Outeuridentifikasie: ’n Forensies-taalkundige ondersoek na Afrikaanse SMS-taal

Lezandra Grundlingh

Lezandra Grundlingh, Departement Afrikaans, Universiteit van Pretoria

Opsomming

Forensiese taalkunde is die studieveld waar linguistiese kennis en metodes gebruik word om ’n verskeidenheid tekste (gesproke en geskrewe) vir regsdoeleindes te ontleed. Outeuridentifikasie is ’n subkategorie van die forensiese taalkunde en word beskryf as die vasstel van ’n outeur se idiolek met die doel om die outeur positief te identifiseer as produseerder van ’n teks (Coulthard 2004:431–47; McMenamin 2010:490–2). Outeuridentifikasie is met ander woorde die proses waardeur die ware outeur van ’n gesproke of geskrewe teks vasgestel word wanneer daar twyfel bestaan oor die identiteit van die outeur van so ’n teks. Die doel van die huidige navorsing was om te bepaal tot watter mate dit moontlik is om vas te stel of ’n spesifieke individu die outeur van ’n spesifieke, beperkte stel (Afrikaanse) SMS’e is. Uit die navorsing blyk dit dat dit nie moontlik is om die outeur van ’n verdagte teks met groot sekerheid te identifiseer indien daar slegs beperkte data tot die navorser se beskikking is nie, maar die resultate is nietemin voldoende om as bewyse gebruik te word.

Trefwoorde: forensiese linguistiek; idiolek; outeuridentifikasie; SMS; stilometrie

Abstract

Author identification: A forensic-linguistic research study in Afrikaans SMS language

Forensic linguistics is a field of study that has gained popularity in many countries around the world (Blackwell 2012). In South Africa forensic linguistics is not a well-known field of study, but academics and postgraduate students are beginning to explore research and study opportunities within this field. Author identification, which is the focus of this article, is only one of the subcategories within forensic linguistics. A very basic definition of forensic linguistics is that it is a section of applied linguistics where a variety of both written and spoken texts is analysed for judicial purposes. The field is roughly divided into two main categories: language use and its judicial implications and the analysis of forensic texts (written or spoken). In the first category forensic linguists consider translating and

interpreting in a courtroom setting, the language use and discourse of a trial and the language rights of individuals in the courtroom or during the course of the trial, among others. The second category includes author identification, speaker identification, profiling and identification of plagiarism (Olsson n.d.:4–5).

Author identification is the analysis of a text with the goal of determining the possible author because there is some uncertainty or dispute about the author of that specific text (or group of texts). Texts that are usually analysed in author identification include ransom notes, e-mail messages, threat letters and blackmail messages. Although author identification analysis has been done on shorter texts such as SMS messages, ransom notes and Facebook messages in the past (Ishihara 2011; McLeod and Grant 2012; Michell 2013) there is still, to some extent, insufficient research in the area of author identification of short (and extremely short) texts. Author identification in Afrikaans SMS messages has never been attempted. It is mainly for this reason that the current article, and the dissertation it is based on, is considered of value to the field of author identification (Thiart 2014).

For the purposes of this study the researcher aimed to answer three questions. First it had to be determined if a generic SMS language exists that could complicate author identification. The presence of a generic SMS language would mean that there are very few individual characteristics present when SMS messages in the corpus are compared. Secondly, it had to be determined whether individual idiolects could be identified within the supposed generic SMS language, and thirdly, to what extent it is possible to identify the author of an SMS text with the limited data available to the forensic linguist.

Thirteen participants between the ages of 18 and 23 were used in the research. The only selection criterion was that the participants had to be mother-tongue speakers of Afrikaans. Each participant was asked to send 5 to 10 SMS messages of between 30 and 50 words each to the researcher. The participants were asked to select messages on their phones that they had already sent, i.e. messages that they had already typed in the past. This was done to ensure that the participants would not type in a different manner when creating new messages that they knew would be used in the analysis. Each participant was given a number in order for the researcher to identify participants and to ensure that they remained anonymous. One participant, Deelnemer 2 (Participant 2), was asked to send a second set of messages to the researcher. This set was labelled Teks X (Text X) and was the “suspect text”. All the other texts in the corpus were compared with Text X in order to determine if it was possible to match the first set of texts from Deelnemer 2 with the “suspect” texts. Based on the statistical analyses and comparisons the researcher would then be able either to identify Deelnemer 2 as the author of Text X with a high percentage of certainty or conclude that it was not possible to match Deelnemer 2 to Text X successfully.

The corpus for the study is small, consisting of only 2 434 words in total. The small corpus is due to the system used to receive the SMS messages from the participants, namely SMSPortal, which places a limit on the number of characters it can read per SMS. This meant that it cut some of the SMS messages and decreased the amount of data available to the researcher.

Both stylometric and stylistic methods were used to analyse the data. WordSmith Tools and Antconc were used to perform statistical analyses on the data and a very basic n-gram analysis was also used to strengthen the results. Both the Pearson's chi-square test and the Yates correction were used in determining the results of the statistical analyses. The limited amount of data that the researcher obtained through the participants is a realistic amount of data that can be expected in a real-life forensic linguistic situation. Even though no actual crime was being investigated, the research gives an accurate indication of what is possible when a forensic linguist has only limited data to analyse and a number of possible authors.

The results of both the stylistic and stylometric analyses answer all three of the research questions mentioned above. Firstly, it was found that no generic SMS language existed among the participants in this study. This indicated that idiolects were present. However, due to the limited data used in the study it was not possible to determine the author of the suspect text with any certainty. Although these results were negative they were still useful in terms of narrowing down the number of suspects from 13 to 11. 11 is still a large number of suspects, but in that group the actual author (Deelnemer 2) was identified as the possible author in most of the analysis results.

The results showed that even though identification of the actual author of the suspect text was not possible in the situation created in the study, the methods used do show potential. As mentioned above, many researchers have proven that, to some extent, successful author identification is possible when a forensic linguist has limited data to analyse. It has to be taken into account, however, that these studies made use of a much larger corpus than was the case in the current study. Other methods should also be tested in a similar small corpus to see if better results can be achieved. It is also important to note that "successful author identification" does not mean that a suspect has been identified with 100% certainty; it simply indicates that the statistical possibility of a suspect's being the author of a specific text is high enough for him or her to be considered as the possible author.

Keywords: author identification, forensic linguistics, idiolect, SMS, stylometry

1. Inleiding

Outeuridentifikasie is een van die hoofvertakkings wat binne die forensiese taalkunde onderskei word. In die laaste paar dekades het outeuridentifikasie gegroei tot 'n interdissiplinêre veld wat van toepassing kan wees op die letterkunde, onderrig (vir plagiaatidentifikasie), nasionale en plaaslike intelligensie, en vanselfsprekend ook op die regswese en -praktyk. Stelselmatig het die fokus van outeuridentifikasie verskuif vanaf die ontleding van handskrif en grafiese eienskappe van tekste na die linguistiese inhoud van tekste wat juridies van belang is. Hierdie tipe tekste sluit byvoorbeeld selfmoordbriewe, tekste wat moontlik plagiaat bevat, afpersingsbriewe en skuldbekentenisse in (Kotzé 2010:186). Met die ontwikkeling en uitbreiding van tegnologie het outeuridentifikasie aangepas om ook outeurs van elektronies geproduseerde tekste te kan identifiseer. Oor die afgelope vyf jaar is verskeie studies rakende outeuridentifikasie van elektroniese tekste onderneem. Daar is gepoog om die outeurs van onder andere aanlyn forums, blogs, tekste

op sosiale netwerke en SMS-boodskappe te identifiseer (Mikros s.j.; Mohan e.a. 2010; Ishihara 2011; McLeod en Grant 2012; Michell 2013).

Outeuridentifikasie in SMS-boodskappe (voortaan ‘SMS’ en ‘SMS’e’ onderskeidelik) is een van die ondersoekvelde wat veral in die buiteland aandag geniet, onder andere in die Verenigde Koninkryk (Grant 2010; McLeod en Grant 2012), asook in Australië (Ishihara 2011). Die stuur van SMS’e is steeds een van die gewildste kommunikasiemiddele ter wêreld, ten spyte van sosiale netwerke soos Facebook en Twitter, wat ook vinnige kommunikasie bewerkstellig. SMS’e en ander vorme van vinnige kommunikasie word ook al hoe meer gebruik in ontvoerings, om kubermisdade te pleeg, en om dwelms of wapens te smokkel. SMS’e word ook gebruik om sulke misdade te probeer verdoesel (Crystal 2008:60–1; Gangs use of the internet and cell phones 2010; Grant 2010:508; Ishihara 2011:47; Blackwell 2012:5).

Ten spyte daarvan dat outeuridentifikasie in SMS’e onder andere gebruik kan word om misdadigers vas te trek, is daar steeds onvoldoende navorsing op hierdie gebied (Ishihara 2011:48).

Een van die redes waarom daar min navorsing oor outeuridentifikasie in SMS’e beskikbaar is, is die feit dat ’n SMS ’n kort teks is met beperkte inhoud. Vergelyk hulle byvoorbeeld met vollengte-dokumente soos dreigbriewe, wat soms uit twee of meer getikte bladsye bestaan (dit wil sê ongeveer 600 tot 700 woorde). Forensiese taalkundiges plaas SMS’e, boodskappe op sosiale netwerke asook selfmoordbriewe en lospryseise in dieselfde kategorie. Hierdie tekste deel die eienskap van bondigheid, en as gevolg hiervan is dit baie moeiliker om die outeur van sulke boodskappe te identifiseer as wanneer vollengte-romans of ander lang tekste ondersoek word. Ten spyte van die problematiese aard van korter tekste (vanuit ’n outeuridentifikasiestandpunt beskou) het navorsing reeds aangetoon dat dit nie onmoontlik is om die outeur van korter tekste te identifiseer nie (Ishihara 2011; MacLeod en Grant 2012).

’n Verdere probleem met SMS-taal, wat ook verband hou met die bondige aard van die tekste, duik op wanneer die linguis die persoonlike styl (idiolek) van die outeur moet identifiseer. Soos later uit die verdere bespreking van idiolek sal blyk, is dit nie eenvoudig om idiolekte te identifiseer nie, en bestaan daar ook twyfel oor die akkuraatheid van afleidings wat op grond van idiolek gemaak word (Grant 2010), veral wanneer korter tekste ontleed word. In die forensiese taalkunde probeer die linguis onder andere die idiolek van ’n moontlike outeur identifiseer omdat die idiolek, wat onbewustelik deur die outeur gebruik word, as ’n identifiserende eienskap van ’n verdagte kan dien.

Hierdie artikel is gebaseer op my meestersgraadverhandeling op die terrein van forensiese taalkunde (Thiart 2014).

Voor die bespreking van forensiese linguistiek voortgesit word, word enkele kernkonsepte kortliks verduidelik om sodoende enige onduidelikheid rondom hierdie konsepte te voorkom.

2. Kernkonsepte

2.1 Leksikale woorde / inhoudswoorde

Leksikale woorde staan ook as inhoudswoorde bekend en is konteksgebonde. Dit beteken dat die algemene woorde wat die outeur van ’n teks gebruik, sal afhang van die konteks van die teks of boodskap.

