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

deur

Lezandra Thiart

‘n Verhandeling voorgelê ter vervulling van die vereistes vir die graad

MA in Linguistiek

in die Departement Afrikaans van die

UNIVERSITEIT VAN PRETORIA

FAKULTEIT GEESTESWETENSKAPPE

STUDIELEIER: Dr. Nerina Bosman

November 2014

Plagiaatverklaring:

Hiermee verklaar ek (Lezandra Thiart) dat die verhandeling: "*Outeuridentifikasie: 'n Forensies-taalkundige ondersoek na Afrikaanse SMS-taal*" my eie werk is. Waar die idees of navorsing van ander individue gebruik is, is die nodige verwysings in die teks bygevoeg en ook by die bibliografie ingesluit.

Lezandra Thiart (28103522)

23/02/2015

Bedankings

Ek bedank graag die volgende individue vir hul hulp en ondersteuning tydens die voltooiing van my meestersgraad.

- Doktor Nerina Bosman sonder wie se hulp en geduld ek nie die meestersgraad sou kon voltooi nie.

- Meneer Hennie Gerber en Meneer Friedel Wolff vir hulle hulp met die statistiese analises en interpretasies.

- Die eksterne nasieners vir hulle waardevolle kommentaar.

INHOUDSOPGAWE

Hoofstuk 1: Agtergrond en probleemstelling……..1

1.1 Inleiding……1

1.2 Agtergrond……2

1.3 Motivering en probleemstelling……4

1.4 Doelstelling.....6

1.5 Navorsingsvrae…….....6

1.6 Navorsingsontwerp…….....8

1.7 Navorsingsmetodologie.....9

1.8 Kernkonsepte……...13

1.9 Uiteensetting van die verhandeling……..17

Hoofstuk 2: Die wye spektrum van die forensiese linguistiek: 'n oorsig.....18

2.1 Forensiese linguistiek: definiëring en afbakening van die vakgebied.....18

2.2 Forensiese linguistiek in Suid-Afrika: 'n opkomende ondersoekterrein.....30

2.3 Outeuridentifikasie.....32

2.4 Outeuridentifikasie en die hof.....46

2.5 Stilometrie.....58

2.6 Idiolek.....67

2.7 Die SMS-kultuur.....75

Hoofstuk 3: Metodologie.....84

3.1 Inleiding.....84

3.2 Dataversameling.....86

3.3 Instrumente.....91

3.4 Analitiese metodes.....99

Hoofstuk 4: Analise en resultate.....110

4.1 Stilistiese analise.....110

4.2 Stilometriese analise.....112

4.3 Generiese taalgebruik in die SMS-korpus.....129

4.4 Opsomming van die resultate.....133

Hoofstuk 5: Gevolgtrekkings.....138

5.1 Doelstelling en navorsingsvrae.....138

5.2 Bruikbaarheid van resultate in die hof.....139

5.3 Die vestiging van forensiese linguistiek as selfstandige vakgebied in Suid-Afrika.....140

5.4 Probleme en beperkings.....141

Bibliografie.....143

Bylae.....153

Hoofstuk 1: Agtergrond en probleemstelling

1.1 Inleiding

Forensiese taalkunde (of 'forensiese linguistiek' soos dit ook bekend staan) is 'n relatiewe jong studierigting binne die toegepaste taalkunde. Dit is 'n rigting wat interdissiplinêre raakpunte tussen taal, misdaad en die reg toon. Dit is in die eerste plek 'n onderafdeling van die breë linguistiek en oorvleuel daarom met ander velde in die linguistiek. Aspekte soos stemherkenning en aksentidentifikasie, wat binne die studieterrein van die fonetiek val, het byvoorbeeld ook bepaalde forensies taalkundige implikasies. Die betekenis van woorde binne 'n bepaalde konteks word ook in forensiese linguistiek ondersoek en kan binne die studieterrein van die semantiek geklassifiseer word.

Laasgenoemde twee voorbeelde van forensiese linguistiek word vandag wyd beoefen. In een onlangse hofsaak in Amerika is die ontleding van 'n telefoonoproep en stemherkenning as deel van die omstandigheidsgetuienis gebruik om vas te stel of die beskuldigde se weergawe van gebeure die waarheid was. In Maart 2012 is George Zimmerman aangekla van die moord op Trayvon Martin. Daar is aangevoer dat Zimmerman Martin doodgeskiet het bloot omdat hy Martin as 'n verdagte persoon beskou het. Zimmerman het egter volgehou dat hy sy geweer afgevuur het om homself te verdedig nadat Martin hom aangeval het. Tydens hierdie insident was Zimmerman besig om met die 911-skakelbordoperateur op sy selfoon te praat. CNN, met behulp van die FBI, het die opnames (deur 911 se oproepsentrum voorsien) geanaliseer en probeer vasstel of dit Zimmerman is wat 'n rassistiese term, verwysende na Martin, gebruik. Hulle het ook probeer vasstel wie die persoon is wat vir 18 sekondes lank tydens die oproep skree. Die resultate van die analises was egter onbeslis. 'n Kenner in die veld van oudioanalises, wat deur die staat aangestel is, kon wel bepaal dat dit Martin is wat in die agtergrond van die oproep praat en onder andere die woorde "I'm begging you" sê (State experts differ on voice analysis, 2013). Zimmerman is onskuldig bevind.

In Suid-Afrika staan forensiese linguistiek nog in sy kinderskoene en forensiese analises soos dié wat by die bogenoemde hofsaak gebruik is, asook forensiese analises wat op outeuridentifikasie

fokus, is baie skaars. Daar is ook geen navorsing oor spesifiek outeuridentifikasie in Afrikaans beskikbaar nie. In hierdie ondersoek staan die uitlig van aspekte wat verband hou met outeuridentifikasie, en dan spesifiek outeuridentifikasie in Afrikaanse SMS-taal, sentraal.

Onderafdelings van die forensiese taalkunde, soos outeuridentifikasie, sprekeridentifikasie asook die taal van regstekste en regsprosesse word reeds vanaf die 1990's in verskeie lande (onder andere in die VSA, Brittanje, Italië en Nederland) bestudeer en die studieveld en beroepsveld bly groei (Broeders, 2001: 64,69; Blackwell, 2012: 1–4). Verskeie buitelandse universiteite bied reeds nagraadse kursusse in forensiese linguistiek aan. Aston Universiteit in Birmingham, Hofstra Universiteit in New York en die Universiteit van Barcelona bied meestersgrade in forensiese linguistiek aan, terwyl Cardiff Universiteit die opsie van 'n meestersgraad of 'n diplomakursus in forensiese linguistiek aanbied. In Suid-Afrika is daar slegs een nagraadse kursus in forensiese linguistiek wat by Noordwes Universiteit aangebied word.

1.2 Agtergrond

Outeuridentifikasie is een van die hoofvertakkings wat binne forensiese taalkunde onderskei word. Aanvanklik is outeuridentifikasiemetodes slegs in die literêre konteks toegepas om die outeurskap van sekere saamgestelde dele van die Ou en Nuwe Testament in die Bybel, asook van sekere werke van Shakespeare, waarvan die outeurskap onder verdenking is, vas te stel (Hockey, s.a.). In die laaste paar dekades het outeuridentifikasie gegroei tot 'n interdissiplinêre veld wat van toepassing kan wees in die letterkunde, onderrig (in verband met plagiaatidentifikasie), nasionale en plaaslike intelligensie en vanselfsprekend ook in die regswese en –praktyk. Stelselmatig het die fokus van outeuridentifikasie verskuif vanaf die analise van handskrif en grafiese eienskappe van tekste na die linguistiese inhoud van tekste wat juridies van belang is. Sulke tekste sal byvoorbeeld selfmoordbriewe, tekste wat moontlik plagiaat bevat, afpersingsbriewe en skuldbekentenisse insluit (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 aanlynforums, blogs, tekste op sosiale netwerke en SMS-boodskappe te identifiseer (Mikros, s.a.; Mohan e.a., 2010; Ishihara, 2011; McLeod en Grant, 2012; Michell, 2013).

Outeuridentifikasie in SMS-boodskappe is een van die ondersoekvelde wat veral in die buiteland aandag ontvang, 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-boodskappe en ander vorme van vinnige kommunikasie word ook al hoe meer gebruik in ontvoerings, om kubermisdade te pleeg, en om met dwelms of wapens te smokkel. SMS-boodskappe word ook gebruik om sulke misdade te probeer verdoesel (Gangs use of the internet and cell phones, 2010; Blackwell, 2012: 5; Crystal, 2008: 60-61; Grant, 2010: 508; Ishihara, 2011: 47).

Crystal (2008) en Grant (2010) beskryf enkele gevalle waar outeuridentifikasie in SMS-boodskappe gebruik is om misdadigers vas te trek. Beide Crystal en Grant verwys na die saak waarin Stuart Campbell in 2002 skuldig bevind is aan die moord op sy vyftienjarige niggie nadat forensiese linguiste vasgestel het dat die SMS-boodskap wat Campbell as alibi gebruik het, 'n vervalsing was. Campbell het die polisie probeer oortuig dat die boodskap deur sy niggie gestuur is, wat sou beteken dat hy onskuldig was. Tog kon die polisie deur middel van vergelykingsverskille wat verband hou met stilistiese eienskappe van die taalgebruik van Campbell en sy niggie, asook ander omstandigheidsgetuienis, vasstel dat Campbell self die SMS'e getik het. Grant verwys ook na 'n onverwante saak waarin David Hodgson op 19 Februarie 2008 skuldig bevind is aan die moord op sy vriendin, Jenny Nicholl. Deur middel van linguistiese analise is bepaal dat Hodgson, eerder as Nicholl, die outeur was van die laaste SMS-boodskappe wat vanaf Nicholl se selfoon aan vriende gestuur is. Blackwell (2012) noem die saak teen Peter Chapman (Maart 2010). Chapman is aangekla van die moord op Ashleigh Hall wat hy op Facebook bevriend het. In hierdie saak is SMS-boodskappe van beide Chapman en Hall se selfone geanaliseer en hierdie analises het bygedra tot Chapman se skuldigbevinding.

Bogenoemde gevalle versterk die argument dat outeuridentifikasie in SMS-boodskappe 'n belangrike bydrae kan lewer in ondersoeke waar SMS'e gebruik is om misdaad of geweld te verdoesel. Ten spyte van die voor die hand liggende voordele wat outeuridentifikasie in SMS-boodskappe inhou, is daar, ook in die buiteland, steeds onvoldoende navorsing in hierdie veld.

Ishihara (2011: 48) merk op "[...] studies specifically focusing on the authorship of SMS messages in forensic contexts are conspicuously sparse".

Een van die redes waarom daar min navorsing oor outeuridentifikasie in SMS-boodskappe beskikbaar is, is die feit dat 'n SMS-boodskap 'n kort teks is met beperkte inhoud. Vergelyk byvoorbeeld vollengte dokumente soos dreigbriewe wat soms uit twee of meer getikte bladsye bestaan (dit wil sê ongeveer 600 tot 700 woorde). Forensiese taalkundiges kategoriseer SMS-boodskappe, *updates* op sosiale netwerke asook selfmoordbriewe en lospryseise in dieselfde kategorie. Hierdie tekste deel almal 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 byvoorbeeld ondersoek word. Ten spyte van die problematiese aard van korter tekste (vanuit 'n outeuridentifikasie-standpunt beskou) het navorsing reeds aangetoon dat dit nie onmoontlik is om die outeur van korter tekste te identifiseer nie (McLeod en Grant, 2012; Ishihara, 2011).

'n Tweede probleem met SMS-taal, wat ook verband hou met die bondige aard van die tekste, duik op wanneer die linguis die persoonlike styl en idiolek van die outeur moet identifiseer. Soos later uit die verdere bespreking van idiolek sal blyk (afdeling 2.6), is dit nie eenvoudig om idiolek of persoonlike skryfstyl 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 geanaliseer 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.

1.3 Motivering en probleemstelling

Forensiese taalkunde is steeds 'n relatief onbekende ondersoekterrein in Suid-Afrika en daar is slegs enkele individue wat in Suid-Afrika in hierdie veld navorsing gedoen en gewerk het of steeds werk: Hubbard (1994, 1995); Moeketsi (1997); Kotzé (2007, 2010); Klopper (2009); Lombard en Carney (2011); Michell (2013) en Carney (2012, 2013, 2014).

Alhoewel daar internasionaal reeds getuienis is van navorsing oor outeuridentifikasie in SMS-boodskappe, is dit baie min as mens dit vergelyk met werk wat gedoen is in verband met outeuridentifikasie van ander, langer tekste. Daar is vandag sowat 42.3 miljoen mense in Suid-Afrika wat toegang het tot selfone (allAfrica, 2012). Daar is tans geen statistiek beskikbaar om aan te dui hoeveel misdade per jaar in Suid-Afrika deur middel van selfone gepleeg word nie, maar daar kan aangeneem word, op grond van die hoeveelheid selfoongebruikers, dat misdade wat gepleeg word met selfone beslis in Suid-Afrika voorkom. In Amerika word beraam dat 80% van die misdade wat gepleeg word verband hou met bendes in die omgewing. Selfone is een van die populêrste toestelle wat deur bendes gebruik word om misdade te pleeg. Volgens 'Gangs use of the Internet and cell phones' (2010) word selfone deur bendes gebruik tydens rooftogte, om transaksies wat die verkoop van dwelms insluit by te staan en vir afpersing. Buck (2012) wys daarop dat kubermisdade, wat identiteitdiefstal en bedrog insluit, al hoe meer deur middel van selfone gepleeg word. In 2011 is daar na beraming sowat 314 000 klagtes van kubermisdade in Amerika aangemeld (Buck, 2012). Vooruitgang op die gebied van outeuridentifikasie van veral 'populêre' vorme van kommunikasie in die sosiale media (en ook spesifiek wat SMS-boodskappe betref) sou ook in Suid-Afrika gebruik kon word om verdagtes van misdade waar selfone gebruik word, vas te trek.

Die kernvrae ten opsigte van outeuridentifikasie in SMS-boodskappe is nog nie binne die Afrikaanse konteks beantwoord nie. In navorsing wat onder andere op Engelse en Singapoerese SMS'e fokus is reeds vasgestel dat die outeurs van die boodskappe met 'n redelike hoë mate van sekerheid vasgestel kan word (Mohan, Baggili en Rogers, 2010; Ishihara, 2011). Daar moet egter in gedagte gehou word dat laasgenoemde studies van groter korpusse gebruik gemaak het as wat in die huidige studie gebruik is.

In ag genome die feit dat daar nog geen navorsing oor spesifiek outeuridentifikasie in Afrikaans gedoen is nie, is sulke navorsing dus van groot belang. Forensies taalkundige navorsing hou ook voordele in vir die breër Suid-Afrikaanse samelewing omdat die metodes wat gevolg word op al die landstale toegepas sou kon word. Daar is reeds verskeie navorsingsmoontlikhede in die veld geskep en verdere navorsing in forensiese taalkunde in Suid-Afrika en in Afrikaans is nou nodig om onsself in die internasionale gemeenskap van forensiese taalkunde te vestig, in ag genome die spoed waarteen dié veld internasionaal groei. Ter ondersteuning van laasgenoemde stelling

geld die opmerking van McMenamin: “The discipline and science of forensic linguistics will not develop the way it should from ‘outside’ study, commentary, and observation” (McMenamin, 2002: 11). McMenamin bedoel hiermee dat forensiese linguiste hulleself moet verdiep in die probleme wat opduik in werklike forensies linguistiese sake. Op grond van sulke probleme moet forensiese linguiste dan hulle metodes en perspektiewe ontwikkel.

In hierdie navorsing is ’n scenario geskep wat problematies is vir die linguis. Die beperkte data wat in die navorsing gebruik is, is nie ideaal nie, maar dit skep die nodige realistiese scenario vir ’n navorser om te bepaal wat wel met die ontleding van beperkte data gedoen kan word en of dit enigsins van nut kan wees in ’n outeuridentifikasie-ondersoek.

’n Laaste motivering vir die studie is dat die klein SMS-korpus wat in die ondersoek saamgestel is, as hulpbron kan dien vir toekomstige navorsing wat met SMS-taal in Afrikaans verband hou.

1.4 Doelstelling

Die doelstelling van die huidige navorsing is 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. ’n Beperkte stel SMS’e is gekies omdat die navorsing poog om outeuridentifikasie binne ’n beperkte korpus, bestaande uit bondige elektroniese tekste, te ondersoek. Dit is nodig om outeuridentifikasieanalises onder minder ideale omstandighede (waar daar verskeie moontlike outeurs is en ’n beperkte hoeveelheid data) te ondersoek aangesien meer en meer individue van elektroniese kommunikasie gebruik maak wat in die meeste gevalle die produksie van slegs bondige tekste moontlik maak.

1.5 Navorsingsvrae

As gevolg van die beperkte omvang van ’n magisterstudie is besluit om op slegs drie navorsingsvrae te fokus wat met behulp van data-insameling en -analise beantwoord moet word. Die navorsingsvrae is geformuleer uit die doelstelling van die studie. Die antwoorde op die navorsingsvrae sal gevolglik lei tot die bereiking, of nie, van die doelstelling.

1.5.1 Kan daar in die data ’n generiese SMS-taal geïdentifiseer word wat outeuridentifikasie sou bemoeilik?

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

1.5.3 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?

Hierdie drie navorsingsvrae is sentraal tot die ondersoek, aangesien dit verband hou met twee belangrike aspekte van outeuridentifikasie, naamlik die insameling van data om as korpus te dien en die vasstel van idiolek wat as ‘stylmerker’ dien vir individuele outeurs om sodoende die identifiseringsproses te vergemaklik. Die konsep van idiolek hou verband met al drie die navorsingsvrae. Die aanwesigheid van ’n generiese SMS-taal sal beteken dat dit moeiliker sal wees om idiolektiese eienskappe in die SMS-boodskappe te identifiseer. ‘Generiese taal’ is natuurlik nie ’n absolute konsep nie, maar ’n relatiewe en buigsame begrip. Alle tekste bevat waarskynlik tot ’n meerdere of mindere mate idiolektiese eienskappe, ook sogenaamde generiese tekste. Die term sal dus in hierdie verhandeling met die nodige omsigtigheid hanteer word. Vergelyk afdeling 1.8.5.

Indien geen generiese taalgebruik aangetoon kan word nie, mag dit beteken dat daar wel genoeg idiolektiese taaleienskappe onder die groep deelnemers bestaan om die deelnemers, tot ’n mate, van mekaar te onderskei . Die derde navorsingsvraag hou verband met die mate waartoe die aanwesigheid van idiolek in die beperkte korpus enigsins tot die positiewe identifisering van die outeur van ’n bepaalde teks kan lei.

1.6 Navorsingsontwerp

Die navorsingsontwerp vir hierdie studie kan as as ’n gemengde metode gedefinieer word omdat dit van kwalitatiewe (stilistiese) en kwantitatiewe (stilometriese) metodes gebruik maak. Om die aard van die gemengde metode te begryp, word daar kortliks na definisies verwys.

Croker (2009: 9) beskryf kwalitatiewe navorsing as volg:

> This research is not necessarily done to predict what may happen in the future or in another setting – what is learned about the phenomenon, participants, or events in the setting can be an end in itself. That is, qualitative research mostly focuses on understanding the particular and the distinctive, and does not necessarily seek or claim to generalize findings to other contexts.

Ivankova en Creswell (2009: 137) verwys na Denzin en Lincoln (2005) wat meen dat kwalitatiewe navorsing die versameling van woorde eerder as syfers (kwantitatiewe navorsing) is en dat die navorser data sonder vooroordeel behoort te analiseer op soek na patrone of temas. Waar dit moeilik of onmoontlik is om vooroordeel heeltemal uit te skakel, behoort navorsers nietemin ’n gepaste kritiese ingesteldheid in hul analises te vertoon. Stilistiese analises van tekste, waar die forensiese linguis self patrone in die teks, asook veronderstelde idiolektiese taalgebruik probeer identifiseer, is ’n voorbeeld van kwalitatiewe navorsing waar vooroordeel nie enige rol behoort te speel nie. In hoofstuk 3 word bespreek hoe die stilistiese analise op die data in hierdie studie uitgevoer is.

Die kwantitatiewe kant van hierdie navorsing hou verband met stilometriese analises waar statistiek gebruik word om sekere gevolgtrekkings te maak. Ivankova en Creswell (2009: 137) beskryf kwantitatiewe navorsing as die insameling van “numeric data, for example, proficiency test scores or multiple choice question [...] responses on questionnaires”. Navorsers probeer dan om die data objektief te analiseer deur ’n verskeidenheid statistiese tegnieke te gebruik. Die resultate lei daartoe dat ’n hipotese aanvaar of verwerp word en daarna kan die resultate gebruik word om veralgemenings te maak ten opsigte van ’n groter bevolking. Die mate waartoe hierdie ondersoek van stilometriese analises en statistiese ontledings gebruik gaan maak om te spekuleer oor gedeelde outeurskap, kan dit dus ook as kwantitatief tipeer.

McMenamin (2002) beskryf kwalitatiewe analises met betrekking tot outeuridentifikasie as die identifisering van kenmerke in ’n teks en die daaropvolgende beskrywing van hierdie kenmerke.

Die forensiese linguis wil met ander woorde vasstel of sekere kenmerke eie is tot die styl van ’n bepaalde skrywer. Kwalitatiewe analise hou in dat hierdie kenmerke getel moet word. Die frekwensie van die kenmerke moet met ander woorde vasgestel word. McMenamin (2002: 76) beskou ’n gemengde metode as voordelig aangesien kwalitatiewe en kwantitatiewe metodes mekaar komplementeer.

Michell (2013) gebruik ’n gemengde metode in sy navorsing en verwys ook na McMenamin (2002) se siening dat die gemengde metode voordelig is vir studies in forensiese linguistiek.

Uit bogenoemde definisies en die verwysings na hoe beide kwalitatiewe en kwantitatiewe navorsing met die huidige studie verband hou en deur ander navorsers ondersteun word, is die rede waarom die gemengde metode gebruik word, duidelik. Die gebruik van meer as een metode maak dit moontlik om ’n navorsingsprobleem meer volledig te ondersoek en om, veral in die geval van die forensiese linguistiek, meer betroubare resultate te produseer.

1.7 Navorsingsmetodologie

Om die data wat deur hierdie navorsing ingesamel is, ten volle te benut, moet ’n deeglike oorsig van die studieterrein gedoen word, die data verkry moet geanaliseer en geïnterpreteer word en die bevindings moet in detail beskryf word. Die uiteindelike beskrywing van die bevindings gee die leser die geleentheid om die data self te interpreteer en te besluit tot watter mate hierdie ontleding ook op ander situasies van toepassing kan wees.

Die navorsingsmetodologie vir hierdie navorsing is tweedelig. Eerstens is ’n breedvoerige literatuuroorsig van die studieterrein onderneem. Die literatuurverkenning is aangevul met ’n oorspronklike empiriese ondersoek.

1.7.1 Literatuurondersoek

Terreinverkenning is noodsaaklik in ’n magisterstudie, aangesien ondersoekers moet kan aantoon dat hulle kennis dra van die ontwikkeling van die vakgebied, van die belangrikste konsepte en begrippe, teoretiese invalshoeke, baanbrekernavorsers en grensverskuiwende navorsing. In hoofstuk 2 van die verhandeling word die resultaat van ’n deeglike literatuuroorsig aangebied.

Die rede vir die literatuuroorsig in hierdie studie is om uit te lig watter navorsing reeds in die veld van forensiese taalkunde en spesifiek outeuridentifikasie gedoen is. Die literatuuroorsig maak dit ook moontlik om aan te dui watter leemtes daar tans in die navorsing, veral in Suid-Afrika en met name in Afrikaans, oor outeuridentifikasie in SMS-boodskappe bestaan. Die literatuuroorsig is verder van waarde omdat sekere afdelings van die navorsingsmetodologie gegrond word op navorsing waaroor bestaande literatuur in die veld al berig het.

1.7.2 Empiriese ondersoek

Empiriese navorsing het ten doel die insameling van primêre data wat deur die ondersoeker gebruik word om die navorsingsvrae te beantwoord en gevolgtrekkings te maak (Dörnyei, 2007: 16). Empiriese navorsing begin by die stel van navorsingsvrae wat deur die loop van die navorsing beantwoord moet word (Litosseliti, 2010: 10). Die data-insamelingmetodes asook die ontledingsmetodes van die huidige navorsing word kortliks in paragraaf 1.7.3 uiteengesit.

1.7.3 Dataversameling

1.7.3 (a) Databasis of korpus

Binne die veld van forensiese taalkunde, en meer spesifiek outeuridentifikasie, is daar twee maniere om data te stoor. Die eerste en bekendste metode is om ’n korpus van die data saam te stel en dan die woorde in die korpus te gebruik vir die data-analises. Een van die kenmerke van ’n korpus is die groot hoeveelhede teks wat gewoonlik gebruik word om die korpus saam te stel. Baker (2010: 6) definieer ’n korpus as ’n “body of language, or more specifically, a (usually) very large collection of naturally occuring language, stored on computer files”. ’n Korpus hoef egter nie uit enorme hoeveelhede teks te bestaan nie, maar gevolgtrekkings wat op grond van analises uit die korpus gemaak word, is meer betroubaar wanneer die korpus verteenwoordigend is van die taalgebruik wat ondersoek word. Die algemene opvatting is dat groter korpusse meer verteenwoordigend is van die spesifieke taalgebruik wat ondersoek word. Die alternatief is om ’n databasis op te stel. ’n Databasis kan makliker vir kleiner studies met ’n spesifieke doel gebruik word, maar is meer beperkend as ’n korpus. Campbell (2005), in Rusko en Garabík (s.a.: 2), beskryf die verskil tussen ’n databasis en ’n korpus soos volg:

> A "database" is an organized collection of information, typically designed for ease of retrieval by computerized methods; a "corpus", on the other hand, is a collection of naturally-occuring spoken or written material in machine-readable form, that are in themselves more-or-less representative of a language for the systematic study of authentic examples of language in use.

Vir die doeleindes van hierdie studie is 'n klein korpus saamgestel sodat ondersoek kon word of daar 'n generiese Afrikaanse SMS-taal en ook as idiolektiese taalgebruik bestaan, al dan nie. Meer detail oor die korpus word in hoofstuk 3 gegee.

Studies wat tot op hede oor outeuridentifikasie van SMS-boodskappe en ander kort elektroniese boodskappe gedoen is, het gebruik gemaak van aansienlik groter korpusse as wat in die huidige studie gebruik is. Die huidige navorsing gebruik egter die beperkte kleiner korpus om twee redes. Eerstens is daar nie 'n korpus van Afrikaanse SMS'e beskikbaar waaruit data gebruik kan word nie. Tweedens is die korpus van die huidige studie baie meer realisties in terme van werklike forensiese ondersoeke. Forensiese linguiste het selde indien ooit toegang tot 'n groot korpus wanneer ondersoeke uitgevoer word:

> Unfortunately, in the real world, we often encounter situations in which our list of candidates might be very large and in which there is no guarantee that the true author of an anonymous text is even among the candidates. Furthermore, the amount of writing we have by each candidate might be very limited, and the anonymous text itself might be short (Koppel e.a., 2012: 284).

Kort boodskaptekste wat deur middel van boodskapdienste soos Whatsapp en BBM gestuur word, wat 'nuwer' en 'goedkoper' maniere is om 'n SMS te stuur, word ook by die studie ingesluit. In hierdie navorsing word die term *SMS* gebruik om te verwys na enige *short message service* wat dit vir individue moontlik maak om vinnig elektroniese boodskappe te stuur wat nie e-posboodskappe, *updates* op sosiale netwerke of bloginskrywings is nie.

1.7.3 (b) Deelnemers

Heigham en Croker (2009: 149) meen dat die neem van steekproewe by kwantitatiewe data verskil van dié by kwalitatiewe data. By kwantitatiewe data is die steekproef gewoonlik baie groot en dit word lukraak uitgevoer, terwyl die neem van steekproewe by kwalitatiewe navorsing baie kleiner en meer doelgerig is omdat die steekproef moet lei tot "an in-depth understanding of the explored phenomenon". Die fokus in die huidige studie val op kwalitatiewe ontleding, maar dit word aangevul deur kwantitatiewe metodes.

Dit is belangrik om die omstandighede waaronder die steekproef gedoen is en 'n beskrywing van die deelnemers aan te teken sodat ander navorsers kan bepaal tot watter mate die huidige navorsingsomstandighede en metodes gedupliseer kan word en ook tot watter mate bevindinge op hul eie navorsing van toepassing is (Heigham en Croker, 2009: 252). Dit is ook belangrik dat die steekproef by kwalitatiewe data 'n bepaalde doel het. Dörnyei (2007: 126) meen die doel van die steekproef is "to find individuals who can provide rich and varied insights into the phenomenon under investigation [...]".

Dörnyei (2007: 127–129) verwys na verskeie strategieë wat gebruik kan word by die neem van 'n steekproef, maar die strategie wat van toepassing is op hierdie studie, is die neem van 'n sogenaamde 'tiperende' steekproef. Dit beteken dat die deelnemers almal min of meer dieselfde kenmerke deel of almal dieselfde ondervinding het wat van toepassing is binne die bepaalde studie. In hierdie geval is al die deelnemers individue binne 'n bepaalde ouderdomsgroep wat onder andere hoofsaaklik deur middel van selfone, en meer spesifiek, SMS-boodskappe, kommunikeer.

1.7.3 (c) Datastel en data-analise

Elke deelnemer het 'n spesifieke getal SMS'e aan die navorser gestuur wat die SMS-korpus vir hierdie studie uitmaak. Die korpus vir die studie is 2434 woorde in totaal. So 'n klein korpus hou bepaalde beperkings in betreffende die analisering van die data. Die datastel, die instrumente wat vir die insameling gebruik is en die instrumente wat by die ontleding van die data gebruik is, word in hoofstuk 3 (die metodologiehoofstuk) bespreek. Hier volg slegs kort beskrywings van elk van die laasgenoemde aspekte.

SMSPortal is vir die ontvang van die SMS'e gebruik aangesien dit maklik is om op die sisteem te registreer. SMS'e word in 'n gebruikersvriendelike formaat aan die navorser gestuur. Alle SMS'e wat deur die deelnemers aan SMSPortal gestuur word, kan per e-pos aan die navorser gestuur word. Dit beteken dat die SMS'e wat ontvang word nie in 'n vreemde formaat gestoor word nie en dadelik gebruik kan word om 'n korpus te bou.

Die stilistiese analise van die data is deur middel van 'n ontledingsblad uitgevoer wat deur die navorser opgestel is. Vir die stilometriese analise van die data is daar van Antconc en WordSmith Tools gebruik gemaak. 'n Eenvoudige n-gramanalise is ook op die data uitgevoer in 'n poging om die resultate te versterk. Die resultate van die stilistiese en stilometriese analises sal in hoofstuk 4 bespreek word.

1.8 Kernkonsepte

1.8.1 Forensiese taalkunde

Forensiese taalkunde is die studieveld waar linguistiese kennis en metodes gebruik word om 'n verskeidenheid tekste (gesproke en geskrewe) te analiseer vir regsdoeleindes. Argumente vir of teen 'n verdagte se betrokkenheid by misdaad op grond van hierdie tekste word gebaseer op stilistiese bewyse.

1.8.2 Outeuridentifikasie

Die vasstel van 'n moontlike outeur se idiolek met die doel om die outeur positief te identifiseer as produseerder van 'n teks staan as outeuridentifikasie bekend (McMennmin, 2010: 490–492; Coulthard, 2004: 431–447). Outeuridentifikasie is met ander woorde die proses waartydens 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.

1.8.3 Funksiewoorde

Funksiewoorde is woorde wat met die struktuur van 'n sin of uitspraak 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. Die plasing van die funksiewoorde in die struktuur van die sin is in sommige gevalle meer ‘idiolekties’ as die gebruik van, byvoorbeeld, sekere selfstandige naamwoorde wat kontekksspesifiek kan wees en met ander woorde meer gereeld sal voorkom wanneer daar oor ’n spesifieke onderwerp geskryf word.

1.8.4 SMS

Die akroniem “SMS” verwys na *short message service* – ’n goedkoop en vinnige manier om boodskappe deur middel van ’n selfoon te stuur (Ishihara, 2011: 47). Dit is belangrik om daarop te let dat die term ‘SMS’ in die volksmond nie noodwendig na die diensgedeelte van die akroniem verwys nie, maar eerder fokus op die *short message*-gedeelte. Dit beteken dat ’n SMS, soos dit hoofsaaklik deur gebruikers verstaan word, as ’n kort, elektroniese boodskap wat deur middel van ’n selfoon gestuur word, omskryf kan 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: 05). Dit beteken dat selfoongebruikers nou op kreatiewer maniere kort boodskappe moet skryf aangesien soveel inligting moontlik soms in een boodskap oorgedra moet word.

Die terme ‘SMS’ en ‘SMS-taal’ word hier gebruik om onderskeidelik te verwys na enige kort, elektroniese **boodskap** wat vanaf ’n selfoon gestuur word (uitsluitend *status updates* op sosiale netwerke en bloginskrywings, maar insluitend Whatsapp- en BBM-boodskappe) en na die **taalgebruik** wat in hierdie kort, eletroniese boodskappe aangetref word.

1.8.5 Generiese SMS-taal

Die term ‘generies’ impliseer dat dit wat deur die term beskryf word, kenmerkend is van ’n groep en nie spesifiek tot een objek of persoon in ’n groep beperk is nie. ‘Generiese taalgebruik’ verwys, binne die grense van hierdie navorsing, na kenmerke soos spelling, die afkorting van woorde, ensovoorts, wat in die meeste gevalle op dieselfde manier in die taalgebruik van die groep aangetref word. Daar is geen standaardvorm in SMS-taal nie, maar daar is wel sekere konvensies wat met die gebruik van SMS-taal verband hou. Daar is, byvoorbeeld, konvensies ten opsigte van die afkorting van woorde in ’n SMS (Olivier, 2013: 501). Hierdie konvensies bewerkstellig effektiewe kommunikasie in SMS-taal en lei daartoe dat daar, in teorie, ’n mate van generiese (Afrikaanse) SMS-taal sou kon wees wat deur SMS-outeurs gebruik word.

Nietemin moet daar in gedagte gehou word dat die SMS-taal van ’n individu, soos enige ander vorm van ’n taal, soms individualistiese (idiolektiese) afwykings toon wat dit onderskei van dit wat as die ‘generiese vorm’ beskou sou kon word.

