginsel voor het onderzoeken van syntactische complexiteit in online dating. Nu er meer inzicht is in het verband tussen het opleidingsniveau van profieleigenaren en de syntactische complexiteit in hun profieltekst, is een volgende stap om te onderzoeken wat voor rol syntactische complexiteit speelt in impressievorming tijdens online dating. Ook zou het interessant kunnen zijn om meer kenmerken van profieleigenaren toe te voegen aan toekomstige analyses. Schijf et al. (2010) toonden bijvoorbeeld al aan dat vrouwelijke middelbare scholieren beter konden spellen dan

mannelijke middelbare scholieren. Daarnaast toonde Hargittai (2006) aan dat het inkomen van internetgebruikers in verband staat met het maken van typografische fouten. Mensen met een hoger inkomen maken minder spelfouten op het internet. Het is relevant om meer inzicht te krijgen in welke factoren naast opleidingsniveau een rol spelen in de aanwezigheid van taalfouten en syntactische complexiteit op online datingsites.

Tot slot vormt de huidige studie het wetenschappelijke fundament van het verband tussen het opleidingsniveau van profieleigenaren en de aanwezigheid van taalfouten en syntactische complexiteit in profielteksten en kunnen de bevindingen van dit onderzoek uitmonden in toekomstig onderzoek naar de rol van taalfouten en syntactische complexiteit in impressievorming op online datingsites.

Referenties

- Antheunis, M. L., Schouten, A. P., & Walther, J. B. (2020). The hyperpersonal effect in online dating: Effects of text-based CMC vs. videoconferencing before meeting face-to-face. *Media Psychology, 23*(6), 820-839.
- Antheunis, M. M. L. (2009) *Online Communication, Interpersonal Attraction, and Friendship Formation*. Universiteit van Amsterdam: Amsterdam.
- Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre?. *Reading and Writing, 22*, 185-200.
- Bernstein, B. (1964). Elaborated and restricted codes: Their social origins and some consequences. *American Anthropologist, 66*(2), 55-69.
- Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5– 35.

- Bonset, H. (2010). Spelling in het onderwijs: hoe staat het ermee, en hoe kan het beter?. *Levende Talen Tijdschrift*, 11(3), 3-17.
- Bulté, B., & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Burgoon, M., & Miller, G. R. (1985). An expectancy interpretation of language and persuasion. In H. Giles & R. Clair (Eds.), *The social and psychological contexts of language*. London, UK: Lawrence Erlbaum Associates, 199-229.
- Corder, S. P. (1974). *Error Analysis: Perspectives on second language acquisition*. London: Longman
- Cragg, L., & Nation, K. (2006). Exploring written narrative in children with poor reading comprehension. *Educational Psychology*, 26, 55–72.
- Crowhurst, M., & Piche, G. L. (1979). Audience and mode of discourse effects on syntactic complexity in writing at two grade levels. *Research in the Teaching of English*, 13, 101–109.
- Daller, H., Hout, R. van, & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24 (2), 197-222.
- Dijk, B. V. (2020). *Work (shop) naar succes* (Master's thesis).
- Donath, J. S. (2002). Identity and deception in the virtual community. In *Communities in cyberspace*. Routledge, 37-68.
- Eek, S. (2009). *Praktijk van online dating in Nederland* (Master's thesis).
- Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, 11(2), 415–441.
- Figueredo, L., & Varnhagen, C. K. (2005). Didn't you run the spell checker? Effects of type of spelling error and use of a spell checker on perceptions of the author. *Reading Psychology*, 26(4-5), 441-458.

- Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T., & Sprecher, S. (2012). Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public Interest*, 13(1), 3-66.
- Fiore, A. T., Taylor, L. S., Mendelsohn, G. A., & Hearst, M. (2008). Assessing attractiveness in online dating profiles. *Proceedings of Computer-Human Interaction* (pp. 797–806). New York, NY: ACM Press.
- Fox, S., Anderson, J. Q., & Rainie, L. (2005). *The Future of the Internet*. Pew Internet & American Life Project.
- Gaies, S. J. (1980). T-unit analysis in second language research: Applications, problems and limitations. *TESOL quarterly*, 53-60.
- Goffman, E. (1956). *The Presentation of Self in Everyday Life*. New York: Doubleday.
- Hancock, J. T., Toma, C., & Ellison, N. (2007). The truth about lying in online dating profiles. *Proceedings of the SIGCHI conference on Human factors in computing systems*. *ACM*, 449-452.
- Hargittai, E. (2006). Hurdles to information seeking: Spelling and typographical mistakes during users' online behavior. *Journal of the Association for Information Systems*, 7(1), 1.ok
- Hendrickson, J. M. (1978). Error correction in foreign language teaching: Recent theory, research, and practice. *The modern language journal*, 62(8), 387-398.
- Hubers, F., & de Hoop, H. (2013). The effect of prescriptivism on comparative markers in spoken Dutch. *Linguistics in the Netherlands*, 30(1), 89-101.
- Hunt, K. W. (1970). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development*, No. 134.
- Jelsma, B. K. (2015). *Productieve taalontwikkeling in het Nederlands van nieuwkomers* (Master's thesis).