2.2 Funksiewoorde

Funksiewoorde is woorde wat met die struktuur van ’n sin of uiting verband hou. Dit sluit onder andere lidwoorde, voegwoorde en voorvoegsels in. Funksiewoorde is nie konteksgebonde nie en is belangrik in outeuridentifikasie, omdat hierdie woorde heeltemal onbewustelik deur die outeur van die teks gebruik word.

2.3 SMS *en* SMS-taal

SMS is ’n afkorting wat staan vir “short message service” en verwys na ’n goedkoop en vinnige manier om boodskappe deur middel van ’n selfoon te stuur (Ishihara 2011:47). ’n SMS kan met ander woorde gedefinieer word as ’n kort elektroniese boodskap wat deur middel van ’n selfoon gestuur word. Die term *SMS-taal* word gebruik om te verwys na die taalgebruik van selfoongebruikers by die stuur van SMS’e. Die tipiese eienskappe van SMS-taal het ontstaan as gevolg van die lengtebeperkings in selfoonboodskappe (Crystal 2008:5). Dit beteken dat selfoongebruikers nou op kreatiewer maniere kort boodskappe moet skryf, aangesien soveel inligting moontlik soms in een boodskap oorgedra moet word.

Die term *SMS* word hier gebruik om te verwys na enige kort elektroniese boodskap wat vanaf ’n selfoon gestuur word (dit sluit WhatsApp- en BBM-boodskappe in, maar sluit boodskappe op sosiale netwerke en bloginskrywings uit). *SMS-taal* verwys na die taalgebruik wat in hierdie kort elektroniese boodskappe aangetref word.

2.4 Vollengte-teks

Vir die doeleindes van die huidige studie verwys *vollengte-teks* na enige teks van 1 000 woorde en meer.

2.5 Woord

Vir die doeleindes van die huidige studie sluit *woord* ook enige afkorting, akroniem, logogram of piktogram in.

3. Forensiese linguistiek

Dit is belangrik om aan te dui waar die forensiese linguistiek binne die veld van forensiese wetenskap inpas, aangesien forensies-linguistiese ontledings en resultate ook tot ’n mate in hofsake gebruik kan word om verdagtes vas te trek.

Die adjektief *forensies* hou volgens MedicineNet.com (2014) verband met “the application of scientific knowledge to legal problems and legal proceedings as, for example, in forensic anthropology, forensic dentistry, forensic experts, forensic medicine (legal medicine), forensic pathology, forensic science etc.”

Die term *forensies* word as ’n sinoniem beskou vir woorde soos *geregtelik* en *juridies* en word tradisioneel met die term *wetenskap* geassosieer. Olsson (s.j.:2) se definisie van *forensiese taalkunde* illustreer duidelik hoe nou verweef die forensiese taalkunde met die regstelsel is:

> Forensic Linguistics is the interface between language, crime and law, where law includes law enforcement, judicial matters, legislation disputes or proceedings in law, and even disputes which only potentially involve some infraction of the law or some necessity to seek a legal remedy.

Forensiese wetenskap word reeds jare lank gebruik om die verdagtes van misdade vas te trek en skuldig of onskuldig te bewys. Forensiese wetenskap sluit onder andere DNS-ontledings, vingerafdrukontledings en bloedvlekontledings in. Handskrifontleding en handtekeningontleding, wat meer ooreenkomste toon met ondersoeke in forensiese linguistiek, word ook as deel van forensiese wetenskap beskou (Jackson en Jackson 2004).

Forensiese linguistiek is ’n baie spesifieke ondersoekveld, aangesien daar slegs op linguistiese aspekte, soos taalgebruik, stemkenmerke en die betekenis van woorde in ’n bepaalde konteks gefokus word.

Alhoewel die forensiese linguistiek verband hou met tradisionele vorme van ontleding, soos handskrifontleding, fokus dit op ander aspekte van die teks, aangesien die eiesoortige aspekte van geskrewe tekste nie in moderne elektroniese tekste teenwoordig is nie.

4. Outeuridentifikasie

Die oorsprong van outeuridentifikasie hou sterk verband met die oorsprong van die forensiese linguistiek as dissipline. Van die eerste forensiese ondersoeke is in der waarheid gevalle waar navorsers gepoog het om outeurs van tekste te identifiseer (Mendenhall 1887; Svartvik 1964; Broeders 2001; Kotzé 2007; Schulstad e.a. 2012):

- Edmond Malone (1787) se ondersoek om te bewys dat William Shakespeare nie vir die teks van *Henry VI* verantwoordelik is nie.

- Augustus de Morgan se 1851-brief waarin hy voorstel dat dit moontlik sal wees om die outeurs van die verskillende Bybelboeke te identifiseer op grond van sins- en woordlengte.
- Mendenhall (1887) se poging om Bacon, Marlowe en Shakespeare as die outeurs van verskeie werke te bevestig of te verwerp deur De Morgan se voorstel te gebruik.
- Mosteller en Wallace (1964) se identifisering van die outeurs van die *Federalist Papers*.
- Jan Svartvik (1968) se studie waarin hy probeer vasstel of Timothy Evans werklik die verklarings aan die polisie gemaak het waarin hy erken dat hy sy vrou en dogter vermoor het.

Bogenoemde forensies-linguistiese ondersoeke is nie onder die subkategorie van outeuridentifikasie erken nie, aangesien die breë veld van forensiese linguistiek nog nie as ’n ondersoekveld gedefinieer was nie. Dit is eers nadat die forensiese linguistiek uitgebrei het om ander ondersoekvelde soos forensiese fonetiek, profielsamestelling en taalgebruik in die regskonteks in te sluit dat subkategorieë soos outeuridentifikasie onder die sambreelterm *forensiese linguistiek* ontstaan het.

5. Idiolek

Idiolek verwys na ’n individu se persoonlike, eiesoortige taalgebruik wat onbewustelik, met ander woorde in die onderbewussyn, gevorm word. Individue is daarom gewoonlik nie bewus van idiolektiese woorde wat hulle gebruik wanneer hulle praat of skryf nie en ook nie bewus van die idiolektiese wyses waarop hulle woorde gebruik nie. Beide McMenamin (2010) en Coulthard (2004) se definisie van *idiolek* kan gebruik word om laasgenoemde stelling te ondersteun:

> An idiolect is a variety of language developed by the individual speaker as a uniquely patterned aggregate of linguistic characteristics observed in his or her language use, often called “individual characteristics” in forensic science. (McMenamin 2010:487)

> The linguist approaches the problem of questioned authorship from the theoretical position that every native speaker has their own distinct and individual version of the language they speak and write, their own idiolect, and […] this idiolect will manifest itself through distinctive and idiosyncratic choices in texts. (Coulthard 2004:431)

5.1 Die problematiek rondom idiolek

Idiolek is nie so ’n eenvoudige konsep as wat dit uit bostaande definisies skyn te wees nie. Wanneer vermoed word dat ’n idiolek of idiolektiese styl teenwoordig is, word die forensiese linguis dikwels met drie vrae gekonfronteer, naamlik:

- Bestaan die idiolek werklik?
- Is die idiolek altyd waarneembaar?
- Is die idiolek ’n akkurate aanwyser van ’n individuele outeur?

Verder moet die forensiese linguis ook idiolektiese ontledings op die algemene taalgebruik van die vermeende outeur rig, eerder as om te fokus op seldsame woorde wat die outeur mag gebruik. Dit beteken dat die linguis 'n idiolek moet kan identifiseer in tekste wat uit gewone en alledaagse taalgebruik bestaan. Alhoewel seldsame woorde wel kan bydra tot die identifisering van 'n outeur, bestaan die moontlikheid dat sulke woorde dalk nie in die bepaalde tekste wat as verdag bestempel is, teenwoordig sal wees nie, aangesien sulke woorde ongewoon in alledaagse gebruik is (Juola 2006:263).

Grant (2010:509) meen tereg dat selfs al kan die bewering dat iets soos 'n idiolek bestaan, ondersteun word, daar steeds geen waarborg is dat 'n individu se idiolek in alle tekste geïdentifiseer kan word nie. By enige individu is daar konstante variasie in die manier waarop die individu praat of skryf. Die variasies wat voorkom, word beskryf as intravariasie (variasie in een persoon se praat- of skryfwyses) en intervariasie (variasie tussen twee of meer individue se praat- of skryfwyses) (Gavaldà-Ferré 2012:262). Crankshaw (2012:2) is van mening dat wanneer variasie op groepsvlak voorkom, soortgelyke variasie ook op individuele vlak sal voorkom. Crankshaw verwys na Anshen (1978:1), wat meen dat twee individue van dieselfde taalgemeenskap nie net verskillende variante van dieselfde taalvorm sal gebruik nie, maar dat een persoon in verskillende kontekste ook verskillende variasies van dieselfde taalvorm sal gebruik.

Hierdie variasie word beskryf in die sogenaamde "uniqueness of utterance principle" (Chomsky 1965; Halliday 1975). Volgens hierdie beginsel sal tekste wat deur twee individue oor dieselfde onderwerp geproduseer word, duidelik van mekaar verskil, maar so ook tekste wat op twee verskillende tye deur dieselfde individu voortgebring is. Die rede hiervoor is dat elke individu by verskillende geleenthede verskillende leksikogrammatiese keuses uitoefen (Crankshaw 2012:3). Dit is hierdie leksikogrammatiese keuses wat lei tot intravariasie wat die identifisering van 'n idiolek kan belemmer.

Alhoewel die bestaan van 'n idiolek of idiolektiese styl betwis word, is daar steeds 'n algemene aanvaarding dat indien iets soos 'n idiolek wel bestaan, dit makliker sal wees om te identifiseer wanneer die linguis 'n beduidende aantal tekste tot sy of haar beskikking het. Ook die lengte van die tekste is van belang – hoe meer inhoud in elke dokument ingesluit word, hoe makliker is dit om individuele taalgebruik te identifiseer. Crankshaw (2012:5) verwys na Coulthard (1998), wat meen dat forensiese linguiste selde tekste wat langer is as 750 woorde ontvang. Volgens Crankshaw (2012:5) kan korter tekste steeds ontleed word, maar nie so suksesvol soos in die geval van langer tekste nie. Dit beteken dat beperkings geplaas word op die manier waarop 'n individu se idiolek voorgestel kan word, en dit bemoeilik 'n diepgaande, volledige voorstelling van die individu se idiolek.

6. Stilometrie

Kotzé (2007:388) definieer stilometrie as

> 'n deeglike kwantitatiewe ontleding, deur middel waarvan die relatiewe frekwensie van identiese woordeskatitems of woordgroepe vergelyk word. Dit word 'n

> kwantitatiewe ontleding genoem omdat dit gebaseer is op die kwantifisering van tekstuele kenmerke as ’n basis vir verdere berekenings, wat beteken dat ieder en elke woord opgeteken en getel moet word. ’n Aantal berekenings word dan op die data uitgevoer, gevolg deur statistiese beduidendheidstoetse.

Stilometrie behels hoofsaaklik twee prosesse, naamlik die seleksie van kenmerke en die daaropvolgende gebruik van ’n klassifikasie-algoritme om hierdie kenmerke statisties te verwerk (Barry en Luna 2012:2). Die linguis moet eerstens besluit watter kenmerke hy/sy in die teks wil selekteer vir verwerking. Hierna word ’n algoritme ingespan om die kenmerke statisties te verwerk en sodoende aan te dui hoe algemeen of vreemd hierdie kenmerke is. Hierdie twee prosesse kan egter ook as twee verskillende metodes in die forensiese linguistiek beskou word. Kotzé (2007:388) meen dat die seleksie van kenmerke as “stilistiese analisering” bekend staan en ’n kwalitatiewe ontleding van die teks behels, terwyl die meet van hierdie stilistiese kenmerke en die statistiese toetse wat daarop uitgevoer word, ’n kwantitatiewe proses is wat dan “stilometrie” genoem word.