1.8.6 Idiolek

Idiolek beskryf ’n individu se eiesoortige taalgebruik wat onbewustelik in hierdie individu se tekste voorkom. Idiolek en dit waarna Crystal (1987: 66) verwys as ‘styl’ stem ooreen wanneer die volgende definisie van Crystal in ag geneem word: “style is viewed as the set of language features that make people distinctive – the basis of their personal identity”. Die kenmerke van ’n individu se idiolek kan byvoorbeeld eiesoortige woordgebruik en sinskonstruksies asook afwykings van ‘tipiese’ taalgebruik insluit.

1.8.7 Stilistiese analises

In stilistiese analises word tekse ondersoek met die oog op die opsporing van kenmerkende herhaalde woorde of woordgroepe wat kollokasies vorm (Kotzé, 2007: 390). Hierdie woorde is kontekssspesifiek en word deur die onderwerp van ’n gesprek of teks bepaal. Wanneer sekere woorde in een teks ook in ander tekste voorkom, word hierdie gelyktydige voorkoms aangeteken.

1.8.8 Stilometriese analises

Stilometriese analises het in die verlede hoofsaaklik betrekking gehad op leksikale woorde. Die gebruik van leksikale woorde word egter dikwels bepaal deur die onderwerp van die bespreking. Dit beteken dat die leksikale woorde van een persoon van situasie tot situasie sal verskil afhangende van die onderwerp van die gesprek of kommunikasie. Kotzé (2007: 390) het byvoorbeeld in sy stilometriese analise van ’n bepaalde teks gefokus op die frekwensie en kollokasie van grammatiese woorde wat nie deur die onderwerp van die teks beïnvloed is nie. Die gebruik van sulke woorde hang af van die manier waarop ’n persoon sy/haar woorde oor ’n bepaalde onderwerp struktureer. Grammatiese woorde sluit onder andere lidwoorde, voegwoorde en voornaamwoorde in. Hierdie woorde word gebruik om die inhoud van ’n teks of bespreking te struktureer sonder dat die spreker daarvan bewus is tydens kommunikasiehandelings (Kotzé, 2007: 390). Daar word aangevoer dat individue nie bewustelik bepaalde konstruksies of

grammatiese woorde gebruik nie, maar dat herhalende konstruksies wel in individue se gesproke en geskrewe taalgebruik voorkom. Hierdie konstruksies kan gebruik word om die outeur van ’n teks te bepaal. Stilometriese analises word gebruik om tipiese strukture in die onbewustelike taalgebruik van individue deur rekenaarprogramme te identifiseer en daarná word statistiese vergelykings getref deur middel van (onder andere) die Chi-kwadraattoets om die moontlikheid van outeurskap tussen twee of meer tekste te bepaal (Kotzé, 2007: 390).

1.8.9 N-gramme

In Rekenaarlinguistiek verwys die term ‘n-gram’ na ’n aangrensende reeks items (*n*) wat in ’n opeenvolging van gesproke of geskrewe teks voorkom. Hierdie items kan foneme, lettergrepe, woorde of letters wees. N-gramme bestaan uit verskillede groottes. ’n N-gram wat uit slegs een item bestaan word ’n ‘uni-gram’ genoem terwyl ’n n-gram wat uit twee itmes bestaan ’n ‘bi-gram’ genoem word. Daar kan ook ‘tri-gramme’, ‘vier-gramme’, ‘vyf-gramme’ ensovoorts aangetref word. Bi-gramme van die woord ‘outeur’ sal byvoorbeeld so lyk: _O, OU, UT, TE, EU, UR, R_

1.8.10 DO (Dieselfde-outeur)- en VO (Verskillende-outeur)-vergelykings

By DO-vergelykings word twee groepe SMS’e wat deur dieselfde outeur geproduseer is, vergelyk, terwyl SMS’e van verskillende outeurs vergelyk word by VO-vergelykings (Ishihara, 2011: 49). Hierdie vergelykings word gebruik om so seker as moontlik te wees oor die outeurskap van SMS-boodskappe. Vergelykings tussen boodskappe van dieselfde outeur en daaropvolgende vergelykings met ’n ander outeur kan aandui dat daar idiolektiese verskille tussen outeur A en outeur B is.

Stilistiese en stilometriese analises asook DO- en VO-vergelykings kan lig werp op die volgende vrae wat volgens Harold Somers (s.a.) as belangrik beskou word:

- Wie het *dié* teks geskryf: A of B?
- Indien A *hierdie* teks geskryf het, het A ook *daardie* teks geskryf?
- Wat is die waarskynlikheid dat A *hierdie* teks geproduseer het?

Die moontlike antwoorde op die eerste vraag: "*Wie* het die teks geskryf?", kan legio wees. Hierdie vraag moet dan ook verkieslik nie as navorsingsvraag gebruik word om antwoorde oor outeurskap te verkry nie.

1.9 Uiteensetting van die verhandeling

Hierdie hoofstuk word afgesluit deur die struktuur van die verhandeling kortliks uiteen te sit.

In **Hoofstuk 2** volg 'n literatuuroorsig wat forensiese linguistiek en die subkategorie van outeuridentifikasie in detail ondersoek. Die aanvaarbaarheid van forensies-linguistiese bewyse in die hof word ondersoek met verwysing na die regstelsels in verskeie lande. Op grond hiervan word redes aangevoer waarom forensies-linguistiese bewyse in outeuridentifikasie ondersoeke wat met elektroniese tekste verband hou as bewyse in die hof toegelaat moet word. Die rol van die forensiese linguis as getuie in hofsake word ook ondersoek. Die hoofstuk bevat ook inligting oor die oorsprong en gebruike van stilometrie in forensiese linguistiek. 'n Oorsig oor die ondersoekveld bekend as adversatiewe stilometrie is ingesluit en hier word kortliks ondersoek hoe individue hulle idiolek in tekste kan verdoesel om sodoende te verseker dat hulle privaatheid in die aanlynomgewing versterk word. Die hoofstuk word afgesluit deur 'n breedvoerige uiteensetting van die eienskappe van SMS-taal en 'n bespreking van die kontroversiële kwessie van die aanwesigheid van idiolek in individue se praat- en skryfstyle.

In **Hoofstuk 3** word die metodologie van die verhandeling bespreek. Beide die stilistiese en stilometriese analises wat in die studie gebruik is word verduidelik en die sagteware wat gebruik is om die data te ontvang en te verwerk, word bespreek.

Die SMS-korpus is ontleed deur van verskillende metodes gebruik te maak en hierdie ontledings en die resultate wat verkry is, word in **Hoofstuk 4** aangebied. Die resultate word bespreek en gevolgtrekkings word op grond van die resultate gemaak.

In **Hoofstuk 5** word die inhoud van die verhandeling saamgevat en die beperkings van die studie word bespreek. Die regsimplikasies van hierdie tipe ondersoek word ook bespreek. Daar word verwys na moontlike soortgelyke ondersoeke wat in die toekoms aangepak kan word asook maniere waarop die huidige studie verbeter kan word ten opsigte van metodologiese verfyning.

Hoofstuk 2: Die wye spektrum van die forensiese linguistiek: ’n oorsig

2.1 Forensiese linguistiek: definiëring en afbakening van die vakgebied

In hierdie hoofstuk word die ondersoekveld van die forensiese linguistiek bespreek. Die oorsprong van die ondersoekveld word nagegaan en daar word spesifiek gefokus op outeuridentifikasie, wat die hooffokus van die huidige navorsing is. Stilometrie, ’n metode wat algemeen in die forensiese linguistiek gebruik word, word ondersoek en beskryf en daar word ook gefokus op die nuwer onderafdeling van stilometrie bekend as “adversatiewe stilometrie”. Laasgenoemde navorsingsveld ontvang in die een-en-twintigste eeu meer aandag aangesien daar groter nood is vir aanlynprivaatheid. Adversatiewe stilometrie beskou outeuridentifikasie vanuit ’n ander invalshoek en poog om vas te stel hoe dit vir ’n individu moontlik sou wees om bewustelik sy of haar identiteit in die aanlynomgewing te vermom deur sy of haar skryfstyl aan te pas. Die term ”idiolek” kom ook aan bod en daar word gepoog om vas te stel of dit werklik moontlik is om aan te voer dat idiolek bestaan en identifiseerbaar is in die tekste van individue. Aangesien die huidige navorsing op outeuridentifikasie in SMS-boodskappe fokus, word die SMS-kultuur en die eienskappe van SMS-boodskappe ook breedvoerig bespreek.

Ten eerste is dit egter belangrik om aan te dui waar die forensiese linguistiek binne die veld van forensiese wetenskap inpas, aangesien forensies-linguistiese analises 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 sinoniem beskou met woorde soos ‘geregtelik’ en ‘juridies’ en word tradisioneel met die term “wetenskap” geassosieer. Forensiese wetenskap word reeds jare lank al gebruik om die verdagtes van misdade vas te trek en skuldig of onskuldig te bewys.

Forensiese wetenskap sluit onder andere DNS-analises, vingerafdrukanalises en bloedvlekanalises in. Handskrifanalises en handtekeninganalises, 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 analise, soos handskrifanalise, fokus dit op ander aspekte van die teks aangesien die eiesoortige aspekte van geskrewe tekste nie in moderne elektroniese tekste teenwoordig is nie. Alle letters word byvoorbeeld eenvormig getik behalwe vir afwykings wat met die vetdruk of skuinsdruk van letters of woorde gepaard gaan.

Die konsep ‘forensiese linguistiek’ kom volgens Coulthard en Johnson (2007: 5) so vroeg as 1949 voor in F.A. Philbrick se boek *Language and the law: the Semantics of Forensic English* wat oor geregtelike Engels handel. Volgens Philbrick (1949: vi) het hy die woord ‘forensies’ gebruik om te verwys na “[...] the English used by advocates and judges in courts of law”. Die woord ‘forensies‘ is egter nie dadelik met betrekking tot taal of taalkunde in gebruik geneem nie. Olsson (2004: 3) en Blackwell (2012: 1) wys daarop dat die term ‘forensiese linguistiek‘ in sy huidige betekenis die eerste keer in 1968 gebruik is met die publikasie van Jan Svartvik se analise van die Timothy Evans verklarings getiteld *The Evans Statements: A case for forensic linguistics*[1]. Ten spyte van meningsverskille oor die eerste gebruik van die term ‘forensiese linguistiek’ is navorsers dit eens dat die term eers tydens die 1990’s erken is toe forensiese linguistiek ook as ’n navorsingsveld aanvaar is. Aanvanklik kon die metodes wat deur linguiste gebruik is om van die vroeë forensies-linguistiese analises uit te voer egter nie onder een akademiese dissipline groepeer word nie aangesien sulke analises herhaaldelik uit die *ontwerp* van ’n nuwe metode, eerder as die *toepassing* van ’n bestaande metode, bestaan het (Blackwell, 2012: 1).

Forensiese linguistiek word beskou as ’n onderafdeling binne die toegepaste taalkunde waar linguistiese kennis en metodes gebruik word om ’n verskeidenheid tekste (beide gesproke en

[1] Jan Svartvik se analise van die *Evans Statements* word by 2.1.1 in meer detail bespreek.

geskrewe) te analiseer vir regsdoeleindes. Olsson (s.a.: 2) se definisie van forensiese taalkunde illustreer duidelik hoe 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.

Leonard (2006: 65) herskryf bogenoemde definisie tot een kompakte sin: "Forensic linguistics applies the well-establish science of linguistics to legal language data". Hierdie definisie maak 'n kernaspek van forensiese linguistiek duidelik, naamlik dat dit 'n toegepaste wetenskap is. Daar moet ook in gedagte gehou word dat forensiese linguiste in verskeie ander dissiplines betrokke kan wees aangesien regskwessies in enige beroep of veld kan voorkom. Dit beteken dat die forensiese linguis in die eerste plek goed onderlê moet wees in die linguistiek en dat hy/sy veral die struktuur en woordeskat, dialekte, sosiolekte, registers ens. van die bepaalde taal waarin hy/sy werk, goed moet ken. In die tweede plek is dit belangrik dat die forensiese linguis ook 'n basiese agtergrond of kennis het in die bepaalde veld of beroep waarin hy/sy ondersoeke doen.

Forensiese linguiste moet daarom nie net oor regskennis in sekere gevalle beskik nie, maar ook verseker dat hulle oor die nodige kennis van die bepaalde dissipline waarin hulle werk beskik. Forensiese linguiste moet ook goed onderlê wees in die taalkunde van die taal waarin hulle werk. McMenamin (2002: 86) meen dat linguiste wat die baanbrekerswerk in forensiese linguistiek gedoen het dikwels sê dat hulle linguistiek beoefen wat net toevallig binne 'n forensiese konteks val en dat 'n forensiese linguis daarom in die eerste plek oor 'n stewige kennis van die linguistiek moet beskik. Volgens Leonard (2006: 65) versterk die forensiese linguis regsake deur streng, wetenskaplik aanvaarde beginsels van linguistiese analise op regsbewyse toe te pas. Die stellings van McMenamin en Leonard probeer nie om regskennis of kennis van hoe daar in regskontekste gefunksioneer word as minder belangrik af te maak nie, maar beklemtoon die belangrikheid van kennis in die veld van taalkunde en toegepaste taalkunde ten opsigte van teorieë en metodes wat gebruik word om taal te analiseer.

Vandag is daar verskeie kategorieë wat binne forensiese linguistiek identifiseer word. **Tabel 1** bevat 'n lys van hierdie kategorieë asook 'n beskrywing van elke kategorie.

Tabel 1: Kategorieë binne die forensiese linguistiek.

Kategorie	Beskrywing en ondersoekfokus
1. Taalgebruik wat verband hou met regsimplikasies	Die taalgebruik in regstekste en regsprosesse is die fokus van ondersoek. Die forensiese linguis is veral gemoeid met die semantiese aspek van die taal en stel ondersoek in na die toeganklikheid van regstaal. ’n Belangrike aspek is die vasstel van die betekenis van woorde binne ’n bepaalde (regs)konteks. Ander aspekte van regstaal en taalkwessies wat met die reg verband hou word ook ondersoek.
1.1 Hoftolking en -vertaling	Die fokus is op tolking en vertaling in die hof. Dit sluit die tolking van die getuienis asook die vertaling van verklarings en ander dokumente in. Hier word veral gefokus op akkuraatheid in tolking en vertaling, die rol van die tolk, die lisensiëring van tolke, ‘beheer’ oor die persoon vir wie daar getolk word, ensovoorts.
1.2 Die taalgebruik en diskoers in die hof	Hier word kwessies soos die volgende ondersoek: die verhouding tussen die regspersone en verdagtes betrokke in die spesifieke saak en die taal (enige intimiderende of manipulerende taalgebruik) wat hulle gebruik. Kwessies rakende mag, partydigheid en kultuur word ook ondersoek.
1.3 Transkripsie van verbale verklarings	Wanneer verklarings wat verbaal afgelê is getranskribeer is vir die hof, word die volledigheid van hierdie verklarings asook die moontlikheid van partydigheid in die transkripsie, ondersoek.

1.4 Taalregte (hierdie kategorie se ondersoeke strek wyer as die hofsaal)	Hier word onder andere die taalregte en taalgebruik van minderheidstale ondersoek – hoe hierdie groepe deur ander tale of dialekte van dieselfde taal gedomineer word asook hoe burokratiese taalgebruik individue onderdruk.
2. Die ondersoek van forensiese tekste (die term 'teks' verwys na beide gesproke en geskrewe tekste)	Forensiese navorsing en ondersoeke wat op tekste uitgevoer word buite die hof hou ook verband met die reg. Forensiese linguiste word dikwels deur die polisie genader om ondersoek in te stel wanneer spesifieke tekste met misdade verband hou. Hierdie ondersoeke neem verskeie vorme aan:
2.1 Outeuridentifikasie	Outeuridentifikasie is die analise van 'n teks met die doel om die moontlike outeur van die teks te bepaal en kan gebruik word om die outeurs van verskeie geskrewe tekste te identifiseer. Tekste sluit, onder andere, die volgende in: dreigemente, afpersbriewe, lasterboodskappe, selfmoordbriewe en lospryseise.
2.2 Profielsamestelling	Hierdie navorsingsveld word as 'n alternatiewe vorm van outeuridentifikasie beskou en behels dat die forensiese linguis deur middel van linguistiese leidrade in 'n bepaalde teks 'n profiel van die moontlike verdagte outeur skep. Deur middel van linguistiese leidrade soos woordgebruik, die gebruik van leestekens en die frekwensie van hoofletters teenoor kleinletters, kan, onder andere, die geraamde ouderdom en die geslag van die outeur vasgestel word.

2.3 Identifisering van plagiaat	Identifisering van plagiaat word veral in die akademiese konteks gebruik. Vandag is daar verskeie rekenaarprogramme wat plagiaat identifiseer, maar in sommige gevalle word die kennis van ’n spesialis steeds ingeroep.
2.4 Sprekeridentifikasie (ook bekend as forensiese fonetiek)	Hierdie subkategorie hou spesifiek verband met die identifisering van die *spreker* van ’n boodskap op grond van onder meer die akoestiese kwaliteite en geluidkenmerke van ’n stem. Sulke boodskappe kan enige van die volgende insluit: telefoonboodskappe, noodoproepe, bandopnames en dreigemente deur middel van telefoonoproepe.
2.5 Forensiese dialektologie	Hierdie afdeling het te make met die nagaan van die linguistiese geskiedenis van asielsoekers. Deur middel van, onder andere, uitspraaktoetse wat aan asielsoekers gegee word, kan daar vasgestel word of die asielsoeker werklik tot ’n bepaalde taalgroep/kultuurgroep behoort.

[Verwerk en aangepas uit Olsson (2004: 4–5)]

Verdere navorsingsvelde soos die ondersoek na die eienaarskap van ’n handelsmerk is moeilik om binne bogenoemde kategorieë te plaas, maar dit word wel as ’n forensies-linguistiese ondersoekveld beskou. Coulthard en Johnson (2007: 121–128) meen daar is vyf kategorieë waaronder linguiste se kennis gegroepeer kan word wanneer hulle genader word om as getuies in hofsake op te tree. Volgens Coulthard en Johnson val navorsing wat verband hou met die eienaarskap van handelsmerke binne die eerste kategorie, naamlik “Morphological meaning and phonetic similarity”. Die ander vier kategorieë is:

- "Syntactic complexity in a letter". Sintaktiese kompleksiteit kan ook op ander lang tekste van toepassing wees. In sulke gevalle moet die linguis probeer vasstel of die kompleksiteit van die sinstruktuur in 'n brief of dokument die leser se begrip van die inhoud belemmer.

- "Lexico-grammatical ambiguity". Hierdie kategorie hou verband met dubbelsinnighede wat in tekste voorkom en gevolglik moet vasgestel word wat die 'werklike' betekenis van sekere woorde of frases binne bepaalde situasies /kontekste is

- "Lexical meaning". Binne hierdie kategorie moet daar bloot vasgestel word wat die betekenis van sekere woorde is. Hier word aspekte soos konteks, kultuur en die oorsprong van woorde in ag geneem.

- "Pragmatic meaning". Die 'manier' waarop daar gekommunikeer word is in hierdie kategorie die hooffokus. Hier is aspekte soos 'gedeelde kennis' tussen die individue wat aan die kommunikasiehandeling deelneem van belang. In hierdie kategorie word vasgestel hoe realisties sekere uitinge sou wees – indien daar beweer word dat iemand iets spesifieks gesê het – en ook watter inligting nodig is vir geldige skuldbekentenis.

2.1.1 Die oorsprong van die vakgebied

Die oorsprong van forensiese linguistiek is vandag steeds 'n kontroversiële kwessie. Tydens die jare sestig, sewentig en tagtig van die vorige eeu is daar reeds, hoofsaaklik in die VSA en Kanada, van linguiste se kennis gebruik gemaak in hofsake waar linguistiek en die reg oorvleuel. Sulke situasies was egter skaars en die metodologieë wat gebruik is, was nog nie in detail ondersoek om die geldigheid en betroubaarheid daarvan vas te stel nie (Turell, 2008: 115).

Oor die algmeen word Jan Svartvik se publikasie, *The Evans Statements: A case for forensic linguistics*, as die vroegste gestaafde gebruik van forensiese linguistiek beskou (Coulthard, 2004: 431; Olsson, 2004: 3, 15; Turell, 2008: 155; Coulthard, 2010: 9; Blackwell, 2012: 1). Die publikasie het in 1968 verskyn en analiseer vier uitlatings wat vermoedelik deur Timothy Evans aan die polisie gemaak is waarin Evans erken dat hy sy vrou en babadogter in 1949 verwurg het. Dit het egter ná Evans se dood aan die lig gekom dat hy onskuldig was aan die moord op sy babadogter en dat sy buurman, John Christie, vir daardie moord (en andere) verantwoordelik

was (Coulthard, 2004: 431, Turell, 2008: 155). Christie is in 1953, drie jaar na Evans se dood, skuldig bevind aan die moord op minstens vyf mense en het ook die doodstraf ontvang. Dertien jaar later, in 1966, is probleme wat reeds in 1953 ontstaan het oor die verklarings wat deur Evans aan die polisie gemaak is, weer ondersoek. Daar is aangevoer dat twee van die verklarings vervals is en dat die taalgebruik in die verklarings nie ooreenstem met die taalgebruik van ’n analfabeet nie (Evans kon nie lees of skryf nie). In Svartvik se studie word die taalgebruik van Evans uit ’n linguistiese oogpunt geanaliseer. Svartvik (1968: 19) het gemeen dat hy bloot die analises doen om vas te stel of die vrae wat rondom die Evans-verklarings ontstaan het geldig is of nie en dat die resultate nie as ’n regskundige gevolgtrekking beskou moet word nie. Svartvik het sekere woordkeuses en frases in die verklarings vergelyk met woordkeuses en frases wat Evans tydens die verhoor gebruik het. Hy het bevind dat die woordkeuses wat in die verklarings aangetref word baie formeel is teenoor die woordkeuses wat Evans gebruik het tydens sy verhoor. Svartvik kon egter nie met sekerheid bepaal of al die verklarings vervals is en of net sommige gedeeltes van die verklarings vervals is nie, alhoewel die analise tot ’n mate daarop dui dat net sekere gedeeltes van die verklarings vervals is.

’n Tweede moontlike oorsprong van forensiese linguistiek word teruggevoer tot die 1851-brief van die Britse logikus, Augustus de Morgan. In hierdie brief wat aan ’n vriend geskryf is (Hockey, s.a: 2), maak De Morgan die voorstel dat dit moontlik sou kon wees om die outeurs van die boeke van die Bybel te identifiseer deur op woordlengte as onderskeidende stylkenmerk te fokus (Kotzé, 2007: 386). Dié voorstel deur De Morgan is volgens Kotzé (2007: 386) waarskynlik waar die forensiese linguistiek sy beslag gekry het. Daar is egter vir sowat 30 jaar ná die brief geen verdere verwikkelings van De Morgan se voorstel aangeteken nie. In 1887 publiseer T.C Mendenhall sy eie bevindings, gegrond op De Morgan se voorstel, nadat hy woord- en sinslengte as merkers van outeurskap in werke van Bacon, Marlowe en Shakespeare gebruik het.

Die gebruik van sinslengte en woordlengte as merkers van outeurskap is later deur Smith (1983) ongeldig bewys nadat Smith bevind het dat wanneer die werke van verskillende skrywers binne dieselfde literêre genre met mekaar vergelyk word, die verspreiding van woordlengtes in die tekste so eenders is dat dit voorkom asof dieselfde persoon die outeur van al die tekste is

(Holmes, 1994: 88). Mendenhall se navorsing het nietemin daartoe gelei dat ander metodes vir outeuridentifikasie ontwikkel is (Holmes, 1994: 88; Holmes, 1998: 112 en Olsson, 2004: 12).

Die analise van die *Federalist Papers* deur Mosteller en Wallace (1964) is een van die bekendste forensies-linguistiese analises en word óók deur sommige navorsers beskou as die eerste werklike forensies-linguistiese analise van ’n teks wat ten doel gehad het om die outeur van die teks te bepaal (Broeders, 2001: 68; Mikros, s.a.: 1 en Stamatatos, s.a.: 1). Die *Federalist Papers* is anoniem tussen 1787 en 1788 gepubliseer en bestaan uit 85 dokumente. Die doel van die dokumente was om New Yorkers te oorreed om die nuwe Amerikaanse grondwet te aanvaar (Holmes, 1998: 112; Coulthard en Johnson, 2007: 169 en Kotzé, 2007: 386). Die dokumente is deur drie outeurs, Alexander Hamilton, James Madison en John Jay, geskryf. Hamilton en Madison het albei beweer dat hulle ’n stel van 12 dokumente, wat deel van die *Federalist Papers* uitmaak, geproduseer het. In ’n poging om die dispuut op te los het Mosteller en Wallace die 12 dokumente ontleed om veronderstelde idiolektiese eienskappe in die dokumente te identifiseer. Eers is ’n stel dokumente van elke outeur geanaliseer om vas te stel of Hamilton en Madison van idiolektiese woorde of sinskonstruksies gebruik maak. Hulle het ook op die frekwensie van funksiewoorde in die onderskeie tekste gefokus. Hierdie analise toon volgens Holmes (1994: 88) sterk ooreenkomste met die navorsing deur Mendenhall (1887), maar word deur Koppel e.a (2009: 5) eerder as ’n “multivariate analysis approach” beskou “[that] augured in a new set of methods for stylometric authorship attribution, based on combining information from multiple textual clues”. Nadat 30 idiolektiese eienskappe van elke outeur se tekste identifiseer is, het Mosteller en Wallace hierdie eienskappe met die 12 kontroversiële dokumente vergelyk en bevind dat Madison die outeur van die 12 dokumente is.

Hierdie bevinding stem ooreen met bevindings deur historici wat Madison ook as die outeur van die 12 dokumente aanwys. Omdat die outeuridentifikasie van die *Federalist Papers* so suksesvol was, is dit algemene gebruik dat nuwe outeuridentifikasiemetodes en -teorieë vandag nog op afskrifte van die *Federalist Papers* getoets word (Holmes, 1998: 112; Coulthard en Johnson, 2007: 169–170 en Stamatatos, s.a.: 20). Stamatatos (s.a.: 20) waarsku egter dat die *Federalist Papers* ’n ideale en seldsame forensies-linguistiese situasie verteenwoordig met min moontlike outeurs en lang tekste. Dit moet in gedagte gehou word wanneer metodes op hierdie dokumente

getoets word aangesien realistiese forensies-linguistiese situasies waarskynlik van laasgenoemde sal verskil.

Nog ’n ontstaandatum van outeuridentifikasie as ondersoekterrein word deur Grieve (2005: 4) en Schulstad e.a. (2012: 1) voorgestel. Hulle meen dat outeuridentifikasiemetodes reeds in die 1700’s gebruik is om die outeurs van literêre tekste te bepaal. Die studies waarna Grieve en Schulstad verwys word ook as die pionierstudies in kwantitatiewe analise (bekend as ‘stilometrie’ in die forensiese taalkunde) beskou. Beide navorsers verwys na die 1787-studie deur Edmond Malone waartydens vasgestel is dat Shakespeare nie die outeur is van enige van die drie gedeeltes waaruit die teks van *Henry VI* bestaan nie. Volgens Grieve (2005: 4) het Malone aangevoer dat Shakespeare nie die outeur van *Henry VI* kan wees nie op grond van die waarneming dat die outeur van die *Henry VI*-teks gereeld sinne aan die einde van die versreël afsluit eerder as om van enjambement (die oorloop van ’n sin van een versreël na die volgende) gebruik te maak. Verder het Malone ook bepaal dat die outeur selde die eindwoorde in versreëls laat rym. Dit is metriese eienskappe wat nie kenmerkend van Shakespeare se verhoogstukke is nie.

In Brittanje word pionierstudies in forensiese linguistiek teruggevoer tot 1985. Dit beteken dat forensiese linguistiek veel later in Brittanje posgevat het as wat die geval in Amerika is. Hierdie pionierstudies het hoofsaaklik in Birmingham plaasgevind waar linguiste in die hof getuig het tydens sake waar handskrifanalises en outeurskapidentifikasie ter sprake was. Hulle getuienis het sowel geskrewe as gesproke tekste betref (Turell, 2008: 115).

Die algemene siening is nietemin dat die studieveld van forensiese linguistiek in die vroeë 1990’s meer formeel gevestig geraak het en as navorsingsveld en studieveld erken is. Tussen 1988 en 1992 het die studieveld baie meer belangstelling gewek en verskeie seminare oor die onderwerp van forensiese linguistiek is in Brittanje en Duitsland gereël. Afgevaardigdes uit Australië, Brasilië, Holland, Griekeland en Oekraïne het hierdie seminare bygewoon. Die studieveld het selfs nog verder uitgebrei en uiteindelik is ’n konferensie in Australië gehou (in 1995) asook ’n konferensie in die VSA (in 1997). Intussen is die International Association of Forensic Linguistics (IAFL) in 1993, die International Association for Forensic Phonetics (IAFP) en die International Journal of Speech, Language and Law in 1994, gestig. Hierdie konferensies en die stigting van die laasgenoemde verenigings dui daarop dat forensiese linguistiek ’n

internasionale studieveld geword het en dat individue uit verskeie lande binne hierdie veld navorsing doen en hulle bevindings gebruik om in die hof getuienis af te lê. Daar was ook ’n merkbare toename in die hoeveelheid forensies-linguistiese publikasies wat gevolg het en die temas waaroor navorsing gedoen is het aansienlik gegroei. Enkele publikasies en artikels is Solan (1993), Levi (1994), McMenamin (1994, 2002), Stygall (1995), Kurzon (1997), Hanlein (1999), Elrich (2001), Foster (2002), Alcaraz en Hughes (2002), Rose (2002), Cotterill (2003), Gibbons (2003), Heffer (2005), Coulthard en Johnson (2007), Kniffka (2007), Eades (1995, 2008) en Shuy (1993, 1998, 2002, 2005) (Johnson en Coulthard, 2010: 1; Turell, 2008: 156).

Die populariteit van die vakgebied het toegeneem, en webtuistes is geskep en aanlynforums gestig waar lede van die forensiese taalkunde-gemeenskap hulle opinies oor verskeie sake en onderwerpe kan lug. Sogenaamde ‘forensies-linguistiese laboratoriums’ is ook aanlyn gestig asook webtuistes vir taal en die reg[2]. Teen die einde van die 20ste eeu is die veld goed gevestig en word voorgraadse en/of nagraadse kursusse by verskeie universiteite in lande soos die VSA, Brittanje, Australië, Sjina, Finland, Duitsland en Japan aangebied (Blackwell, 2012: 3; McMenamin 2002: 87). In die 21ste eeu neem belangstelling in die veld, ook in Suid-Afrika, toe.

2.1.2 Formele taalkunde teenoor toegepaste taalkunde

Soos reeds vermeld, word forensiese linguistiek binne die veld van toegepaste taalkunde gekategoriseer. Daar is egter steeds eienskappe van die formele taalkunde wat ’n rol speel tydens ’n forensies-linguistiese ondersoek. McMenamin (2002: 62) verwys na forensiese linguistiek as onder andere “one of many developing disciplines in applied linguistics, which draws on the scientific study of language to solve forensic problems”. Om hierdie rede is dit nodig om aan te dui hoe die velde van toegepaste taalkunde en formele taalkunde van mekaar verskil.

Toegepaste taalkunde is gewortel in tweedetaalonderrigteorieë. As gevolg van ontwikkelings in die onderrig van taalvaardighede, het toegepaste taalkunde uitgebrei om ook ander studievelde in te sluit. Omdat taalonderrig later uit ’n ander oogpunt as die tradisionele Grammatika-vertaalmetode beskou is, het teoretici begin om kennis uit ander vakgebiede, soos sosiologie en

[2] Byvoorbeeld: http://www.forensiclinguistics.net/cfl_fl.html; http://www.languageandlaw.org en http://www.iula.upf.edu/forensiclab

die sosiale wetenskappe, op taalonderrig toe te pas. Wei (2011: 7) beskryf hedendaagse toegepaste taalkunde as:

> [...] a broad field of study of language learning and language use by different learner and user groups as well as wider social issues such as language planning, language ideology and language and social (dis)advantage. It is no longer focused on applying any specific linguistic theory or model, but on developing a critical perspective on language in everyday social life. Methodologically, applied linguistics has adopted a broad-based discourse analysis, complemented by multimodality analysis.

Van die baie studievelde benewens tweedetaalonderrig wat vandag binne toegepaste taalkunde geklassifiseer word, is onder andere taalbeplanning, interkulturele stereotipering, *forensiese taalkunde*, mediadiskoers en dowestudies.

Wei (2011: 9) waarsku dat die fokus op die taalstruktuur self nie verlore moet gaan nie en voer aan dat toegepaste taalkunde juis onderskei kan word van ander domeine in die sosiologie, ekonomie, politiek en regswetenskap omdat dit spesifiek op ‘taal’ fokus. Hy stel onomwonde:

> At its core applied linguistics needs a coherent theory of language, whether this comes from linguistics or from some other discipline, a set of rigorous descriptive tools to handle language, and a body of research relevant to language practice.

Hierdie fokus op taal vorm die kern van die forensiese linguistiek waar kennis van formele taalkunde gebruik moet word. Soos die naam ‘toegepaste taalkunde’ aandui, word kennis van formele *taalkunde* tydens forensies-linguistiese ondersoeke *toegepas* om resultate te produseer. Olsson (s.a.: 2) wys daarop dat hierdie ‘toepassing’ van kennis in forensiese linguistiek nie dieselfde beteken as in, byvoorbeeld, die toegepaste statistiek nie. By laasgenoemde word ’n teorie toegepas “underpinning a particular science to the practice of that science”, terwyl daar in die forensiese linguistiek taalkundige kennis toegepas word binne ’n bepaalde konteks – in hierdie geval is dit die regskonteks.

McMenamin (2002: 57) verskaf ’n bruikbare beskrywing van wat as **formele taalkunde** beskou kan word:

> Linguistics is about understanding the system of language. The aims of linguistic science are theoretical insofar as linguists discover the underlying rules and patterns of language and then describe them in the languages of the world. Linguists look for language characteristics that are present in all languages (universals), as well as features found only in certain language families or individual languages.

Formele taalkunde hou verband met die mikrosisteem van taal en verwys na die bestudering van die taalsisteem. Dit sluit aspekte soos die leksikon, sintaksis, morfologie, fonologie, fonetiek en pragmatiek van die taal in (Johnson en Johnson, 1999). Binne die studie van formele taalkunde probeer die linguis onder andere vasstel watter patrone binne die spesifieke taal voorkom en wat die taalreëls van ’n bepaalde taal is. Hierdie kennis van ’n taalsisteem maak dit vir die linguis moontlik om in die subkategorie van outeuridentifikasie, byvoorbeeld, verskeie afwykings en variasies tussen verskillende outeurs se skryfstyle te identifiseer. Die kennis van formele taalkunde word egter ook in ander subkategorieë van die forensiese linguistiek toegepas, soos byvoorbeeld forensiese fonetiek.