- Johnson, R. I. (1917). The persistency of error in English composition. *The School Review*, 25(8), 555-580.
- Kim, A. Y., & Escobedo-Land, A. (2015). OkCupid data for introductory statistics and data science courses. *Journal of Statistics Education*, 23(2).
- Klaassen, P. B. L. (2021). De ontwikkeling van syntactische complexiteit in het schrijfproces.
- Klooster, W. (2001). *Grammatica van het hedendaags Nederlands: Een volledig overzicht*. Den Haag: Sdu Uitgevers.
- Klooster, W. (2001). *Grammatica van het hedendaags Nederlands: Een volledig overzicht*. Den Haag: Sdu Uitgevers.
- Kreiner, D. S., Schnakenberg, S. D., Green, A. G., Costello, M. J., & McClin, A. F. (2002). Effects of spelling errors on the perception of writers. *The Journal of general psychology*, 129(1), 5-17.
- Kyle, K. (2006). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. (Doctoral dissertation).
- Lardiere D. (2009). Some thoughts on the contrastive analysis of features in second language acquisition. *Second Language Research*, 25, 173–227.
- Le Dorze, G., & Bédard, C. (1998). Effects of age and education on the lexico- semantic content of connected speech in adults. *Journal of Communication Disorders*, 31(1), 53-71.
- Lea, M., & Spears, R. (1992). Paralanguage and social perception in computer-mediated communication. *Journal of Organizational Computing and Electronic Commerce*, 2(3-4), 321-341.
- Liu, Y., & Ginther, D. W. (2002). Instructional Strategies for Achieving a Positive Impression in Computer-Mediated Communication (CMC) Distance Education Courses. *Proceedings of the Annual Mid-South Instructional Technology Conference*, 6.

- Loban, W. (1976). Language development: Kindergarten through grade twelve (Research Report No. 18). Urbana, IL: *National Council of Teachers of English*.
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474-496.
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96, 190-208.
- McKenna, K. Y. A., Glean A. S. & Gleason M. E. J. (2002). Relationship formation on the internet: What's the big attraction? *Journal of Social Issues*, 58(1), 9-31.
- Morales, L. (2009). Nearly half of Americans are frequent Internet users. *Washington DC: Gallup Poll*.
- Mulder, K., & Hulstijn, J.H. (2011). Linguistic skills of adult native speakers, as a function of age and level of education. *Applied Linguistics*, 32(5), 475-494.
- Norrish, J. (1983). *Language learners and their errors*. London: Macmillan Press. P. 7
- Park, N., Jin, B., & Jin, S. A. A. (2011). Effects of self-disclosure on relational intimacy in Facebook. *Computers in Human Behavior*, 27(5), 1974-1983.
- Queen, R., & Boland, J. E. (2015). I think your going to like me: Exploring the role of errors in email messages on assessments of potential housemates. *Linguistics Vanguard*, 1, 283–293.
- Richards, J. C. & Schmidt, R. (2002). *Dictionary of language teaching and applied linguistics* (3rd Ed.). London: Longman.
- Richardson, K., Calnan, M., Essen, J., & Lambert, L. (1976). The linguistic maturity of 11-year-olds: Some analyses of the written compositions of children in the National Child Development Study. *Journal of Child Language*, 3, 99–115.

- Rosen, L. D., Cheever, N. A., Cummings, C., & Felt, J. (2008). The impact of emotionality and self-disclosure on online dating versus traditional dating. *Computers in Human Behavior, 24*(5), 2124-2157.
- Rosenfeld, M. J., Thomas, R. J., & Hausen, S. (2019). Disintermediating your friends: How online dating in the United States displaces other ways of meeting. *Proceedings of the National Academy of Sciences, 116*(36), 17753-17758.
- Schijf, T., van der Leij, A., van Berkel, A., Bekebrede, J., & Zijlstra, B. (2010). Spellingvaardigheid van brugklassers. *Levende Talen Tijdschrift, 11*(2), 3-12.
- Schoot, F. (z.d.). De invloed van opleidingsniveau en soort datingsite op het taalgebruik in online datingprofielen.
- Smith, A. W., & Duggan, M. (2013). *Online dating & relationship*. Washington, DC: Pew Research Center.
- Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication, 33*(2), 149-183.
- Surkyn, H., Sandra, D., & Vandekerckhove, R. (2022). Adolescents and verb spelling: The impact of gender and educational track on rule knowledge and linguistic attitudes. *Dutch Journal of Applied Linguistics, 11*, 1-19.
- Tesak, J., Ahlsen, E., Györi, G., Koivuselkä-Sallinen, P., Niemi, J., & Tonelli, L. (1995). Patterns of ellipsis in telegraphese: a study of six languages. *Folia Linguistica, 29*(3-4), 297-316.
- Tinder Users by Country 2023*. (z.d.). Retrieved from:
<https://worldpopulationreview.com/country-rankings/tinder-users-by-country>