Alhoewel stilometrie reeds vanaf die 1700’s gebruik is om die outeurs van literêre tekste te bepaal, meen Schulstad e.a. (2012:1) dat stilometrie nie vandag net vir literêre of historiese doeleindes gebruik word nie. Volgens Schulstad het stilometrie in moderne forensiese linguistiek ’n veel wyer toepassing:

> [I]t also has forensic applications. [...] More recent studies have used stylometry to determine the authorship of e-mails and online messages to counteract cybercrime. In addition to identifying an author, stylometry can also be used to detect multiple authors in a text (plagiarism) or to assign an author to a sociolinguistic category such as gender.

Soos reeds genoem, is die aanname in stilometrie dat die kern van die individuele styl van elke outeur vasgevang kan word deur ’n sekere aantal kwantitatiewe kriteria (Somers s.j.). Hierdie kwantitatiewe kriteria word ook diskrimineerders genoem. Alhoewel ’n groot aantal stylaspekte onbewustelik deur ’n outeur gekies word, is die realiteit dat ander aspekte wel bewustelik deur omstandighede en die onderwerp van die teks beïnvloed word. Daar is, met ander woorde, aspekte van elke outeur se styl wat maklik is om na te boots. Rekenaargebaseerde stilometrie maak dit makliker om bewuste stylmerkers van onbewuste stylmerkers in verskillende outeurs se werke te onderskei, en daarom is hierdie vorm van stilometrie so gewild (Somers s.j.).

Dit is belangrik om kennis te neem van die feit dat stilometrie vandag deur kunsmatige intelligensie oorheers word. Dit beteken dat die menslike element van stilometrie al minder van toepassing is (Brennan e.a. 2012:1). Die stelselmatige verskuiwing in stilometrie na ’n sterker rekenaargebaseerde benadering kan moontlik in die toekoms vertroue in hierdie metode versterk en daartoe bydra dat resultate wat uit hierdie ontledings verkry word, makliker as geldige bewyse in die hof aanvaar word.

6.1 Korter tekste en stilometrie

Kort tekste is problematies in stilometriese ontledings, aangesien daar nie genoeg linguistiese data is wat verwerk kan word by die ontledings van baie kort tekste nie (Barry en Luna 2012:4–5). Dit beteken dat die linguis van ’n groot hoeveelheid teks gebruik moet maak om enigsins moontlike sukses in ’n stilometriese ontleding te behaal. Stamatatos e.a. (2001:193) meen een rede waarom sommige stilometriese ontledings onsuksesvol op die korter duur is, is dat die meeste stilometriese ontledings ontwerp is om lang literêre tekste te ontleed. Verder voer Stamatatos e.a. (2001:196, 208) aan dat tekslengtes van minder as 1 000 woorde nie geskik is vir stilometriese ontledings wat op die leksikale eienskappe van die outeur se taalgebruik fokus nie. “Leksikale eienskappe” verwys na die algemene woorde sowel as die funksiewoorde in ’n outeur se teks.

Dit is met ander woorde nodig om ’n stilometriese metode te gebruik wat wel akkurate resultate met korter tekste kan lewer, sodat die ontleding van korter tekste moontlik en sinvol sal wees. Die huidige navorsing poog om ’n forensies-linguistiese situasie te skep wat so na as moontlik aan die werklikheid is, waar die forensiese linguis wat die outeurskap van SMS’e probeer bepaal, nie ’n groot aantal tekste tot sy/haar beskikking sal hê nie. Selfs wanneer die forensiese linguis toegang het tot al die SMS’e op die verdagte se selfoon, is die lengte van die boodskappe steeds problematies, aangesien ’n groot aantal SMS’e nodig is om gelykstaande aan ’n vollengte-teks (ongeveer 1 000 woorde) te wees. Chaski (2001:4) meen dat korter tekste en min data ’n algemene verskynsel in die forensiese linguistiek is, en beskryf dokumente wat algemeen van forensiese belang is, as kort dokumente van beperkte omvang wat nie op enige manier aangevul kan word nie.

Ten spyte van die gebrek aan data wat in verskeie outeuridentifikasiesituasies aangetref word, is dit steeds in sommige gevalle moontlik om die outeurs van die verdagte teks te bepaal.

7. SMS-taal

SMS’e word, soos reeds genoem, in die forensiese linguistiek as bondige tekste met unieke eienskappe gekategoriseer. SMS’e was aanvanklik tot net 140 karakters per boodskap beperk, en dit het beteken dat SMS-gebruikers op nuwe en vernuwende maniere hul taal moes aanpas om te verseker dat soveel moontlik inligting in die bestek van 140 karakters ingepas kon word (Mobile Pronto 2010). SMS-taal het mettertyd ook verder ontwikkel en idiolektiese eienskappe van elke outeur begin insluit, aangesien daar verskeie maniere is om woorde in ’n SMS voor te stel. SMS-gebruikers kan onder andere van afkortings, akronieme en emotikons in hul boodskappe gebruik maak.

7.1 Die invloed van Engels

Die invloed van Engels op Afrikaans is nie ’n onlangse verskynsel nie. Carstens (2003:314) wys daarop dat Afrikaans reeds sedert die begin van die 19de eeu sterk deur Engels beïnvloed word. Carstens (2003:315) is ook van mening dat die feit dat die

Afrikaanssprekende vandag Engels gebruik as “kontakmiddel met die buiteland”, nie weggeredeneer kan word nie. Afrikaansprekendes hoor en lees nie net daagliks Engels nie, maar moet ook gereeld in sosiale of formele situasies in Engels met ander individue kommunikeer.

Carstens (2003:315) meen daar is veral drie terreine in Afrikaans waar die invloed van Engels opvallend is. Hierdie terreine is (1) leenwoorde, (2) direkte oornames en (3) anglisismes. In Afrikaanse SMS-taal word veral direkte oornames uit Engels gebruik. Dit vind plaas omdat taalvermenging nie in hierdie situasie taboe is nie (Thiart 2014:77).

Enkele Engelse woorde wat gereeld in Afrikaanse SMS’e voorkom, is *cheers*, *worry*, *weird* en *awesome*.

7.2 Logogramme

’n Logogram kan beskryf word as ’n verskynsel waar ’n letter, syfer, simbool of teken ’n hele woord of selfs ’n frase verteenwoordig. Logogramme in SMS-taal is ’n vernuwende manier om soveel moontlik binne die beperkings van ’n SMS te skryf (Thiart 2014:78). Algemene logogramme in Afrikaanse SMS-taal is onder andere :

• @ = by (die): Sien jou @ die fliek / Sien @ Menlyn.
• x = “soen”. Hierdie logogram word gewoonlik gebruik om ’n boodskap af te sluit. Word ook gebruik as plaasvervanger vir “ek is” of “ek’s” in die teks; byvoorbeeld: x moeg (Ek is/Ek’s moeg).
• k = ek: k is kwaad.
• zzzzz – slaap. Kan aandui dat die outeur moeg is of nou gaan slaap, byvoorbeeld: Gaan nou zzzzzz.

7.3 Verkortings

Olivier (2013: 490) verwys na Heyns (2009) en Saal (2012), wat daarop wys dat daar verskeie maniere is waarop verkorting in SMS’e gebruik word. Hulle noem onder andere twee maniere van verkorting. In die eerste geval word woorde gewoonlik verkort deur ’n sillabe weg te laat, en gevolglik word woorde soos *bib*, *foon* en *prof* aangetref. ’n Verdere vorm van verkorting wat plaasvind, is weglatings (“omissions”). Weglatings (wat soortgelyk is aan sinkopee in formele taalkunde) beteken dat ’n letter (klank) uit ’n woord weggelaat word (Thiart 2014:80). By weglatings is dit gewoonlik die vokale wat weggelaat word om die woord te verkort. Dit verhoog ook die spoed waarteen ’n SMS getik kan word. Algemene weglatings in Afrikaanse SMS-taal is onder andere in die volgende woorde te sien:

• skt = skat
• ltr/latr = later
• wanr = wanneer
• nj = en jy.

’n Derde verskynsel wat onder verkorting gekategoriseer kan word, is die gebruik van “voorletterwoorde”. Hierdie woorde word geskep deur slegs die eerste letter van elke woord

in ’n sin of naam saam te voeg om ’n nuwe “woord” te vorm, byvoorbeeld: KFC (Kentucky Fried Chicken). Voorletterwoorde kom meer gereeld in SMS-taal voor en word deurentyd geskep. Alhoewel akronieme (letterwoorde) ook soms in SMS’e voorkom, word daar hoofsaaklik van bekende akronieme gebruik gemaak. Akronieme word gevorm deur afgekorte dele van woorde saam te voeg tot ’n nuwe woord (byvoorbeeld: Unisa (Universiteit van Suid-Afrika)). Die invloed van Engels by voorletterwoorde is weer eens opmerklik. Daar is verskeie Engelse frases wat ook deur middel van voorletterwoorde verkort word en in Afrikaanse SMS-taal voorkom:

• brb = be right back
• gtg = got to go (“g2g” kom ook algemeen voor)
• wmj? = wat maak jy?
• hgd? = hoe gaan dit?

Die eienskappe van SMS-taal wat hier bespreek is, is slegs enkeles wat in Afrikaanse SMS-taal en SMS-taal in die algemeen aangetref word. Omdat hierdie eienskappe van persoon tot persoon verskil en oor ’n tydperk in dieselfde persoon se skryfstyl varieer, is dit soms moeilik om die vermeende idiolek van ’n outeur te identifiseer. Dit is waarom forensiese linguiste groot hoeveelhede data van elke individu benodig wanneer kort tekste ontleed word.

8. Metodologie

8.1 Inleiding

Die metodes wat in die huidige navorsing gebruik is, word gesamentlik as ’n gemengde metode beskou, aangesien beide ’n kwalitatiewe en ’n kwantitatiewe metode gebruik is om die data te ontleed (Dörnyei 2007:44; Angouri 2010:29–30). Angouri (2010:29–30) voer aan dat dit voordelig is om van gemengde metodes gebruik te maak in navorsing in die sosiale en geesteswetenskappe, en haal onder andere Greene (1989) aan wat skryf: “[C]ombining the two paradigms [d.i. kwantitatiewe en kwalitatiewe metodes – L.T.] is beneficial for constructing comprehensive accounts and providing answers to a wider range of research questions.” Dörnyei (2007:166) waarsku egter dat die gemengde metode nie as ’n “‘anything goes’ disposition” beskou moet word nie en dat dit belangrik is om te onthou dat die navorser moet verseker dat die navorsingsmetodologie en interpretasie van die data konsekwent is.

In hierdie studie is die gemengde metode gebruik, omdat beide kwalitatiewe en kwantitatiewe ontledings nodig was om die navorsingsvrae te beantwoord:

1. Kan daar in Afrikaans ’n generiese SMS-taal geïdentifiseer word wat outeuridentifikasie bemoeilik?
2. Is dit moontlik om binne die veronderstelde generiese SMS-taal individuele, idiolektiese taal by SMS-gebruikers te identifiseer?