2.2 Forensiese linguistiek in Suid-Afrika: ’n opkomende ondersoekterrein

Ondersoek in die veld van forensiese linguistiek in Suid-Afrika is tans baie skaars, alhoewel daar tekens is dat die forensiese linguistiek wel ’n opkomende ondersoekterrein is. Daar is reeds genoem dat slegs enkele individue in die verlede navorsing in hierdie veld gedoen het of tans in die veld werk. Ten spyte van vooruitgang soos die kursus in forensiese linguistiek wat vanaf 2011 by Noordwes-Universiteit (NWU) onder leiding van prof. Ernst Kotzé aangebied is, is daar nie enige merkwaardige groei in forensiese linguistiek in Suid-Afrika te bespeur nie. Die enkele navorsers en akademici wat reeds navorsing in forensiese linguistiek gepubliseer het word kortliks genoem.

Proff. Hubbard en Kotzé werk beide op dié terrein. Hubbard het verskeie artikels gepubliseer met die forensiese linguistiek as fokus. Hierdie artikels sluit *Errors in court: a forensic application of error analysis* (1994), *Linguistic fingerprinting?* en *A case study in forensic stylometrics* (1995)

in. Hubbard het ook ’n (ongepubliseerde) referaat gelewer oor stilometrie: *Stylometric and error analysis in the context of a style shift in abusive e-mail texts* (2009). Slegs enkele van Hubbard se artikels en publikasies fokus uitsluitlik op outeuridentifikasie en die analises is op Engelse tekste gebaseer. Kotzé se artikels is getiteld *Die vangnet van die woord: forensies-linguistiese getuienis in ’n lastersaak* (2007) en *Author identification from opposing perspectives in forensic linguistics* (2010). Beide Kotzé se artikels handel oor outeuridentifikasie in Engelse tekste.

Moeketsi het ook ’n bydrae gelewer tot die navorsingsveld van forensiese taalkunde in Suid-Afrika en het in 1997 en 1999 die volgende artikel en boek, onderskeidelik, gepubliseer: *Of African languages and forensic linguistics: the South African multicultural criminal courtroom* en *Discourse in a multilingual and multicultural courtroom: a court interpreter’s guide*. Die fokus van hierdie navorsing is hoofsaaklik die taalgebruik in Suid-Afrikaanse howe. Moeketsi probeer vasstel of kommunikasie tussen verdagtes en regspersone effektief geskied in ’n multikulturele konteks, asook hoe die taalgebruik in howe verdagtes kan intimideer.

Taylor (1998) publiseer ’n artikel getiteld *Addressing the insane language of the law* wat fokus op taalgebruik wat verband hou met regsimplikasies. In 2006 word taalgebruik en verwante regsimplikasies ook deur Reddy en Potgieter ondersoek in hulle artikel *‘Real men stand up for the truth’: discursive meanings in the Jacob Zuma rape trial,* en dieselfde onderwerp word ook in 2011 en 2012, onderskeidelik, herhaal met die publisering van Lombard en Carney se artikel *Die wenslikheid van Afrikaans as vaktaal vir regstudente* en Carney se artikel *’n Forensies-semantiese beskouing van die woordgebruik ‘onkoste’ in die hofsaak Commissioner for South African Revenue Service vs. Labat Africa Limited*.

In 2002 verskyn Thetela se artikel *Sex discourses and gender constructions in Southern Sotho: a case study of police interviews of rape/sexual assault victims*. Thetela fokus spesifiek op geslagstudies in die regstelsel, maar bestudeer hierdie verskynsel vanuit ’n Afrika- (spesifiek Suid-Sotho-) perspektief. Thetela beskryf die navorsing self as ’n fokus op die seks-diskoers of “text and talk about sex” en gaan na hoe geslagsverhoudings en identiteit in sosiale stelsels deur middel van so ’n diskoers gevorm word.

Sanderson (2007) ondersoek die navorsingsveld van handelsmerkdispute wat ook min aandag tot op datum ontvang het. Die artikel is getiteld *Linguistic analysis of competing trademarks*.

Klopper (2009) ondersoek in sy artikel, *The case for cyber forensic linguistics*, forensiese taalkunde uit die oogpunt van die rekenaarwetenskap, wat ook ’n onbekende navorsingsveld in Suid-Afrika is.

Michell (2013) se meesterverhandeling is gerig op ’n soortgelyke onderwerp as die huidige navorsing en ondersoek outeurskapidentifikasie op die sosiale netwerk Facebook. Die titel van die verhandeling is *Investigating the use of forensic stylometric analysis to determine authorship on a publicly accessible social networking site (Facebook)*.

Uit bogenoemde is dit duidelik dat daar wel navorsing oor forensiese taalkunde in Suid-Afrika gedoen is, maar die meeste navorsing fokus op die regsimplikasies van taal. Verder is die meerderheid studies in Engels en is daar slegs een studie oor outeurskap-identifikasie wat in Afrikaans verskyn. Daar is nog geen studies oor outeurskapidentifikasie in Afrikaanse SMS-taal beskikbaar nie. Die leemte in forensies-linguistiese navorsing in Afrikaans is duidelik sigbaar en juis om hierdie rede is die huidige navorsing van belang.

2.3 Outeuridentifikasie

2.3.1 Die ontstaan van ’n subkategorie

Die oorsprong van outeuridentifikasie hou sterk verband met die oorsprong van forensiese linguistiek as dissipline. Die ‘eerste’ forensiese ondersoeke wat in 2.1.1 beskryf is, is in der waarheid gevalle waar navorsers gepoog het om die outeur(s) van teks(te) identifiseer:

- 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.

Dié eerste 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 ná forensiese linguisiek uitgebrei het om ander ondersoekvelde soos forensiese fonetiek, taalregte en taalgebruik in die regskonteks in te sluit dat subkategorieë, soos outeuridentifikasie, onder die sambreelterm ‘forensiese linguistiek’ ontstaan het.

Die metodologiese vordering in outeuridentifikasie is opvallend. Alhoewel metodes soos dié wat deur Mendenhall (1887) gebruik is later ongeldig bewys is (Smith, 1983), het verskeie navorsers intussen probeer om nuwer en akkurater statistiese metodes te ontwikkel. Baie van die aanvanklike metodes is egter nooit as geldig beskou nie (Juola, 2006: 240). Hierdie metodes word deur Holmes (1994) en Stamatatos (s.a.) uiteengesit en enkeles word hier onder kortliks bespreek. Daar moet in gedagte gehou word dat die verskeidenheid metodes vir outeuridentifikasie wat tot dusver ontwikkel is toegeskryf kan word aan die onvermoë van forensiese linguiste om vas te stel wat die perfekte korpusgrootte of die ideale tekslengte vir suksesvolle outeuridentifikasie in elke geval is. Daar is wel spekulasies in navorsing oor wat as die ideale korpus beskryf kan word. Stamatatos (s.a: 21) meen ’n ideale analise kan slegs plaasvind wanneer daar absolute beheer is ten opsigte van die genre en die onderwerp wat gebruik word. Dit is egter ’n onrealistiese verwagting, veral in terme van beheer oor die onderwerp. Die omstandighede en tekste beskikbaar in werklike gevalle laat gewoon nie sulke beheer toe nie (Juola, 2006: 246–247).

2.3.2 Enkele metodes in outeuridentifikasie

Die identifisering van ’n outeur berus op drie soorte bewyse, naamlik eksterne, linguistiese en verklarende bewyse. Eksterne bewyse sluit onder andere die outeur se handtekening en handskrif in. Verklarende bewyse verwys na die bedoeling van die outeur met die skryf of produksie van ’n teks. Binne laasgenoemde groep bewyse word daar ook vasgestel hoe ’n bepaalde teks met

ander tekste wat deur dieselfde outeur geproduseer is, vergelyk. Linguistiese bewyse fokus op die woorde en woordpatrone in die teks of dokument (Corney, 2003: 13–14).

2.3.2.1 Woordlengte en hoeveelheid lettergrepe

Van die eerste ondersoeke wat in outeuridentifikasie gedoen is, het hoofsaaklik op woordlengte, woordverspreiding en die hoeveelheid lettergrepe in woorde gefokus. Soos reeds genoem het Mendenhall (1887) op die eienskap van woordlengte gefokus om moontlike outeurs van bepaalde tekste te identifiseer. Mendenhall se navorsing word in twee referate, naamlik "The Characteristic Curves of Composition" (1887) en "A Mechanical Solution to a Literary Problem" (1901) bespreek. Tydens die eerste studie (1887) het Mendenhall drie eksperimente op drie verskillende groepe tekste en outeurs uitgevoer. Elkeen van hierdie drie eksperimente is uitgevoer om 'n bepaalde aanname ten opsigte van woordlengte as merker van outeurskap te toets. Mendenhall het bevind dat woordlengte nie 'n betroubare metode vir outeuridentifikasie is nie. Hy meen dat hierdie metode moontlik slegs gebruik kan word in sekere gevalle van outeureliminering ("authorship elimination") (Grieve, 2005: 10). Volgens Holmes (1994: 88) is woordlengte as 'n merker van outeurskap onbetroubaar aangesien die leksikon van 'n outeur kontekssspesifiek is. Dit beteken dat die woorde wat in die teks gebruik word en gevolglik die woordlengtes binne die teks van konteks tot konteks sal verskil. Holmes (1994: 88) beskryf hierdie verskynsel soos volg:

> [...] when works which are of various literary genres are compared (or works written during different eras), the differences observed are likely to exceed greatly any distinguishing characteristics which may reliably identify authors. Furthermore, when works in the same literary form by different contemporaneous authors are compared, their word-length distributions may appear so similar that they seem to have been written by the same hand.

Holmes (1994: 88) verwys na Smith (1983) wat Mendenhall se metode vir outeuridentifikasie as so onbetroubaar beskryf dat enige student wat in die veld van outeuridentifikasie studeer die metode heeltemal moet ignoreer. Volgens Grieve (2005: 11) het C.S. Brinegar wel in 1963 die basis van Mendenhall se navorsing in woordlengte as merker van outeurskap gebruik om te probeer vasstel of Mark Twain die ware outeur van die Quintus Curtius Snodgrass-briewe is. Brinegar het Mendenhall se metode effens aangepas deur van die t-toets en die Chi-

kwadraattoets gebruik te maak om die woordspektra met mekaar te vergelyk, eerder as om slegs ’n visuele vergelyking (stilistiese analise) te tref.

Verskeie navorsers het ook die moontlikheid ondersoek dat die gemiddelde hoeveelheid lettergrepe per woord en die verspreiding van hierdie woorde in die teks ’n deurslaggewende faktor kan wees om moontlike outeurs en selfs verskillende teksgenres te identifiseer (Fucks, 1952; Fucks en Lauter, 1965; Brainerd, 1974 en Bruno, 1974). Brainerd (1974) het gevind dat die verspreiding van lettergrepe per woord’n redelik betroubare manier is om outeurskap in sekere literêre genres te bepaal, maar slegs indien outeurs se skryfwyses baie eenvormig is. Brainerd waarsku egter dat sukses met dié metodologie in een genre nie noodwendig beteken dat sukses ook in ’n ander genre behaal sal word nie aangesien skryfstyle verander soos daar van een genre na die volgende beweeg word.

2.3.2.2 Sinslengte

Grieve (2005: 12) verwys na navorsing deur H.T. Eddy (1887) waarin Eddy voorstel dat sinslengte en die verspreiding van sinslengte binne ’n bepaalde teks moontlik ’n meer betroubare manier sal wees om die outeur van ’n teks te bepaal. William Benjamin Smith (1888) stem volgens Grieve (2005: 13) saam met die aanname wat deur Eddy gemaak is, maar fokus in sy navorsing eerder op die gemiddelde hoeveelheid sinne per bladsy. Die analisering van sinslengte as ’n metode om outeurskap te bepaal, is ook deur Yule in 1939 voorgestel (Grieve, 2005: 14). Yule het begin deur die verspreiding van sinslengte in tekste van Francis Bacon, Samuel Taylor Coleridge en Charles Lamb te ondersoek. Hy het bevind dat sinslengte wel ’n karakteriserende eienskap van ’n outeur se skryfstyl is en het op grond van hierdie bevindings gepoog om die tegniek op twee gevalle van omstrede outeurskap toe te pas. Alhoewel Yule bevind het dat sinslengte nie altyd betroubaar is nie, het sy navorsing belangrike vrae laat ontstaan oor wat die definisie van ’n ‘sin’ in statistiese analises is. Grieve (2005: 14-17) verwys na verskeie navorsers wat na Yule se ondersoek ook die moontlikheid van sinslengte as merker van outeurskap oorweeg het.[3] Sommige van hierdie metodes, soos die “compound Poisson distribution for

[3] Die navorsers na wie Grieve verwys, is onder andere: Williams (1940), Wake (1957), Sichel (1974), Morton (1965) en Kjetsaa (1979).

representing sentence-length distributions" is volgens Holmes (1994: 89) 'n redelik betroubare metode.

2.3.2.3 Funksiewoorde

Grieve (2005: 18, 32) verwys na W.B. Smith se 1888 studie as die eerste outeuridentifikasiestudie wat van funksiewoorde gebruik gemaak het om die outeurs van beplaade tekste te identifiseer. Mosteller en Wallace (1964) se studie oor die outeurskap van die *Federalist Papers* word deur Grieve (2005: 34) geïdentifiseer as een van die bekendste voorbeelde van outeuridentifikasie waar, onder andere, die frekwensie van fuksiewoorde in die tekste gebruik is op die outeurs van mekaar te onderskei. Die *Federalist Papers*-studie word beskou as die eerste nie-tradisionele outeurskapindentifikasiestudie (teenoor sogenaamde tradisionele outeurskapidentifikasiestudies wat uitsluitlik op menslike kundigheid gegrond was) aangesien Mosteller en Wallace ook van statistiese toetse gebruik gemaak het om die ondersoek te voltooi (Stamatatos, s.a: 1). Dit is, volgens Stamatatos (s.a: 1), as gevolg van laasgenoemde studie dat navorsing in outeuridentifikasie (van geskrewe tekste) tot in die laat 1990's uitgebrei het om 'n verskeidenheid metodes in te sluit wat Stamatatos beskryf as pogings om die kenmerke wat gebruik word vir die kwantifisering van skryfstyl te definieer. Hierdie tak van die navorsing in outeuridentifikasie staan as 'stilometrie' bekend. Stilometrie word in paragraaf 2.5 bespreek.

Hoewel verskeie studies wat funksiewoorde en kollokasies as merkers vir outeurskapidentifikasie ondersoek tot 'n mate suksesvol was om outeurs te identifiseer, is daar steeds twyfel oor die betroubaarheid van hierdie metode in outeurskapidentifikasie. Holmes (1994: 91) verwys na Oakman (1980) wat tereg meen dat die geval van die *Federalist Papers* die ideale scenario vir outeuridentifikasie is (verskeie lang tekste met min moontlike outeurs) en dat funksiewoorde in sulke gevalle tot positiewe resultate sal lei, maar dat dieselfde metode nie in ander scenario's noodwendig effektief sal wees nie.

2.3.2.4 Rekenaargesteunde metodes: CUSUM-tegniek

Stamatatos (s.a: 1) meen die CUSUM-tegniek (of QSUM-tegniek) is die bekendste voorbeeld van 'n nie-tradisionele, rekenaargesteunde metode van outeurskapidentifikasie wat tydens die 1990's ontwikkel is. Hierdie tegniek is aanvanklik as so betroubaar beskou dat data wat daardeur gelewer is in die howe gebruik kon word en as deskundige getuienis beskou is. Die CUSUM-

tegniek is egter later deur navorsing as onbetroubaar bewys en die data daaruit verkry is as ongeldig bestempel (Holmes en Tweedie, 1995). Die hoofrede waarom die CUSUM-tegniek misluk het nadat dit aanvanklik as betroubaar bestempel is, is die gebrek aan objektiewe evaluasies vir outeuridentifikasiemetodes tydens die vroeë 1990's. Stamatatos (s.a: 2) meen dat daar beperkings op evaluasieprosesse was wat tot onbetroubare evaluering van metodes gelei het. Hierdie beperkings was onder andere:

- Omslagtige data. Sommige 'proefanalises' het volledige boeke ingesluit en die styl van hierdie tekste was nie homogeen nie.
- 'n Klein getal moontlike outeurs (gewoonlik twee of drie).
- Moeilike vergelykings tussen metodes. Tydens die eerste metodologiese toetse en vergelykings was geskikte maatstafdata nie beskikbaar nie.

Sedert die ontwikkeling en uitbreiding van elektroniese tekste soos e-pos, aanlynforums, SMS-boodskappe, blogs en sosiale netwerke is groot vordering in outeuridentifikasiemetodes gemaak. Tegnologiese ontwikkeling en die enorme hoeveelheid tekste wat vandag elektronies beskikbaar is, het daartoe aanleiding gegee dat areas soos inligtingherwinning, masjienleer (*machine learning*) en natuurlike taalprosessering (*natural language processing*) ook gegroei het. Die dekades sedert 2000 kan beskou word as die nuwe era van outeurskapidentifikasie met tegnologie wat die kern van hierdie ondersoeke vorm (Stamatatos, s,a: 2). Outeuridentifikasie het met ander woorde beweeg van 'n rekenaargesteunde veld tot 'n rekenaargesentreerde veld.

2.3.2.5 Rekenaargesteunde metodes: N-gramanalise

Met die ontwikkeling van rekenaarprogramme vir outeuridentifikasiedoeleindes is die n-gram-benadering tot outeuridentifikasie ontwikkel (vergelyk paragraaf 3.4.2.4 en 4.2.3 vir 'n bespreking van die n-gramanalise wat in die huidige navorsing gebruik is). N-gramanalises is oorspronklik gebruik in teks- en taalkategorisering. Dieselfde metode wat in n-gramanalises gebruik word om tekste of tale te kategoriseer kan aangepas word om die veronderstelde idiolek van individue te onderskei. N-gramanalise is 'n baie gewilde metode omdat dit op enige taal of dokument (ongeag hoe 'vreemd' die struktuur mag wees) toegepas kan word solank daardie dokument uit 'n relatief groot hoeveelheid teks bestaan. Volgens Cavnar en Trenkle (1994: 1) is

n-gramanalise gebaseer op die berekening en vergelyking van frekwente n-gram profiele. Ten eerste word die n-gramsisteem gebruik om profiele te bereken uit die opleidingsdata (*training set data*). Hierdie profiele word in verskillende kategorieë geplaas afhangende van die tipe kategorisering. Kategorisering van tekste, tale en outeurs sal vanselfsprekend uit verskillende stelle kategorieë bestaan. Hierna bereken die sisteem ’n profiel vir ’n spesifieke dokument wat geklassifiseer moet word. Laastens word die ‘afstand’ tussen die dokument se profiel en elk van die kategorieë se profiele bereken om vas te stel watter dokumentprofiel en kategorieprofiel die meeste met mekaar ooreenstem. Hierdie twee profiele het met ander woorde die kortste afstand en gevolglik die grootste hoeveelheid ooreenkomste.

’n Belangrike aspek van die n-gramanalise is dat die lengte van ‘*n*’ vasgestel moet word omdat dit bepaal hoeveel leksikale, tematiese en kontekstuele informasie in elke teks vasgevang kan word. Die navorser kan byvoorbeeld van bi-gramme, tri-gramme, ensovoorts gebruik maak om die analises uit te voer. Die n-gramanalise berus met ander woorde op die aanname dat teks bloot uit ’n reeks karakters bestaan. Hierdie karakters plus letterfrekwensies, getalle, leestekens ensovoorts, kan deur middel van ’n n-gramanalise getel word. Die lengte van ‘*n*’ bepaal met ander woorde hoe groot die ‘stukke’ is wat geanaliseer word. Hoe langer die n-gram, hoe beter kan leksikale, tematiese en kontekstuele informasie vasgestel word. Langer n-gramme verhoog ook die dimensies van die analise aangesien honderde en selfs duisende onderskeidende kenmerke geproduseer kan word. Korter n-gramme maak dit moontlik om “sub-word information” vas te stel (Stamatatos, s.a: 6–7). Sulke informasie verwys na enige inligting wat met die lettergrepe in die teks verband hou. Cavnar en Trenkle (1994: 2) stel die verskillende n-gramme vir die woord “text”soos volg voor:

bi-grams: _T, TE, EX, XT, T_

tri-grams: _TE, TEX, EXT, XT_, T_

quad-grams: _TEX, TEXT, EXT_, TX_, T__

Uit die literatuur blyk dit dat William Ralph Bennett (1976) die eerste navorser was wat n-gramanalise as ’n moontlike identifikasiemetode in outeuridentifikasie voorgestel het. Keselj, Fuchun, Cercone en Thomas (2003) het na Bennett se studie in hul eie navorsing verwys en bevind dat n-gramanalises tot die suksesvolle identifisering van outeurs lei wanneer die 1500 tot

2000 mees algemene 6-gramme of 7-gramme vergelyk is (Grieve, 2005: 50). Grieve (2005: 50) waarsku dat die volgende aspekte van die laasgenoemde groep navorsers se studie in ag geneem moet word voordat daar enige finale uitsprake oor die sukses van die n-gramanalise gemaak word:

> While an impressive degree of accuracy was achieved in all of these studies, it must be acknowledged that their results are based on questionable experimental design. Specifically, the sets of possible authors that these researchers considered [...] span such a wide range of dialects, registers, eras and subjects that it is impossible to predict if their method would be capable of distinguishing between a more stylistically and thematically homogeneous set of possible authors [...].

Clement en Sharp (2003) asook Khmelev en Tweedie (2001) het die n-gramanalise in hul navorsing gebruik met wisselende resultate. In die Khmelev en Tweedie studie is bi-gramme (2-gramme) gebruik en ’n Markov-model is opgestel waar elke moontlike opeenvolging van twee letters uit die moontlikheid bestaan dat die tweede letter op die eerste letter volg binne die skryfstyl van ’n bepaalde outeur. Hierdie metode was tot 74.4% suksesvol om tussen 45 outeurs te onderskei, maar die datastel is nietemin te onbetroubaar om algemene gevolgtrekkings ten opsigte van die resultate te maak. Om suksesvolle algemene gevolgtrekkings te maak op grond van die resultate van navorsing wat van die n-gramanalise gebruik gemaak het, moet die eksperimentele ontwerp van die studies korrek uitgevoer word. Dit beteken onder andere dat die datastelle wat gebruik word nie te klein of te uiteenlopend moet wees nie aangesien daar geen algemene gevolgtrekkings op grond van die resultate van die studie gemaak kan word nie.

Ten spyte van die swak ontwerp van sommige studies meen Grieve (2005: 51) dat n-gramanalise wel ’n goeie aanwyser van outeurskap is aangesien n-gramme sensitief is vir stylaspekte wat woorde, kollokasies, leestekengebruik en skryftekens insluit.

2.3.3 Samevatting

Die term ‘stilistiese- en stilometriese analises’ word vandag gebruik om byna al die bogenoemde metodes in outeuridentifikasie saam te vat. Tegnologiese vooruitgang het daartoe gelei dat meer betroubare statistiese analises met behulp van rekenaars gedoen kan word en daarom kan die forensiese linguis ’n verskeidenheid kenmerke in ’n enkele teks identifiseer en daarna statistiese toetse op die teks uitvoer om vas te stel of sy/haar berekeninge of hipoteses korrek is. Dit is

vandag algemeen vir forensiese linguiste om verskeie styleienskappe van die teks stilisties én stilometries te ondersoek en op grond van hul kennis van formele taalkunde gissings oor die teks of die outeur te maak wat daarna statisties getoets word.

Volgens Sierra e.a. (2013) sluit stilometrie onder andere "word and character n-grams, punctuation marks, function words, vocabulary richness, part of speech frequencies, word collocations, grammatical errors (and) word, sentence and paragraph lengths" in. Verder definieer hulle stilometrie baie eenvoudig as "Study (of the) the style of an author through the identification of its features of style (style marker)." Sowel stilistiese as stilometriese analises is ook in die huidige ondersoek uitgevoer.

Ten spyte van al die metodes wat in die verlede getoets is en die vooruitgang wat gemaak is met betrekking tot outeuridentifikasie het die drie scenario's waarmee forensiese linguiste hulle bemoei nie veel verander nie. Hierdie drie scenario's bly relatief konstant, wat beteken dat forensiese linguiste op oplossings vir spesifieke probleme in hulle navorsing kan fokus. Die drie bekendste outeuridentifikasie-scenario's is die volgende:

In die eerste geval moet die linguis kan bepaal of een outeur verantwoordelik is vir al die tekste in 'n stel. Met ander woorde, die linguis moet bepaal of die teks van 'n outeur ooreenstem met ander tekste deur die outeur wat reeds deel uitmaak van 'n stel tekste. Tweedens moet die linguis een verdagte teks kan vergelyk met die tekste van 'n groep moontlike outeurs indien daar nie net een verdagte outeur is nie. Laastens moet die forensiese linguis 'n teks met 'n verdagte kan verbind in die geval waar die verdagte reeds deur eksterne metodes (metodes buite die taalkunde) as 'n moontlike outeur geïdentifiseer is (McMenamin, 2002: 95). Die derde scenario is die algemeenste van die drie. Wanneer die polisie 'n forensiese linguis nader is 'n verdagte gewoonlik reeds geïdentifiseer en die polisie roep dan die forensiese linguis se hulp in sodat 'n bewyslewerende saak teen die verdagte opgestel kan word (Grant, 2010: 514).

Sierra e.a. (2013) meen dat daar 'n vierde scenario is wat tydens forensies-linguistiese ondersoeke kan opduik. In hierdie scenario is daar geen verdagtes nie, net 'n stel tekste. Dit beteken dat die forensiese linguis eienskappe van die teks, onder andere die taalgebruik en styl, moet gebruik om informasie oor die outeur van die teks te verkry en sodoende 'n profiel van die

moontlike outeur te skep (*profiling*). Profielsamestelling[4] word ook deur Olsson (2004: 101–105) as ’n scenario in outeuridentifikasie beskou.

2.3.4 Enkele navorsers in die veld van outeuridentifikasie

Soos duidelik blyk uit die voorafgaande inligting is daar uiteenlopende navorsing wat tot op hede in forensiese linguistiek gedoen is. In outeuridentifikasie is ’n verskeidenheid ondersoeke reeds deur navorsers onderneem.

Met toenemende belangstelling in die idee dat elke individu oor ’n unieke skryf - en praatstyl beskik, is outeuridentifikasiemetodes selfs meer verfyn en aangepas in ’n poging om aan te dui of idiolek wel bestaan, al dan nie. Die insluiting van die konsep ‘idiolek’ in outeuridentifikasie-studies het merkbaar toegeneem vanaf die laat 1990’s. Idiolek is deur verskeie forensiese linguiste as ’n belangrike konsep binne die veld van forensiese linguistiek beskou en gevolglik het ’n groot aantal studies rondom idiolek gesentreer. Idiolek as tema in outeuridentifikasiestudies blyk duidelik sentraal te wees in die werk van verskeie bekende forensiese linguiste.

2.3.4.1 ’n Fokus op idiolek

Malcolm Coulthard (2004) ondersoek in een van sy artikels, getiteld: *Author identification, idiolect and linguistic uniqueness*, die moontlikheid om outeuridentifikasie te gebruik vir die doeleiendes van plagiaatidentifisering. Hier beklemtoon Coulthard (2004: 433) juis die belangrikheid van idiolek en meen dat idiolek gebruik moet word om die teenwoordigheid van plagiaat te bevestig. Met ander woorde, idiolek kan daartoe lei dat ’n duidelike verskil opgemerk word tussen die outeur se taalgebruik en die taalgebruik uit ’n teks wat deur die outeur oorgeskryf of oorgetik is sonder om veranderinge aan te bring, of in sommige gevalle, met minimale veranderinge aangebring. Behalwe vir die identifisering van idiolek meen Coulthard (2004: 434) ook dat daar klem gelê moet word op die persentasie individuele woorde (of leksikale tipes en tekens) wat ooreenstem in twee of meer tekste. Hierdie persentasie is ’n goeie aanduiding van watter teks die oorspronklike teks is en watter tekste plagiaat bevat.

[4] Vergelyk ook punt 2.2 in Tabel 1 (p. 22)

Gerald McMenamin ondersoek ook outeuridentifikasie en spesifiek die rol wat idiolek in outeuridentifikasie speel. McMenamin ondersoek onder andere in sy publikasie *Forensic Linguistics: Advances in forensic stylistics* (2002) die area van forensiese stilistiek. Hier word geskrewe tekste geanaliseer deur op patrone en variasie in die taalgebruik van 'n bepaalde dokument te let (2002: 178–189). Sulke patrone en variasie maak volgens McMenamin (2002: 179) deel uit van die outeur se idiolek en deur middel van idiolek kan outeurskap dan aan verskillende skrywers toegeken word. McMenamin (2002: 139) verwys na 'n skaal wat deur die Scientific Working Group for Forensic Document Examination (SWGDOC) ontwikkel is. Hierdie skaal bestaan uit nege punte wat elkeen 'n persentasie ooreenkomste tussen die verdagte teks en die bekende teks aandui. Die nege op die skaal beteken dat daar aan al die kriteria voldoen word en gevolglik is daar 'n positiewe identifikasie. Dit wil sê al die kenmerke van die verdagte teks stem ooreen met die bekende teks. Met ander woorde die outeur van die bekende teks is ook die outeur van die verdagte teks. Punte ses, sewe en agt beteken steeds dat die outeur van die verdagte teks positief geïdentifiseer is, maar daar is verskeie aspekte van die verdagte teks en die bekende teks wat nie ooreenstem nie. Punt vyf beteken dat daar geen gevolgtrekking gemaak kan word nie terwyl punte twee, drie en vier aandui dat daar 'n hoë moontlikheid is dat die outeur van die bekende teks nie die outeur van die verdagte teks is nie. Punt een dui daarop dat die outeur van die bekende teks nie die outeur van die verdagte teks kan wees nie.

Soos Coulthard, wys McMenamin (2002: 184) daarop dat daar verskeie vrae ontstaan wanneer forensiese stilistiek gebruik word wat daartoe lei dat die betroubaarheid van 'n forensies-stilistiese analise bevraagteken word. Hierdie vrae sluit onder andere in of stilistiek gevestig genoeg is om tot betroubare resultate te lei in forensiese taalkunde en of 'n norm vir variasie in taalgebruik bestaan en bepaal kan word. McMenamin meen dat metodes om te verseker dat outeuridentifikasie meer wetenskaplik is deurentyd ontwikkel en getoets word en dat enige sosiale, geografiese of situasionele norm vasgestel en gebruik kan word om variasie in geskrewe taal te verduidelik.

John Olsson het ook verskeie ondersoeke in outeuridentifikasie voltooi en benader die konsep van idiolek vanuit 'n ander invalshoek. Die term 'linguistiese vingerafdruk' (wat dit vir forensiese linguiste moontlik maak om die outeurs van tekste te identifiseer) word soms as sinoniem vir 'idiolek' gebruik (Coulthard, 2004; Brennan e.a., 2012: 3) en so 'n term is volgens

Olsson (2004: 31) problematies. Alhoewel daar deur Olsson melding gemaak word van verskeie navorsers wat hierdie term gebruik, meen Olsson dat die idee van so ’n linguistiese vingerafdruk ’n mite is en voeg hy by “the proof of its existence is notable for its absence” (Olsson, 2004: 31). Die argument teen die gebruik van die term ‘linguistiese vingerafdruk’ word verder gevoer in paragraaf 2.6 aangesien daar verskeie navorsers is wat Olsson se siening ondersteun.

Chaski (2005) meen ook dat idiolek bestaan en dat dit deur middel van stilometriese analises vasgestel kan word. Chaski se stilometriese analises is gegrond op “readily computable and countable language features, e.g. word length, phrase length, sentence length, vocabulary frequency (and) distribution of words of different lengths”. Chaski fokus egter hoofsaaklik op outeuridentifikasie in tekste wat as digitale getuienis beskou word. Dit is met ander woorde rekenaargeproduseerde tekste wat met misdade soos moord, finansiële misdade, dreigemente en identiteitsdiefstal geassosieer word. In hierdie gevalle doen die outeur sigself voor as iemand anders en gebruik hy of sy digitale tekste om misdade te pleeg. Outeuridentifikasie in SMS-boodskappe kan ook tot ’n mate as digitale getuienis geklassifiseer word aangesien misdadigers selfoonboodskappe vir dieselfde doeleindes kan gebruik, naamlik om hulle identiteit te beskerm wanneer hulle misdade gepleeg het. Na aanleiding van drie sake waarna Chaski (2005: 2) verwys kom sy tot die gevolgtrekking dat outeuridentifikasie nie as ’n aparte ondersoekveld kan funksioneer nie, maar eerder saam met reeds gevestigde forensiese prosedures soos die biometriese analise van die sleutelbordgebruiker gebruik moet word om sekerheid oor die outeur te verkry. Die doel van Chaski (2005) se ondersoek is uiteindelik om ’n kwantitatiewe metode te toets wat volgens haar 95% akkuraatheid behaal met outeurskapidentifikasie. Hierdie metode staan as die sintaktieseanalise-metode bekend. Chaski (2005: 3) meen die verskil tussen die sintaktieseanalise metode en ander stilometriese analises is die feit dat eersgenoemde oor “linguistic sophistication and foundation in linguistic theory” beskik.

2.3.4.2 ’n Fokus op kort tekste

Outeuridenitifikasie in digitale tekste is nie ongewoon nie, maar outeuridentifikasie in kort digitale tekste soos SMS-boodskappe en ander nuwe vorme van kommunikasie soos *tweets* en *status updates* is minder algemeen. Soos Chaski is daar verskeie navorsers wat op digitale, en spesifiek korter digitale tekste, fokus. Tim Grant (2010, 2012) is een navorser wie se werk oor outeuridentifikasie in kort tekste verskeie moontlikhede en probleme uitlig en bespreek. Grant se

artikel getiteld *Txt 4n6: Idiolect free authorship analysis?* (2010) bespreek die kwessie van betroubaarheid in outeurskapidentifikasie, spesifiek met betrekking tot idiolek. Grant fokus met ander woorde gelyktydig op twee problematiese aspekte van outeuridentifikasie, naamlik die lengte van tekste (kort digitale tekste) en die aanwesigheid van idiolek. Grant argumenteer dat idiolek nie so maklik waarneembaar is soos sommige navorsers aanvoer nie en dat die rede vir idiolek in 'n bepaalde teks altyd verklaarbaar moet wees. Dit is volgens Grant (2010: 9) juis die verklaring van idiolek wat problematies is vir die linguis. Om hierdie probleem op te los, stel Grant (2010: 13) voor dat die linguis 'n gekombineerde benadering volg waar kennis uit die formele taalkunde en kognitiewe taalkunde geneem word om idiolek in 'n bepaalde teks te verduidelik. Hierdie menings en voorstelle van Grant word later in meer detail bespreek.