- Tummers, J., & Deveneyns, A. (2016). Lexicale rijkdom in het professioneel hoger onderwijs: Aanzet tot sociolinguïstische staalkaart. In D. van Mierop, L. Buysse, R. Coesemans & P. Gillaerts (Eds.), *De macht van de taal* (pp. 257-273). Leuven: ACCO.
- Van Bree, C. (2000). Taalbewustzijn, taalverandering en regionale taalvariatie. *De toekomst van de variatielinquïstiek. Een bundel artikelen aangeboden aan Jo Daan bij gelegenheid van haar negentigste verjaardag. Themanummer Taal & Tongval*, 22-46.
- Van der Zanden, T., Mos, M., & Schouten, A. (2018). Taalaccommodatie in online datingprofielen: Effecten van opleidingsniveau en type datingsite op taalgebruik. *Tijdschrift voor Taalbeheersing*, 40(1), 83-106.
- Van der Zanden, T., Schouten, A. P., Mos, M. B., & Kraemer, E. J. (2020). Impression formation on online dating sites: Effects of language errors in profile texts on perceptions of profile owners' attractiveness. *Journal of Social and Personal Relationships*, 37(3), 758-778.
- Van Hout, R. W. N. M. (1999). *Taal en toeval*. Nijmegen/Tilburg: KUN/KUB.
- Verheijen, L. (2016). Linguistic characteristics of Dutch computer-mediated communication. CMC and school writing compared. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana, Slovenia*.
- Vermeer, A. (2016). Lexicale rijkdom, frequentielagen en tekstmoeilijkheid. *Dutch Journal of Applied Linguistics*, 5(1), 18-33.
- Walther, J. B., & Burgoon, J. K. (1992). Relational communication in computer-mediated interaction. *Human communication research*, 19(1), 50-88.
- Walther, J. B., & D'Addario, K. P. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review*, 19, 323–345.

- Walther, J.B. and Parks, M.R. (2002). Cues Filtered Out, Cues Filtered In: Computer-Mediated Communication and Relationships. In: Knapp, M.L. and Daly, J.A., Eds., *Handbook of Interpersonal Communication*, 3rd Edition, Sage, Thousand Oaks, 529-563.
- Wotipka, C. D., & High, A. C. (2016). An idealized self or the real me? Predicting attraction to online dating profiles using selective self-presentation and warranting. *Communication Monographs*, 83(3), 281-302.
- Zdorenko, T., & Paradis, J. (2012). Articles in child L2 English: When L1 and L2 acquisition meet at the interface. *First language*, 32(1-2), 38-62.

Bijlagen

Bijlage 1. Beschrijving van het aselect genereren van de deelsteekproef

In de huidige studie werd een deelsteekproef van 300 datingprofielen gegenereerd waarin rekening werd gehouden met bepaalde criteria. Ten eerste werd in het complete corpus een selectie van opleidingsniveaus zoals beschreven in de methodesectie doorgevoerd. Daarnaast werd slechts de leeftijdscategorie 18-35 jaar oud meegenomen, en werden alle profielen waarvan hun profieltekst (essay0) minder dan 100 of meer dan 250 woorden bevatte geëxcludeerd. Na deze filters bleven 33.256 profielteksten over. Uit deze overgebleven profielteksten werd onderscheid gemaakt tussen 3 groepen opleidingsniveaus: laagopgeleid, middelhoogopgeleid en hoogopgeleid (zie de verdeling in de methodesectie). Hierna zijn nieuwe bestanden in Excel aangemaakt voor datingprofielen uit elke opleidingsgroep, waarin de datingprofielen van mannen en vrouwen gescheiden werden. Vervolgens kon een willekeurige selectie worden gemaakt aan de hand van de formule =INTEGER(ASELECT()*x. In een extra kolom in het databestand kreeg elk profiel een willekeurig getal en zijn de eerste 50 datingprofielen per opleidingsniveau en per geslacht geselecteerd die de deelsteekproef van 300 datingprofielen hebben gevormd.

Bijlage 2. Codeboek voor het coderen van de subcategorieën taalfouten in de deelsteekproef van OkCupid

1. Onbepaald lidwoordfout

A of an

Regel: Als een woord met een klinker of klinkerklank begint moet het onbepaald lidwoord ervoor 'an' zijn. Als een woord met een medeklinker of medeklinkerklank begint moet het lidwoord ervoor 'a' zijn.