3. Tot watter mate is dit moontlik om die outeur van ’n vermeende SMS-teks te identifiseer met die beperkte data wat tipies ter beskikking is?

Die huidige navorsing is ’n simulasiestudie, wat beteken dat dit nie ’n ontleding is van SMS’e wat met werklike misdade verband hou nie. ’n Denkbeeldige situasie is geskep wat soortgelyk is aan die werklike situasies waarmee forensiese linguiste gekonfronteer kan word. Die denkbeeldige situasie bestaan uit ’n paar moontlike vermeende outeurs wat die verdagte teks (Teks X) kon geskryf en gestuur het. Die navorser moet vasstel of dit moontlik is om met ’n hoë persentasie sekerheid die ware outeur van Teks X te identifiseer.

Omdat die navorsing poog om ’n situasie te skep wat werklike forensies-linguistiese scenario’s naboots, is dit nodig om soveel moontlik metodes (binne die beperkte aard van die studie) te ondersoek, aangesien daar verkieslik honderd persent sekerheid moet wees oor die outeur van ’n verdagte teks wanneer ’n outeuridentifikasie-ontleding afgehandel is. So ’n mate van sekerheid is nie tot op hede moontlik nie. Ontledings wat as moontlike bewyse in die hof gebruik word, moet nietemin steeds met ’n hoë mate van sekerheid gepaard gaan wanneer ’n moontlike outeur van ’n spesifieke teks geïdentifiseer word. ’n Gemengde metode sal, in hierdie geval, teoreties ’n hoër mate van sekerheid oor die vermeende outeur van ’n teks tot gevolg te hê.

Die probleme rondom die ontleding van SMS’e in outeuridentifikasie het hoofsaaklik met die bondigheid van die boodskappe te doen. Die forensiese linguis moet daarom hipoteties oor ’n groot getal SMS’e van een outeur beskik wanneer hy of sy poog om idiolektiese eienskappe te identifiseer wat die outeur en die bepaalde boodskappe met mekaar verbind. In hierdie navorsing is die uiteindelike doel egter om vas te stel of dit enigsins moontlik is om die outeur van ’n boodskap te identifiseer wanneer die forensiese linguis slegs oor enkele SMS’e van elke outeur beskik en die getal moontlike outeurs meer as twee of drie is. So ’n scenario kan moontlik voorkom wanneer daar bepaal moet word watter lid van ’n bepaalde bende of watter persoon in ’n besigheid (of afdeling van ’n besigheid) ’n bepaalde SMS gestuur het. Navorsing wat outeuridentifikasie in klein getalle kort tekste ondersoek, is reeds in die verlede uitgevoer, maar hierdie studies het steeds van aansienlik meer tekste gebruik gemaak as wat in die huidige navorsing gebruik is (Chaski 2005; Mohan e.a. 2010; MacLeod en Grant 2012).

Dit is belangrik om daarop te wys dat enige resultate wat uit die huidige navorsing verkry is, nietemin as van belang beskou kan word. Selfs negatiewe resultate is steeds van waarde. “Negatiewe resultate” beteken bloot dat dit onwaarskynlik is dat ’n outeur geïdentifiseer kan word wanneer daar ’n groep verdagtes is en die forensiese linguis slegs enkele SMS’e (5 tot 10) van elke vermeende outeur tot sy/haar beskikking het. So ’n resultaat gee natuurlik aanleiding tot verdere navorsing, waar ander metodes as dié wat in die huidige navorsing gebruik is, getoets kan word om die sukses van sulke metodes in ’n minder ideale, realistiese scenario te bepaal.

8.2 Deelnemers

In die huidige studie bestaan die korpus uit 13 deelnemers tussen die ouderdomme van 18 en 23 jaar. Aanvanklik is beplan dat 30 deelnemers aan die navorsing sou deelneem, maar

slegs 14 individue het ingestem, en uit daardie groep het slegs 13 deelnemers se data aan die kriteria voldoen. Om ingesluit te word in die huidige studie, moes die deelnemer ’n moedertaalspreker van Afrikaans wees, en moes hy/sy 5 tot 10 SMS’e van tussen 30 en 50 woorde elk aan die navorser stuur. Nadat etiekklaring deur die Universiteit van Pretoria vir die studie verleen is, is die potensiële deelnemers ingelig dat deelname aan die studie vrywillig is. Ingeligte toestemming is ook deur elke deelnemer onderteken. Elke deelnemer is ook deur ’n nommer (D1 tot D14) geïdentifiseer om te verseker dat die anonimiteit van die deelnemers gehandhaaf word.

Volgens ’n studie wat in 2012 aanlyn gepubliseer is (PewInternet 2012), is die ouderdomsgroep 18 tot 29 die groep wat die meeste van SMS’e vir kommunikasie-doeleindes gebruik maak, en is individue tussen 18 en 22 jaar die mees aktiewe groep selfoongebruikers. Die huidige navorsing het egter ook 23-jarige deelnemers toegelaat.

Dit is belangrik om daarop te let dat boodskapdienste soos WhatsApp en BBM ook gebruik word om boodskappe te stuur. Dieselfde taalgebruik wat in SMS’e voorkom, word ook aangetref in WhatsApp en BBM, aangesien hulle bloot nuwer en goedkoper maniere is om ’n SMS te stuur. In hierdie navorsing is die term *SMS* gevolglik gebruik om te verwys na enige kortboodskapdiens (“short message service”) wat dit vir individue moontlik maak om vinnig elektroniese boodskappe te stuur en wat nie ander elektroniese kommunikasie soos e-posboodskappe, boodskappe (“updates”) op sosiale netwerke en bloginskrywings insluit nie. Hierdie studie maak van die ontvang en stuur van die gewone SMS gebruik, omdat die insamelingsmetode dit toelaat.

8.3 Datastel

Die denkbeeldige situasie wat vir die navorsing geskep is, vereis ’n beperkte hoeveelheid data wat tot die forensiese linguis se beskikking is. Om hierdie rede is daar nie gepoog om vollengte-tekste (ongeveer 1 000 woorde, afhangende van die teksdigtheid en leesbaarheid) in die navorsing te gebruik nie.

Die datastel in hierdie studie bestaan uit 2 434 woorde. Elke deelnemer het 5 tot 10 SMS’e aan die navorser gestuur met ’n lengte van tussen 30 en 50 woorde elk. Die deelnemers is gevra om self die SMS’e te selekteer wat hulle aan die navorser wil stuur. Dit beteken dat die deelnemers nie vir die doeleindes van hierdie navorsing nuwe SMS’e moes tik wat aan die navorser gestuur word nie. Een deelnemer (Deelnemer 2) is gevra om ’n ekstra stel SMS’e te stuur wat as die verdagte teks (Teks X) sou dien. Teks X word gebruik om te bepaal of dit moontlik is om die outeur van die verdagte teks vas te stel, en dit is die teks waarmee al die ander tekste vergelyk word.

Om rekord te hou van elke deelnemer se teks is die selfoonnommer van elke deelnemer wat SMS’e na ’n kortkode (“short code”) gestuur het, op die sisteem geregistreer. Die selfoonnommers is gebruik om te verseker dat die deelnemers se SMS’e onder die korrekte nommer (1 tot 14) gestoor is.

Die datastel is gebruik om vas te stel of daar idiolektiese taalgebruik onder die 13 deelnemers bestaan. Die datastel is ook gebruik om vas te stel of daar as alternatief vir

idiolektiese taalgebruik, generiese SMS-taaleienskappe onder hierdie groep Afrikaanse SMS-gebruikers bestaan. Daar kan vervolgens vasgestel word of die outeur van ’n verdagte teks deur middel van die beskikbare data geïdentifiseer kan word.

8.4 Datastel-insameling

Vir die ontvangs van die SMS’e is daar van SMSPortal gebruik gemaak. SMSPortal is ’n massa-SMS-diens wat dit moontlik maak om ’n groot aantal SMS’e te ontvang en uit te stuur.

Die grootste beperking op die data wat deur SMSPortal ontvang is, is die feit dat die sisteem die inhoud van die SMS’e net tot op 60 karakters kan registreer. Dit beteken dat die SMS’e, wat alreeds as kort tekste geklassifiseer word, selfs nog meer verkort is.

9. Instrumente en programmatuur

Drie instrumente is vir die data-ontleding gebruik: Antconc en WordSmith Tools is programmatuur wat vir verskeie linguistiese ontledings gebruik kan word, terwyl die n-gramontleding wat in die navorsing gebruik is, tradisioneel gebruik word in taalidentifiseringsprogramme en -sagteware.

9.1 Analitiese metodes

9.1.1 Stilistiese ontleding

Die SMS-data is op twee wyses ontleed. Die eerste ontleding is ’n beskrywende, kwalitatiewe, stilistiese ontleding van die teks. Stilistiese ontledings in die letterkunde het tradisioneel op die estetiese kwaliteit van uitdrukkings en die ooreenstemming van taalgebruik met bepaalde taalreëls gefokus. Moderne linguistiese stilistiese ontledings hou egter, in kontras met laasgenoemde, verband met die wetenskaplike interpretasie van stylmerkers soos dit waargeneem en beskryf word in die taalgebruik van verskillende groepe en individue (McMenamin 2010:488). Stylmerkers kan beskou word as die waarneembare resultaat van die onbewuste keuses wat ’n outeur tydens die skryfproses maak.

Die stilistiese ontleding verteenwoordig die menslike element van die data-ontleding en is in wese teksontleding. Met teksontleding poog die navorser om taaldata in te samel wat so natuurlik moontlik onder die bepaalde omstandighede bekom is; met ander woorde, die data moet so goed as moontlik die natuurlike taalgebruik, of taalgebruik binne ’n natuurlike omgewing, verteenwoordig. In die geval van hierdie ondersoek word veronderstel dat die SMS-taal wat die deelnemers gebruik, wel tot ’n groot mate natuurlike taaldata verteenwoordig, omdat die data wat die deelnemers aan die navorser gestuur het, onder normale omstandighede geproduseer is en die deelnemers bloot die SMS’e wat hulle wou stuur, moes kies uit die SMS’e wat reeds op hul selfone beskikbaar was. Die teksontleding berus hoofsaaklik op die navorser se eie interpretasie en taalkennis (Lazaraton 2009:247–

50). Tydens hierdie proses besluit die linguis watter kenmerke hy/sy in die teks wil selekteer vir verwerking.

In die huidige navorsing is ’n kleurgids gebruik om die onderskeie elemente in die tekste te merk. Deur middel van hierdie metode kan die linguis ’n oorsig kry van die linguistiese veranderlikes in ’n teks of in verskeie tekste. In die huidige navorsing is die volgende eienskappe in die onderskeie SMS’e gemerk:

• Leestekens
• Aanspreekvorme
• Lagtekens / Piktogramme
• Logogramme
• Ongewone gebruik van hoofletters en kleinletters
• Verkortings
• Niestandaardspellings
• Engelse woorde/sinne
• Funksiewoorde
• Individuele eienskappe (herhalende woorde, frases en/of leestekengebruik).

Elke deelnemer se teks en Teks X is stilisties deur middel van die kleurgids ontleed (figuur 1). Daarna is die 17 kenmerke van Teks X (wat tydens die stilistiese ontleding verkry is) getabuleer. Elke deelnemer se teks is met die bepaalde kenmerke vergelyk en ’n persentasie van ooreenkoms is uitgewerk. Dit beteken dat die hoeveelheid ooreenstemming tussen elke deelnemer en Teks X bepaal is. Die resultate van hierdie vergelykings word in tabel 4 aangedui.