McLeod en Grant (2012) ondersoek die problematiek van bondige tekste asook die algemene opvatting dat outeuridentifikasiemetodes aangepas moet word vir korter tekste. Die meerderheid outeuridentifikasiemetodes wat in vorige navorsing gebruik is, is gegrond op langer, en hoofsaaklik literêre, tekste. Sulke metodes is egter nie geskik om korter tekste te analiseer nie. McLeod en Grant (2012: 221) het 'n uitbreiding op die *Jaccard's co-efficient*-metode gebruik nadat hulle verskeie navorsers se metodes ondersoek en bespreek het (McLeod et al., 2012: 211–213). Laasgenoemde metode is tot die Delta-S-metode herdoop. McLeod en Grant (2012: 221) meen: "the approach reported on here advances the state of the art in terms of the size of the message to which authorship analysis can be applied." Hulle erken egter dat toekomstige verbeteringe moontlik en nodig is om die metode selfs meer akkuraat te maak.

Mohan, Baggili en Rogers (2010) ondersoek ook outeuridentifikasie in SMS-boodskappe, maar maak gebruik van n-gramanalise. Mohan e.a. (2010) probeer in hul studie vasstel of die beperktheid van die teks in SMS-boodskappe steeds geanaliseer kan word soos wat die geval is by langer tekste. Die studie fokus op die vasstel van outeurskap in hoofsaaklik twee gevalle:

> a) For cases in which suspects have been identified or b) For cases in which suspects have not been identified. In the first, the suspects are limited to the forensic case under investigation. In the second, the investigator may have a database that contains a number of SMS messages with their respective N-grams, in which the examiner may try to find a correlation between the messages and an author.

Hulle maak spesifiek van n-gramanalise gebruik omdat n-gramme, soos reeds bespreek, nie taalspesifiek is nie en gebruik kan word om tale wat uit simbole eerder as letters bestaan, soos Japanees, te analiseer. In die studie is bevind dat n-gramanalise ’n akkuraatheid van tussen 65% en 72% het. Selfs wanneer die hoeveelheid SMS-boodskappe relatief min is en die hoeveelheid moontlike outeurs redelik baie, kan hierdie metode steeds ’n akkuraatheid van sowat 70% tot gevolg hê. Volgens Mohan e.a. (2010: 9) is verfyning van hul tegniek ook nodig en moet verdere studies in outeuridentifikasie in SMS-boodskappe aangepak word, aangesien navorsers nog verskeie veranderlikes in SMS-boodskappe moet ondersoek.

Ishihara (2011) se ondersoek na outeuridentifikasie in SMS-boodskappe kombineer die n-gramanalise en die *Log-likelihood Ratio (LR)-based evidence evaluation*. Daar is volgens Ishihara (2011: 48) nog geen empiriese navorsing gedoen ten opsigte van outeuridentifikasie in SMS-boodskappe binne die raamwerk van LR nie en daarom kan die akkuraatheid van laasgenoemde raamwerk vir outeuridentifikasie in SMS-boodskappe nog nie bepaal word nie. Ishihara ondersoek die tekort aan navorsing binne die LR-raamwerk en bepaal deur middel van navorsing of outeurs van SMS’e wel geïdentifiseer kan word en hoe akkuraat hierdie resultate is. Tweedens probeer Ishihara bepaal hoe sterk, of tot watter graad, die bewyse vir individuele outeurs in SMS’e opgemerk kan word, en derdens word ondersoek ingestel na die betroubaarheid van bewyse afhangende van die hoeveelheid SMS’e beskikbaar vir ontleding. Ishihara (2011: 54–55) se gevolgtrekking is dat die akkuraatheid van outeuridentifikasie in SMS-boodskappe afhang van die hoeveelheid boodskappe wat beskikbaar is vir analise. Verder meen Ishihara, soos ook opgemerk in voorafgaande navorsing, dat daar steeds verbeterings in die veld van outeuridentifikasie van SMS-boodskappe nodig is. Daar word voorgestel dat verskillende metodes tydens verskillende fases van outeurskapidentifikasie toegepas word, byvoorbeeld: “[...] focus on specific words/expressions which are high in idiosyncrasy, pre-process of messages prior to modelling, different modelling techniques, different LR calculations”.

Ishihara (2011: 55) wys ook daarop dat sy studie slegs gebruik maak van SMS-boodskappe wat deur Singapoerese getik is en daarom sal die databasis wat opgestel is nie gebruik kan word in navorsing wat op ander tale fokus nie, tensy laasgenoemde ’n vergelykende studie is.

Uit die bespreking is dit duidelik dat studies in outeuridentifikasie oor ’n wye spektrum van tekste (digitale en gedrukte tekste) strek. Die bespreking sluit uiteraard nie al die studies in wat

tot op hede in outeuridentifikasie gedoen is nie, maar dit dui daarop dat twee kernkonsepte wat in die huidige navorsing van belang is, idiolek en bondige (digitale) tekste, reeds verskeie kere ondersoek is. Uit die literatuur is dit duidelik dat die term ‘idiolek’ tot verskeie probleme lei aangesien dit nie moontlik is om in elke scenario aan te voer dat idiolek teenwoordig is nie. Die hoeveelheid tekste en die lengte van die tekste bepaal ook hoe akkuraat gevolgtrekkings ten opsigte van die teenwoordigheid van idiolek is. Klein hoeveelhede kort tekste kan daartoe lei dat die navorser onder die indruk kom dat idiolek wel in ’n bepaalde datastel teenwoordig is terwyl die ‘idiolek’ wat die navorser identifiseer bloot die gevolg van toeval is en moontlik sal verdwyn sodra daar meer tekste tot die navorser se beskikking is. Sukses is egter behaal in outeuridentifikasiestudies van bondige tekste, veral met n-gram- en *log-likelihood*-metodes. Dit is belangrik om te onthou dat elke scenario verskil en dat metodes in sekere scenario’s meer suksesvol sal wees as in ander. Daarom word daar aanbeveel dat outeuridentifikasiestudies, veral wanneer kort tekste geanaliseer word, van meer as een metode gebruik maak om die akkuraatheid van die resultate te verhoog.

2.4 Outeuridentifikasie en die hof

Een van die grootste probleme met forensies-linguistiese metodes is die feit dat verskeie howe (onder andere in die VSA en sommige dele van Australië en Brittanje) nie bewyse wat deur middel van hierdie metodes verkry word as geldig beskou sonder motivering nie. Dit beteken dat bewyse wat deur die forensiese linguis verkry is bekend gemaak kan word, maar dit beteken nie dat hierdie bewyse in die hof gebruik gaan word of erken sal word nie. Selfs al tree die forensiese linguis self as ’n deskundige getuie op kan sy/haar getuienis steeds verontagsaam word. Chaski (2001: 1) beklemtoon dat forensies-linguistiese bewyse eers as geldig oorweeg word indien die metodes wat gebruik is om, byvoorbeeld, die outeur van ’n teks te identifiseer, empiries geëvalueer is. Dit is egter slegs één van die vereistes vir geldige bewyse in die hof.

2.4.1 Die rol van die (forensiese) linguis in hofsake

Coulthard (2010: 473) meen daar is twee soorte getuies in ’n hofsaak: die eerste soort getuie is iemand wat persoonlik by die saak of die individue betrokke is terwyl die tweede soort getuie

geen persoonlike betrokkenheid by die saak het nie. Hierdie getuie staan as ’n deskundige getuie bekend. Wanneer forensiese linguiste getuig, getuig hulle as deskundige getuies.

Linguiste kan om verskeie redes deur die hof of ’n ondersoekspan genader word. Coulthard (2010: 474) noem sake waar linguiste ingeroep is as deskundige getuies. Hierdie sake wissel van handelsmerkdispute (Shuy, 2002 en Gibbons, 2003) en die ‘eienaarskap’ van woorde in ’n plagiaatsaak (Turell, 2004) tot die vasstelling van die outentisiteit van twee dokumente in ’n moordsaak (Coulthard, 2002). Olsson (s.a.: 3) verwys na die dienste wat linguiste aan die hof lewer as “linguistic intelligence work”. Sulke werk sluit ook die analisering van SMS-boodskappe, dreigemente en losprysbriewe in. Linguiste kan ook gevra word om veronderstelde selfmoordbriewe te analiseer om vas te stel of die teks wel ’n werklike selfmoordbrief is. Die polisie kan die linguis ook vra om ’n opinie te lewer oor ’n bepaalde teks of ’n oudioband. Dit is onwaarskynlik dat enige opmerkings wat die linguis maak ten opsigte van die geloofwaardigheid van die teks(te) as betroubare bewyse in die hof aanvaar sal word. Om hierdie rede word linguistiese analises gewoonlik tot die eerste fase (die ondersoekende fase) van die ondersoek beperk (Olsson, s.a.: 3).

Tydens die verhoorfase van ’n ondersoek word linguiste gewoonlik genader om analises uit te voer betreffende outeuridentifikasie, betekenisinterpretasies, analises van dreigemente of die herkoms en samestelling van tekste. Die aard van die saak, krimineel of siviel, sal die aanvaarbaarheid van sulke ‘bewyse’ bepaal (Olsson, s.a.: 3).

Indien ’n verdagte of beskuldigde skuldig bevind word, word appèl gewoonlik deur die verdediging aangeteken. Tydens hierdie fase kan die linguis ingeroep word om ’n dispuut oor die bewoording, interpretasie of outeurskap in ’n verklaring of skulderkenning op te klaar. Dit kan ook gebeur dat daar ’n nuwe interpretasie van ’n forensiese teks (selfmoordbrief of losprysbrief) is wat ná die skuldigbevinding duidelik geword het en wat die linguis moet verifieer (Olsson, s.a.: 3).

Solan (2010: 400) waarsku dat die forensiese linguis bewus moet wees van twee aspekte van hofsake waarmee hy/sy moontlik gekonfronteer sal word:

> [...] the brutality of the adversarial system, including snide and personal attacks on individuals working within standard scientific paradigms, and a broad concern that the forensic identification sciences lack adequate scientific foundation.

In ’n poging om laasgenoemde probleme vir die forensiese linguis op te klaar, stel Solan (2010: 404) voor dat forensiese linguiste vaardigheidstoetse aflê. Die metodologieë wat gebruik word in die ontwikkeling van sulke vaardigheidstoetse moet buite die hof getoets word aangesien dit die omgewing is waar forensiese linguiste die meeste van hul tyd spandeer. Vaardigheidstoetse sal die ontwikkeling van protokol en die opstel van verslae wat die kundigheid van die forensiese linguis en die geldigheid van die bewyse beklemtoon (soos by 2.4.2.1 bespreek) komplementeer. Die ontwikkeling van geldige vaardigheidstoetse is volgens Solan (2010: 405) problematies aangesien dit moeilik is om materiaal te ontwikkel wat relevant is tot werklike, alledaagse forensiese probleme. In die geval waar ’n dokument slegs twee moontlike outeurs het is dit redelik eenvoudig om een van die outeurs as die ware outeur van die teks te identifiseer, aangesien die verwerping van een outeur as die ware outeur van die teks in hierdie geval beteken dat die ander outeur die ware outeur van die teks is. In ander gevalle word daar van die forensiese linguis verwag om vas te stel of ’n verdagte ’n bepaalde teks geproduseer het. In hierdie geval is daar nie ander verdagtes nie en die vraag is gevolglik hoeveel moontlike outeurs verwerp moet word voordat die linguis met sekerheid kan sê dat die verdagte wel die teks geproduseer het of nie. Sulke probleme sal eers opgelos moet word voordat vaardigheidstoetse betekenisvol sal wees.

2.4.2 Forensies-linguistiese bewyse: ’n kort oorsig

2.4.2.1 Die Verenigde State van Amerika

Volgens die Amerikaanse howe word forensies-linguistiese bewyse dikwels nie in ag geneem nie omdat hierdie bewyse nie die ‘Daubert-toets’ slaag nie. Volgens die Daubert-toets moet bewyse onder andere aan die volgende vereistes voldoen voordat dit as geldig beskou kan word: (1) Die deskundige getuie moet oor bevredigende kennis van die veld (in hierdie geval word na forensiese linguistiek verwys) beskik. Hierdie kennis moet uit ondervinding sowel as opleiding en onderrig in die veld bestaan. Verder moet die getuie ook as ’n persoon van aansien in sy/haar

akademiese- of ander portuurgroep beskou word. (2) Die tegniek wat gebruik is om die bewyse te verkry moet empiries getoets wees. Dit moet ook weerlegbaar wees (Chaski, 2001: 1). (3) Die tegniek moet reeds blootgestel wees aan portuurbeoordeling en publikasie. (4) Die deskundige moet die foutgrens van hierdie tegniek kan aandui. (5) Die tegniek moet met duidelikheid aan die hof verduidelik kan word sodat die basiese konsep agter die tegniek deur almal (diegene in die hof teenwoordig sowel as die algemene publiek) begryp kan word (Olsson, 2004: 41–42). Nie alle howe in die VSA erken die Daubert-toets of neem al die vereistes van die Daubert-toets in ag nie, maar in gevalle waar die Daubert-toets wel erken word, moet bewyse steeds aan die meerderheid vereistes voldoen om as geldig beskou te word. Chaski (2001: 1–2) verwys na die hofsaak *United States v. Van Wyk, 83 F. Supp. 2d 515, D.N.J., 2000* waar die hof bevind het dat die staat se bewyse slegs aan een van die Daubert-toetsvereistes voldoen. Om hierdie rede het die hof besluit om nie die staat se getuienis te aanvaar nie (Chaski, 2001: 2):

> Although Fitzgerald employed a particular methodology that may be subject to testing, neither Fitzgerald nor the Government has been able to identify a known rate of error, establish what amount of samples is necessary for an expert to be able to reach a conclusion as to probability of authorship, or pinpoint any meaningful peer review. Additionally, as Defendant argues, there is no universally recognized standard for certifying an individual as an expert in forensic stylistics.

Die Daubert-toets vervang die Frye-toets wat aanvanklik in howe gebruik is om die geldigheid van bewyse te toets. Die Frye-toets het ook bekendgestaan as die "general acceptance test" en hierdie toets het op die uitgangspunt berus dat "providing a method had acceptance from the scientific community it could be held to be valid in a court of law" (Olsson, 2004: 41).

In die VSA is daar tans drie groepe tegnieke wat in outeuridentifiaksie gebruik word en waarvan die resultate in die hof aanvaar word indien die tegnieke empiries getoets is. Hierdie tegnieke word deur Chaski (2001: 2–3) soos volg opgesom:

> In the first group there are two techniques – syntactically classified punctuation and syntactic analysis of phrase structure – which withstand the scrutiny of experimental testing and statistical analysis. [...] In the second group are several techniques – sentential complexity, vocabulary richness, readability, content analysis – which quantify linguistic patterns, and are amenable to statistical testing. [...] In the third group are 'forensic stylistic' techniques – spelling errors, puntuation errors, word form errors, grammatical errors – which are rooted in handwriting identification and prescriptive grammar.

Op 1 September 2008 het die Council for the Registration of Forensic Practitioners forensiese linguistiek as ’n spesialiteitsveld erken. Die erkenning het gehelp om vertroue in forensiese linguistiek te bevorder (Mitchell, 2008). Mitchell (2008) verwys na dr. Tim Grant (professor in forensiese linguistiek by Aston universiteit) wat meen dat dié erkenning aandui dat forensiese linguistiek al hoe meer as ’n wetenskaplike veld beskou word.

In 2009 stel die National Research Council of the National Academics (NRC) ’n verslag saam wat handel oor die status van forensiese identifikasie. Hierdie verslag is toepaslik op alle forensiese prosedures en is getiteld: *Strengthening forensic science in the United States: A path forward*. Solan (2010: 398) verwys na die volgende bevindings in die verslag (NRC 2009: S-7) wat ook van belang is vir forensies-linguistiese bewyse:

> Two very important questions should underlie the law’s admission of and reliance upon forensic evidence in criminal trials: (1) the extent to which a particular forensic discipline is founded on a reliable scientific methodology that gives it the capacity to accurately analyze evidence and report findings and (2) the extent to which practitioners in a particular forensic discipline rely on human interpretation that could be tainted by error, the threat of bias, or the absence of sound operational procedures and robust performance standards. … Unfortunately, these important questions do not always produce satisfactory answers in judicial decisions pertaining to the admissibility of forensic science evidence proffered in criminal trials.

Die verslag stel voor dat baie navorsing in ’n bepaalde terrein nodig is om die beperkings van die metodes of prosedures vas te stel en sodat die moontlikheid van variasie en vooroordeel vasgestel kan word (vergelyk die verwysing na Grant (2008) se databasis onder 2.4.2.2). Solan (2010: 405) verwys na Olsson (2003) wat meen dat groot hoeveelhede navorsing in die veld van outeuridentifikasie van SMS-boodskappe gedoen moet word om die resultate van outeuridentifikasie in hierdie spesifieke ondersoekterrein meer geloofwaardig te maak.

Solan (2010: 405) meen dat die diversiteit wat tussen verskillende individue se skryfstyle en selfs binne een persoon se skryfstyl voorkom van groot belang is in sake wat met outeuridentifikasie te make het:

> This diversity should be taken into account in determining – based on research from corpora that are already available or which are gathered for research purposes – how predicative of co-authorship and non-authorship these conventions are.

Verder word voorgestel dat daar in die subkategorieë van die forensiese linguistiek gefokus word op die tipe dokumente wat tydens ondersoeke geanaliseer word en dat die argumente vir of teen ’n bepaalde posisie gestandardiseer word sodat daar eenstemmigheid kan wees in die regsgemeenskap (Solan, 2010: 405). Indien daar forensiese linguiste is wat voortdurend by een bepaalde soort forensiese ondersoek betrokke is, sal dit vir die toekoms van forensiese linguistiek van groot waarde wees indien so ’n persoon sy/haar ondervindings en benaderings kan opteken vir toekomstige gebruik deur ander forensiese linguiste in dieselfde veld.

Volgens die NRC-verslag (NRC 2009: S-6) is laasgenoemde navorsing nodig, maar skiet dit tans tekort veral in forensiese dissiplines wat staatmaak op “subjective assessments of matching characteristics”. Die verslag stel voorts voor dat sulke forensiese dissiplines streng protokol ontwikkel wat subjektiewe interpretasies sal verhoed. Navorsing en evaluasieprogramme wat aan streng reëls en kriteria voldoen, moet ook ontwikkel word (Solan, 2010: 389).

Solan (2010: 398–399) waarsku dat enigiemand wat in forensiese velde werk en forensiese analises doen versigtig moet wees vir die sogenaamde “confirmation bias”. So ’n vooroordeel kom gewoonlik voor wanneer die polisie byvoorbeeld die forensiese linguis inlig dat hulle seker is die verdagte is skuldig en dat die linguis dit net moet bevestig. In so ’n geval is dit veral belangrik dat linguiste objektief bly wat betref hulle analises en gevolgtrekkings, aangesien die gevolgtrekkings wat gemaak word aan ondersoek onderworpe sal wees.

2.4.2.2 Australië

In die Australiese howe word daar nie verwag dat forensies-linguistiese bewyse aan die Daubert-vereistes voldoen nie, maar die howe beklemtoon dat bewyse wel relevant en betroubaar moet wees. Hierdie reël staan as “the relevance and reliability rule” bekend (Olsson: 2004: 43). Deskundige getuies moet verseker dat hulle die bevindinge kan kwalifiseer – deur aan te dui waar hulle glo die bewyse onvolledig of onakkuraat is – en getuies moet verder redes kan verskaf vir hulle opinies. Dit beteken dat die Australiese howe wel die opinies van deskundige getuies in ag neem. Olsson (2004: 43) noem egter dat daar steeds verskeie howe in Australië is wat skepties is oor linguistiese bewyse omdat daar ’n algemene opvatting bestaan dat elke individu oor die nodige linguistiese kennis beskik om hul eie opinies te vorm en dat die studie van taal nie

werklik ’n tegniese ondersoekterrein is nie. Laasgenoemde is ’n algemene probleem regoor die wêreld en juis daarom meen Grant (2008) dat dit nodig is om ’n gespesialiseerde taaldatabasis op te stel waaruit linguistiese data gekwantifiseer kan word. Grant het self ’n gespesialiseerde taaldatabasis opgestel met meer as 8 000 tekste en elke teks is deur middel van statistiese metodes geanaliseer. Grant (2008) maak van hierdie databasis gebruik omdat hy glo dat dit belangrik is vir die forensiese linguis om sy/haar kundigheid in die vakgebied te demonstreer. Hierdie kundigheid moet duidelik meer as die vermoë van die gemiddelde jurielid wees.

2.4.2.3 Engeland en Wallis

Die regstelsel in Engeland en Wallis verskil van die regstelsel van die VSA en Australië. Tydens die vroeë 1990’s is Lord Woolf aangestel as ‘opsiener’ oor die hervorming van die siviele hofstelsel in Engeland en Wallis. Die rede vir die hervorming was dat daar bekommernis ontstaan het dat die reg besig was om ontoeganklik te word vir lede van die publiek. Volgens Woolf beteken die insluiting van kenners in ’n hofsaak dat die koste van so ’n hofsaak opgejaag word en dit lei daartoe dat die publiek nie meer enige kwessies deur middel van hofsake wil laat uitklaar nie (Olsson, 2004: 43). Woolf het voorgestel dat slegs een kenner vir albei partye aangestel word en dat, in gevalle waar daar nie op een kenner besluit kan word nie, die hof self die kenner moet aanstel. Die rede hiervoor is volgens Woolf dat kenners ‘ekonomies’ moet wees en dat enige meningsverskille wat tussen twee of meer kenners kan ontstaan so gou as moontlik geëlimineer moet word. Woolf het nooit enige wetenskaplike standaarde vir bewyse voorgestel soos in die geval met die Daubert-toets nie (Olsson, 2004: 44).

2.4.2.4 Suid-Afrika

In Suid-Afrika is daar ook, soos in Amerika en Australië, vereistes wat met die geldigheid van elektroniese bewyse verband hou. Hierdie vereistes is van belang vir die huidige navorsing aangesien outeuridentifikasie van Afrikaanse SMS-boodskappe binne hierdie regsomgewing uitgevoer sal word. Volgens Watney (2009: 2) word die geldigheid van elektroniese bewyse deur drie stappe binne die raamwerk van die Suid-Afrikaanse bewysreg bepaal.

Ten eerste moet vasgestel word watter tipe elektroniese bewyse gebruik word. Die bewyse word in een van die volgende kategorieë verdeel: dokumentêre bewyse of egte bewyse. **Dokumentêre bewyse** beteken dat die bewyslewerende gewig deur die betroubaarheid van ’n *persoon* bepaal

word. Met ander woorde, die data is deur ’n persoon ingevoer op die rekenaar/masjien en die rekenaar/masjien het nie die data geprosseseer of verander nie. Die bewyslewerende gewig van **egte bewyse** word deur die betroubaarheid van die *masjien* se hardeware en sagteware bepaal. Dit beteken dat die data deur ’n persoon ingevoer is op die rekenaar/masjien en die rekenaar/masjien het die data verander na ’n ander formaat as dit waarin die data oorspronklik ingevoer is (Krige, 2012). Egte bewyse word ook beskou as dokumente wat deur ’n rekenaar/masjien gegenereer is sonder die persoonlike betrokkenheid van ’n individu. Met ander woorde, die persoon het net die elektroniese sisteem geaktiveer en die sisteem neem dan self data op en stoor dit. Data word gegenereer deur middel van die sisteem se sagteware en nie ’n persoon wat data invoer nie (Krige, 2012). Laasgenoemde twee kategorieë verwys slegs na bewyse wat as ‘dokumente’ geklassifiseer kan word. Dit sluit enige boek, kaart, plan, tekening, foto, pamflet, lys, brief, rekord en plakkaat in (Watney, 2009: 5; Krige, 2012). Watney (2009: 5) noem nietemin dat die term ‘dokument’ ook kan verwys na “any device by means of which information is recorded or stored.” Volgens Streicher (2010: 2) word SMS’e ook as dokumente beskou volgens die Wet op Elektroniese Kommunikasie en Transaksies. SMS-boodskappe kan as toelaatbare bewyse in beide kriminele en siviele sake gebruik word. Sulke bewyse het egter tot op hede net op die inhoud van die SMS-boodskap gefokus nadat daar bepaal is dat die SMS-boodskap van ’n bepaalde selfoon gestuur is. Die gebruik van outeuridentifikasie om die identiteit van die outeur van ’n SMS-boodskap(pe) te bevestig is nog nie in Suid-Afrikaanse howe as bewys gebruik nie.

In die tweede plek is dit van belang om die aanvanklike geldigheid of outentisiteit van elektroniese bewyse te bepaal. Watney (2009: 8) beskryf laasgenoemde vereiste soos volg:

> The South African law of evidence requires that anyone who wants to use a document as evidence has to satisfy the court that it is authentic, in other words, that the document is what it claims to be. Due to its high degree of volatility, electronic evidence can easily be manipulated, altered or damaged after its creation and therefore authenticity must be proved.

Derdens word die bewyslewerende gewig van die bewyse vasgestel. Die bewyslewerende gewig van elektroniese bewyse word deur die volgende kriteria bepaal: (1) die betroubaarheid van die manier waarop die data gegenereer, gestoor en gekommunikeer is; (2) die betroubaarheid van die manier waarop die integriteit/volledigheid van die data behou is; (3) die wyse waarop die

skepper van die data geïdentifiseer is; en (4) enige ander relevante faktore (Watney, 2009: 10). Watney (2009: 10) verwys na Hofman (2006) wat meen dat die hof, behalwe vir die laasgenoemde kriteria, ook van kenners gebruik moet maak om die tegniese prosedures agter die verkryging, verwerking en stoor van die elektroniese data te verduidelik sodat besluite rondom die bewyslewerende gewig van die bewyse makliker geneem kan word.

Die geldigheid van die bewyse word egter eers werklik in ag geneem wanneer daar reeds bepaal is dat die bewyse wettiglik verkry is. Situasies waar bewyse op ’n onkonstitusionele wyse verkry is kom nogtans voor en daar is gevalle waar sulke bewyse wel toegelaat word aangesien regverdigheid, volgens Watney (2009: 3), in sommige gevalle vereis dat “evidence, albeit obtained unconstitutionally, nevertheless be admitted.”

Ten spyte van laasgenoemde stappe is daar ongelukkig tans, binne die Suid-Afrikaanse reg, slegs enkele prosedures in plek wat die versameling, bewaring en aanbieding van elektroniese bewyse vir die doeleindes van strafsake monitor (Watney, 2009: 2). Dit beteken dat die Suid-Afrikaanse reg, betreffende elektroniese bewyse, soms belemmer word.

2.4.3 Die toelaatbaarheid van SMS-boodskappe as elektroniese bewyse

Die insluiting van selfoonbewyse, en spesifiek SMS-boodskappe, in hofsake is nie so eenvoudig soos uit voorafgaande bespreking mag blyk nie. Dit wil voorkom asof die grootste probleem met die toelaatbaarheid van SMS-boodskappe as elektroniese bewyse die feit is dat dit soms baie moeilik is om die outentisiteit van die inhoud en die identiteit van die outeur te bepaal (Davis, 2013). In verskeie Amerikaanse state is laasgenoemde probleem juis die huidige fokus en word verskillende voorstelle gemaak om die toelaatbaarheid van SMS-boodskappe as bewyse in hofsake te bevorder, aangesien die aantal misdade wat met selfoondata verbind word toeneem (Stevenson, 2008; Costello, 2013). In Pennsylvania het die appèlhof besluit dat daar nie aparte standaarde benodig word vir SMS-boodskappe of IM-boodskappe (*Instant Messages*) wat as bewysstukke gebruik word nie. Die appèlhof verdedig die rede vir hul besluit soos volg (Stevenson, 2008):

> We see no justification for constructing unique rules for admissibility of electronic communications such as instant messages; they are to be evaluated on a case-by-case basis as any other document to determine whether or not there has been an adequate foundational showing of their relevance and authenticity.

Die ‘case-by-case’-metode as ’n basis vir die bepaling van toelaatbaarheid van SMS-boodskappe en ander elektroniese boodskappe blyk ’n populêre opsie in Amerikaanse howe te wees. Volgens die *Federal Rules of Evidence 901 (b)(4)* kan die outentisiteit van bewyse bepaal word deur “distinctive characteristics” of deur die “appearance, contents, substance, internal patterns, or other characteristics of the item” te bestudeer (Costello, 2013). Dit beteken dat die bestudering van SMS-boodskappe voortdurend eiesoortige eienskappe sal oplewer wat dan volgens die ‘case-by-case’-metode geëvalueer kan word. Costello (2013) noem dat eiesoortige eienskappe van SMS-boodskappe en ander elektroniese boodskappe reeds in hofsake in Noord-Carolina gebruik is om die outentisiteit van die boodskappe te bepaal. Een van die mees algemene eiesoortige eienskappe waarop gefokus word, is die inhoud van die SMS-boodskap wat gewoonlik met die misdaad verband hou. Verder is daar ook in die meerderheid van die gevalle omstandigheidsgetuienis wat die saak teen die verdagte ondersteun.

Costello (2013) meen dat die proses om SMS-boodskappe as geldige bewyse te gebruik in twee fases geskied. In die eerste plek moet tegniese hulp ingeroep word om sonder twyfel te bevestig dat die spesifieke SMS-boodskap van ’n bepaalde nommer gestuur is en tweedens word omstandigheidsgetuienis gebruik as bewysmiddel dat die verdagte wel die persoon is wat die SMS-boodskap gestuur het.

In Sri Lanka word SMS-boodskappe ook as toelaatbare bewyse beskou en word dit in die kategorieë van ‘primêre bewyse’ (die oorspronklike SMS’e) en ‘sekondêre bewyse’ (fotokopieë of skermbeelde van die oorspronklike SMS’e) verdeel (Abeywickrema, 2008). Volgens die Wet op Elektroniese Transaksies in Sri Lanka word die terme ‘kommunikasie’, ‘databoodskap’ en ‘elektroniese dokument’ soos volg gedefinieer (Abeywickrema, 2008):

> Section 26 of the Electronic Transactions Act interprets the word ‘communication’ as “any statement, declaration, demand, notice or request, including an offer and the acceptance of an offer that a person is required to make or chooses to make in connection with an electronic transaction within the meaning of this Act”. Under the same section ‘data message’ means information generated, sent, received or stored by electronic, magnetic, optical or other similar means. ‘Electronic’ means information generated, sent, received or stored by electronic, magnetic, optical, or similar capacities regardless of the medium. The term ‘electronic document’ is interpreted as to include documents, records, information, communications or transactions in electronic form.

Hierdie definisies maak dit volgens die Sri Lankaanse howe moontlik om SMS-boodskappe as geldige bewyse in hofsake toe te laat.

Davis (2013) waarsku egter teen die toelating van SMS-boodskappe as geldige bewyse in die hof sonder die nodige ondersoek van die boodskappe en die selfoon self. Die bewysreël bepaal gewoonlik dat dokumente as ongeldig beskou moet word indien dit nie die oorspronklike dokumente is nie. Fotokopieë en selfs skermbeelde van die SMS-boodskappe wat as bewyse in die saak gebruik word kan maklik gemanipuleer word deur programme soos Photoshop. Verder is fotokopieë nie altyd 'n betroubare weergawe van al die data op die bepaalde selfoon nie, aangesien slegs die boodskappe wat van belang is vir die saak gewoonlik gefotokopieer word. Dit beteken dat die verdediging kan aanvoer dat die inligting wat die ander kant van die saak stel of die onskuld van die verdagte bewys, uitgelaat is. In so 'n geval sal die selfoon moontlik aan 'n visenteringslasbrief onderworpe wees. Gelukkig is daar met die hulp van moderne tegnologie maniere om wel die betroubaarheid van die SMS-boodskap te verifieer. Davis (2013) stel voor dat 'n forensiese ondersoek altyd eers op die betrokke selfoon en SMS-boodskappe uitgevoer word voordat hierdie bewyse in die hof gebruik word. Die parlement in Suid-Afrika het egter besluit dat laasgenoemde probleme nie noodwendig in alle gevalle tot die ongeldigheid van bewyse lei nie (Hershensohn, 2005: 8):

> In terms of section 15 of the *Electronic Communications and Transactions Act 25 of 2002*, it is provided that the rules of evidence must not be applied to deny the admissibility of a data message purely because it is constituted by a data message, or on the grounds that it is not in its original form, if it is the best evidence that the person adducing it can obtain.

Nietemin moet elektroniese bewyse, soos reeds bespreek, wel aan sekere kriteria voldoen om as geldig beskou te word.[5]

2.4.4 Die geldigheid van outeuridentifikasie in SMS-boodskappe as bewyse in die hof

Die vraag is of outeuridentifikasie wat op SMS-boodskappe uitgevoer word as geldige bewyse in die hof beskou moet word. Daar is reeds genoem dat SMS'e wel as dokumente beskou word en daarom as bewyse in beide kriminele en siviele sake toegelaat kan word. Outeuridentifikasie wat op SMS-boodskappe gebaseer is, sal met ander woorde 'n analise van dokumente wat reeds as

[5] Hier word terugverwys na die kriteria vir die bewyslewerende gewig van bewyse wat wel van belang is vir elektroniese dokumente, soos deur Watney (2009: 10) bespreek. Vergelyk p. 52

geldige bewyse in die hof aanvaar is behels. Die resultate van ’n outeuridentifikasie-ondersoek kan ook as omstandigheidsgetuienis beskou word aangesien die resultate van so ’n ondersoek die vermoede dat ’n persoon die outeur van ’n bepaalde SMS-boodskap(pe) is kan ondersteun of uitskakel. Costello (2013:2) verwys na Sonia Escobio O’Donnell, wat meen dat omstandigheidsgetuienis wat daarop dui dat ’n spesifieke SMS-boodskap deur ’n bepaalde persoon of van ’n bepaalde toestel gestuur is van deurslaggewende belang is, aangesien dit nie aanvaar kan word dat die outeur van die boodskap en die eienaar van die selfoon dieselfde persoon is nie. Alhoewel moderne tegnologie dit moontlik maak om vas te stel of ’n bepaalde SMS-boodskap vanaf ’n spesifieke selfoon gestuur is, deur ’n spesifieke selfoon ontvang is of op ’n selfoon uitgewis is, is dit nie ’n waarborg dat die verdagte die outeur van daardie bepaalde SMS-boodskap is nie. Dit beteken dat outeuridentifikasie as ’n addisionele en baie belangrike analise beskou moet word. Hunt en Zabel (2012) meen ook dat addisionele analisering van ’n SMS-boodskap nodig is aangesien die inhoud van so ’n boodskap nie outomaties met die eienaar van die selfoon verbind kan word nie:

> How do you show that a text message from Joe Smith is a text message from Joe Smith? The answer is not as simple as “it came from Smith's phone number.” This is no different than what is required to authenticate a handwritten letter. A letter from Mary Jones may bear her signature, but that signature could be forged. A court would likely require the proponent to produce something beyond the letter itself as evidence such as a witness who could identify her signature.