Let op op de letter 'h':

Honest = je spreekt de 'h' niet uit, dus 'an'

Happy = je spreekt de 'h' wél uit, dus medeklinkerklank, dus 'a'

Fout: I'm a honest lady → I'm an honest lady

2. Ellipsen van onderwerp/gezegde

In een zin is het onderwerp en/of de persoonsvorm weggelaten. Dit is een vorm van spreektaal dus is discutabel of het echt een taalfout is. Op datingprofielen wordt heel veel gebruik gemaakt van ellipsen, dus je gaat deze vaak tegenkomen.

Voorbeelden (NL):

[ik] ben een gezellige meid.

[ik hou van] gezellige gesprekken, uitgaan en uit eten.

Gecodeerd als fout:

[I am] not really sure what else to say other then if your in need of someone to watch your back in this upcoming zombie apocalyptic event then look no further.

[I] joined the gym recently and enjoying the time i spend their 3-4 times a week and 1 with my trainer.

What i'm looking for in a guy is somebody that enjoys creative extra-curricular activities that gets the off the couch and out of their house. In other-words, [I'm looking for] people that choose to live their lives.

My name is isaac i have a big heart , [I am] injoying life as it is trying 2 find a new job tho cuz the one i have sucks nucking futs.

i am very laid back, easy going, [i] like to have fun and [I am] a drama free kind of guy

[it is] better if you know the answer already

Opmerking:

Bij opsommingen in een lijstje zoals hieronder, niet gecodeerd op ellipsen. Het zijn namelijk geen hele zinnen.

age 22

the basics-

height: 6'3

eye color: blue

weight 210

birthday: nov 30, 1988

relationship status:single

professional gamer for halo 3 (2005-2007) i'm apart of the us army (5 years)

soccer 16 years

rugby 3 years

opmerking:

“i was born and raised in shanghai, a self-made entrepreneur, came to usa for business since 2007, now travel between china and usa. “

Het onderwerp staat alleen helemaal aan het begin van de zin. Ik ga dan na of dit onderwerp vóór elk stukje opnieuw geplaatst kan worden.

[i am] a self-made entrepreneur = persoonsvorm ontbreekt = ellips

Wanneer je hier 1 keer [i] invoegt, kloppen ook de twee zinsdelen in de zin. Dus voor deze hele zin codeer ik 1 ellipsfout.

3. Apostroffout / apostrof ontbreekt

Apostroffout is wanneer er samenvoegingen zijn waarin de apostrof ontbreekt:

Im → i'm

Ill → i'll

Youll → you'll

Lets → let's

Thats → that's

Er wordt voor gekozen om alleen deze categorie binnen apostroffouten mee te pakken:

Examples: don't = do not I'm = I am shouldn't = should not didn't = did not, that's

En bijv geen bezittelijke termen: james' car.

Deze regels kunnen namelijk voor Brits & Amerikaans Engels verschillen en dat valt niet te controleren.

Opmerking

It's possibilities = niet gecodeerd

Er wordt alleen gecodeerd bij een vergeten apostrof en niet bij een verkeerde apostrof

Let op: its possibilities (zijn mogelijkheden) is correct. Its = bezittelijk. It's = het is

Let op:

In het voorbeeld 'i would love it when your adventurous' → dit wordt niet gecodeerd als apostroffout maar als lexicaal homofone spelfout.

4. Lexicaal homofone spelfout

Het geschreven foute woord kan zowel een ander bestaand woord zijn dat hetzelfde klinkt (lexicaal homofoon) maar niet klopt in de context

Too als het to moet zijn → homofone fout omdat beide woorden bestaan

I use to b e → moet zijn 'used' = homophone fout

Voorbeelden zijn:

Your & you're

Woman & women

There & they're & their

To & too

Whole & hole

5. Niet-homofone spelfout & overige spelfouten

of het geschreven en het bedoelde woord klinken hetzelfde maar het geschreven woord resulteert in een niet-bestaand woord (non-homofone spelfout).

hole heartily → non-homofone fout. Moet eigenlijk zijn 'wholeheartedly'. Omdat dit 1 woord is, codeer ik ook maar 1 fout.

Ganna → non-homofone fout. Het klinkt hetzelfde als 'gonna' maar is geen bestaand woord.

U → spreektaal voor 'you', codeer ik NIET als fout. Is gewoon spreektaal. Geen idee wat mijn begeleider hiervan vindt maar ik schrijf er dan wel iets over in mijn discussie.

Cuz → spreektaal voor 'cause', codeer ik NIET als fout.

Favorate → klinkt hetzelfde maar is geen bestaand woord dus non-homofone fout