Figuur 1. ’n Voorbeeld van hoe die tekste deur middel van die kleurgids gemerk is vir ontleding.

Nadat so ’n ontleding afgehandel is, kan die navorser waarnemings maak oor die onderskeie deelnemers se tekste. Dit is nie in alle gevalle moontlik om so ’n stilistiese ontleding met die hand uit te voer nie, maar die klein hoeveelheid data in die huidige navorsing het hierdie stilistiese metode toegelaat.

9.1.2 Stilometriese ontleding

In die tweede fase van die ontleding, naamlik die stilometriese ontleding, is die programme Antconc en WordSmith Tools, asook n-gramontleding, gebruik om die tekste statisties te verwerk. Hierdie gedeelte van die ontleding verteenwoordig die kwantitatiewe aspek van die gemengde metode. Die doel van die stilometriese ontleding is om statisties te probeer vasstel watter woorde oorwegend deur ’n spesifieke individu gebruik word, hoe belangrik hierdie woorde in die bepaalde teks is en ook in watter patrone die woorde in die individu se sinne voorkom. Slegs twee van Antconc se funksies (die sleutelwoordlys (Keyword list) en die woordelys (Word list) is vir die doeleindes van die huidige navorsing gebruik.

9.1.2.1 Antconc: Die Word list-funksie en die Keyword list-funksie

Die Word list-funksie word gebruik om ’n lys te genereer van al die woorde in ’n bepaalde korpus, asook die frekwensie waarvolgens die woorde in die korpus voorkom. Woorde wat die meeste in die korpus voorkom en gevolglik die hoogste frekwensie het, verskyn boaan die lys. Die Word list-funksie het ook die vermoë om woorde op grond van “stam”-vorme te tel (Anthony 2004:10).

Die lyste van verskillende tekste kan met mekaar vergelyk word om sodoende vas te stel of daar ooreenkomste is tussen die frekwensies van sekere woorde in verskillende tekste. ’n Woordelys van elke deelnemer se teks, asook die verdagte teks (Teks X), is gegenereer, en die teks van elke deelnemer is met Teks X vergelyk. In tabel 1 en tabel 2 word die frekwensie van die eerste 10 algemene woorde en 11 funksiewoorde van elke deelnemer en die verdagte teks met mekaar vergelyk. Daar is op 11 funksiewoorde besluit, aangesien daar ’n variasie op die woord “maar” in die verdagte teks voorkom. Om hierdie rede is “ma” (die variasie op “maar”) ook ingesluit.

Vergelyking: 10 mees frekwente *woorde*

	Teks X	D1	D2	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14
ons	6	1	7	2	2	3	0	0	1	2	5	5	7	4
en	4	4	3	4	11	6	4	4	3	2	1	3	7	4
moet	4	1	1	2	3	0	1	1	0	0	1	0	0	0
ek	3	7	5	17	17	8	9	1	12	3	4	4	7	8
kry	3	1	1	3	2	1	1	1	1	1	0	0	0	2
lekker	3	0	1	5	0	0	0	0	0	0	1	0	2	0
wat	3	1	1	1	2	2	1	0	0	1	1	1	2	0
die	2	3	4	3	3	10	4	2	3	6	3	5	7	0
dit	2	3	4	8	1	2	3	1	1	1	4	1	2	0
gehad	2	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabel 1: ’n Vergelyking tussen die 10 mees frekwente woorde in elke deelnemer se teks met dié van Teks X.

Vergelyking: 11 mees frekwente *funksiewoorde*

	Teks X	D1	D2	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14
ons	6	1	7	2	2	3	0	0	1	2	5	5	7	4
en	4	4	3	4	11	6	4	4	3	2	1	3	7	4
moet	4	1	1	2	3	0	1	1	0	0	1	0	0	0
ek	3	7	5	17	17	8	9	1	12	3	4	4	7	8
wat	3	1	1	1	2	2	1	0	0	1	1	1	2	0
die	2	3	4	3	3	10	4	2	3	6	3	5	7	0
dit	2	3	4	8	1	2	3	1	1	1	4	1	2	0
ma	2	4	0	0	1	0	0	0	0	0	0	0	0	0
maar	2	0	2	5	1	3	1	1	0	0	2	0	1	0
n	2	1	1	1	8	0	2	4	2	3	2	2	3	0
sal	2	0	4	2	1	3	3	3	2	0	1	3	2	0

Tabel 2: ’n Vergelyking tussen die 11 mees frekwente funksiewoorde in elke deelnemer se teks met dié van Teks X.

Dit is duidelik dat Deelnemer 14 in albei tabelle die minste ooreenkoms met Teks X toon, terwyl dit blyk dat Deelnemers 2 en 11 die meeste ooreenkoms met Teks X toon.

Die Word list-funksie is ook gebruik om te bepaal of daar ’n generiese SMS-taal onder die groep deelnemers bestaan. Aangesien elke woord (op grond van spelling) apart gelys word, is dit maklik om die variante van een woord vas te stel en sodoende te bepaal of daar wel ’n generiese SMS-taal onder die groep deelnemers bestaan.

Alhoewel die Word list-funksie nuttig is om, onder andere, die mees frekwente woorde in die teks te identifiseer, kan hierdie funksie nie aandui hoe belangrik ’n bepaalde woord binne die korpus tekste is nie (Anthony 2004:10). Om laasgenoemde inligting te bekom moet die Keyword list-funksie van Antconc ingespan word (Anthony 2005:733). Om die Keyword list-funksie suksesvol te gebruik, word daar twee korpusse of tekste benodig. Hierdie funksie gebruik statistiese toetse soos die chi-kwadraat- of logwaarskynlikheidstoetse om aan te dui watter woorde ongewoon frekwent in die korpus voorkom in vergelyking met die woorde in ’n verwysingskorpus (Anthony 2014:7). Teks X is as die verwysingskorpus gebruik, aangesien elke deelnemer se teks met Teks X vergelyk is.

Die Keyword list-funksie genereer ’n lys van woorde soortgelyk aan dié van die Word list-funksie, maar hier word ook die “keyness” of “sleutelwaarde” van elke woord binne die korpus/teks uitgelig. Die sleutelwaarde van elke woord word bepaal deur die ongewoon hoë frekwensie van daardie woord binne die korpus. As ’n bepaalde woord ’n hoë frekwensie het, beteken dit dat daardie woord aansienlik meer in die korpus voorkom as in die verwysingskorpus. As die woord ’n laer gemiddelde sleutelwaarde het in vergelyking met die sleutelwaarde van ander tekste wat met dieselfde verwysingskorpus vergelyk is, beteken dit dat die woord nie soveel meer kere in die korpus voorkom as in die verwysingskorpus nie. Volgens Kotzé (2010:189) bestaan daar ’n sterk moontlikheid dat twee verskillende outeurs

vir die tekste verantwoordelik is wanneer die sleutelwaarde van die woorde baie hoog is. Hoe laer die sleutelwaarde van die woorde in ’n teks, hoe groter is die kans dat dieselfde outeur albei tekste geproduseer het.

9.1.2.2 WordSmith Tools

WordSmith Tools (WST) se Keyword list-funksie is gebruik om vas te stel of die resultate wat deur Antconc verkry is, met die resultate in WST ooreenstem.

Dit is belangrik om in ag te neem dat die beperkings en verstellings nie dieselfde is by Antconc as wat in WST die geval is nie. Die rede hiervoor is dat Antconc nie van ’n outomatiese p-waarde van 0,05% gebruik maak nie. Nadat die resultate in Antconc verkry is, moet die navorser self besluit watter persentasies van belang is. Antconc maak ook in die sleutelwoordlys van die logwaarskynlikheidsverhouding gebruik, eerder as die chi-kwadraat-statistiek. In WST is die p-waarde 0,000001. Dit kan verhoog word na 0,01 of 0,05. Die opsie om ’n chi-kwadraat-toets of ’n logwaarskynlikheidstoets te gebruik, is beskikbaar in WST. Die verskil tussen die resultate van hierdie twee statistiese toetse is egter so klein dat dit glad nie die resultate en gevolglik die identifikasie van ’n moontlike outeur beïnvloed het nie.

Tydens die ontleding van die data in WST is daar bepaal dat die p-waarde in WST op 0,1 gestel moet word voordat enige resultate van die beperkte data verkry kan word. WST genereer aansienlik minder resultate as Antconc en daar is slegs enkele verskille in die rangorde van woorde wat in die sleutelwoordlyste van elke program aangedui word, asook klein verskille in die sleutelwaarde van die woorde.

Dit is nie werklik nodig om albei programme vir die ontleding van die data te gebruik nie, aangesien die WST-ontleding nie enige groot verskille in die resultate tot gevolg het nie.

9.1.2.3 N-gramontleding

Die laaste tipe stilometriese ontleding wat uitgevoer is, het gebruik gemaak van n-gramontleding in ’n poging om die resultate te verbeter. Die klein hoeveelheid data is egter problematies, en het daartoe gelei dat slegs ’n baie eenvoudige n-gramontleding van die data uitgevoer is. Die metode wat gebruik is, is Cavnar en Trenkle (1994) se metode vir taalherkenning. Die n-gramontleding werk deur n-gramprofiele te bereken en te vergelyk. In die eerste plek word die profiel van die opleidingsdata bepaal (dit is die korpus waarmee die ander tekste vergelyk gaan word) en daarna word die profiel vir elke dokument wat geklassifiseer moet word, bepaal. Laastens word die “afstand” tussen die opleidingsdata en elke dokument wat geklassifiseer moet word, vasgestel. Die dokumente met die kortste afstande is, met ander woorde, die dokumente wat die naaste aan die opleidingsdata is en gevolglik die grootste getal ooreenkomste met die opleidingsdata toon. Wolff (2014) meen ander metodes sou ook oorweeg kon word, maar waarsku dat van hierdie metodes moontlik te kompleks is om te implementeer, aangesien die hoeveelheid data in die huidige studie so beperk is.

9.1.2.4 Die chi-kwadraattoetse (Pearson chi-kwadraattoets en die Yates-korreksie)

9.1.2.4.1 Die Pearson chi-kwadraattoets

Chi-kwadraattoetse is reeds verskeie kere deur navorsers gebruik om die frekwensies van woorde en leestekens in verskillende tekste met mekaar te vergelyk (Hubbard 1995, Chaski 2001 en Kotzé 2007). Die bekendste chi-kwadraattoets is Pearson se chi-kwadraattoets, wat ook as Pearson se chi-kwadraattoets vir onafhanklikheid bekend staan. Die chi-kwadraattoets word gebruik om twee of meer frekwensies met mekaar te vergelyk om vas te stel wat die moontlikheid (weergegee in persentasievorm) is dat enige verskille tussen die frekwensies bloot toevallig is, of nie. Dit kan in die forensiese linguistiek gebruik word om die "probability of success" (waarskynlikheid van sukses) dat twee tekste deur dieselfde outeur geproduseer is, te bepaal. Hierdie waarskynlikheid van sukses word deur 'n p-waarde aangedui. Volgens Kotzé (2007:391) moet die graad van waarskynlikheid op 0,05 (5%) gestel word. Dit laat toe vir soveel moontlik gevalle van beduidende verskille tussen die dokumente. Die resultate wat in die chi-kwadraattoets verkry word, bepaal of die nulhipotese (H_0) aanvaar of verwerp word.