Enige forensiese analises van bewyse moet aan streng vereistes voldoen en dit is juis hierdie vereistes wat die toelaatbaarhied van outeuridentifikasie as ’n geldige metode in hofsake kan belemmer. Die grootste probleem is sekerlik die feit dat geen outeuridentifikasiemetode honderd persent akkuraat is nie (Holmes, 1994; Holmes, 1998; Chaski, 2005; Ishihara, 2011; McLeod en Grant, 2012). Dit is ’n ernstige probleem, aangesien verdagtes nie veronderstel is om gevonnis te word tensy die bewyse teen die verdagte sonder twyfel aandui dat hy of sy skuldig is nie. Die tweede probleem is dat daar so ’n groot verskeidenheid metodes is wat gebruik kan word om die moontlike outeur van ’n teks of SMS-boodskap te bepaal dat dit moeilik is om vas te stel watter metode werklik die beste resultate lewer. Outeuridentifikasie kan, om laasgenoemde redes, nie as die deurslaggewende bewys in ’n saak beskou word nie, maar kan eerder as ’n bykomende analise of omstandigheidsgetuienis beskou word. Dit moet in gedagte gehou word dat outeuridentifikasie van SMS-boodskappe, ten spyte van die afwesigheid van ’n honderd persent

sekerheid, reeds in die verlede bygedra het tot skuldigbevindings in verskeie sake (Crystal, 2008; Grant, 2010; Blackwell, 2012).

Soos reeds genoem, is dit belangrik om ’n betroubare metode te gebruik tydens outeuridentifikasie-analises van tekste. Betroubare metodes dra by tot die geldigheid van bewyse en dit versterk ook hierdie bewyse as omstandigheidsgetuienis in sekere gevalle. Stilometrie is die gekose analisemetode in die huidige ondersoek. Alhoewel die akkuraatheid van sommige stilometriese metodes betwis word, bly stilometrie steeds ’n gunsteling analitiese metode onder forensiese linguiste, aangesien stilometrie uit die inkorporasie van verskeie ander metodes tydens ’n enkele ondersoek bestaan en in verskeie gevalle al tot hoë persentasies sukses in, spesifiek, outeuridentifikasiestudies gelei het. In 2.5 word stilometrie as metode bespreek en die probleme en beperkings van die metode(s) word ook uitgelig.

2.5 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 bestaan hoofsaaklik uit 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). Hierdie twee komponente beteken dat die proses van stilometriese analise tweeledig is. 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. Laasgenoemde 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 analise van die teks is, terwyl die meet van hierdie stilistiese kenmerke en die statistiese toetse wat daarop uitgevoer word ’n kwantitatiewe proses is wat dan ‘stilometrie’ genoem word. Dit is belangrik dat die stilistiese en stilometriese analise van ’n teks

of tekste mekaar komplementeer. Die stilistiese analise is ’n belangrike eerste stap aangesien dit die kenmerke identifiseer wat deur die stilometriese analise geanaliseer word. Die tipe kenmerke moet bowendien aan sekere vereistes voldoen om van waarde vir die stilometriese analise te wees. Grieve (2005: 2) beskryf die stilistiese analise van ’n teks vir die doeleindes van outeuridentifikasie en beklemtoon die belangrikheid van stilometriese analises as deel van die metodologie:

> A successful quantitative authorship attribution depends on the investigator’s selection of a set of textual measurements whose values are relatively consistent across each possible author’s writing sample and relatively variable across the set of possible authors [...].

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 gebruik word vir literêre of historiese doeleindes nie. Volgens Schulstad het stilometrie in moderne forensiese linguistiek ’n veel wyer gebruiklikheid:

> [...] it 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.

Holmes (1998) verwys in sy artikel na die wisselvallighede van die stilometriese metode in die verlede en noem verskeie metodes wat binne die kategorie van stilometrie geplaas word en gebruik word om die akkuraatheid van ’n stilometriese analise te vermeerder. Een van die akkuraatste stilometriese analises wat in die verlede gebruik is, is die identifisering van funksiewoorde (voorsetsels, voegwoorde, lidwoorde ens.). Hierdie woorde word onbewus deur die outeur in ’n teks gebruik en hulle is nie konteksgebonde nie. Tydens so’n stilometriese analise word daar vasgestel hoeveel keer sekere funksiewoorde in, byvoorbeeld, elke 1 000 woorde voorkom en ’n frekwensie word daarvolgens uitgewerk. Die frekwensie van een teks kan dan met die frekwensies in ander tekste vergelyk word om vas te stel wat die moontlikheid is dat die outeur van ’n teks(te) ook ’n ander teks(te) geproduseer het.

Hierdie tegniek is onder andere in 1962 deur Ellegard gebruik om die outeurskap van die *Junius Letters* vas te stel en dit is ook in 1964 deur Mosteller en Wallace toegepas om die outeurs van die *Federalist Papers* te identifiseer. Juola het ook onder andere ’n soortgelyke tegniek gebruik

om vas te stel dat J.K. Rowling die skrywer is van die 'debuut'roman *The Cuckoo's Calling* (Brown-Jackson, 2013). Rowling het onder die skuilnaam 'Robert Galbraith' geskryf maar is deur middel van forensies-linguistiese tegnieke, wat die volgorde van aangrensende woorde, die volgorde van karakters, 'n berekening van die algemeenste woorde in die teks en die outeur se voorkeur vir lang of kort woorde insluit, as die outeur uitgewys (Zax, 2014).

Volgens Holmes (1998: 113–115) lewer die gebruik van "principal components analysis" en sogenaamde "neural networks" in stilometrie die beste resultate op. Die probleem is egter dat daar vir sulke analises groot hoeveelhede teks benodig word voor akkurate resultate moontlik is. 'n Tweede probleem is dat die aantal verdagte outeurs in 'n ondersoek ook die akkuraatheid van sekere metodes kan beïnvloed. Volgens Luyckx en Daelemans (2011: 37) kan 'n groot aantal verdagte outeurs selfs gewoonlik betroubare metodes soos die identifisering van funksiewoorde beïnvloed:

> […] it may be correct to claim that distributions of function words are important markers of author identity, but the distribution of a particular function word, while useful to distinguish between one particular pair of authors, may be irrelevant when comparing another pair of authors.

Die ideale outeuridentifikasiesituasie is met ander woorde een waar daar 'n beperkte aantal verdagte outeurs is, maar groot hoeveelhede teks.

Soos reeds genoem, is die aanname in stilometrie dat die kern van die individuele styl van elke outeur vasgevang kan word deur 'n sekere hoeveelheid kwantitatiewe kriteria (Somers, s.a.). 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. Met ander woorde, daar is aspekte van elke outeur se styl wat maklik is om na te maak.

Rekenaargesentreerde 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 populêr (Somers, s.a.). Dit is belangrik om kennis te neem van die feit dat stilometrie vandag deur kunsmatige intelligensie gedomineer 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 rekenaargesentreerde benadering kan moontlik vertroue

in hierdie metode in die toekoms versterk en daartoe bydra dat resultate wat uit sulke analises verkry word, makliker as geldige bewyse in die hof aanvaar word.

Alhoewel stilometrie algemeen in outeuridentifikasie gebruik word, moet navorsers bewus wees van hierdie metode se tekortkominge. Daar is, soos reeds bespreek, nie een kombinasie van tekskenmerke wat 100% akkuraat is wanneer dit in ’n stilometriese analise gebruik word nie. ’n 1998-opname oor stilometrie het bevind dat daar reeds meer as 1 000 kenmerke was wat in stilometriese analises gebruik is (Schulstad e.a., 2012: 2 en Zechner, s.a.: 3). Hierdie kenmerke sluit leksikale, sintaktiese, strukturele, idiosinkratiese en kontekssspesifieke kenmerke in. Leksikale kenmerke verwys byvoorbeeld na die aantal woorde in die woordeskat, terwyl sintaktiese kenmerke sinstrukture en die gebruik van sekere funksiewoorde insluit. Strukturele kenmerke verwys na die manier waarop die teks gestruktureer is en neem aspekte soos die insluiting van paragrawe, paragraaflengte en die gebruik van inkeping in ag. Kontekssspesifieke kenmerke verwys na die gebruik van kontekssspesifieke woorde terwyl idiosinkratiese kenmerke interessante eienskappe wat met die spelling van woorde of ander taalfoute verband hou, insluit (Schulstad e.a., 2012: 2). Om te probeer verseker dat stilometrie so betroubaar as moontlik is, het verskeie navorsers bevind dat die linguis veral ’n fokus op funksiewoorde (grammatikale woorde) en sintaktiese kenmerke by die analise moet insluit, maar ook semantiese kenmerke en leksikale kenmerke as diskrimineerders moet oorweeg (Holmes, 1998; Luyckx en Daelemans, 2011: 39; Koppel en Schler, 2004: 490). Schulstad e.a. (2012: 2) verwys na ’n studie deur Grieve (2005) waarin bevind is dat die herhaling van funksiewoorde, leestekens, bi-gramme (2-gramme) en tri-gramme (3-gramme) die meeste sukses behaal as merkers van outeurskap tydens ’n stilometriese analise. Grieve spekuleer dat die rede vir die sukses van laasgenoemde kombinasie gegrond is op die feit dat funksiewoorde en leestekens aandui hoe sinne gekonstrueer word, terwyl n-gramme eerder beïnvloed word deur die inhoud van die teks.

Navorsers moet ook in ag neem dat suksesvolle stilometriese analises gewoonlik van groot hoeveelhede data gebruik maak. “Ideale” data word deur Luyckx en Daelemans (2011: 38) beskou as lang tekste vir elke outeur of verskeie kort tekste vir elke outeur. Die sukses van ’n stilometriese analise neem met ander woorde af soos wat die hoeveelheid teks tot die forensiese linguis se beskikking afneem.

Alhoewel die moontlike sukses van ’n stilometriese analise van die omstandighede waarbinne die analise plaasvind en die hoeveelheid teks afhang, kan die linguis aan homself/haarself die volgende geslote vrae stel. Hierdie vrae kan die linguis van hulp wees in die proses van outeuridentifikasie:

- Wie het hierdie teks/SMS/dokument geskryf/gepubliseer – outeur A of outeur B?
- Indien outeur A *hierdie* dokument(e) geskryf/geproduseer het, het hy/sy ook *daardie* dokumente geskryf/geproduseer?
- Hoe waarskynlik is dit dat outeur A hierdie dokument geskryf/geproduseer het?

Stilometrie het twee subkategorieë, naamlik, stilokronometrie (*stylochronometry*) en adversatiewe stilometrie (*adversarial stylometry*). Volgens Corney (2003: 21) het stilokronometrie te make met die vasstel van die kronologiese volgorde van ’n bepaalde outeur se werk. Corney meen daarom dat terwyl stilometrie gebaseer is op die aanname dat elke outeur oor ’n unieke styl beskik, stilokronometrie aanneem dat so ’n unieke styl met die verloop van tyd sal verander. Corney (2003: 21) verwys na navorsers wat reeds studies in stilokronometrie voltooi het en maak op grond van die resultate die volgende opmerking:

> The results of these various studies seem to indicate that an author’s style can and does change over a period of time. In these cases the period of time in question was more than ten years. These results should be kept in mind for any forensic investigations, and the known writings of any particular investigated author should be sampled from a period of time which is relatively short in this context, such as one or two years.

Die tweede subkategorie, adversatiewe stilometrie, word in groter detail in die volgende gedeelte bespreek.

2.5.1 Adversatiewe stilometrie (*Adversarial stylometry*)

Adversatiewe stilometrie kan beskou word as ’n ander perspektief of invalshoek van stilometrie. Adversatiewe stilometrie daag die aananme uit dat die outeur van ’n bepaalde teks *nie* sy/haar skryfstyl bewustelik verander voor hy/sy die teks produseer nie (Brennan e.a., 2012). Die algemene veronderstelling is dat die outeur van ’n teks altyd sy/haar werklike skryfstyl gebruik wanneer hy/sy ’n teks produseer en nie bewustelik sy/haar skryfstyl verander nie. Met ander

woorde, binne adversatiewe stilometrie probeer linguiste vasstel watter tegnieke deur onopgeleide persone gebruik kan word om hul eie skryfstyl te verdoesel en ook tot watter mate hierdie tegnieke bestaande stilometriese metodes kan flous. Brennan e.a. (2012: 2) definieer adversatiewe stilometrie soos volg:

> We define adversarial stylometry as the notion of applying deception to writing style to affect the outcome of stylometric analysis. This new problem space in the field of stylometry leads to new questions such as what happens when authorship recognition is applied to deceptive writing? Can effective privacy-preserving countermeasures to stylometry be developed? What are the implications of looking at stylometry in an adversarial context?

Brennan e.a. (2012: 2) meen dat adversatiewe stilometrie ’n belangrike navorsingsveld is, veral met betrekking tot privaatheid en sekuriteit. Huidige sisteme wat anonimiteit in die aanlynomgewing bevorder, fokus hoofsaaklik op privaatheid ten opsigte van die plekbepaling van ’n outeur, maar byna geen aandag word geskenk aan privaatheid wat met die inhoud van die data verband hou nie (Rao en Rohatgi, 2000: 86). Die uitgangspunt is dat aanlynprivaatheid ’n basiese reg van rekenaar- en selfoongebruikers is en dat daar metodes moet wees om die identiteit van gebruikers te beskerm. Brennan e.a. (2012: 2) meen:

> Writing style as a marker of identity is not addressed in current circumvention tools, nor is it addressed in the security and privacy community at large. Given the high accuracy of even basic stylometry systems this is not a topic that can afford to be overlooked.

Dit blyk dat studies in adversatiewe stilometriese metodes daarop gemik is om individue die kennis te gee om hulself te beskerm teen ‘negatiewe’ aanlyngedrag soos boeliegedrag en wraakneming. Dié navorsing word veral deur verskeie joernaliste, besighede, aktiviste en wetstoepassers verwelkom en gebruik (Brennan e.a., 2012: 04; Kacmarcik en Gamon, 2006: 1). Alhoewel dit waar is dat individue wat deur middel van die internet boodskappe of ander vorme van teks publiseer die reg tot privaatheid het, moet die realiteit dat sommige individue tegnieke in adversatiewe stilometrie kan gebruik om hul identiteit tydens onwettige aktiwteite te verdoesel, ook in ag geneem word.

Uit die navorsing blyk dat drie tegnieke deur die outeur van ’n teks gebruik kan word om sy/haar skryfstyl te vermom. Die outeur van ’n teks kan in die eerste plek bewustelik probeer om ’n teks op so ’n manier te skryf dat sy/haar persoonlike skryfstyl nie herkenbaar is nie. Hierdie tegniek

word **verduistering** genoem. Tweedens kan die outeur deur middel van **nabootsing** probeer om die skryfstyl van ’n ander outeur na te maak. Dit beteken dat outeur A tekste van outeur B bestudeer en daarna ’n teks produseer wat so na as moontlik aan outeur B se taalgebruik en skryfstyl is, of wat ten minste groot hoeveelhede van laasgenoemde eienskappe insluit. Die derde tegniek wat ’n outeur kan gebruik is die **masjienvertaling** van tekste. Die outeur kan die woordorde en woordkeuses binne die teks verander deur middel van ’n vertalingsprogram. Die outeur vertaal met ander woorde die teks in ’n ander taal, soos Duits, en vertaal die teks daarna weer terug in die taal waarin die oorspronlike teks geskryf is. Die outeur kan ook die teks in twee tale vertaal voordat dit weer terugvertaal word na die oorspronklike taal. Brennan e.a. (2012: 8) het egter bevind dat laasgenoemde tegniek redelik onbetroubaar is by sekere taalgroepe omdat die sinstruktuur van die vertaalde teks ook soms verander en dit beteken dat die teks wat op die einde geproduseer word nie meer koherent is nie.

Daar is ook verskeie probleme met die toets van adversatiewe stilometriese metodes aangesien sommige individue sukkel om hul skryfstyl te verdoesel en dit beteken dat die effektiwiteit van adversatiewe stilometriese metodes eerder van die individu as van die metode afhang. Brennan e.a. (2012: 20) stel voor dat ’n generiese skryfstyl ontwikkel moet word om hierdie probleem op te los. Die program Anonymouth (https://psal.cs.drexel.edu/anonymouth) is ’n moontlike oplossing vir die probleem, maar die program is tans skaars uit die ‘prototipe’-fase van sy ontwikkeling en verbeteringe en aanpassings is nodig vir die program om optimaal in ’n verskeidenheid situasies te funksioneer. Anonymouth ondersteun die vermommingsproses van ’n outeur se skryfstyl deur voorstelle te maak oor moontlike veranderings wat die outeur in sy/haar skryfstyl kan aanbring. Hierdie voorstelle is gegrond op analises van hedendaagse stilometriese metodes (Brennan e.a., 2012: 20; Caliskan Islam e.a., 2013). Kacmarcik en Gamon (2006: 444) stel ook voor dat daar eerder ’n algemene metode moet wees wat gebruik kan word om ’n outeur se skryfstyl te verdoesel. In hulle navorsing, wat van statistiek gebruik maak, word aangevoer dat anonimiteit verkry kan word deur 14 veranderinge per 1 000 woorde aan te bring. Verder meen Kacmarcik en Gamon (2006: 449) dat die “unmasking”-tegniek wat deur Koppel en Schler (2004) gebruik is om die outeur van ’n bepaalde teks te verifieer, aangepas kan word om as ’n adversatiewe stilometriese metode gebruik te word. Met ander woorde, wanneer daar bepaal word (deur middel van “unmasking”) watter eienskappe van die outeur hom/haar as outeur van

die teks weggee, kan die outeur daardie bepaalde eienskappe verander of aanpas om sy/haar skryfstyl te verdoesel.

Uit die voorafgaande bespreking blyk dit dat adversatiewe stilometrie in sommige gevalle hedendaagse stilometriese metodes maklik kan omseil en juis om hierdie rede stel Brennan e.a. (2012: 21) voor dat stilometriese metodes in die toekoms in situasies getoets word waar adversatiewe stilometrie toegepas is om die effektiwiteit van stilometriese metodes in die ergste gevalle te toets:

> The analysis of stylometry techniques and their weaknesses to adversarial writing demonstrates that we must test stylometry methods for their resistance to adversaries in situations where their presence is likely.

Rao en Rohatgi (2000: 94) beweer dat die "principal component analysis" wat die basis van hul navorsing uitmaak baie effektief sal wees in 'n adversatiewe situasie mits die metode tesame met "misspelling lists and the classical stylometry measures mentioned" gebruik word. Die klassieke stilometriese metodes waarna Rao en Rohatgi verwys, is onder andere die analisering van die frekwensie van funksiewoorde in 'n teks. Kacmarcik en Gamon (2006: 451) is van mening dat Koppel en Schler (2004) se "unmasking" tans die veiligste stilometriese metode is om in 'n adversatiewe situasie te gebruik aangesien dié metode redelik gehard is teen die verdoeseling van 'n outeur se skryfstyl.

2.5.2 Korter tekste en stilometrie

Kort tekste is problematies in stilometriese analises aangesien daar nie genoeg linguistiese data is wat verwerk kan word tydens die analises 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 analise te behaal. Stamatatos e.a. (2001: 193) meen een rede waarom sommige stilometriese analises onsuksesvol op die korter duur is, is die feit dat die meeste stilometriese analises ontwerp is om lang literêre tekste te analiseer. Verder voer Stamatatos e.a. (2012: 196, 208) aan dat tekslengtes van minder as 1 000 woorde nie geskik sal wees vir stilometriese analises wat op die leksikale eienskappe van die outeur se taalgebruik fokus nie:

> It seems that a text-length shorter than 1,000 words is not adequate for representing sufficiently the characteristics of the idiosyncratic style of an author

> by using either lexical measures, the presented set of style markers, or a combination of them.

Dit is met ander woorde nodig om ’n stilometriese metode te gebruik wat wel akkurate resultate met korter tekste kan lewer sodat die analise van korter tekste moontlik en sinvol sal wees. Die huidige ondersoek poog om ’n forensies-linguistiese situasie te skep wat so na as moontlik aan die werklikheid is. Dit beteken dat die forensiese linguis wat die outeurskap van SMS-boodskappe probeer bepaal, nie ’n groot hoeveelheid tekste tot sy/haar beskikking sal hê nie. Selfs wanneer die forensiese linguis toegang het tot al die SMS-boodskappe op die verdagte se selfoon is die lengte van die boodskappe steeds problematies, aangesien ’n groot hoeveelheid SMS-boodskappe nodig is om gelykstaande aan ’n vollengte teks (d.w.s omtrent 1 000 woorde in lengte) te wees. Chaski (2001: 4) meen ook dat korter tekste en min data ’n algemene verskynsel in die forensiese linguistiek is:

> [...] forensically significant documents are often short and cannot be amplified; indeed, even known documents are often short in length and limited in quantity.

Ten spyte van die gebrek aan data wat in verskeie outeuridentifikasiesituasies aangetref word, is dit steeds, in sommige gevalle, moonlik om die outeurs van die verdagte teks te bepaal. Hubbard (1995: 7) verwys na ’n studie deur Morton (1978) waar bevind is dat kollokasies van frekwente woorde met ander frekwente woorde (*van die, in die, saam met,* ens.) een van die betroubaarste diskrimineerders tussen verskillende outeurs is. Tweedens word beweer dat woordpare ook as diskrimineerder kan optree en derdens word die frekwensie waarteen sekere woorde gereeld gebruik word ook as ’n baie belangrike diskrimineerder beskou. Hubbard (1995: 7) noem dat hy van die derde diskrimineerder in sy navorsing gebruik gemaak het aangesien die tekste waarmee hy gewerk het ook relatief kort was en daarom nie genoeg teks bevat het om genoegsame kollokasies en woordpare in te sluit nie. Hubbard (1995: 8–9) het die tekste wat met mekaar vergelyk moet word by ’n bepaalde woordtelling afgesny sodat al die tekste so lank soos die kortste teks sou wees (783 woorde). Die rede hiervoor is dat Hubbard wou vasstel hoe akkuraat ’n stilometriese analise is op tekste wat min of meer dieselfde lengte as die verdagte teks is. Hubbard het van die Chi-kwadraattoets gebruik gemaak om die frekwensies van woorde tussen verskillende tekste en tussen gedeeltes van dieselfde teks te bepaal. Volgens Hubbard (1995: 9–10) het hierdie metode genoeg sukses behaal om as ’n betroubare metode binne die omstandighede van die studie beskou te word, maar hy waarsku nietemin dat die resultate van

stilometriese analises in korter tekste versigtig geïnterpreteer moet word aangesien die inhoud van korter tekste in ander omstandighede tot onbetroubare resultate kan lei.

Op grond van die voorafgaande bespreking blyk dit nodig om te bepaal of dit moontlik is om die betroubaarheid van die resultate wat verkry is uit ’n klein korpus te verhoog deur nie net die frekwensies van woorde in die onderskeie SMS-boodskappe in ag te neem nie, maar ook op aspekte soos spelling, funksiewoorde, die insluiting van Engelse woorde en logogramme of simbole en die gebruik van leestekens te fokus. Dié eienskappe word nie noodwendig bewustelik deur die outeur van die SMS beheer nie en daarom is dit moontlik dat sulke eienskappe sal aanleiding gee tot veronderstelde idiosinkratiese skryfstyle in SMS-boodskappe. Soos reeds genoem is idiolek ’n baie komplekse konsep en is dit nie noodwendig so maklik om idiosinkratiese skryfstyle te identifiseer of die teenwoordigheid van idiosinkratiese skryfstyle te ondersteun nie. Idiolek en die problematiek rondom die konsep word daarom in die volgende gedeelte bespreek.

2.6 Idiolek

Dit is duidelik dat outeuridentifikasie in tekste, en veral by SMS-boodskappe, nie ’n maklike taak is nie. Verskeie probleme met outeuridentifikasie, soos die bondigheid van die tekste, is reeds in hierdie navorsing aangeraak, maar die mees kontroversiële aspek, naamlik idiolek, word nou meer breedvoerig ondersoek. ’n Begrip wat in die literatuur gebruik word wanneer die konsep ‘idiolek’ bespreek word, is ‘generiese taalgebruik’. Dit is dus nodig om eers hierby stil te staan.

2.6.1 Wat is generiese taalgebruik?

‘Generiese taalgebruik’ is ’n konsep wat nou verband hou met ‘idiolek’. Soos wat reeds by 1.8.5 genoem is, beskryf die term ‘generies’ iets wat op ’n groep, eerder as ’n individu, van toepassing is. Die veronderstelling dat daar iets soos generiese taalgebruik is, laat die moontlikheid oop dat die afwesigheid van generiese taalgebruik kan lei tot die afleiding dat idiolek (wat dan in afwykings van die generiese taalgebruik opgespoor kan word) bestaan. Generiese taalgebruik - soos wat dit aangetref word in geskrewe tekste - verwys na eienskappe soos spelling, sinskonstruksies en afkortings van woorde wat op vaste reëls binne ’n bepaalde taal gegrond is. Hierdie reëls word deur die meerderheid taalgebruikers van ’n bepaalde taal nagekom. Tot ’n

mate kan die reëls egter aangepas word om individualistiese taalgebruik te skep. Generiese taalgebruik dien met ander woorde as ’n maatstaf of norm waarteen idiolektiese taalgebruik gemeet kan word. Een probleem wat egter opduik met die definiëring van ‘generiese taalgebruik’ en ‘idiolektiese taalgebruik’ is die feit dat eienskappe van die taalgebruik wat aanvanklik idiolekties was so algemeen word dat dit mettertyd voorkom asof hierdie eienskappe eintlik aan ’n tipe generiese vorm van die taal behoort. Verdere probleme met die konsep van ‘idiolek’ volg in 2.6.3.

2.6.2 Wat is 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. Daar is verskeie definisies van idiolek (hier word slegs na vier verwys) maar hierdie definisies is steeds gegrond op die sentrale idee van ‘individualiteit’ en ‘uniekheid’:

> An *idiolect* (Bloch, 1948: 7) 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)

Volgens Turell (2010: 217) moet die term ‘idiolek’ nie in forensiese kontekste gebruik word nie. Turell meen dat daar eerder na ‘idiolektiese styl’ verwys moet word om te beklemtoon dat “each person favours certain linguistics features which constitute their individual use of language”. Barber (2004) voer aan dat “[A]n idiolect, if there is such a thing, is a language that can be characterized exhaustively in terms of intrinsic properties of some single person at a time, a person whose idiolect it is at that time”.

2.6.3 Die problematiek rondom idiolek

Idiolek is nie so ’n eenvoudige konsep soos blyk uit bogenoemde definisies nie. Wanneer veronderstel word dat idiolek of idiolektiese styl teenwoordig is, word die forensiese linguis dikwels met drie vrae gekonfronteer, naamlik: Bestaan idiolek werklik? Is idiolek altyd waarneembaar en is idiolek ’n akkurate aanwyser van ’n individuele outeur? Verder moet die forensiese linguis ook idiolektiese analises op die *algemene* taalgebruik van die verdagte outeur doen eerder as om te fokus op seldsame woorde wat die outeur mag gebruik. Dit beteken dat die linguis 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 individuele 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 intra-variasie (variasie in een persoon se praat- of skryfwyses) en inter-variasie (variasie tussen twee of meer individue se praat- of skryfwyses) (Gavaldà-Ferré, 2012: 262). Intra-variasie dui met ander woorde op variasie in een persoon se praat- en skryfwyses maar waar verskeie elemente (woorde, frases of uitdrukkings) steeds tussen tekste (deur dieselfde outeur) ooreenstem, terwyl inter-variasie dui op variasie tussen tekste deur verskillende outeurs met slegs enkele, en in sommige gevalle geen, ooreenkomste tussen elemente in hierdie tekste nie. 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:

> […] not only do any two members of what should reasonably be the same speech community use different variation of the same linguistic form, but so does each individual member.

Hierdie variasie word beskryf in die sogenaamde ‘uniqueness of utterance principle’ (Chomsky, (1965) en 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 geproduseer is. Die rede hiervoor is dat elke individu by verskillende geleenthede verskillende leksiko-grammatiese keuses uitoefen (Crankshaw, 2012: 3). Dit is hierdie leksiko-grammatiese keuses wat lei tot intra-variasie wat die identifisering van idiolek kan belemmer.

Ten spyte van laasgenoemde probleme is die aanname steeds dat daar ondanks variasie in individue se praat-en skryfwyses sekere woorde of frases is wat algemeen deur die individu gebruik word en dat sulke woorde en frases gebruik kan word om idiolek te identifiseer. Gavaldà-Ferré (2012: 262) verwys na Rose (2002) wat meen dat "[T]he greater the ratio of between-speaker to within-speaker variation, the easier the identification". Gavaldà-Ferré (2012: 270) se studie het bevind dat intra-variasie stadige variasie toon terwyl inter-variasie vinniger variasie toon. Dit beteken dat die spreker/skrywer se eie taalgebruik wel oor 'n tydperk 'n ander vorm aanneem, maar dat hierdie 'nuwer vorm' steeds verskeie ooreenkomste toon met die vorige/ouer vorms:

> […] the results for the experiments that have been conducted show two main important factors to be considered. On one hand, intra-speaker comparisons give results that are closer to one, and therefore show slow variation. These results confirm the hypothesis that a speaker's 'idiolectal style' seems to remain quite stable despite the course of time and a long term situation of language contact. On the other hand, the inter-speaker variation has proved to be higher than intra-speaker variation which confirms the proposal formulated in section 1 that although there is intra-speaker variation, each speaker has a unique 'idiolectal style' that separates them from the rest of speakers from the same community.

Volgens Coulthard (2004: 431) is dit moontlik om, op grond van sy bogenoemde definisie van idiolek, asook navorsing deur onder andere Gavladà-Ferré, onder die vals indruk te verkeer dat daar die moontlikheid bestaan "to devise a method of *linguistic fingerprinting* – in other words that the linguistic 'impressions' created by a given speaker/writer should be usable, just like a signature, to identify them". So 'n metode is egter onmoontlik en Coulthard (2004: 432) verwys daarom na die konsep van linguistiese vingerafdrukke as, sou dit sinoniem vir idiolek wees, as 'n misleidende metafoor.

Crankshaw (2012: 2) is ook skepties oor die gebruik van die term 'linguistiese vingerafdruk' as sinoniem vir 'idiolek'. Volgens Crankshaw (2012: 4) en Coulthard (2004: 432) is 'n idiolek nie so uniek of onveranderlik soos 'n vingerafdruk nie:

> The physical fingerprint renders a person uniquely identifiable, as each sample is both 'identical and exhaustive', containing all of the information necessary for an individual's identification. In contrast to this, even an extremely large sample of linguistic data can only ever give partial information of a person's idiolect.

Daar moet ook in ag geneem word dat eksterne faktore 'n individu se taalgebruik sal beïnvloed en verander. Hierdie faktore kan onder andere sosiale klas, vlak van opvoeding, ouderdom en geslag insluit. Crankshaw (2012: 3) verwys na McMenamin (2002: 53) wat aandui dat laasgenoemde faktore lei tot "subtle differences in their (die individue – LT) internalized grammar which then manifests itself in a person's speech, writing and responses to others".

Alhoewel die bestaan van idiolek of idiolektiese styl betwis word, is daar steeds 'n algemene aanvaarding dat indien idiolek bestaan dit makliker sal wees om te identifiseer wanneer die linguis 'n substansiële hoeveelheid tekste tot sy of haar beskikking het. Die lengte van die tekste is ook van belang – hoe meer inhoud in elke dokument ingesluit word, hoe makliker is dit om individuele taalgebruik te identifiseer. Dit is egter die geval dat forensiese linguiste selde tekste wat langer is as 750 woorde ontvang. Volgens Crankshaw (2012: 5) kan korter tekste steeds geanaliseer word, maar nie so volledig 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 verhoed 'n diepgaande, volledige voorstelling van die individu se idiolek.

Grant (2010: 509–510) skryf dat die linguis nie slegs 'n kennis van idiolek en idiolektiese patrone moet dra nie, maar ook, en dit is belangriker as laasgenoemde, metodologies moet kan bewys dat hierdie patrone bestaan:

> A theory of idiolect must provide an explanation as to why one individual's production is consistent across texts, and must also explain why that individual's language is distinctive as compared with that of other individuals.

Barber (2004) sluit by laasgenoemde stelling aan wanneer hy skryf dat "[A]lthough the properties of *x*'s idiolect are tied stipulatively to intrinsic properties of *x*, this in itself does not mean that two distinct individuals could not share an idiolect, or have a pair of significantly overlapping idiolects".

'Bewyse' van idiolek in 'n bepaalde teks is belangrik aangesien die mening dat idiolek in 'n bepaalde teks of tekste teenwoordig is nie genoeg is om aan te voer dat die teks die produk van 'n bepaalde outeur is nie. Dit beteken dat idiolek duidelik bewysbaar of identifiseerbaar moet

wees. Die forensiese linguis moet kan aandui – deur tekste met ander te vergelyk en ook deur vergelykings tussen tekste van dieselfde outeur – dat daar ’n idiolektiese taalgebruik teenwoordig is. Idiolek is, volgens Grant (2010: 509), ’n moeilike konsep om te verduidelik. Grant verwys na die kompleksiteit van idiolek wanneer hy skryf: “(C)onsistency and distinctiveness may, of themselves, be evidence that an idiolect exists, but they do not constitute an explanatory theory of idiolect”. Dit beteken dat ’n teks of tekste kenmerke van “consistency and distinctiveness” mag bevat wat gebruik kan word om ’n outeur te identifiseer, maar dit beteken nie dat die linguis kan verduidelik waarom hierdie eienskappe idiolekties van die spesifieke outeur is nie.

2.6.4. Teorieë oor die teenwoordigheid van idiolek

Tans is daar twee uiteenlopende sienings wat met idiolek verband hou. Die een is ’n kognitiewe teorie gebaseer op die uitgangspunt dat die linguis oor kennis moet beskik van die kognitiewe meganismes wat die produksie van ’n teks moontlik maak. Die tweede teorie is dat ’n stilistiese begrip van taal en taalproduksie genoeg is om konsekwensie en eiesoortigheid in geskrewe dokumente te verklaar (Grant, 2010: 510).

2.6.4.1 Die kognitiwistiese teorie

Volgens die kognitiwistiese teorie word ’n individu se taalproduksie bepaal deur die individu se linguistiese bevoegdheid. ‘Linguistiese bevoegdheid’ word deur Grant (2010: 510) beskryf as die kognitiewe vermoë om taal te produseer. Dit word weerspieël in taalgebruik. Kognitiwistiese teorieë maak dit tot ’n sekere mate moontlik om aspekte soos sintaktiese kompleksiteit vas te stel en die leksikon van ’n individu te meet. Hierdie waarnemings en metings kan in beperkte mate gebruik word om variasie tussen individue en groepe te demonstreer. Grant (2010: 510) verwys na voorbeelde waar kwantitatiewe- en rekenaarlinguistiek, by veral langer tekste, gebruik is om die eienskappe van individue se taalproduksie wiskundig te beskryf (Holmes, 1998; Grant, 2007; Chaski, 2001 en Grant, 2008). Hoewel kognitiwistiese teorieë kan beskryf waarom daar patrone en vaste kenmerke binne ’n individu se taalproduksie is, bied dit egter nie die nodige verduideliking vir idiolek nie.

2.6.4.2 Die stilistiese teorie

Die stilistiese teorie word beskou as teenoorstellend tot die kognitiwistiese teorie ten opsigte van beskouings en studies wat met idiolek verband hou. Teoretici binne die veld van stilistiek glo dat stilistiek noodsaaklik is om die taalverskille tussen individue te begryp. Hierdie siening word nie deur kognitiwistiese teoretici ondersteun nie, aangesien hulle glo dat die stilistiese teorie nie op ’n standvastige linguistiese teorie gegrond is nie. Grant (2010: 512) meen egter dat “understanding language variation stylistically, as the interaction between habit and context, does not imply a lack of linguistic theory so much as an alternative linguistic theory”. Grant (2010: 513) skryf verder dat idiolek nie slegs deur die kognitiwistiese teorie of sosiolinguistiese studies bepaal kan word nie. Daar moet eerder op ’n verenigde teorie van idiolek gefokus word waar kognitiwistiese en stilistiese teorieë saamsmelt sodat die sterkpunte uit elke teorie benut kan word. So ’n verenigde teorie sal moontlik beter bewyse en verklarings kan oplewer vir die aanwesigheid van idiolek:

> […] although cognitivist theories can provide convincing explanations for some aspects of language production these theories hold less power in and of themselves in explaining individual variation. Conversely, while stylistic approaches to the linguistic individual do concentrate on providing explanations for language variation between individuals they are perhaps less interested in explaining how these might be realized psychologically.
>
> Grant (2010: 214)

Die stilometriese metodes wat algemeen in outeuridentifikasie gebruik word (onder andere: die bepaling van die frekwensie van funksiewoorde, algemene sinslengte, die keuse van sinonieme, algemene lengte van paragrawe en woorde, ensovoorts) maak nie eksplisiet melding van idiolek nie. Met ander woorde, die navorsers beskryf nie hul navorsing as ondersoeke na die idiolek van individue in ’n poging om die outeurs van tekste te identifiseer nie, maar tog is dit presies wat in hierdie navorsing ondersoek word. Dit is tog so dat stilometriese metodes eienskappe van outeurs se skryfstyle ondersoek wat hulle van mekaar onderskei om sodoende te bepaal watter outeur ’n bepaalde teks geproduseer het. Dié eienskappe van ’n outeur se skryfstyl is juis eienskappe wat nie bewustelik deur die outeur beheer word nie. Om hierdie rede is dit, myns insiens, tog ’n geldige argument om aan te voer dat idiolek bestaan en ’n belangrike rol speel in outeuridentifikasie.

Die waarneembaarheid van idiolek is egter ’n ander saak. Hoewel stilometriese metodes idiolek gebruik om outeurs te identifiseer moet daar nie aangeneem word dat idiolek altyd duidelik in verskillende outeurs se skryfstyle waarneembaar is nie. Soos wat uit die voorafgaande argumente blyk, sal dit uiteraard makliker wees om verskillende outeurs se idiolek te identifiseer indien die forensiese linguis van ’n groot aantal tekste gebruik kan maak. Hoe meer teks die linguis tot sy/haar beskikking het, hoe makliker sal dit wees om nie net aan te dui dat idiolek in een outeur se skryfstyl teenwoordig is nie, maar ook hoe die idiolek van twee outeurs van mekaar verskil.

’n Belangrike aspek van outeuridentifikasie wat reeds deur navorsers in die forensiese linguistiek bespreek is (Mosteller en Wallace, 1964; Holmes, 1998) en wat van belang kan wees in die identifisering van idiolek is ’n fokus op funksiewoorde in ’n teks. Dit beteken dat die hooffokus van ’n ondersoek na idiolek nie net die opvallende woorde of gebruike van woorde in ’n teks is nie, maar eerder die gebruik van woorde wat nie kontekssspesifiek is nie. Myns insiens kan daar veral in SMS-taal ook opgelet word na die woorde wat uit ’n boodskap gelaat word (byvoorbeeld: voornaamwoorde, lidwoorde, hulpwerkwoorde, ensovoorts) aangesien dit in hierdie bepaalde genre ’n redelik algemene verskynsel is (Olivier, 2013: 490).

Idiolek vorm die basis van navorsing in outeuridentifikasie. Die feit dat dit moontlik is om outeurs van mekaar te onderskei, ondersteun die argument dat elke individuele outeur oor ’n eie idiolek beskik. Die gebrekkige akkuraatheid van sommige metodes in outeuridentifikasie (wat baie lae persentasies van sekerheid behaal ten opsigte van die sekerheid waarmee outeurs geïdentifiseer word), ondersteun egter die tweede argument wat aanvoer word dat idiolek ’n baie subtiele verskynsel is en dat die forensiese linguis, veral met betrekking tot regsake, met redelike sekerheid moet kan bewys dat idiolek teenwoordig is en dat een persoon se idiolek van ’n ander persoon se idiolek onderskei kan word.

In 2.7 word die algemene eienskappe van Afrikaanse SMS-taal bespreek om aan te dui watter eienskappe in SMS-taal tot idiolektiese taalgebruik aanleiding kan gee.

2.7 Die SMS-kultuur

Uit die bespreking oor idiolek is dit duidelik dat dit nie in alle gevalle maklik is om idiolek te identifiseer nie. Die SMS-kultuur het die 'normale' of 'tradisionele' gebruik van taal beïnvloed en gevolglik het nuwer maniere van skryf (spesifiek in elektroniese tekste) ontstaan. Selfs al het 'n groot persentasie van die bevolking die 'selfoontaal' verwelkom en gebruik hulle dit, is daar steeds variasies wat voorkom soos elke individu sy/haar selfoontaal aanpas soos dit vir hom/haar gemaklik is. In hierdie gedeelte word die kenmerke van Afrikaanse SMS-taal oor die algemeen bespreek en in hoofstuk 4 sal gefokus word op die kenmerke wat onder die groep deelnemers in die huidige studie voorkom.

2.7.1 Die ontstaan van die SMS-boodskap

SMS-boodskappe is in die laat 1980's ontwerp deur 'n Finse ingenieur, Matti Makkonen, maar dit het eers in die vroeë 1990's posgevat toe teksboodskappe per selfoon die roepradio's (*pagers*) vervang het wat toe nog grootliks in gebruik was. Die SMS-diens het stadig begin aangesien selfoonmaatskappye eers betroubare maniere gesoek het om geld te eis vir dié nuwe diens, maar nadat prosedures in plek gestel is om hierdie diens te akkommodeer is 'n eksponensiële groei in SMS-verbruikers opgemerk (Crystal, 2008: 04; Mobile Pronto, 2010). Teen 2001 is daar reeds 12.2 miljard SMS'e in Brittanje gestuur. Hierdie syfer het teen 2004 verdubbel en in 2007 het navorsers die syfer op 45 miljard beraam. Hierdie wêreldwye verskynsel het ook oorgespoel na Suid-Afrika en volgens Chris Rawlinson (2011) se blog, Infographic: Cellphone usage + SA mobile stats, is Suid-Afrika gelys as vyfde in die wêreld wat die gebruik van selfoondata betref. Die VSA is sewende en Rusland het die eerste plek beklee. Verder berig 'n Unicef-navorsingsverslag (2012) dat Suid-Afrika die meeste huishoudings het met meer as een selfoon (57, 7%) in vergelyking met ander Afrikalande. Unicef (2012) verwys ook na 'n RIA (Research ICT Africa) huishoudelike opname wat in 2010 plaasgevind het wat aandui dat 70,6% van mense 16 jaar en ouer wat in stedelike gebiede woon, 'n selfoon besit, terwyl 48,9% van mense in landelike gebiede van selfone gebruik maak.

'n Rede vir die gewildheid van SMS'e kan gevind word in, onder andere, die einste eienskap wat outeuridentifikasie by SMS-boodskappe bemoeilik, naamlik bondigheid. SMS'e maak dit moontlik om vining en maklik 'n kort boodskap aan enigiemand met 'n selfoon te stuur. Omdat

SMS'e aanvanklik beperk was tot slegs 140 karakters, het dit beteken dat die outeur van 'n SMS op nuwe en innoverende maniere sy/haar taal moes 'aanpas' om sodoende te verseker dat soveel moontlik inligting in die bestek van 140 karakters ingepas kon word (Mobile Pronto, 2010).

Laasgenoemde 'aanpassings' het tot vele debatte gelei oor die moontlikheid dat SMS-taal die taalvermoë van 'n selfoongebruiker verlaag, maar hierdie argumente is, onder andere deur Crystal (2008, 2011), ongeldig bewys. Crystal (2008: 9) skryf die volgende ter ondersteuning van sy argument dat SMS-taal nie nadelig is nie:

> All the popular beliefs about texting are wrong, or at least debatable. Its graphic distinctiveness is not a totally new phenomenon. Nor is its use restricted to the young generation. There is increasing evidence that it helps rather than hinders literacy. And only a very tiny part of the language uses its distinctive orthography.

Vir hierdie navorsing word die argumente vir en teen SMS-taal egter nie in ag geneem nie. SMS-taal (of die gebrek daaraan in sommige gevalle) is vir die doeleindes van hierdie navorsing noodsaaklik aangesien die innoverende gebruik van taal tydens die stuur van SMS'e, al dan nie, deel uitmaak van elke outeur se SMS-idiolek en dit is die identifisering van hierdie idiolek wat outeuridentifikasie vergemaklik.

Daar moet beklemtoon word dat idiolek (ook SMS-idiolek) nie noodwendig bestaan uit opvallende verskille of selfs konstante variasies tussen individue nie. Idiolek is in die meeste gevalle subtiel en 'n individu se idiolek kan ook met tyd verander en aangepas word. Crankshaw (2012: 3) meen "uniqueness in writing has been documented to be not only something that is found from one individual to the next, but also within an individual themselves". Dit beteken dat die individu se styl van tyd tot tyd varieer en gevolglik sal daar verskille wees in een persoon se huidige skryfstel en daardie selfde persoon se skryfstyl oor 'n paar jaar. Crankshaw verwys verder na Halliday (1975) en Coulthard (2004) wat die "uniqueness of utterance principle" beskryf. Hierdie beginsel is gegrond op die waarneming dat wanneer twee individue oor dieselfde onderwerp skryf, hulle twee unieke tekste sal produseer, maar wanneer elk later weer oor die onderwerp skryf, sal dieselfde individu 'n opvallend verskillende teks produseer. Dit beteken dat tekste tussen twee individue verskil, maar so ook twee tekste oor dieselfde onderwerp wat deur dieselfde individu geproduseer is. Om hierdie rede is dit belangrik dat die

forensiese linguis, soos Grant (2010:9) noem, altyd moet aantoon waarom hy/sy ’n spesifieke verskynsel as idiolekties bestempel.

Afrikaanse SMS-taal (AST) het behalwe vir bondigheid ook ander kenmerkende eienskappe wat die identifisering van veronderstelde idiolek in SMS’e bemoeilik. ’n Bespreking van die algemene kenmerke wat in AST aangestref word volg in 2.7.1.

2.7.1 Afrikaanse SMS-taal

2.7.1.1 Die invloed van Engels

Ponelis onderskei in ’n berig wat heet *Die taal wat ons praat* (18/12/2009) tussen twee uiteenlopende variëteite van Afrikaans, naamlik die klassieke variëteit en die demotiese variëteit. Volgens Ponelis vertoon die demotiese variëteit die grootste invloed van Engels en word hierdie variëteit gebruik as “omgangstaal en spreektaal, in die familiekring, die buurt en die vriendekring”. Demotiese Afrikaans moet egter nie met Afrikaanse omgangstaal verwar word nie aangesien baie mense ’n omgangstaal gebruik wat glad nie demoties is nie. Die belangrikste struktuurkenmerk van demotiese Afrikaans is die “ontsettend hoë mate van Engelse invloed” (Ponelis, 2009).

Die invloed van Engels op Afrikaanse individue se taalgebruik is opmerkbaar verskillend, maar hierdie invloed kan nie misgekyk word nie. Veral wanneer SMS-taal gebruik word, word daar in sommige gevalle ’n verskeidenheid Engelse terme aangetref aangesien taalvermenging nie in hierdie situasie taboe is nie. Sommige van die Engelse terme wat gereeld in Afrikaanse SMS-boodskappe voorkom is:

- Hi
- Nice
- Cool
- Awesome
- Crazy
- Weird

- Got to go
- Be right back
- Friends for ever

Bogenoemde is enkele voorbeelde van Engelse woorde en frases wat in AST gebruik word. Engelse woorde en frases is egter nie die enigste kenmerkende aspek van AST nie en soms word vollengte sinne en woorde op verskeie maniere aangepas in ’n SMS-boodskap. Hier volg kort besprekings van ander kenmerke wat in AST aangetref word.

2.7.1 2 Logogramme

’n Logogram word beskryf as ’n verskynsel waar ’n letter, syfer, simbool of teken ’n hele woord of selfs ’n frase verteenwoordig. Logogramme, soos al die prossese wat volg, is ’n innoverende manier om soveel moontlik binne die beperkings van ’n SMS te skryf. Algemene logogramme in AST is onder andere[6]:

- @ – 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’ → ‘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.

  Byvoorbeeeld: *Gaan nou **zzzzzz*** (slaap).
- n – en: *ek **n** jy moet gaan*.
- c – see: *ons gaan **c** toe.*
- 6 – ses: *dit was ’n suk**6***

[6] Verskeie van die voorbeelde wat gebruik is, is van die internetbron http://blessieland.webs.com/101smsafkortings.htm verkry. Die lys is deur Leon van der Vyver saamgestel.

By logogramme word die *klank* van ’n letter ook soms gebruik om ’n hele woord te verteenwoordig, byvoorbeeld: x (ek’s), n (en). Tot ’n mindere mate kan k (ek) ook in hierdie kategorie geplaas word. Met ander woorde, die outeur buit die fonetiese eienskappe van ’n letter uit sodat ’n enkele letter ’n woord verteenwoordig.

2.7.1.3 Piktogramme/Beeldskrif

Die term ‘piktogram’ of beeldskrif verwys na gevalle in ’n taal waar beelde of prente gebruik word om ’n woord of konsep te verteenwoordig. Hiërogliewe is ’n bekende voorbeeld van beeldskrif waar tekeninge of afbeeldings gebruik is om woorde, klanke en sillabes voor te stel. In hedendaagse SMS-taal het ’n soortgelyke fenomeen ontstaan in die vorm van *smileys* (lagtekens).

Die term *smiley* verwys na ’n gestileerde voorstelling van ’n glimlaggende gesig. Hierdie voorstelling was aanvanklik bloot ’n dubbelpunt met ’n enkel hakie en het soos volg op die skerm vertoon: **:)**. Die voorstelling het ook gevarieer en is soms as **:-)** aangebied. In 1953 het die vorm verander na ’n ronde, geel, twee dimensionele gesiggie (Wikipedia 2013):

Die term *smiley* verwys vandag nie meer net na die geel glimlaggende gesig nie, maar sluit eerder alle prentjies in wat op ’n soortgelyke gestileerde manier voorgestel word: ☺ ☹ **;-)** ens. Die term *smiley* word vandag vervang met die term *emojis* of *emoticons* (emotikons). Emotikons is klein prentjies wat die outeur by ’n SMS-boodskap kan insluit en óf as vervanger van ’n woord kan optree óf bydra tot die inhoud van die boodskap (versiering of beklemtoning). Dié prentjies word ook gebruik om emosie in ’n boodskap uit te druk. Emotikons word met ander woorde gebruik as ’n illustrasie van iemand se gemoedstoestand of om die stemming of aard van ’n situasie uit te beeld. Emotikons is meer kompleks as basiese lagtekens en het daarom ook wyer toepassings (Wikipedia, 2013).

Ter illustrasie van die bogenoemde funksies van emotikons volg hier enkele voorbeelde:

2.7.1.3 (a) Vervanging

- Ek is ☺ *of* *☺ (*Net die piktogram) *Ek is gelukkig*.
- Is jy ☹ *of* *☹ ? (*Net die piktogram) *Is jy ongelukkig / kwaad?*
- Sy **:'-(** *Sy huil*.
- Ek **:-D** nou lekker! *of* ***:-D** (*Net die piktogram) *Ek lag nou lekker!/*Ek lag.*

2.7.1.3 (b) Versiering of beklemtoning

- *Dankie dat jy hier was, dit was lekker!* ☺
- *Geniet jou familie-naweek* **;-)** *Hier kan die knipoog aandui dat die outeur sarkasties is en weet dat die ontvanger nie uitsien na die naweek nie – die konteks sal die interpretasie bepaal.
- Ek kan nie glo wat ek hoor nie! **:-o**
- Jy beter nie uitpraat nie. **:-x**
- Ek is lief vir jou <3 (hart) / **:-*** (soen)

2.7.1.4 Verkortings

Daar is verskeie maniere waarop verkorting in SMS'e gebruik word. Olivier (2013: 490) verwys na Heyns (2009) en Saal (2012) wat onder andere twee maniere van verkorting noem. 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 as sinkopee in formele taalkunde) beteken dat 'n letter (klank) of uit 'n woord weggelaat word. By weglatings is dit gewoonlik die vokale wat uitgelaat word om die woord te verkort en sodoende ook die spoed waarmee 'n SMS getik word te vermeerder. Algemene weglatings in AST is onder andere die volgende:

- skt – *skat*
- ltr/latr – *later*
- wanr – *wanneer*
- nj – *en jy*
- prt – *praat*
- hkm? – *hoekom?*

'n Derde verskynsel wat onder verkorting gekategoriseer kan word, is die gebruik van "initialisms". "Initialisms" (inisiale) is 'n term wat deur Crystal (2008: 41–45) gebruik word om 'woorde' te beskryf wat gevorm word deur die eerste letters van twee of meer woorde te kombineer. Die verskil tussen inisiale en akronieme is dat inisiale meer gereeld in AST voorkom, gewoonlik nie in die standaardvorm van die taal aanvaar word nie en dat 'nuwe' inisiale deurentyd geskep word. Akronieme kom egter ook soms in SMS'e voor, maar daar word gewoonlik van bekende akronieme gebruik gemaak (byvoorbeeld: FB, TUKS, KFC ens.). By inisiale word die invloed van Engels weereens opgemerk. Daar is verskeie Engelse frases wat ook deur middel van inisiale verkort word en in AST voorkom.

- brb – *be right back*
- gtg – *got to go* ('g2g' kom ook algemeen voor)
- wmj? – *wat maak jy?*
- hgd? – *hoe gaan dit?*
- lib – *lê in bed*
- lol – *laugh out loud*
- vmi – *vertel my iets*
- wwjw? – *wat wil jy weet?*

Klisis word ook as ’n meganisme gebruik om woorde in SMS’e te verkort. Hier is die fonetiek weereens van belang aangesien sommige SMS-outeurs die woorde wat reeds in die taal geassimileer word, skryf (tik) soos hulle klink, byvoorbeeld: isi (*is nie* of *is die*), ini (*in die*), vani (*van die*), kani (*kan nie*), weti (*weet nie*), waars (*waar is*) ens.

2.7.1.5 Niestandaardspellings

Een van die kenmerkendste eienskappe van ’n SMS is die feit dat die spelling in die meeste gevalle afwyk van die standaardvorm. Die rede hiervoor is óók omdat niestandaardspellings woorde verkort en daarom meer woorde en karakters per SMS moontlik maak. Enkele voorbeelde van niestandaardspellings in AST is:

- hys – *huis*
- leka – *lekker*
- ni – *nie*
- di – *die*
- awsum – *awesome*
- grute – *groete*
- hu – *hoe*

Dit is soms baie moeilik om tussen ‘niestandaardspellings’ en ‘verkortings’ te onderskei. Wanneer veral kort woorde soos ‘die’ en ‘nie’ afgekort word na ‘di’ en ‘ni’word kategorisering bemoeilik. Myns insiens is woorde wat reeds kort is en waarvan enige verdere verkorting nie ’n groot verskil aan die lengte van die boodskap sal maak nie, niestandaardspellings eerder as verkortings, soos byvoorbeeld: ‘di’, ‘ni’, ‘ello’ (hello), ‘oki/ok’ (okay) ensovoorts. Daar is egter steeds in sommige gevalle nie ’n duidelike onderskeid nie, selfs al word die laasgenoemde reël toegepas, en gevolglik berus die besluit uiteindelik op die individu wat die teks analiseer.

2.7.1.6 Gebruik van hoofletters en kleinletters

Die afwesigheid van hoofletters in SMS'e is nie vreemd nie. Die rede hiervoor is moontlik dat die outeur by sommige selfone 'n aparte sleutel moet druk om die hoofletterfunksie van die selfoon te aktiveer (behalwe by die begin van 'n boodskap waar hoofletters altyd outomaties geaktiveer word). In 'n poging om tyd te spaar tik outeurs dan gewoonlik bloot die hele SMS in kleinletters. Die meeste selfone het egter vandag 'n ingeboude funksie wat sorg dat die letter na 'n punt outomaties 'n hoofletter word. Die outeur moet wel steeds hoofletters in enige ander gedeelte van die teks self aktiveer. Hoofletters word veral vir beklemtoning in die teks gebruik. 'n Voorbeeld van 'n SMS-boodskap met minimale hooflettergebruik (en ter verdere illustrasie van bogenoemde verskynsels) sal soos volg lyk:

Hi mev. huganit? xwl net weet of d hysi bd see d nawk vanaf Sat8 beskb of oop v bespr is. ds net x n my girlfriend. Danku.

(Seegogga se Bloggie, 2010)

Engelse woorde en frases, logogramme, piktogramme, verskeie vorme van verkortings, niestandaardspelling en die gebruik van hoofletters en kleinletters is slegs enkele eienskappe van Afrikaanse SMS-taal wat gebruik word om die veronderstelde idiolek van 'n outeur te identifiseer. Hierdie eienskappe verskil van persoon tot persoon en kan ook in een persoon se skryfstyl oor 'n tydperk varieer. Om hierdie rede is dit soms baie moeilik om die idiolek van een persoon te identifiseer as daar nie genoeg data is om die waarnemings van die navorser te ondersteun nie. In die volgende hoofstuk word die metodes bespreek wat in die huidige navorsing gebruik is om die moontlike idiolek van die deelnemers aan die studie te identifiseer en sodoende die outeurs van die tekste tot 'n mate van mekaar te onderskei.

Hoofstuk 3: Metodologie

3.1 Inleiding

Die metodes wat in die huidige navorsing gebruik word, word gesamentlik as ’n gemengde metode beskou aangesien beide ’n kwalitatiewe en kwantitatiewe metode gebruik word om die data te analiseer (Dörnyei, 2007: 44; Angouri, 2010: 29–30). Volgens Dörnyei (2007: 164) is daar drie redes waarom navorsers van die gemengde metode gebruik maak. Die huidige navorsing ondersteun die twee menings dat ’n gemengde metode eerstens die begrip van ’n komplekse saak uitbrei en tweedens die resultate van die studie meer betroubaar en verifieerbaar maak (Dörnyei, 2007: 164). Angouri (2010: 29–30) meen 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 meen: “combining 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 word die gemengde metode gebruik omdat beide kwalitatiewe en kwantitatiewe analises nodig is om die navorsingsvrae te beantwoord:

1) Kan daar in Afrikaans ’n generiese SMS-taal geïdentifiseer word wat outeuridentifikasie sou 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 verdagte SMS-teks te identifiseer met die beperkte data wat tipies ter beskikking is?

Die huidige navorsing is ’n pseudo-studie wat beteken dat dit nie ’n analise 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 verdagte outeurs wat die verdagte teks (Teks X) wat met ’n bepaalde misdaad verbind word, kon geskryf en gestuur het. Die navorser moet vasstel of dit moontlik is om die ware outeur van Teks X te identifiseer.

Omdat die navorsing poog om ’n situasie te skep wat naby aan werklike forensies-linguistiese scenario’s is, 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-analise afgehandel is. So ’n mate van sekerheid is nie tot op hede moontlik nie. Analises 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. Om laasgenoemde rede word ’n gemengde metode ingespan om die analises uit te voer. ’n Gemengde metode sal, in hierdie geval, teoreties ’n hoër mate van sekerheid oor die outeur van ’n teks tot gevolg te hê.

In vorige navorsing is reeds bepaal dat die ideale forensies-linguistiese scenario, ten opsigte van outeuridentifikasie, uit min moontlike outeurs en’n groot hoeveelheid tekste (of verskeie lang tekste) met ’n homogene onderwerp bestaan. So ’n perfekte scenario is egter onmoontlik in die werklikheid (Luyckx en Daelemans, 2011; Juola, 2006; Stamatatos, s.a.). In sommige gevalle kan ’n soortgelyke ‘ideale’ scenario wel voorkom wanneer die polisie reeds ’n verdagte geïdentifiseer het voor hulle die forensiese linguis nader om die tekste te analiseer. Dit beteken dat die forensiese linguis twee of drie verdagte outeurs (die hoofverdagte en een of twee ander moontlike outeurs) se tekste met mekaar vergelyk. In so ’n geval is die proses van outeuridentifikasie eenvoudiger, afhangende van die hoeveelheid teks en die lengte van die teks wat beskikbaar is.

Die probleme rondom die analisering van SMS-boodskappe in outeuridentifikasie is hoofsaaklik die bondigheid van die boodskappe. Die forensiese linguis moet daarom hipoteties oor ’n groot hoeveelheid SMS-boodskappe van een outeur beskik wanneer hy of sy poog om idiolektiese eienskappe te identifiseer wat die outeur en die bepaalde boodskappe met mekaar sal 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-boodskappe van elke outeur beskik en die getal moontlike outeurs meer as tien is. So ’n scenario

kan moontlik voorkom wanneer daar bepaal moet word watter persoon binne ’n bende ’n bepaalde SMS gestuur het, watter persoon in ’n klas of graad ’n bepaalde SMS gestuur het of watter persoon in ’n besigheid (of afdeling van ’n besigheid) ’n bepaalde SMS gestuur het. Dit is slegs drie scenario’s waar daar van die forensiese linguis verwag sal word om een outeur uit ’n groot groep moontlike outeurs te identifiseer. Navorsing wat outeuridentifikasie in ‘klein hoeveelhede’ kort tekste ondersoek is reeds uitgevoer in die verlede, 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).

Enige resultate wat uit die huidige navorsing verkry word, word egter as van belang beskou. Selfs negatiewe resultate sal steeds van waarde wees. Negatiewe resultate sal bloot beteken dat dit onwaarskynlik is om ’n outeur te identifiseer wanneer daar sowat 13 verdagtes is en die forensiese linguis slegs enkele SMS-boodskappe (5 tot 10) van elke verdagte outeur tot sy/haar beskikking het. So ’n resultaat sal natuurlik ook aanleiding kan gee 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.

3.2 Dataversameling

Soos reeds aangedui, poog die huidige navorsing om ’n realistiese forensies-linguistiese scenario te skep met verskeie verdagte outeurs en slegs enkele tekste van elke outeur tot die forensiese linguis se beskikking. So ’n scenario verteenwoordig egter slegs *een moontlike* scenario waar die hoofdoel van die ondersoek outeuridentifikasie is. Besonderhede in verband met die seleksie van die deelnemers asook die prosedure van dataversameling word volgende bespreek.

3.2.1 Deelnemers

Ivankova en Creswell (2009: 149) meen dat die neem van steekproewe by kwantitatiewe data verskil van dié by kwalitatiewe data. By kwantitatiewe data is die steekproef gewoonlik baie groot en dit word lukraak uitgevoer terwyl die neem van steekproewe by kwalitatiewe navorsing, soos die geval in die huidige navorsing, baie kleiner en meer doelgerig is omdat die steekproef moet lei tot “an in-depth understanding of the explored phenomenon”. In die huidige studie

bestaan die steekproef uit 13 deelnemers tussen die ouderdom 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 13 deelnemers se data aan die kriteria voldoen. Een van die deelnemers het slegs twee SMS-boodskappe aan die navorser gestuur en gevolglik kon hierdie deelnemer se teks nie gebruik word nie aangesien dit te kort was. Een van die kriteria vir deelname was naamlik dat elke deelnemer ten minste vyf SMS-boodskappe aan die navorser moet stuur.

Dörnyei (2007: 127–129) verwys na verskeie strategieë wat gebruik kan word by die neem van ’n steekproef. Die strategie wat van toepassing is op hierdie studie is die neem van ’n sogenaamde ‘homogene’ steekproef. Dit beteken dat die deelnemers almal min of meer dieselfde kenmerke deel of almal dieselfde ondervinding het wat van toepassing is binne die bepaalde studie. In hierdie geval is al die deelnemers individue binne ’n bepaalde ouderdomsgroep wat onder andere hoofsaaklik deur middel van selfone, en meer spesifiek, SMS-boodskappe, kommunikeer. Dörnyei (2007: 127) meen dat’n homogene steekproef effektief is aangesien dié strategie die navorser toelaat om ’n diepgaande analise uit te voer wat lei tot die identifisering van algemene patrone binne ’n groep individue wat almal soortgelyke kenmerke deel. In die huidige navorsing is die groep deelnemers homogeen ten opsigte van ouderdom en die feit dat almal moedertaal Afrikaanssprekende studente is wat aan die universiteit van Pretoria studeer. Beide mans en vrouens het aan die navorsing deelgeneem, maar daar is nie rekord gehou van die hoeveelheid mans teenoor die hoeveelheid vrouens wat deelgeneem het nie aangesien die geslag van die deelnemers nie van belang is in die huidige navorsing nie.

Daar moet egter in gedagte gehou word dat die resultate wat uit hierdie studie verkry word nie veralgemeen kan word in die groter gemeenskap nie, aangesien 13 individue nie verteenwoordigend is van die totale gemeenskap nie.

Volgens ’n studie wat in 2012 aanlyn gepubliseer is (PewInternet, 2012), is die ouderdomsgroep 18 tot 29 die groep wat die meeste van SMS-boodskappe vir kommunikasiedoeleindes gebruik maak. Volgens dié studie gebruik 97% van die individue in hierdie groep SMS-boodskappe om te kommunikeer. 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 laasgenoemde bloot ‘nuwer’ en ‘goedkoper’ maniere is om ’n SMS te stuur. In hierdie navorsing word die term ‘SMS’ gevolglik gebruik om

te verwys na enige “short message service” wat dit vir individue moontlik maak om vinnig elektroniese boodskappe te stuur en wat nie ander elektroniese kommunikasie soos e-posboodskappe, *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.

Die groep van 13 deelnemers wat aan hierdie studie deelgeneem het, is tussen die ouderdomme 18 en 23. Volgens navorsing is individue tussen 18 en 22 jaar die mees aktiewe groep selfoongebruikers (PewInternet, 2012), maar vir die huidige navorsing is 23-jarige deelnemers ook toegelaat. Daar word aangeneem dat hierdie ouderdomsgroep selfoongebruikers ook gekonfyt is in die gebruik van SMS-taal. Elke deelnemer word deur ’n nommer (1 tot 14) geïdentifiseer om te verseker dat die anonimiteit van die deelnemers gehandhaaf word.

3.2.2 Datastel: ’n SMS-korpus

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 (omtrent 1000 woorde) in die navorsing te gebruik nie.

Die datastel in hierdie studie, soos in hoofstuk 1 genoem, bestaan uit 2434 woorde. Elke deelnemer het 5 tot 10 SMS’e aan die navorser gestuur. Die SMS’e van elke deelnemer word in **Bylaag 1** ingesluit. Eers is daar gehoop dat die deelnemers volledige SMS’e aan die navorser sou stuur wat tussen 60 en 70 woorde lank is, maar SMSPortal se sisteembeperkings (soos bespreek onder 3.2.3.1) het dit onmoontlik gemaak. Die sisteem het die datastel aansienlik laat krimp.

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.

Tabel 2: Die hoeveelheid woorde per teks vir elke deelnemer

Deelnemers	**Hoeveelheid woorde per teks**
Teks X	126
Deelnemer 1	171
Deelnemer 2	174
Deelnemer 4	227
Deelnemer 5	324
Deelnemer 6	213
Deelnemer 7	164
Deelnemer 8	103
Deelnemer 9	150
Deelnemer 10	121
Deelnemer 11	128
Deelnemer 12	153
Deelnemer 13	204
Deelnemer 14	176

Die datastel bestaan met ander woorde uit een korpus wat saamgestel is uit die 5 tot 10 SMS'e van elke deelnemer. 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. Die rede vir die gebruik van persoonlike SMS'e wat reeds deur die deelnemers geproduseer is, is om te verhoed dat die deelnemers hul SMS-styl verander omdat hulle weet die SMS'e word deur 'n navorser gelees. Soos reeds gesê, is enige kort teks wat met behulp van 'n diensverskaffer deur middel van 'n selfoon gestuur word, as geldige data in hierdie studie gebruik. Moderne slimfone het die opsie om boodskappe vanuit toepassings soos Whatsapp en BBM te knip en plak in 'n SMS in. Dit beteken dat die deelnemers boodskappe uit ander toepassings aan die navorser per SMS kon stuur, selfs al gebruik die deelnemers nie noodwendig meer die tradisionele SMS as daaglikse kommunikasiemiddel nie.