Die nulhipotese is dat daar geen verhouding tussen die veranderlikes is nie (dit beteken dat daar geen verhouding tussen die waargeneemde frekwensies (O)[1] en die verwagte frekwensies (E)[2] is nie, en gevolglik dat daar geen beduidende verskil tussen die twee waardes/frekwensies is nie). Die twee veranderlikes is, met ander woorde, onafhanklik van mekaar. Daarteenoor is die alternatiewe hipotese (H_a) dat daar wel 'n verhouding tussen die veranderlikes bestaan (daar is 'n verhouding tussen O en E) en dat die veranderlikes afhanklik is van mekaar (met ander woorde, daar is 'n beduidende verskil tussen die twee waardes/frekwensies).

Indien $p < 0,05$ / $p=0,05$, word die nulhipotese aanvaar. Indien $p > 0,05$, word die nulhipotese verwerp. By outeuridentifikasie beteken dit die volgende:

$p < 0,05$ / $p=0,05$ beteken dat die moontlikheid dat daar 'n verhouding tussen die veranderlikes bestaan, slegs 5% of minder is. Daar is, met ander woorde, beduidende verskille tussen die tekste, wat daarop dui dat die tekste heel waarskynlik deur verskillende outeurs geproduseer is. 'n Verskil of afwyking op 'n waarskynlikheidsvlak van 5% of minder kan aan die toeval toegeskryf word.

$p > 0,05$ beteken dat die verskille tussen die tekste minder is en daarom bestaan die moontlikheid dat dieselfde outeur verantwoordelik is vir albei tekste. Daar is egter steeds 'n kans dat die veranderlikes in die tekste genoeg van mekaar verskil om aan te dui dat die tekste deur twee of meer verskillende outeurs geproduseer is. Indien $p=0,10$, beteken dit dat daar slegs 'n 10%-moontlikheid is dat daar 'n verhouding tussen die veranderlikes bestaan. Daar is met ander woorde steeds 'n 90%-kans dat daar geen verhouding tussen die veranderlikes is nie.

'n Hoër p-waarde dui daarop dat daar 'n kleiner moontlikheid van toeval is en 'n laer p-waarde dui op 'n groter moontlikheid van toeval.

Die “gewone” chi-kwadraattoets moet volgens Chaski (2001:9) gebruik word slegs indien nie meer as 20 persent van die verwagte frekwensies minder as 5 is en geen verwagte frekwensies minder as 1 is nie. In die huidige navorsing is die data beperk en voldoen gevolglik nie aan die genoemde kriteria nie. Antconc is gebruik om die frekwensies van die woorde in elke deelnemer se teks, asook in Teks X, te bepaal. Nadat die frekwensies van die 10 mees frekwente algemene woorde en die 11 mees frekwente funksiewoorde verkry is, was dit duidelik dat byna 70% van die verwagte frekwensies minder as 5 sou wees, en sommige verwagte frekwensies minder as 1.

Die Pearson chi-kwadraat is nietemin gebruik om die data te ontleed, maar dit is duidelik uit tabel 5 en tabel 6 dat daar op grond van die chi-kwadraatwaardes en p-waardes wat uit die Pearson chi-kwadraattoets verkry is, nie enige definitiewe gevolgtrekkings gemaak kan word nie.

9.1.2.4.2 Die Yates-korreksie

Aangesien die Chi-kwadraattoets geen bruikbare resultate oplewer nie (soos aangedui in tabel 4 en 5) is daar besluit om van die Yates-korreksie gebruik te maak.

Daar is besluit om die data eerder in 2x2-tabelle te verdeel en die Yates-korreksie op elke tabel toe te pas. Die Yates-korreksie stel voor dat 0,5 van die verskil tussen die waargeneemde en verwagte frekwensies vir elke waarde afgetrek word voordat die chi-kwadraatformule gebruik word (Goehring 1981). Hierdie korreksie word volgens Statistics How To (2015) soos volg omskryf:

> The Yates correction is a correction made to account for the fact that both Pearson’s chi-square test and McNemar’s chi-square test are biased upwards for a 2x2 contingency table. [...] Chi-square tests are biased upwards when used on 2x2 contingency tables. The reason for this is that the statistical chi-square distribution is continuous and the 2x2 contingency table is dichotomous.

Dit beteken dat die Pearson en McNemar Chi-kwadraattoetse soms die statistiese resultate groter maak as wat dit moet wees in ’n 2x2-tabel, aangesien so ’n tabel tweeledig is. Met ander woorde, die tabel bevat twee veranderlikes.

Volgens How2stats (2011) word die Yates-korreksie algemeen in die literatuur gebruik, maar daar is oortuigende bewyse dat die korreksie heeltemal te konserwatief is, selfs wanneer dit in klein hoeveelhede data gebruik word. How2stats (2011) verwys na verskeie navorsers wat meen dat die Yates-korreksie, as gevolg van laasgenoemde waarneming, eintlik glad nie gebruik moet word nie. Navorsers na wie How2stats (2011) verwys, is Camilli en Hopkins (1978, 1979), Feinberg (1980), Larntz (1978) en Thompson (1988).

Die Yates-korreksie is op die data toegepas, maar het nie tot meer bruikbare resultate gelei nie. Die resultate van die Yates-korreksie word by tabel 6 en tabel 7 ingesluit.

Die voorafgaande bespreking van stilistiese en stilometriese ontledings wat op die data toegepas is, maak dit duidelik dat die beperkte hoeveelheid data in die huidige studie

problematies is. Alhoewel daar woordelyste en sleutelwoordlyste gegenereer kan word, is die persentasies van veral die sleutelwaardes baie klein. Uit die woordelyste wat gegenereer is, lyk dit asof daar nie ’n generiese taal onder die groep deelnemers bestaan nie. Dit blyk egter dat dit wel moontlik is om die deelnemers van mekaar te onderskei op grond van hul skryfstyle en taalgebruik.

10. Resultate

Die resultate van die vier verskillende ontledings wat op die data toegepas is, word soos volg opgesom:

10.1 Resultate van die stilistiese ontleding.

Uit die stilistiese ontleding blyk dit dat daar een deelnemer is wat die hoogste persentasieooreenkoms (verskynselooreenkomste) met Teks X het. Hierdie resultate word in tabel 3 saamgevat. Deelnemer 2 (D2) deel 64,7% van die kenmerke in Teks X. D1, D4 en D5 kan ook as moontlike outeurs oorweeg word, aangesien hulle ook ’n ooreenkomspersentasie bo 50% het. Aangesien die datastel in die huidige navorsing beperk is, is daar besluit dat slegs tekste met ooreenkomspersentasies van 50% en hoër oorweeg sou word as geskryf deur die moontlike outeur van die verdagte teks. D1, D4 en D5 deel ’n ooreenkomspersentasie van 58,8% met die verdagte teks. Deur hierdie stilistiese ontleding te gebruik, was die navorser in staat om D2 as een van die moontlike outeurs te identifiseer, maar die persentasie waarmee D2 met Teks X ooreenstem, is te laag om met sekerheid enige gevolgtrekkings te maak.

Deelnemers	Hoeveelheid ooreenkoms met Teks X
Deelnemer 1	10/17= 58,8%
Deelnemer 2	11/17 = 64,7%
Deelnemer 4	10/17 = 58,8%
Deelnemer 5	10/17 = 58,8%
Deelnemer 6	2/17 = 11,8%
Deelnemer 7	4/17 = 23,5%
Deelnemer 8	6/17 = 35,3%
Deelnemer 9	5/17= 29,4%
Deelnemer 10	2/17 = 11,7%
Deelnemer 11	4/17 = 23,5%
Deelnemer 12	5/17 = 29,4%
Deelnemer 13	7/17 = 41,2%
Deelnemer 14	4/17 = 23,5%

Tabel 3: Die resultate van die stilistiese ontleding

10.2 Resultate van die Pearson chi-kwadraattoets en die Yates-korreksie op die data van die 10 mees frekwente algemene woorde en die 11 mees frekwente funksiewoorde.

Uit die resultate van die Pearson chi-kwadraattoets op die 10 mees frekwente funksiewoorde (tabel 4) lyk dit asof D2 (46,83%) en D8 (47,85%) die hoogste p-waardes het en daarom as twee moontlike outeurs van die verdagte teks geïdentifiseer kan word. Albei die waardes is egter te laag om enige definitiewe gevolgtrekkings te maak. Uit die resultate van die Pearson chi-kwadraattoets op die 11 mees frekwente funksiewoorde (tabel 5) is D2 (64,69%) en D11 (72,64%) as die twee deelnemers geïdentifiseer wat die hoogste p-waardes het. D2 en D11 het albei ’n hoë p-waarde, maar daar moet onthou word dat beperkte data die werklikheid kan verdraai, en daarom moet enige resultate wat deur middel van toetse op beperkte data verkry is, baie versigtig geïnterpreteer word.

	Chi^2	p
D1	12,47	0,1881
D2	8,67	0,4683
D4	18,085	0,0342
D5	20,341	0,0159
D6	19,18	0,0237
D7	17,728	0,0385
D8	8,563	0,4785
D9	21,1	0,0122
D10	12	0,2133
D11	9,41	0,4003
D12	12,81	0,1714
D13	14,301	0,1120
D14	16,052	0,0658

Tabel 4: Die resultate van Pearson se chi-kwadraattoets vir die 10 mees frekwente algemene woorde.

	Chi^2	p
D1	12,59	0,2475
D2	7,815	0,6469
D4	19,258	0,0371
D5	50,98	0,0001
D6	16,74	0,0803
D7	15,058	0,1300
D8	11,782	0,2999
D9	19,93	0,0299
D10	13,372	0,2036
D11	6,99	0,7264
D12	10,22	0,4214

D13	11,3103	0,3339
D14	18,32	0,0498

Tabel 5: Die resultate van die Pearson chi-kwadraattoets vir die 11 mees frekwente funksiewoorde.

Nadat die Pearson chi-kwadraattoets op die data uitgevoer is, is die Yates-korreksie op dieselfde data uitgevoer in ’n poging om meer betroubare resultate te verkry. Vir die Yates-korreksie is daar slegs van 2x2-tabelle gebruik gemaak, aangesien die data saamgegroepeer is om die frekwensies vir elke deelnemer en Teks X in albei gevalle te verhoog.

Die Yates-korreksie is deur middel van die statistiese pakket R, weergawe 3.02, gedoen (Gerber 2014). Die funksie chisq.test() is gebruik. Uit tabel 6 en tabel 7 is dit duidelik dat die persentasies wat deur die Yates-korreksie verkry is, hoër is as die persentasies wat deur die Pearson chi-kwadraattoets verkry is. In tabel 6 is deelnemers 4 (62,78%), 8 (57,03%) en 14 (41,78%) as die deelnemers met die hoogste p-waardes geïdentifiseer. Die persentasies is egter steeds te laag om met enige sekerheid aan te voer dat enige van hierdie deelnemers wel die outeur van die verdagte teks is. In tabel 7 is deelnemers 6 (100%), 10 (100%), 7 (95,5%) en 1 (97,96%) as die vier deelnemers met die hoogste p-waarde geïdentifiseer. Die uiters hoë p-waardes beklemtoon die feit dat die enige resultate in die huidige navorsing met die nodige versigtigheid geïnterpreteer moet word. Ten eerste is twee 100%-ooreenstemmings uiteraard onmoontlik, aangesien dit nie vir een deelnemer moontlik sal wees om met die beperkte data 100%-ooreenstemming met die verdagte teks te kan toon nie; en dit is nog minder moontlik dat twee individue 100%-ooreensteming met Teks X sal toon. Die beperkte data en die resultate wat tot dusver uit die statistiese toetse verkry is, maak ook die persentasies by D7 en D1 hoogs onwaarskynlik.