Die korpus is gebruik om vas te stel of daar *idiolektiese* taalgebruik onder die 13 deelnemers bestaan. Die korpus is ook gebruik om vas te stel of daar, alternatiewelik tot idiolektiese taalgebruik, *generiese* SMS-taal onder hierdie groep Afrikaanse SMS-gebruikers bestaan. Daarna is vasgestel of die outeur van 'n verdagte teks met die data beskikbaar geïdentifiseer kan word.

Die selfoonnommer van elke deelnemer wat SMS'e na die *short code* gestuur het, is op die sisteem geregistreer. Die selfoonnommers is gebruik om te verseker dat elke deelnemer se SMS'e onder die korrekte nommer (1 tot 14) gestoor is. Die SMS-teks van elke deelnemer is uit die e-pos geknip en in 'n Word-dokument geplak. Geen veranderinge is tydens hierdie fase aan die tekste[7] aangebring nie.

3.2.3 Datastel: Insameling

3.2.3.1 SMSPortal

Vir die ontvang van die SMS'e is daar van die *BulkSMS*-sisteem, SMSPortal[8], 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 diens het twee opsies waarvan die *keyword*- (trefwoord) opsie in hierdie geval gebruik word. Die trefwoordopsie maak dit moontlik om SMS'e wat na 'n bepaalde nommer (*short code*) gestuur word te selekteer op grond van die eerste woord in die SMS. In die huidige navorsing is die woord "Forensic" as die trefwoord geregistreer. Dit beteken dat die SMS'e wat na die bepaalde nommer gestuur word met die woord 'Forensic' moet begin sodat die sisteem die SMS'e kan herken. Die SMS'e is vervolgens per e-pos aan die navorser gestuur sodat die SMS-teks in 'n formaat is wat bloot in 'n ander dokument in geknip en geplak kan word vir analisering.

Die e-possisteem waarmee die SMS'e na die navorser gestuur is, kon al die lagtekens en ander piktogramme identifiseer wat in 'n gewone SMS gebruik word. Toepassings soos Whatsapp en BBM het egter 'n groter verskeidenheid lagtekens en piktogramme wat by 'n boodskap gevoeg kan word. Wanneer 'n boodskap wat sulke lagtekens/piktogramme bevat in 'n SMS in geknip en geplak word, registreer sulke tekens nie meer nie en word hierdie bepaalde lagtekens/piktogramme bloot met 'n blokkie in die SMS-formaat vervang: □. Hierdie blokkie word as 'n vraagteken (?) in die e-pos oorgedra. Dit beteken dat elke 'vreemde' lagteken of piktogram wat in die SMS oorgekopieer is as 'n blokkie verskyn in die SMS en as 'n vraagteken in die teks wat vir analisering gebruik word. Die feit dat nie alle tekens deur die SMS-stelsel 'gelees' kan word nie is net in sommige gevalle problematies. Wanneer vraagtekens in die

[7] Die begrip 'teks' word hier gebruik om te verwys na elke deelnemer se SMS'e. Deelnemer 1 se teks sal met ander woorde uit al die SMS'e bestaan wat Deelnemer 1 ingestuur het vir ontleding.
[8] http://www.smsportal.co.za/

middel van ’n sin voorkom of nie in die konteks pas nie kan die navorser aanneem dat daardie bepaalde vraagtekens lagtekens of piktogramme voorstel:

- *Hi, gister se braai***?** *was sooo lekker!*
- *Ek het die nuutste boek gekry***?**

Wanneer die vraagtekens egter aan die einde van sinne voorkom is dit moeiliker om te bepaal of die outeur ’n lagteken of piktogram bygevoeg het en of hy/sy net bloot van meer as een vraagteken gebruik maak:

- *Wanneer het jy dit gehoor***??** (Die navorser kan nie met sekerheid weet of een van die twee vraagtekens nie dalk ’n lagteken is wat verbasing of skok aandui nie)

Die *soort* lagteken of piktogram is nie vir die huidige navorsing van belang nie. Die navorser moet slegs kan bepaal of ’n lagteken of piktogram in die boodskap gebruik is.

’n Verdere beperking van SMSPortal is die feit dat die sisteem slegs SMS’e tot op 60 karakters kan lees. In SMS’e tel enige woord, spasie of leesteken as ’n karakter en gevolglik is byna al die SMS’e wat deur die deelnemers ingestuur is afgesny in die middel van ’n sin. Hierdie beperking het die data wat uiteindelik beskikbaar was verminder.

3.3 Instrumente

Drie instrumente is vir die data-analise gebruik. Antconc en WordSmith Tools is aanlyn sagteware wat vir verskeie linguistiese analises gebruik kan word, terwyl die n-gramanalise wat in die navorsing gebruik is tradisioneel gebruik word in taalidentifiseringsprogramme en sagteware.

3.3.1 Antconc[9]

Antconc 3.4.1 is gratis aanlyn sagteware wat gebruik word vir analises in korpuslinguistiek. Die sagteware is ontwerp deur Laurence Anthony (Ph.D) van Waseda Universiteit in Tokyo, Japan en word soos volg beskryf:

> *Antconc* is a freeware, multiplatform tool for carrying out corpus linguistics research and data-driven learning. (Anthony, 2014: 1)

Antconc is die eerste keer in 2002 bekendgestel en was aanvanklik ’n eenvoudige KWIC (Key Word in Context)-konkordansieprogram. Antconc 1.0 is gratis aanlyn beskikbaar gestel en deur verskeie gebruikers afgelaai en getoets. Die kommentaar van die gebruikers, beide positief en negatief, is gebruik om die sagteware op te gradeer en die funksies uit te brei (Anthony, 2005: 730). Antconc word in sommige gevalle as ’n alternatief tot WordSmith Tools gebruik aangesien die sagteware gratis beskikbaar is en versoenbaar is met Windows, Macintosh en Linuxbedryfstelsels. Antconc benodig geen addisionele sagteware wanneer dit op ander sisteme as Windows geïnstalleer word nie en Antconc is ook in staat om tekste in byna enige taal te prosesseer, insluitende Asiatiese tale soos Sjinees en Japanees (Wilkinson, 2012).

Antconc-sagteware is gebruik om die data statisties te analiseer en te interpreteer aangesien dit maklik bekombaar is en, soortgelyk aan Wordsmith Tools, het die nuutste weergawe van Antconc ook verskeie funksies wat gebruik kan word om analises op ’n teks of korpus uit te voer. Die *Keyword list*- en *Word list*-funksies van Antconc is in die huidige navorsing gebruik.

3.3.2 WordSmith Tools

WordSmith Tools (WST) is ontwikkel deur Mike Scott van die Universiteit van Liverpool. Michell (2013: 72) verwys na Guillén Nieto e.a. (2008) se beskrywing van WST as:

> [...] an organic integrated suite of programs for examining the manner in which grammatical and lexical features act in a text.

In ’n bespreking van WST 5 wys Prinsloo en Prinsloo (2011) daarop dat WST hoofsaaklik gebruik word om patrone in tekste te identifiseer. In die huidige navorsing is daar hoofsaaklik van Antconc gebruik gemaak om die data te analiseer, maar daar is ook gepoog om die data deur

[9] Beskikbaar by: http://www.antlab.sci.waseda.ac.jp/software.html

WST te analiseer sodat die resultate van die twee analises met mekaar vergelyk kon word. WST sluit ook *Keyword*- en *Word List*-funksies is, maar slegs die *Keyword*-funksie in WST is gebruik aangesien die *Word List*-funksie van albei sagteware dieselfde resultate oplewer. Vergelyk paragraaf 3.4.2.3.

3.3.3 Die n-gramanalise

Die n-gramanalise in die huidige navorsing is gebaseer op die n-gramanalise vir taalherkenning wat deur Cavnar en Trenkle (1994) gebruik is[10]. Hierdie metode is ontwikkel om teks kategorisering in groot hoeveelhede elektroniese dokumente te vergemaklik. Die n-gramanalise wat deur Cavnar en Trenkle ontwikkel is is verdraagsaam teenoor grammatiese foute en spelfoute in elektroniese dokumente. Om dié rede kan hierdie metode ook die unieke aard van SMS-taal analiseer sonder te veel probleme. Die n-gramanalise werk deur n-gramprofiele te bereken en te vergelyk. In die eerste plek word die profiel van die opleidingsdata bepaal (dit is die korpus waarteen die ander tekste vergelyk gaan word) en daarna word die profiel vir elke dokument wat geklassifiseer moet word bepaal. Laastens word ‘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 hoeveelheid ooreenkomste met die opleidingsdata toon.

In die huidige navorsing is die beperkte data problematies. Aangesien die opleidingsdata wat gebruik is in sommige gevalle kleiner is as die tekste wat daarmee vergelyk word, beteken dit dat die resultate moontlik anders sou lyk indien daar meer data tot die navorser se beskikking was. Met die beperkte data was dit nie te min moontlik om bruikbare resultate te verkry. Die resultate word op p. 110 bespreek.

3.3.4 Die Chi-kwadraattoetse (Pearson Chi-kwadraattoets en die Yates korreksie)

3.3.4.1 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 die sogenaamde “Pearson’s chi-square test”

[10] Die n-gramanalise word meer volledig op p. 37 en p. 127 bespreek.

wat ook as die "chi-square test for independance" 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" (moontlikheid van sukses) dat twee tekste deur dieselfde outeur geproduseer is, te bepaal. Hierdie 'moontlikheid van sukses' word deur 'n *p-waarde* aangedui. Die graad van waarskynlikheid "dat die resultaat van die vergelyking *verkeerd* kan wees, met ander woorde dat die verskynsel wat waargeneem word slegs toevallig voorgekom het" moet volgens Kotzé (2007: 391) op 0.05 (5%) gestel word by die Chi-kwadraattoets om "toe te laat vir soveel as moontlik gevalle van beduidende verskille tussen die dokumente wat vergelyk word [...]". Die resultate van die Chi-kwadraattoets bepaal of 'n nulhipotese aanvaar of verwerp word.

Die nulhipotese (H_0) is dat daar geen verhouding tussen die veranderlikes is nie (geen verhouding tussen die waargeneemde frekwensies (O)[11] en die verwagte frekwensies (E)[12] 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.

Indien $p < 0.05$ / $p=0.05$ word die nulhipotese aanvaar. Indien $p > 0.05$ word die alternatiewe hipotese aanvaar. In 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.

$p > 0.05$ 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%

[11] Die **waargenome frekwensies** (*observed frequencies* - O) verwys na die data wat die navorser ingesamel of waargeneem het.
[12] Die **verwagte frekwensies** (*expected frequencies* - E) verwys na die frekwensies wat die navorser sal voorspel in elke sel van die tabel (Easton en McColl, s.a.)

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.

Dit is nie nodig om die Chi-kwadraattoets elektronies uit te voer nie. Die toets kan met die hand gedoen word, maar dit is tydrowend wanneer die navorser ’n groot hoeveelheid data het om te analiseer en die moontlikheid van berekeningsfoute neem toe. ’n Eenvoudige uiteensetting van die Chi-kwadraattoets wat in die huidige navorsing gebruik is, word ingesluit om die berekening van die resultate te verduidelik.

Die Chi-kwadraattoets vir onafhanklikheid maak van die vergelyking $(f_o-f_e)^2 / f_e$ gebruik om die p-waarde te bepaal. In hierdie vergelyking dui f_o op die frekwensie van die waargeneemde data (O/o) en f_e dui op die frekwensie van die verwagte waardes (E/e). ’n Alternatiewe, en meeralgemene, manier om die vergelyking te skryf, lyk soos volg:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad [13]$$

Hierdie twee weergawes van die vergelyking lei tot dieselfde resultate aangesien albei interpreteer word as:

$$\text{Chi}^2 = \frac{(\text{Observed frequency } (O) - \text{Expected frequency } (E))^2}{\text{Expected frequency}}$$

[13] Die sigma (∑) dui daarop dat die som van elke *i* wat bereken word gegee moet word. Met ander woorde laasgenoemde formule word vir elke sel in die tabel gedoen en daarna word die totaal uitgewerk.

Resultate vir die Chi-kwadraattoets word verkry nadat die volgende vier stappe in die proses voltooi is.

1. Die data wat vergelyk word, word in tabelvorm ingesleutel. Die syfers in elke kolom is die hoeveelheid kere wat elke woord in 'n bepaalde teks voorkom. In die tabel hier onder word twee tekste ter illustrasie gebruik.

	Teks x	D1	**TOTAAL**
Ons	*6*	*1*	**7**
En	*4*	*4*	**8**
moet	*4*	*1*	**5**
Ek	*3*	*7*	**10**
Kry	*3*	*1*	**4**
lekker	*3*	*0*	**3**
Wat	*3*	*1*	**4**
Die	*2*	*3*	**5**
Dit	*2*	*3*	**5**
TOTAAL	**30**	**21**	**51**

2. Die totale vir elke kolom en ry word uitgewerk asook die groottotaal. In hierdie geval is die groottotaal 51.

3. Daarna word die grade van onafhanklikheid [*degrees of freedom* (*df*)] bepaal. Die formule vir die *df* lyk soos volg:

   df = (die aantal rye minus een) x (die aantal kolomme minus een)

   Dus: (R-1)*(K-1)= 8*1 = **8** of (9-1)*(2-1)= 8*1= **8**

4. Hierna word die verwagte waarde vir elke waargeneemde waarde bepaal. Die verwagte waarde word verkry deur die totaal van die ry met die totaal van die kolom te vermenigvuldig en daarna deur die groottotaal vir die data (51) te deel.

   Die verwagte waarde vir die eerste waargeneemde waarde (6) sal met ander woorde soos volg bepaal word:

   (7*30)/N = 4.117 (afgerond na 4.12).

   Met hierdie formule word die verwagte waardes vir elke waargeneemde waarde in die tabel bepaal.

5. Laastens word die Chi-kwadraattoets gebruik om die Chi-kwadraatwaardes te bepaal.

6. Die som van die hoeveelhede wat deur elkeen van hierdie formules gegenereer word vorm uiteindelik die Chi-kwadraatwaarde van 11.1. Wanneer ’n p-waarde van 0.05 met ’n *df* van 8 op die Chi-kwadraattabel (’n distribusietabel van Chi-kwadraatwaardes) nagegaan word, is dit duidelik dat, indien daar ’n beduidende verskil tussen die twee tekste is, die Chi-kwadraatwaarde ten minste 15.507 moet wees. Dit beteken dat die twee tekste nie beduidend van mekaar verskil nie. Die Chi-kwadraatwaarde van 11.1 en die *df* van 8 gee ’n p-waarde van 0.194 (19.4%):

   $[(6 - 4.12)^2 / 4.12)] + [(b – b_1)^2 / b_1)] + [(c - c_2)^2 / c_2)]$ ens. = 0.86 + b+ c ens. = (byvoorbeeld) 0.194

   Die p-waarde van 19.4% bevestig die laasgenoemde waarneming: p>0.05 beteken dat die verskille tussen die tekste minder is en daarom bestaan die moontlikheid (’n 19.4% moontlikheid) dat die tekste deur dieselfde outeur geproduseer is.

Chaski (2001: 9) verwys na verskeie navorsers wat waarsku dat die volgende in ag geneem moet word voordat die Chi-kwadraattoets toegepas word:

> First, in applying the chi-square statistic we have to consider the size of our observed frequencies because these are used to calcultate the expected frequencies. The practical rule allows no more than 20 per cent of the expected frequencies generated during the calculation to be less than 5, while none of the expected frequencies can be less than 1.

In die huidige navorsing is die data beperk en voldoen gevolglik nie aan die laasgenoemde kriteria nie. Antconc is gebruik om die frekwensies van die woorde in elke deelnemer se teks asook Teks X te bepaal. Nadat die frekwensies van die 10 mees frekwente algemene woorde en die 11 mees frekwente funksiewoorde verkry is (vergelyk p. 112 en p.113) is dit duidelik dat byna 70% van die verwagte frekwensies minder as 5 sal wees en sommige verwagte frekwensies sal minder as 1 wees. Nietemin is die Chi-kwadraatwaardes vir elke vergelyking tussen Teks X en die deelnemers vir die 10 mees algemene woorde en die 11 mees algemene funksiewoorde uitgewerk en by **Tabel 6** en **Tabel 7** (p.114 en p.115) ingesluit. Uit die tabel is dit duidelik 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.

3.3.4.2 Die Yates korreksie

Op grond van die resultate in **Tabel 6** en **Tabel 7**, is daar besluit om Yates se korreksie te gebruik aangesien die data in die huidige navorsing duidelik te beperk is om enigsins bruikbare

resultate met die gewone Chi-kwadraattoets te behaal. 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. 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 dit bevat twee veranderlikes.

Die Yates korreksie voorkom gewoon 'n oorskatting van die statisitese gewig van die resultate van klein hoeveelhede data. Die korreksie word gewoonlik gebruik in gevalle waar een verwagte frekwensie minder as 5 is. Hierdie korreksie verlaag die Chi-kwadraatwaarde en verhoog die p-waarde. Om hierdie rede moet die resultate uit so 'n toets steeds met die nodige versigtigheid geïnterpreteer word. 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 waarna How2stats (2011) verwys is: Camilli en Hopkins (1978, 1979); Feinberg (1980); Larntz (1978) en Thompson, (1988).

Nietemin is die Yates korreksie in die huidige navorsing gebruik om te toets of dit enigsins tot meer bruikbare resultate sal lei. Uit **Tabel 8** en **Tabel 9** (p.119 en p.120) is dit egter duidelik dat die resultate van die Yates korreksie nie méér bruikbaar is as die resultate wat deur die Pearson Chi-kwadraattoets verkry is nie.

3.4 Analitiese metodes

3.4.1 Stilistiese analise

Die SMS-data is op twee maniere geanaliseer. Die eerste analise is ’n beskrywende, kwalitatiewe, stilistiese analise van die teks. McMenamin (2010: 488) beskryf stilistiese analises as die “study of style in language”. Stilistiese analises in die letterkunde het tradisioneel op die estetiese kwaliteit van uitdrukkings en die ooreenstemming van taalgebruik met bepaalde taalreëls gefokus. Moderne linguistiese stilistiese analises 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 is die waarneembare resultaat van die onbewuste keuses wat ’n outeur tydens die skryfproses maak. Hierdie onbewuste keuses word ’n gewoonte en dieselfde keuses kom gevolglik herhaaldelik in ’n outeur se skryfstyl voor:

> Stylistic variation is reflected as class characteristics observed in the writing of distinct social and geographical groups, and also as individual features observed in the idiolect of single writers who share a language or dialect.
>
> (McMenamin, 2010: 489)

Die stilistiese analise verteenwoordig die ‘menslike’element van die data-analise en is in wese teksanalise. Met teksanalise 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 stuur onder ‘normale’ omstandighede geproduseer is en die deelnemers bloot die SMS’e wat hulle wil stuur moes selekteer uit die SMS’e wat reeds op hul selfone beskikbaar was. Die teksanalise berus hoofsaaklik op die navorser se eie interpretasie en taalkennis (Lazaraton, 2009: 247–250). Tydens hierdie proses besluit die linguis watter kenmerke in die teks hy/sy wil selekteer vir verwerking. McMenamin (2010: 490) verwys na Wachal (1966) se modelle vir stilistiese analise en meen dat hierdie drie modelle steeds met sukses in die forensiese linguistiek gebruik kan word. Die eerste model staan as die ‘ooreenkomsmodel’ (*resemblance model*) bekend en word gebruik wanneer eksterne faktore die

hoeveelheid moontlike outeurs beperk. Dit gebeur wanneer daar inligting in ’n bepaalde teks voorkom waarvan net enkele individue kennis dra. Die tweede model is die ‘konsekwentheidsmodel’ (*consistency model*). Hierdie model word gebruik om te bepaal of verskeie tekste deur dieselfde outeur geskryf is. Die konsekwentheidsmodel word gereeld in outeuridentifikasie gebruik wanneer die forensiese linguis een of twee tekste het waarvan die outeur(s) bekend is en ander tekste waarvan die outeur(s) bepaal moet word. Die forensiese linguis sal gevolglik die konsekwentheid van eienskappe in die tekste probeer vasstel om sodoende te bepaal wie die moontlike outeur van die verdagte teks(te) is. Die derde model word die ‘bevolkingsmodel’ (*population model*) genoem. Die bevolkingsmodel word gebruik wanneer die hoeveelheid moontlike outeurs baie groot is. In so ’n geval word die ooreenkomsmodel bloot herhaaldelik gebruik totdat daar net een of twee moontlike outeurs oorbly. Vir die huidige navorsing is die konsekwentheidsmodel gebruik en daar is gebruik gemaak van ’n ontledingsblad wat die navorser self opgestel het. Hierdie blad bevat kleure wat verskillende eienskappe in die teks verteenwoordig. Byvoorbeeld:

- pers verteenwoordig logogramme
- blou verteenwoordig lagtekens (*smileys*)
- geel verteenwoordig leestekens

Die navorser gebruik die kleurgids om die onderskeie elemente in die tekste te merk. Nadat so ’n analise afgehandel is, kan die navorser waarnemings maak oor die onderskeie deelnemers se tekste. Dit is nie in alle gevalle moontlik om so ’n stilistiese analise met die hand uit te voer nie, maar die klein hoeveelheid data in die huidige navorsing het hierdie stilistiese metode toegelaat. Deur middel van hierdie metode kan die linguis ’n oorsig kry van die ‘linguistiese veranderlikes’ in ’n teks of in verskeie tekste. Linguistiese veranderlikes word deur McMenamin (2010: 491) gedefinieer as “the isolation of structural linguistic units that carry significance with respect to group or individual writing style”.

In die huidige navorsing is die volgende eienskappe in die onderskeie SMS’e gemerk:

- Leestekens
- Aanspreekvorme

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

Figuur 3.1 is ’n voorbeeld van een deelnemer se teks nadat ’n stilistiese analise daarop uitgevoer is. Dit is belangrik om daarop te let dat een eienskap miskien onder meer as een kleur gekategoriseer kan word. Die gebruik van ’n logogram wat eers in pers gemerk is kan later in ’n ander kleur gemerk word as die navorser besef dat die individu altyd dieselfde logogram gebruik of oorwegend van ’n bepaalde logogram gebruik maak.

Figuur 3.1: 'n Voorbeeld van 'n stilistiese analise deur gebruik te maak van 'n kleurgids.

Elke deelnemer se teks en Teks X is stilisties deur middel van die kleurgids geanaliseer. Daarna is die 17 kenmerke van Teks X (wat tydens die stilistiese analise 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 (vergelyk **Tabel 3** op p. 111). Dit is belangrik om daarop te let dat daar in sommige gevalle nie volledige data was om definitiewe gevolgtrekkings te maak nie. In Teks X het 'n onvolledige ellips, [..], byvoorbeeld voorgekom, maar party deelnemers het glad nie van 'n ellips gebruik gemaak nie en gevolglik was dit onmoontlik om te bepaal of die bepaalde deelnemer 'n volle ellips of 'n onvolledige ellips gebruik het. Die volledige tabel waaruit die persentasieooreenkomste bepaal is, word in **Bylaag 2** ingesluit.

3.4.2 Stilometriese analise

In die tweede fase van die analise, naamlik die stilometriese analise, is die sagtewareprogramme Antconc en WordSmith Tools asook n-gramanalise gebruik om die tekste statisties te verwerk. Hierdie gedeelte van die analise verteenwoordig die kwantitatiewe aspek van die gemengde

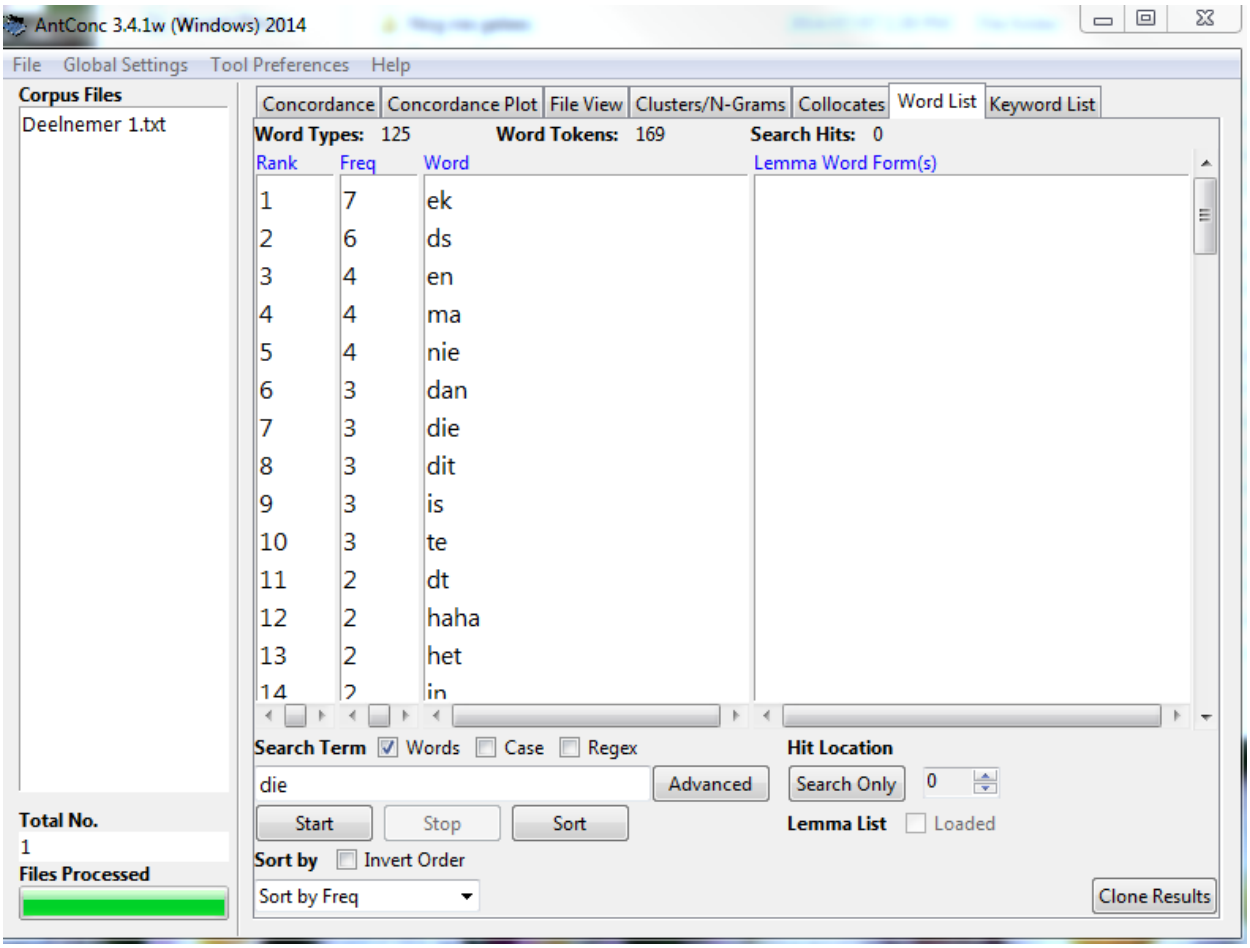
metode. Die doel van die stilometriese analise is om statisties te probeer vasstel watter woorde oorwegend deur ’n spesifieke individu gebruik word, hoe belangrik hierdie woorde in die bepaalde teks is en ook in watter patrone die woorde in die individu se sinne voorkom. Soos reeds aangedui is slegs twee van Antconc se funksies (*Keyword list* en *Word list*) vir die doeleindes van die huidige navorsing gebruik.

3.4.2.1 Antconc: Die *Word list*-funksie

Die *Word list*-funksie word gebruik om ’n lys te genereer van al die woorde in ’n bepaalde korpus asook die frekwensie waarmee die woorde in die korpus voorkom. Woorde wat die meeste in die korpus voorkom en gevolglik die hoogste frekwensie het, verskyn bo-aan die lys. Die *Word list*-funksie het ook die vermoë om woorde op grond van ‘stamvorme’ te tel (Anthony, 2004: 10). Volgens Anthony (2005: 732) is so ’n funksie baie handig aangesien dit interessante areas vir verdere ondersoek asook probleemareas in ’n korpus kan uitlig.

Dit is baie belangrik dat die teks of korpus wat gebruik word vir die *Word list* ‘skoon’ is. Hiermee word bedoel dat onvolledige woorde of simbole wat die sagteware as woorde beskou uitgehaal moet word voordat die teks opgelaai word. In die geval van die huidige navorsing het die tekste verskeie woorde gehad wat afgesny is aangesien die sisteem nie altyd die volledige SMS kon lees nie. Elke teks moes eers skoongemaak word van enige los letters of onvolledige woorde voordat die analises gedoen is en resultate gegenereer is. Al die tekste is hierna omgeskakel na ‘plain text’-formaat sodat Antconc die tekste kon lees. **Figuur 3.2** is ’n voorbeeld van ’n *Word List* wat in Antconc gegenereer is.

Figuur 3.2: 'n Voorbeeld van 'n Word list in Antconc



Die lyste van verskillende tekste kan met mekaar vergelyk word om sodoende vas te stel of daar ooreenkomste is tussen die frekwensie van sekere woorde in verskillende tekste. 'n *Word list* 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 4** en **Tabel 5** (p. 112 en p. 113) word die tabel gegee waarin die frekwensie van die eerste tien algemene woorde en 11 funksiewoorde van elke deelnemer en die verdagte teks met mekaar vergelyk is. 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.

Die *Word list*-funksie is ook gebruik om te bepaal of daar 'n generiese SMS-taal onder die groep deelnemers bestaan. Soos in 4.2.1 bespreek word, is die *Word list*-funksie van Antconc gebruik

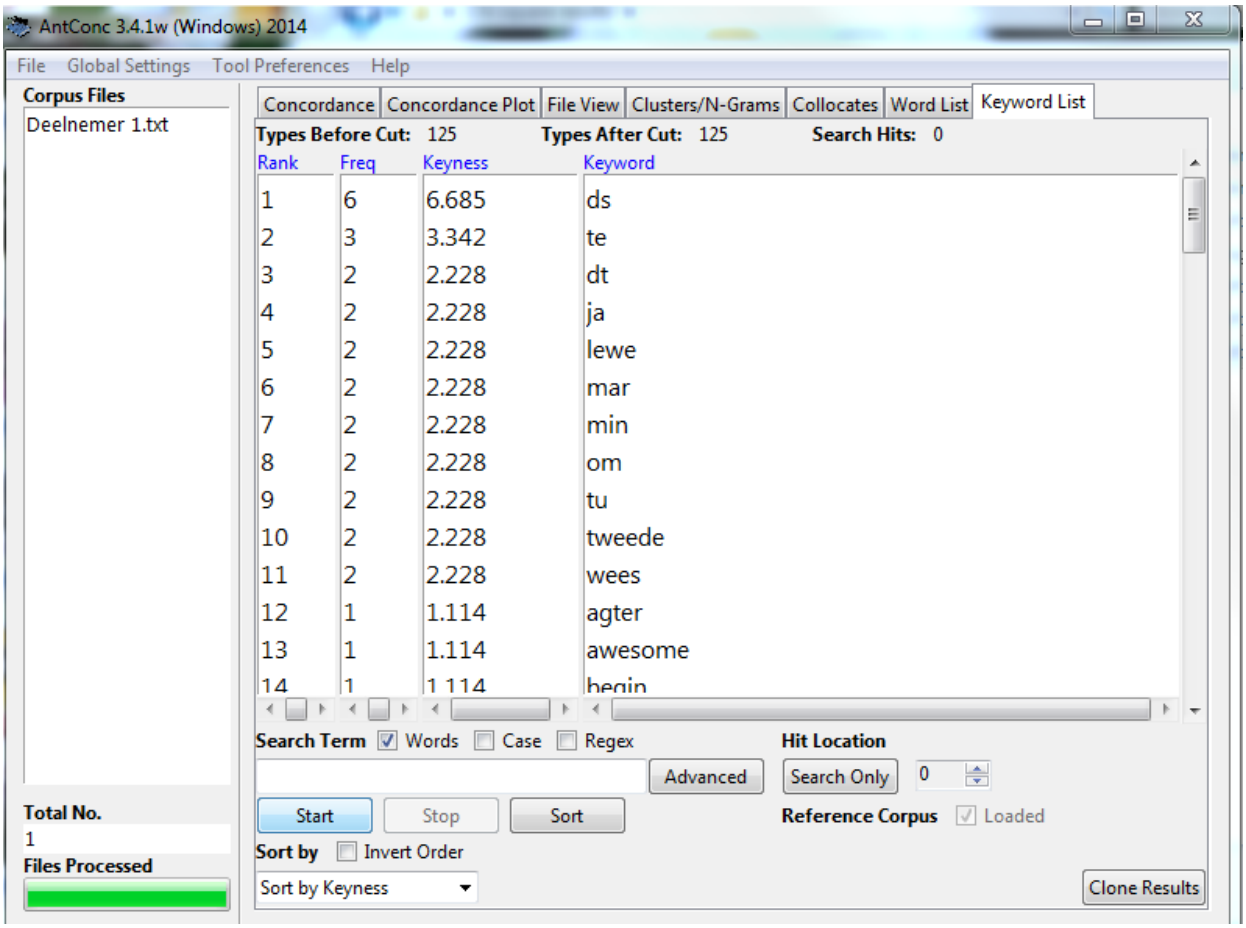
om een woordelys van al die deelnemers se tekste te produseer. Aangesien elke woord (op grond van spelling) apart gelys word, is dit maklik om die variasies op een woord vas te stel en sodoende te bepaal of daar wel ’n generiese SMS-taal onder die groep deelnemers bestaan.

3.4.2.2 Antconc: Die *Keyword list*-funksie

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 ook ingespan word (Anthony, 2005: 733). Om die *Keyword list*-funksie suksesvol te gebruik word daar twee korpusse of tekste benodig. Hierdie funksie gebruik die Chi-kwadraat- of *log-likelihood* statistiese toetse 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.