	Yates-korreksie (Chi-kwadraatwaarde)	p-waarde
D1	2,388	0,1223
D2	2,0913	0,1481
D4	0,2351	0,6278
D5	4,571	0,03252
D6	4,6545	0,03097
D7	2,7978	0,09439
D8	0,3222	0,5703
D9	2,388	0,1223
D10	1,3846	0,2393
D11	2,1843	0,1394
D12	3,902	0,04823
D13	3,3174	0,06855
D14	0,6565	0,4178

Tabel 6: Die resultate van die Yates-korreksie op die 10 mees frekwente algemene woorde.

	Yates-korreksie (Chi-kwadraatwaarde)	*p-waarde*
D1	0,001	0,9796
D2	0,2835	0,5944
D4	1,0546	0,3044
D5	0,464	0,4958
D6	0	1
D7	0,0032	0,955
D8	2,1507	0,1425
D9	2,8839	0,08947
D10	0	1
D11	0,2064	0,6496
D12	0,2064	0,6496
D13	0,2684	0,6044
D14	7,7143	0,005479

Tabel: 7: Die resultate van die Yates-korreksie op die 11 mees frekwente funksiewoorde.

Dit is duidelik dat die Yates-korreksie nie tot meer betroubare resultate lei nie, en dat daar geen definitiewe gevolgtrekkings gemaak kan word oor die moontlike outeur van die verdagte teks op grond van die resultate nie.

Die volgende stap was om die persentasie sleutelwaarde van die woorde in elke teks in vergelyking met Teks X te bepaal.

10.3 Resultate van die Keyword-list funksie in Antconc.

Ten eerste is die sleutelwaardes van die eerste woord van elke deelnemer se teks ondersoek. Hier is D13, D4 en D6 as die deelnemers met die laagste sleutelwaardes (en daarom die hoogste moontlikheid as outeurs van die verdagte teks) geïdentifiseer.

Tydens die tweede toets is die sleutelwaardes van die eerste tien woorde van elke deelnemer se teks uitgewerk. Die resultate het aangedui dat D13, D11 en D2 die deelnemers met die laagste sleutelwaardes is.

Die laaste toets het die sleutelwaardes van die eerste 20 woorde in elke deelnemer se teks ondersoek en daar is vasgestel dat D13, D2 en D11 die laagste sleutelwaardes het.

10.4 Resultate van die N-gramontleding.

In die huidige navorsing is daar van 1-, 2-, 3- en 4-gramme gebruik gemaak, met ’n kort profiellengte van 50 n-gramme, wat in inkremente van 10 vermeerder is tot ’n profiellengte van 400. Slegs drie konfigurasies se resultate word in ag geneem, aangesien die omvang van die studie beperk is en daar moontlik honderde konfigurasies getoets kan word met kombinasies van die genoemde aantal n-gramme en profiellengtes.

Die eerste ontleding (bestaande uit 1-, 2-, 3- en 4-gramme, en profiellengtes van 50 tot 400) het aangedui dat D4 die waarskynlikste outeur van Teks X is. D2 en D5 is ook onderskeidelik as moontlike outeurs gelys.

Tydens die tweede ontleding (bestaande uit 1-, 2- en 3-gramme, en profiellengtes van 50 tot 400) is D2 as die waarskynlikste outeur van Teks X geïdentifiseer. D4 en D5 is ook as moontlik outeurs gelys.

Die derde ontleding (bestaande uit 1- en 2-gramme, en profiellengtes van 50 tot 400) het bevind dat D2 weer eens die waarskynlikste outeur van Teks X. D5 en D4 is weer eens gelys, maar het van posisie verander.

11. Antwoorde op die navorsingsvrae

Op grond van die resultate kan die navorsingsvrae soos volg beantwoord word:

1. Kan daar onder die groep deelnemers ’n generiese SMS-taal geïdentifiseer word wat outeuridentifikasie sou bemoeilik?

Uit die ontleding van die data om die teenwoordigheid van ’n generiese SMS-taal onder die groep deelnemers te identifiseer, blyk dit dat daar geen generiese SMS-taal onder die groep deelnemers bestaan nie.

Om die teenwoordigheid van ’n generiese SMS-taal te identifiseer, is die deelnemers gevra om elkeen dieselfde teks (deur die navorser voorsien) as SMS’e te tik en aan die navorser te stuur. Hierdie SMS’e is tot een teks gekombineer en deur middel van die Word list-funksie in Antconc ontleed. ’n Lys van al die woorde wat in ’n bepaalde teks verskyn, is sodoende gegenereer. Die woorde is op grond van spelling geïdentifiseer, wat beteken dat elke unieke spelling van ’n woord as ’n aparte woord in die lys verskyn. Uit die woordelys wat gegenereer is, is dit duidelik dat daar geen generiese SMS-taal onder die groep deelnemers bestaan nie (op grond van die variante van die woorde wat in die teks voorkom). Indien daar wel ’n generiese SMS-taal onder die groep teenwoordig was, sou dit beteken dat daar geen spellingvariasie by enige woorde kon voorkom nie, aangesien elke deelnemer dan elke afsonderlike woord op dieselfde wyse as die ander deelnemers sou spel.

2. Is dit moontlik om binne die veronderstelde generiese SMS-taal individuele, idiolektiese taal by SMS-gebruikers te identifiseer?

Tot ’n mate is dit wel moontlik om ’n idiolektiese SMS-taal onder die groep deelnemers te identifiseer. Uit beide die stilistiese en stilometriese ontledings is dit duidelik dat daar deelnemers is wie se skryfstyle met dié van die outeur van die verdagte teks sowel ooreenstem as daarvan verskil. Aangesien hierdie deelnemers van mekaar onderskei kan word, is dit moontlik om aan te voer dat die idiolektiese eienskappe van die deelnemers verskil. Alhoewel idiolektiese taalgebruik onder die groep deelnemers teenwoordig is, is die

hoeveelheid data egter te min om met enige hoë persentasie van sekerheid aan te voer dat een deelnemer se idiolek opvallend van ’n ander deelnemer se idiolek verskil.

3. Tot watter mate is dit moontlik om die outeur van ’n verdagte SMS-teks te identifiseer met die beperkte data wat tipies ter beskikking is?

Uit voorafgaande bespreking en op grond van die resultate wat uit die verskeie ontledings verkry is, is dit moontlik om die gevolgtrekking te maak dat die outeur van die verdagte SMS-teks nie met groot sekerheid bepaal kan word nie. Die hoofrede vir die onbesliste resultate is die beperkte hoeveelheid data wat tot die navorser se beskikking was. Indien die navorser genoeg data tot sy of haar beskikking het, kan hy of sy meer beslissende gevolgtrekkings maak oor die idiolektiese eienskappe van elke vermeende outeur.

In die huidige navorsing is die data voldoende om aan te dui

- dat daar nie ’n generiese SMS-taal onder die groep deelnemers bestaan nie en
- dat idiolektiese eienskappe gevolglik teenwoordig is.

Die data is egter te beperk om vas te stel tot watter mate die idiolekte van die deelnemers van mekaar verskil. Dit is moontlik dat byvoorbeeld D11 konstant as een van die moontlike outeurs in die navorsing geïdentifiseer word, maar dat wanneer meer data beskikbaar is, dit sal aandui dat D11 geen idiolektiese ooreenkomste met die verdagte teks se outeur toon nie. Die moontlikheid bestaan ook dat deelnemers wat in die huidige navorsing dieselfde persentasie-ooreenkoms met die verdagte teks deel, in werklikheid geen idiolektiese eienskappe met mekaar deel nie.

Dit blyk dat die navorsing suksesvol was om die navorsingsvrae van die huidige studie te beantwoord. Dit is egter nodig om die resultate verder, in ’n opvolgstudie, binne die groter konteks van die forensiese linguistiek te ondersoek.

12. Gevolgtrekking

Uit die huidige navorsing, asook navorsing oor outeuridentifikasie wat reeds gedoen is, is dit moontlik om verskeie gevolgtrekkings te maak oor die bruikbaarheid van die resultate wat in die huidige navorsing verkry is. Hierdie studie beklemtoon ook die groeipotensiaal vir die navorsingsveld van outeuridentifikasie in Suid-Afrika.

Die doelstelling van die huidige studie, naamlik om vas te stel of dit moontlik is om die outeur van ’n beperkte teks vas te stel, is bereik. Daar is vasgestel dat dit, op grond van die resultate wat uit beide die stilistiese en stilometriese ontledings verkry is, nie moontlik om met sekerheid die ware outeur van die verdagte SMS-teks te identifiseer nie.

Die beperkte data het die positiewe identifisering van die ware outeur van die verdagte teks beslis bemoeilik, maar die resultate wat verkry is (alhoewel negatief) kan steeds as bruikbaar beskou word. Resultate soos dié wat in die huidige navorsing verkry is, sou

waarskynlik wel as omstandigheidsgetuienis in ’n hof gebruik kon word, aangesien die ontledings dit wel moontlik gemaak het om die aantal waarskynlike outeurs van 13 tot 11 te verminder. Deur die resultate in elkeen van die ontledings met mekaar te vergelyk en die moontlike outeurs aan te toon, kan D1, D2, D4, D5, D6, D7, D8, D10, D11, D13 en D14 as die 11 moontlike outeurs van die verdagte teks geïdentifiseer word. Uiteraard maak dit nie ’n groot verskil in ’n werklike situasie nie, aangesien 11 verdagtes steeds ’n hoë getal is.

Dit is belangrik om daarop te let dat alhoewel één deelnemer nie as die moontlike outeur van die verdagte teks geïdentifiseer is nie, D2 (die ware outeur van die verdagte teks), in die meerderheid gevalle wel as een van die moontlike outeurs van Teks X geïdentifiseer is.

Die resultate dui daarop dat daar potensiaal is ten opsigte van suksesvolle outeuridentifikasie wanneer die forensiese linguis slegs kort tekste tot sy of haar beskikking het en daar verskeie outeurs is wat die ware outeur van die verdagte teks kan wees.

13. Probleme en beperkings

13.1 Beperkte data

Soos blyk uit navorsing wat reeds in outeuridentifikasie gedoen is, is dit duidelik dat daar verskeie kere scenario’s opduik waar min data tot die navorser se beskikking is en tot probleme kan lei met veral die statistiese toetse wat daarop toegepas word (Chaski 2001; Stamatatos e.a. 2001; Barry en Luna 2012). In die huidige navorsing is die beperkte hoeveelheid data ook problematies. Dit het daartoe gelei dat daar nie in die ontledings wat op die data uitgevoer is, ’n hoë mate van sekerheid verkry kan word oor die outeur van die verdagte teks nie.

Dit is nodig om reeds bestaande metodes (soos n-gramontledings en logwaarskynlikheidsverhoudings) aan te pas ten einde groter sukses in die ontleding van beperkte data te behaal. Dit is ook belangrik om databasisse op te stel sodat daar genoeg data is om vergelykende studies mee aan te pak en sodat die data in die databasis gebruik kan word in masjienleertegnieke (soos SVM en n-gramontledings). Ten spyte van die beperkte data wat tot die navorser se beskikking was, was dit nietemin moontlik om tot ’n mate idiolekte onder die huidige groep deelnemers te identifiseer.

13.2 Vaardighede en kennis van die forensiese linguis

Dit is belangrik om daarop te let dat die forensiese linguis nie in alle gevalle ’n forensieslinguistiese ontleding, wat van statistiek gebruik maak, alleen kan uitvoer nie. Forensiese linguiste is soms afhanklik van kenners op ander vakgebiede, soos byvoorbeeld statistiek en rekenaarwetenskap, aangesien ’n forensiese linguis self nie noodwendig ’n kenner van laasgenoemde vakgebiede is nie. In die huidige navorsing was dit byvoorbeeld sinvol om die hulp van ’n rekenaarwetenskaplike in te roep, aangesien ek nie self genoeg kennis dra van die vakgebied nie. Dit word egter aanbeveel dat forensiese linguiste die nodige kennis in statistiek en rekenaarwetenskap verwerf.

Bibliografie

Angouri, J. 2010. *Quantitative, qualitative or both? Combining methods in linguistic research*. In Litosseliti (red.) 2010.

Anthony, L. 2004. Antconc: A learner and classroom friendly, multi-platform corpus analysis toolkit. In *Proceedings of IWLeL 2004: An interactive workshop on language e-learning*.http://www.laurenceanthony.net/research/iwlel_2004_anthony_antconc.pdf (5 Augustus 2014 geraadpleeg).

—. 2005. Antconc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. *In 2005 IEEE International Professional Communication Conference Proceedings*. http://www.ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1494244 (11 Julie 2014 geraadpleeg).

—. 2014. Antconc: Readme. http://www.antlab.sci.waseda.ac.jp/software/antconc341/AntConc_readme.pdf (16 Junie 2014 geraadpleeg).

Barry, K. en K. Luna. 2012. Stylometry for online forums. http://cs229.stanford.edu/proj2012/BarryLuna-StylometryforOnlineForums.pdf (18 Julie 2013 geraadpleeg).

Blackwell, S. 2012. History of forensic linguistics.http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0508/full (16 Januarie 2013 geraadpleeg).

Brennan, M., S. Afroz en R. Greenstadt. 2012. Adversarial stylometry: circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security,* 15(3):1–22.

Broeders, A.P.A. 2001. Forensic speech and audio analysis forensic linguistics. A review. 13th Interpol Forensic Science Symposium. 16–19 Oktober. Frankryk: Lyon. http://www.taracentar.hr/attachments/interpol_forensic.pdf (16 Januarie 2013 geraadpleeg).

Camilli, G. en K.D. Hopkins. 1978. Applicability of chi-square to 2 * 2 contingency tables with small expected frequencies. *Psychological Bulletin*, 85:163–7.

—. 1979. Testing for association in 2 * 2 contingency tables with very small sample sizes. *Psychological Bulletin*, 86:1011–4.

Carstens, W.A.M. 2003. *Norme vir Afrikaans: enkele riglyne by die gebruik van Afrikaans*. Pretoria: Van Schaik Uitgewers.

Cavnar, W.B. en J.M. Trenkle. 1994. N-gram-based text categorization. odur.let.rug.nl/vannoord/TextCat/textcat.pdf (15 September 2014 geraadpleeg).

Chaski, C.E. 2001. Empirical evaluations of language-based author identification techniques. www.iula.opf.edu/materials/050520spassova.pdf (1 Augustus 2014 geraadpleeg).

—. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1):1–14. https://www.utica.edu/academic/institutes/ecii/publications/articles/B49F9C4A-0362-765C-6A235CB8ABDFACFF.pdf (5 November 2014 geraadpleeg).

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press. http://babel.ucsc.edu/~hank/aspects_ch3.pdf (10 Oktober 2013 geraadpleeg).

Coulthard, M. 2004. Author identification, idiolect and linguistic uniqueness. *Applied Linguistics,* 25:431–47.

Coulthard, M. en A. Johnson (reds.). 2010. *The Routledge handbook of forensic linguistics*. New York: Routledge.

Crankshaw, R. 2012. The validity of the linguistic fingerprint in forensic investigation. *Diffusion*, 5(2). atp.uclan.ac.uk/buddypress/diffusion/?p=1228 (22 April 2013 geraadpleeg).

Crystal, D. 2008. *Txtng: The gr8 db8*. Oxford: Oxford University Press.

Dörnyei, Z. 2007. *Research methods in applied linguistics*. Oxford: Oxford University Press.

Easton, V.J. en J.H. McColl. s.j. Statistics glossary (v1.1). http://www.stats.gla.ac.uk/steps/glossary/categorical_data.html (18 Julie 2014 geraadpleeg).

Feinberg, S. E. 1980. *The analysis of cross-classified categorical data*. Cambridge: MIT.

Gangs use of the internet and cell phones. I look both ways. ilookbothways.com/2010/06/14/gangs-use-of-the-internet-and-cell-phones (8 Oktober 2014 geraadpleeg).

Gavaldà-Ferré, N. 2012. The study of inter- and intra-speaker variation towards an index of idiolectal similitude. In *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*, Aston Universiteit, Birmingham.

Gerber, H. 2014. Persoonlike kommunikasie, April.

Goehring, H.J. 1981. *Statistical methods in education*. Virginia: Information Resources Press.

Grant, T. 2010. Txt 4n6: Idiolect free authorship analysis? In Coulthard en Johnson (reds.) 2010.

Halliday, M.A.K. 1975. *Learning how to mean: explorations in the development of language*. Londen: Edward Arnold.

Heigman, J. en R.A. Crocker (reds.). 2009. *Qualitative research in applied linguistics*. Londen: Palgrave Macmillan.

How2stats. 2011. Yates' correction. www.how2stats.net/2011/09/yates-correction.html (13 Februarie 2015 geraadpleeg).

Ishihara, S. 2011. A forensic authorship classification in SMS messages: A likelihood ratio based approach using N-gram. In *Proceedings of Australasian Language Technology Association Workshop*.

Jackson, A.R.W en J.M. Jackson. 2004. *Forensic science*. Londen: Pearson Education Limited.

Juola, P. 2006. Authorship attribution. *Foundations and trends in information retrieval,* 1(3):233-334.

Kotzé, E.F. 2007. Die vangnet van die woord: forensies-linguistiese getuienis in 'n lastersaak. *South African Linguistics and Applied Language Studies*, 25(3):385–99.

—. 2010. Author identification from opposing perspectives in forensic linguistics. *South African Linguistics and Applied Language Studies,* 28(2):185–97.

Larntz, K. 1978. Small sample comparisons of exact levels for chi-square goodness of fit statistics. *Journal of the American Statistical Association*, 73: 253-263.

Lazaraton A. 2009. Discourse analysis. In Heigman en Crocker (reds.) 2009.

Litosseliti, L. (red.). 2010. *Research methods in linguistics*. Londen: Continuum International Publishing Group.

McLeod, N. en T. Grant. 2012. Whose tweet? Authorship analysis of micro-blogs and other short-form messages. In *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*. Birmingham: Aston Universiteit.

McMenamin, G.R. 2010. Theory and practice of forensic stylistics. In Coulthard en Johnson (reds.) 2010.

MedicineNet.com. 2014. Definition of forensic. www.medicinenet.com/script/main/mobileart.asp?articlekey=10604 (9 Oktober 2014 geraadpleeg).

Mendenhall, T.C. 1887. The characteristic curves of composition. *Science*, 9(214):237–49. http://www.jstor.org/stable/1764604 (3 Junie 2014 geraadpleeg).

Michell, C.S. 2013. Investigating the use of forensic stylistics and stylometric techniques in the analysis of authorship on a publicly accessible social networking site (Facebook). Ongepubliseerde Meestersgraadverhandeling, Universiteit van Suid-Afrika.

Mikros, G.K. s.j. Authorship attribution in Greek blogs. http://users.uoa.gr/~gmikros/Pdf/ AA%20and%20GI%20in%20Greek%20blogs_Qualico12.pdf (7 Junie 2013 geraadpleeg).

Mobile Pronto. 2010. *The history of SMS text messaging*. http://www.mobilepronto.org/en-us/the-history-of-sms-html (29 Januarie 2013 geraadpleeg).

Mohan, A., I.M. Baggili en M.K. Rogers. 2010. Authorship attribution of SMS messages using an N-grams approach. http://www.cerias.purdue.edu/assets/pdf/ bibtex_archive/2010-11-report.pdf (9 Junie 2014 geraadpleeg).

Olivier, J. 2013. *Die mate van konsekwentheid in SMS-Afrikaans*. *LitNet Akademies*, 10(2). http://www.litnet.co.za/Article/die-mate-van-konsekwentheid-in-sms-afrikaans (1 Augustus 2014 geraadpleeg).

Olsson, J. 2004. *Forensic linguistics: an introduction to language, crime and the law*. Londen en New York: Continuum.

—. s.j. Forensic linguistics. Proefhoofstuk. www.eolss.net/Sample-Chapters/CO4/E6-91-13.pdf (17 Januarie 2013 geraadpleeg).

PewInternet. 2012. Cell phone activity 2012. http://www.pewinternet.org/Reports/ 2012/Cell-Activities/Additional-Demographic-Analysis/Demographics.aspx (15 Maart 2013 geraadpleeg).

Schulstad, I., M. Boga., C. Jordan en K. Pally. 2012. Evaluation of a stylometry system on various length portions of books. http://www.csis.pace.edu/~ctappert/srd2012/d5.pdf (23 Mei 2014 geraadpleeg).

Somers, H. s.j. Stylometry and authorship. [Powerpoint]. Universiteit van Manchester: Skool van Rekenaarwetenskap.http://personalpages.manchester.ac.uk/staff/harold.somers/LELA30922/Authorship.ppt (17 April 2013 geraadpleeg).

Stamatatos, E., N. Fakotakis en G. Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35:193-214.

Statistics How To. 2015. Yates correction: What is it used for in statistics. http://www.statisticshowto.com/what-is-the-yates-correction (13 Februarie 2015 geraadpleeg).

Svartvik, J. 1968a. The Evans statements: A case for forensic linguistics. Deel 1 van 2.http://www.thetext.co.uk/Evans%20Statements%20Part%201.pdf (13 Junie 2014 geraadpleeg).

—. 1968b. The Evans statements: A case for forensic linguistics. Deel 2 van 2. http://www.thetext.co.uk/Evans%20Statements%20Part%202.pdf (13 Junie 2014 geraadpleeg).

Thiart, L. 2014. Outeuridentifikasie: ’n Forensies-taalkundige ondersoek na Afrikaanse SMS-taal. Ongepubliseerde MA-verhandeling, Universiteit van Pretoria.

Thompson, B. 1988. Misuse of chi-square contingency-table test statistics. *Educational and Psychological Research*, 8(1): 39-49.

Wolff, F. 2014. Persoonlike kommunikasie, 18 September.

Eindnotas

[1] *Waargenome frekwensies* (*observed frequencies* – O) verwys na die data wat die navorser ingesamel of waargeneem het.

[2] *Verwagte frekwensies* (*expected frequencies* – E) verwys na die frekwensies wat die navorser sal voorspel in elke sel van die tabel (Easton en McColl s.j.).