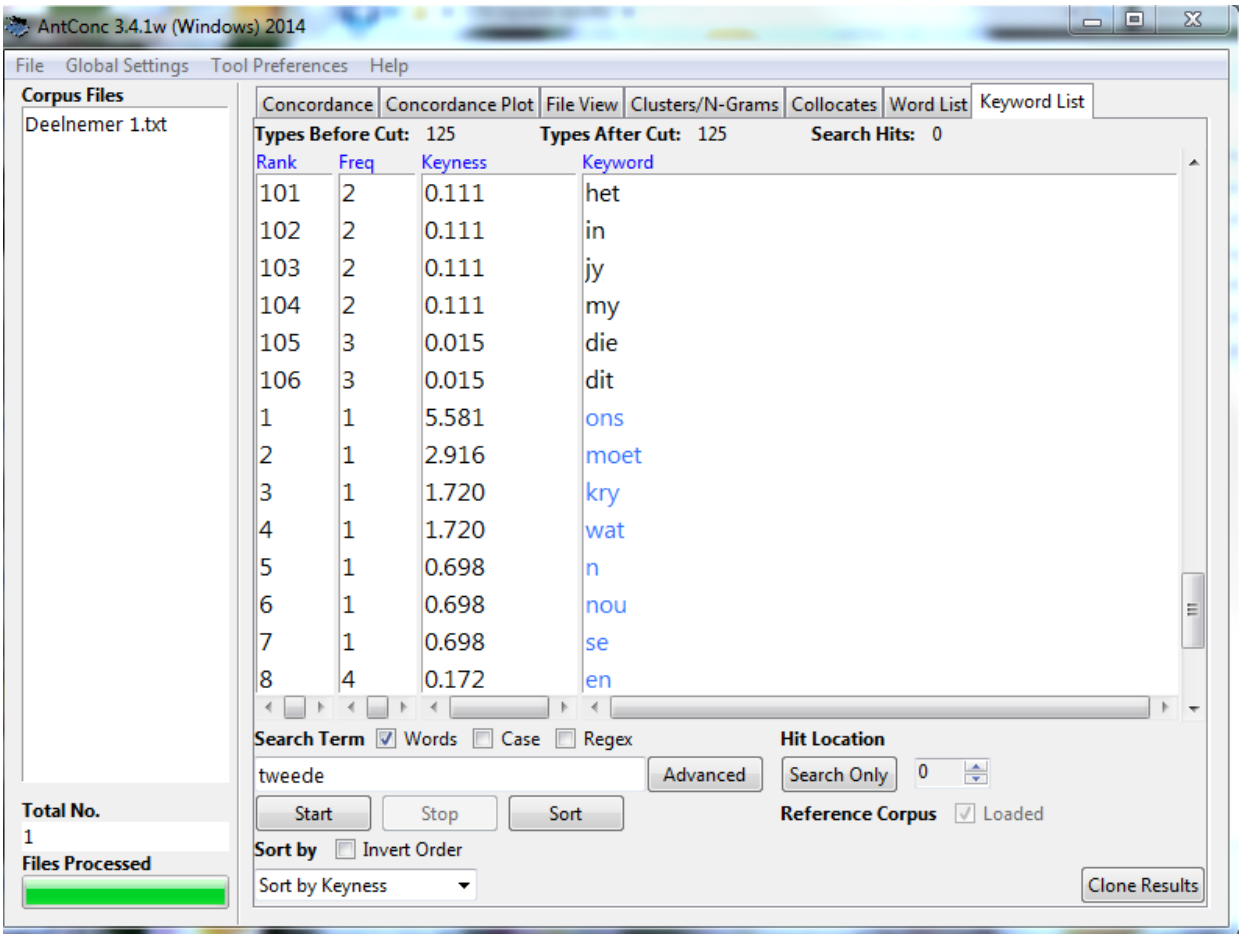
Die *Keyword list*-funksie genereer ’n lys van woorde soortgelyk aan die *Word list*-funksie, maar hier word die *keyness* of ‘sleutelwaarde’ van elke bepaalde woord binne die korpus/teks ook uitgelig. Die sleutelwaarde van elke woord word bepaal deur die ongewone hoë frekwensie van daardie woord binne die korpus in vergelyking met die frekwensie daarvan in die verwysingskorpus. As die woord ’n 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. Met ander woorde hoe laer die sleutelwaarde van die woorde in ’n teks. hoe groter is die kans dat dieselfde outeur albei tekste geproduseer het. **Figuur 3.3** is ’n voorbeeld van ’n *Keyword list* in Antconc.

Figuur 3.3: 'n Voorbeeld van 'n Keyword list wat in Antconc gegenereer is



Met die *Keyword list*-funksie is dit ook moontlik om die woorde te sien wat *negative key* is (**Figuur 3.4**). Dit is die woorde wat meer frekwent in die verwysingskorpus voorkom as in die korpus wat daarmee vergelyk word.

Figuur 3.4: Woorde wat* negative key *is word in Antconc in blou gemerk



3.4.2.3 WordSmith Tools (WST)

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 waardes van belang is. Antconc maak ook in die *Keyword list* van die *log-likelihood ratio* gebruik, eerder as die Chi-kwadraat-statistiek. 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.

In WST moet die p-waarde op 0.1 gestel 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 posisies van woorde wat in die *Keyword lists* van elke program aangedui word asook klein verskille in die sleutelwaarde van die woorde. Enkele vergelykende resultate is by **Figuur 4.5** ingesluit. Dit is met ander woorde nie werklik nodig om al twee programme vir die analisering van die data te gebruik nie, aangesien die WST-analise nie enige groot verskille in die resultate tot gevolg het nie.

3.4.2.4 N-gramanalise

Die laaste tipe stilometriese analise wat uitgevoer is, het gebruik gemaak van n-gramanalise 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-gramanalise op die data uitgevoer is. Die metode wat gebruik is, is Cavnar en Trenkle (1994) se metode vir taalherkenning. Hierdie metode is reeds in **hoofstuk 2** bespreek (p. 37). Wolff (persoonlike kommunikasie, 18/09/2014) meen ander metodes sou ook oorweeg kon word, maar sulke metodes is moontlik te kompleks om te implementeer aangesien die hoeveelheid data in die huidige studie so beperk is. Die resultate van die n-gramanalise word onder 4.2.3 in detail bespreek.

Die voorafgaande bespreking van stilistiese en stilometriese analises wat op die data uitgevoer is maak dit duidelik dat die beperkte hoeveelheid data in die huidige studie problematies is. Alhoewel daar woordlyste en *keyword lists* gegenereer kan word, is die sleutelwaardes wat verkry word baie laag. Uit die woordlyste wat gegenereer is, lyk dit asof daar nie ’n generiese taal onder die groep deelnemers bestaan nie, maar hierdie waarneming word in hoofstuk 4 verder ondersoek. Dit blyk ook dat dit wel moontlik is om die deelnemers van mekaar te onderskei deur middel van hul skryfstyle en taalgebruik. Uit die resultate wat by **Tabel 3** (p.111) ingesluit is, is dit duidelik dat daar verskille is in die frekwensie van woorde wat deur die deelnemers gebruik is. Hierdie waarneming is egter nie genoeg om te bevestig dat daar definitiewe idiolektiese verskille tussen die deelnemers is nie. In hoofstuk 4 volg die bespreking van die resultate wat uit die stilistiese en stilometriese analises verkry is. Op grond van dié resultate is gevolgtrekkings

gemaak oor die moontlikheid van suksesvolle outeuridentifikasie in soortgelyke scenario's as wat in die huidige navorsing getoets is.

Hoofstuk 4: Analise en resultate

In hierdie hoofstuk word die resultate van die ontledings wat uitgevoer is, aangebied. Die doel van die analises was om die volgende navorsingsvrae te beantwoord:

1. Kan daar in Afrikaans ’n generiese SMS-taal geïdentifiseer word wat outeuridentifikasie sou 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 verdagte SMS-teks te identifiseer met die beperkte data wat tipies ter beskikking is?

Uiteindelik het die beantwoording van die navorsingsvrae ook bepaal of die doel van die studie (wat in die derde navorsingsvraag saamgevat word) bereik is, al dan nie: Is dit moontlik om die outeur van ’n verdagte SMS-teks te identifiseer indien die navorser beperkte data tot sy of haar beskikking het en daar verskeie moontlike outeurs in die bepaalde situasie is? (Vergelyk ook paragraaf 1.4 en 1.5)

4.1 Stilistiese analise

4.1.1 Die kleurgids

In die stilistiese analise is elke deelnemer se teks met behulp van ’n kleurgids, wat deur die navorser opgestel is, geanaliseer. Die verdagte teks (Teks X) is ook met behulp van die kleurgids geanaliseer. Vir kontroleringsdoeleindes weet die navorser in hierdie studie dat D2 die ware outeur van Teks X is. Dit is belangrik dat die navorser tydens ’n pseudo-studie (soos die geval in die huidige navorsing) sekerheid het oor die outeur van ’n verdagte teks. Die navorser moet immers sy of haar resultate kan verifieer voordat die sukses van die navorsing bepaal kan word.

Sewentien kenmerke is in Teks X geïdentifiseer en getabuleer (Bylaag 2). Elkeen van die deelnemers se tekste is met Teks X se kenmerke vergelyk en die ooreenkomste is gemerk. Die totale hoeveelheid (persentasie) ooreenkoms tussen elkeen van die deelnemers se tekste en Teks X is uitgewerk.

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.

4.1.1.1 Persentasieooreenkoms tussen die deelnemers en Teks X.

Uit die stilistiese analise blyk dit dat daar **een deelnemer** is wat die hoogste persentasieooreenkoms 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. D1, D4 en D5 deel ’n ooreenkomspersentasie van 58.8% met die verdagte teks. Deur hierdie stilistiese analise te gebruik, was die navorser in staat om D2 as een van die moontlik outeurs te identifiseer, maar die persentasie waarmee D2 met Teks X ooreenstem, is te laag om met sekerheid enige gevolgtrekkings te maak.

Tabel 3: Die persentasie ooreenkomste tussen Teks X en elke deelnemer se teks

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%

Die resultate van die stilistiese analisa toon dus aan dat hierdie analise nie tot ’n bevredigende mate en sonder twyfel die moontlike outeur van die verdagte teks kan uitwys nie.

4.2 Stilometriese analise

4.2.1 Resultate: Die *Word List*-funksie

Die *Word List*-funksie is gebruik om ’n aparte woordelys vir elke deelnemer en Teks X op te stel. Daar is besluit om die eerste 10 algemene woorde met die hoogste frekwensies in elke deelnemer se teks met dié in Teks X te vergelyk. ’n Vergelyking met Teks X is op die eerste 10 algemene woorde uitgevoer om vas te stel of die ware outeur van die verdagte teks in albei toetse geïdentifiseer kan word. Hierna is dieselfde vergelyking met Teks X uitgevoer met die eerste 11 mees frekwente funksiewoorde (lidwoorde, voorvoegsels, voegwoorde ensovoorts) in die deelnemers se tekste. Die 11 mees frekwente funksiewoorde is gekies aangesien die outeur van Teks X beide ‘maar’ en die vorm ‘ma’ in die teks gebruik het. Die vergelykings van die tekste vir die tien mees frekwente (algemene) woorde en die 11 mees frekwente funksiewoorde met Teks X is by **Tabel 4** en **Tabel 5** ingesluit. Vir die statistiese ontleding van die vergelykings wat tussen die deelnemers en Teks X gedoen is, is die beide die Pearson Chi-kwadraattoets en die Yates korreksie gebruik.

Tabel 4: ’n Vergelyking tussen die 10 mees frekwente woorde in elke deelnemer se teks met Teks X.

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 5: 'n Vergelyking tussen die 11 mees frekwente funksiewoorde in elke deelnemer se teks met 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

Uit die tabelle blyk dit dat daar verskeie deelnemers is wat ooreenkomste toon met Teks X in terme van die frekwensie van sekere algemene woorde en funksiewoorde. Dit is moontlik om af te lei dat Deelnemer 14 in albei tabelle die minste ooreenkoms met Teks X toon terwyl dit blyk asof Deelnemers 2 en 11 die meeste ooreenkomste met Teks X toon. Verdere statistiese toetse egter is uitgevoer om vas te stel of hierdie waarneming geldig is.

4.2.1.1 Die Chi-kwadraattoets

Soos by 3.3.4 bespreek, is die Pearson Chi-kwadraattoets en die Yates korreksie in die huidige navorsing gebruik om te bepaal of die resultate verkry uit die beperkte data enigsins gebruik kan word om die moontlike outeur van die verdagte teks te bepaal. Die resultate vir die Pearson Chi-kwadraattoets word eerste bespreek.

4.2.1.1.1 Chi-kwadraattoets: 10 mees frekwente (algemene) woorde en die 11 mees frekwente funksiewoorde.

Die Chi-kwadraattoets vir albei die groepe frekwensies (vir die 10 mees algemene woorde en 11 funksiewoorde) is uitgevoer deur eerstens elke deelnemer se teks met Teks X in aparte tabelle te vergelyk. Daarna is die verwagte frekwensies vir elke waargeneemde frekwensie uitgewerk en

die Chi-kwadraatwaarde vir elke vergelyking is vasgestel. Die aanlynsagteware Graphpad is gebruik om die p-waardes vir elke deelnemer in die twee groepe te bepaal. Graphpad kan gebruik word om ’n verskeidenheid statistiese berekeninge uit te voer, maar vir die doeleindes van die huidige navorsing is daar slegs van die opsie gebruik gemaak wat die p-waarde bepaal wanneer die Chi-kwadraatwaarde en die *degrees of freedom* (*df*) ingelees word.

In **Tabel 6** en **Tabel 7** is die resultate vir die 10 mees frekwente algemene woorde en die 11 mees frekwente funksiewoorde opgesom.

Tabel 6: Die resultate van Pearson Chi-kwadraattoets vir die 10 mees frekwente algemene woorde.

	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 7: Die resultate van die Pearson Chi-kwadraattoets vir die 11 mees frekwente funksiewoorde.

	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	01300
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

Soos reeds genoem moet die p-waarde by Chi-kwadraattoetse verkieslik op 0.05% gestel word om toe te laat vir soveel moontlik gevalle van beduidende verskille tussen die dokumente (Kotzé, 2007: 391). ’n P-waarde van 0.05% is daarom vir elkeen van die groepe gebruik. Die *degrees of freedom* vir die eerste groep frekwensies (die 10 mees algemene woorde) is 9. Dit beteken dat wanneer die df en die p-waarde van 0.05 op ’n Chi-kwadraatdistribusietabel vergelyk word die Chi-kwadraatwaarde waarop albei ontmoet 16.919 is. Vergelyk **Figuur 4.1**. Dit beteken dat by 0.05% met 9 df moet die Chi-kwadraatwaarde 16.919 wees vir die tekste om beduidend van mekaar te verkil.

Elke Chi-kwadraatwaarde kan met die Chi-kwadraatwaarde van 16.919 vergelyk word. Hoe kleiner die Chi-kwadraatwaarde, hoe groter is die ooreenstemming tussen die verdagte teks en die deelnemer se teks. Hoe groter die Chi-kwadraatwaarde hoe kleiner is die ooreenstemming tussen die verdagte teks en die deelnemer se teks.

Figuur 4.1: Die Chi-kwadraatdistribusietabel waarop die Chi-kwadraatwaarde van 16.919 gekry is.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819

Dieselfde metode is by die Chi-kwadraatberekeninge vir die 11 mees frekwente funksiewoorde gevolg. Hier is die df 10 aangesien daar van 11 woorde gebruik gemaak word (dit wil sê daar is een ekstra ry in die tabelle. Hier het die Chi-kwadraatwaarde in die distribusietabel op 18.307 uitgewerk. Vergelyk **Figuur 4.2** Dit beteken dat by 0.05% met 10 df die Chi-kwadraatwaarde 18.307 moet wees vir die tekste om beduidend van mekaar te verskil. Dieselfde vergelyking as laasgenoemde kan ook in hierdie geval gedoen word en sal beteken dat Chi-kwadraatwaardes baie laer as 18.307 aandui dat daardie deelnemer se teks meer ooreenkomste met die verdagte teks het terwyl Chi-kwadraatwaardes van 18.307 en hoër aandui dat die deelnemer se teks en die verdagte teks nie veel ooreenkomste deel nie.

Figuur 4.2: Die Chi-kwadraatdistribusietabel waarop die Chi-kwadraatwaarde van 18.307 gekry is.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819

Uit die **Tabel 6** en **Tabel 7** is dit duidelik dat die data in die huidige ondersoek te min is om enige bruikbare resultate te verkry. Uit die resultate in **Tabel 6** blyk dit asof D2 en D8 die enigste twee moontlike outeurs kan wees, maar die persentasies is so laag (46.83% en 47.85% onderskeidelik) dat dit nie bruikbaar is nie. In **Tabel 7** kan D2 (64.69%) en D11 (72.64%) as die moontlike outeurs van die verdagte teks oorweeg word. 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. Die navorser weet in hierdie geval dat D2 die werklike outeur van die verdagte teks is, maar in ’n werklike situasie sou hierdie resultate eers verder ondersoek moet word voor enige definitiewe gevolgtrekking gemaak kan word.

Soos wat die geval was met die stilistiese analise, kon ook die Chi-kwadraattoets nie Deelnemer 2 bo alle twyfel aanwys as die outeur van die verdagte teks nie.

4.2.1.1.2 Die Yates korreksie.

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.

Vir die Yates korreksie is die data vir die 10 mees frekwente algemene woorde en die 11 mees frekwente funksiewoorde soos volg verdeel:

Die 10 mees frekwente algemene woorde is verdeel in leksikale woorde (konteksgebonde: ‘kry’, ‘lekker’ en ‘gehad’) en funksiewoorde (nie-konteksgebonde: ‘ons’, ‘en’, ‘moet’, ‘ek’, ‘wat’, ‘die’, ‘dit’). Wanneer die frekwensies vir Teks X en byvoorbeeld D1 opgetel word lyk die 2x2 tabelle soos volg:

	Teks X	D1
Leksikaal	8	1
Funksie	24	20

Die 11 mees frekwente funksiewoorde is verdeel in eerste persoon persoonlike voornaamwoorde (‘ek’, ‘ons’) en die res (‘en’, ‘moet’, ‘wat’, ‘die’, ‘dit’, ‘ma’, ‘maar’ ‘’n’, ‘sal’). Die 2x2 tabelle lyk soortgelyk aan die bogenoemde een.

	Teks X	D1
1ste Persoon	9	8
Die res	23	17

Hierna is Yates korreksie op elkeen van die 2x2 gebeurlikheidstabelle gedoen. Soos genoem by 3.3.4.2 word daar in die Yates korreksie 0.5 afgetrek by die verskil tussen elke waargeneemde frekwensie en verwagte frekwensie. Die Yates korreksie is deur middel van die statistiese pakket R, weergawe 3.02 gedoen. Die funksie chisq.test () is gebruik[14]. Die resultate van die Yates korreksie word in **Tabel 8** en **Tabel 9** opgesom.

[14] H. Gerber, persoonlike kommunikasie, 18/02/2015

Tabel 8: Die resultate van die Yates korreksie op die 10 mees frekwente algemene woorde.

	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: 9: Die resultate van die Yates korreksie op die 11 mees frekwente funksiewoorde.

	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

Die resultate wat deur die Yates korreksie verkry is is hoër as die resultate wat deur die Pearson Chi-kwadraattoets verkry is. Die resultate moet egter steeds baie versigtig geïnterpreteer word. In **Tabel 8** het deelnemers 4, 8 en 14 die hoogste p-waardes (62.78%, 57.03% en 41.78% onderskeidelik). 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. Die resultate in **Tabel 9** lyk effens anders. Hier is vier deelnemers met baie hoë p-waardes. D6 (100%), D10 (100%), D7 (95.5%) en D1 (97.96%). Ten eerste is twee een honderd persent ooreenkomste uiteraard onmoontlik aangesien dit nie vir een deelnemer moontlik sal wees om met die beperkte data een honderd persent ooreenstemming met die verdagte teks te kan toon nie, dit is nog minder moontlik dat twee individue een honderd persent ooreensteming met Teks X sal hê. Die beperkte data en die resultate wat tot dusver uit die statistiese toetse verkry is maak die persentasies by D7 en D1 ook hoogs onwaarskynlik.

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.

4.2.2 Resultate: Die *Keyword list*-funksie

Die *Keyword list*-funksie word gebruik om woorde te identifiseer wat ongewoon frekwent in die korpus voorkom in vergelyking met dieselfde woorde in die verwysingskorpus. In die *Keyword list*-funksie word twee korpusse benodig wat met mekaar vergelyk word. Vir die huidige navorsing is Teks X as die verwysingskorpus geselekteer en elke deelnemer se teks is met dié van Teks X vergelyk. Elke deelnemer se teks is eerstens deur middel van die *Keyword list*-funksie van Antconc geanaliseer.

Die *Keyword list* gee die sleutelwaardes van elke woord in die teks. Hoe hoër die sleutelwaarde van die woorde hoe groter is die moontlikheid dat die bepaalde tekste aan verskillende outeurs behoort. Laer sleutelwaardes verhoog die moontlikheid dat een outeur vir al die tekste verantwoordelik is. Die sleutelwaardes van drie verskillende analises van die data is oorweeg.

4.2.2 (a) Sleutelwaardes: Die eerste woord.

In die eerste plek is die sleutelwaarde van die eerste woord van elke deelnemer se teks ondersoek. Volgens die resultate het D13 die laagste sleutelwaarde vir die eerste woord gevolg deur D4 en D6. Hierdie metode is duidelik baie onbetroubaar aangesien die eerste woord in die lys nie ’n akkurate aanduiding is van die algehele sleutelwaarde van die deelnemer se woorde nie.

4.2.2 (b) Sleutelwaardes: Die eerste tien woorde.

Hierna is die algehele sleutelwaarde van die eerste tien woorde van elke deelnemer uitgewerk deur die gemiddeld van hierdie resultate te bepaal. Volgens hierdie resultate (**Figuur 4.3**) blyk dit dat D13 steeds die laagste sleutelwaarde het met D11 en D2 in die tweede en derde plek, onderskeidelik.

Figuur 4.3: Gemiddelde sleutelwaardes van die eerste 10 woorde van elke deelnemer

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		D1	D2	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	
2	Keyness v	6.685	6.566	4.321	7.863	4.682	5.376	4.762	4.817	5.919	4.763	4.806	2.943	8.677	
3	Keyness v	3.342	2.189	3.545	3.932	2.809	3.253	3.175	3.831	3.289	2.88	4.806	2.943	5.423	
4	Keyness v	2.228	2.189	3.529	3.276	2.809	3.225	3.175	3.831	2.959	2.88	3.605	2.943	5.423	
5	Keyness v	2.228	2.189	2.659	3.276	2.809	3.225	3.175	3.831	2.959	2.757	3.605	1.962	5.423	
6	Keyness v	2.228	2.189	2.659	2.621	2.494	3.225	3.175	2.554	2.959	1.904	3.605	1.962	4.339	
7	Keyness v	2.228	2.189	2.659	2.621	2.486	3.225	3.175	2.554	2.959	1.904	2.403	1.962	4.339	
8	Keyness v	2.228	2.189	2.158	1.966	2.486	2.15	3.175	2.408	2.959	1.904	2.403	1.962	4.339	
9	Keyness v	2.228	2.189	1.772	1.966	1.873	2.15	1.587	1.606	2.959	1.378	2.403	1.962	3.306	
10	Keyness v	2.228	1.793	1.772	1.966	1.873	2.15	1.587	1.277	2.959	1.378	2.403	1.394	3.254	
11	Keyness v	2.228	1.102	1.772	1.966	1.873	2.15	1.587	1.277	2.475	1.378	2.403	1.394	3.254	
12															
13		2.7851	2.4784	2.6846	3.1453	2.6194	3.0129	2.8573	2.7986	3.2396	2.3126	3.2442	2.1427	4.7777	
14															

D2 is in die tweede geval as een van die moontlike outeurs geïdentifiseer, maar beklee slegs die derde posisie. Dit beteken dat D13 eerder as die moontlike outeur verdink sou word.

4.2.2 (c) Sleutelwaardes: Die eerste 20 woorde.

Die algehele sleutelwaardes van die eerste 20 woorde vir elke deelnemer is ook uitgewerk (**Figuur 4.4**). Uit hierdie resultate blyk dit dat D13 steeds die laagste sleutelwaarde het en daarom die moontlike outeur van die verdagte teks moet wees, maar D2 is in hierdie geval in die tweede posisie as moontlike outeur geïdentifiseer terwyl D11 na die derde posisie geskuif het.

Gevolglik lyk dit asof die *Keyword list*-funksie nie met die huidige hoeveelheid data akkuraat genoeg is om D2 as die moontlike outeur van Teks X te identifiseer nie.

Figuur 4.4: Gemiddelde sleutelwaardes van die eerste 20 woorde van elke deelnemer

14														
15		D1	D2	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14
16	Keyness v	6.685	6.566	4.321	7.863	4.682	5.376	4.762	4.817	5.919	4.763	4.806	2.943	8.677
17	Keyness v	3.342	2.189	3.545	3.932	2.809	3.253	3.175	3.831	3.289	2.88	4.806	2.943	5.423
18	Keyness v	2.228	2.189	3.529	3.276	2.809	3.225	3.175	3.831	2.959	2.88	3.605	2.943	5.423
19	Keyness v	2.228	2.189	2.659	3.276	2.809	3.225	3.175	3.831	2.959	2.757	3.605	1.962	5.423
20	Keyness v	2.228	2.189	2.659	2.621	2.494	3.225	3.175	2.554	2.959	1.904	3.605	1.962	4.339
21	Keyness v	2.228	2.189	2.659	2.621	2.486	3.225	3.175	2.554	2.959	1.904	2.403	1.962	4.339
22	Keyness v	2.228	2.189	2.158	1.966	2.486	2.15	3.175	2.408	2.959	1.904	2.403	1.962	4.339
23	Keyness v	2.228	2.189	1.772	1.966	1.873	2.15	1.587	1.606	2.959	1.378	2.403	1.962	3.306
24	Keyness v	2.228	1.793	1.772	1.966	1.873	2.15	1.587	1.277	2.959	1.378	2.403	1.394	3.254
25	Keyness v	2.228	1.102	1.772	1.966	1.873	2.15	1.587	1.277	2.475	1.378	2.403	1.394	3.254
26	Keyness v	2.228	1.094	1.772	1.883	1.873	2.15	1.587	1.277	1.48	1.378	2.403	1.394	3.254
27	Keyness v	1.114	1.094	1.772	1.311	1.873	2.15	1.587	1.277	1.48	1.378	2.191	1.123	2.169
28	Keyness v	1.114	1.094	1.772	1.311	1.873	2.15	1.587	1.277	1.48	1.378	1.392	0.981	2.169
29	Keyness v	1.114	1.094	1.772	1.311	1.873	2.15	1.587	1.277	1.48	1.378	1.202	0.981	2.169
30	Keyness v	1.114	1.094	1.772	1.311	1.845	2.15	1.587	1.277	1.48	1.378	1.202	0.981	2.169
31	Keyness v	1.114	1.094	1.772	1.311	1.243	1.724	1.587	1.277	1.48	1.378	1.202	0.981	2.169
32	Keyness v	1.114	1.094	1.772	1.311	1.243	1.445	1.587	1.277	1.48	1.378	1.202	0.981	2.169
33	Keyness v	1.114	1.094	1.772	1.311	0.936	1.075	1.587	1.277	1.48	1.378	1.202	0.981	2.169
34	Keyness v	1.114	1.094	1.772	1.311	0.936	1.075	1.587	1.277	1.48	1.378	1.202	0.981	2.169
35	Keyness v	1.114	1.094	1.63	1.311	0.936	1.075	1.587	1.277	1.48	1.378	1.202	0.981	2.169
36														
37		2.00525	1.7862	2.2212	2.25675	2.04125	2.36365	2.22215	2.0378	2.3598	1.8453	2.3421	1.5896	3.5276
38														

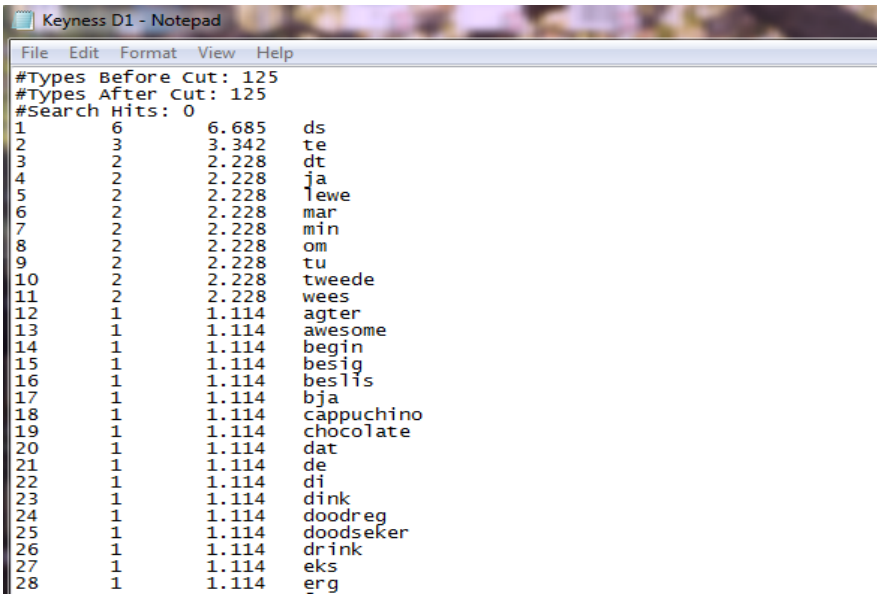
4.2.2.1 WordSmith Tools: *Keyword list*.

Soos reeds bespreek (vergelyk paragraaf 3.3.4) lewer die *Keyword list* in WST nie veel resultate op nie. Alhoewel daar enkele verskille is in die sleutelwaardes van die woorde in WST, word dit nie in die huidige navorsing as problematies beskou nie aangesien die posisie van een woord in die lys nie verandering in die resultate van die statistiese toetse tot gevolg sal hê nie. Die resultate wat deur WST gegenereer is, word bloot as addisionele inligting beskou aangesien Antconc in die huidige navorsing gebruik word. In **Figuur 4.5** word enkele vergelykings tussen die resultate uit Antconc en WST, met betrekking tot die *Keyword lists*, getref.

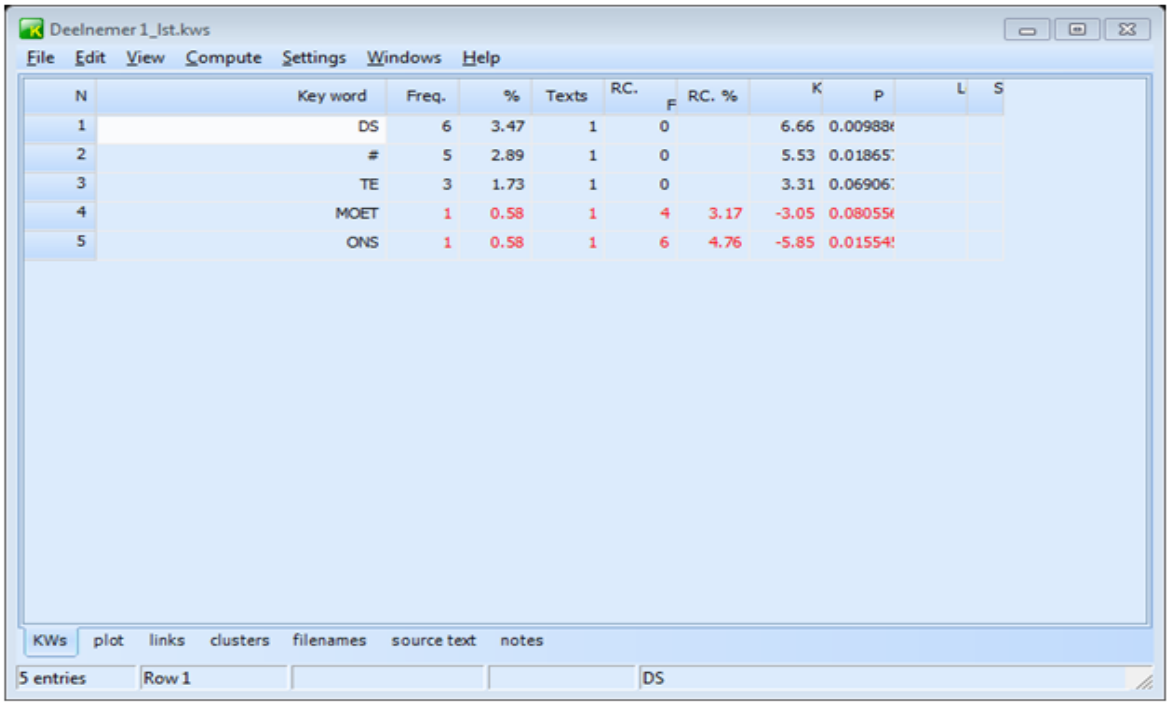
Figuur 4.5: Vergelyking van enkele resultate tussen die **Keyword lists** *wat deur Antconc en WordSmith Tools gegenereer is*

1. Vergelyking: Sleutelwaardes in Deelnemer 1 se teks

Sleutelwaardes aangedui in Antconc:

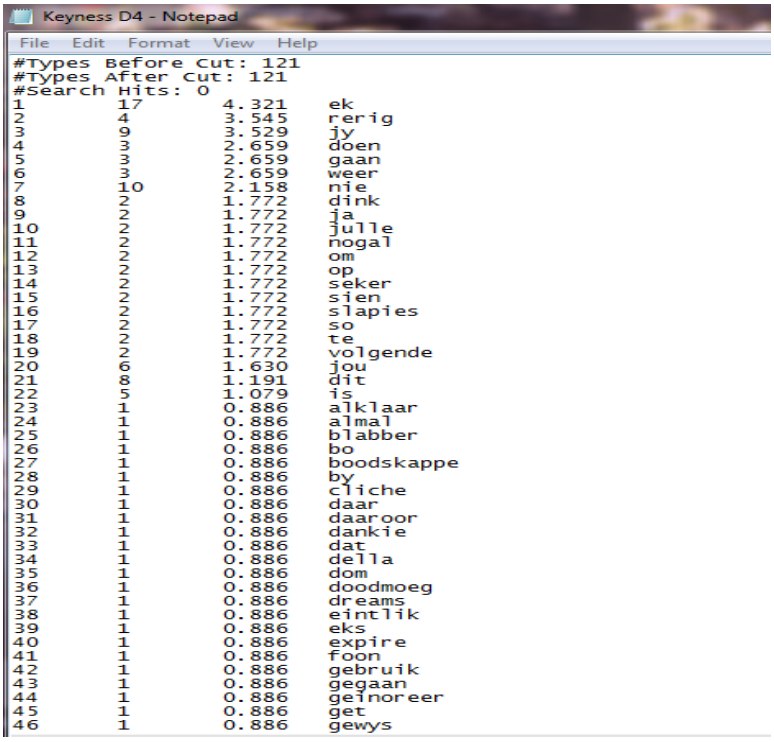


Sleutelwaardes aangedui in WordSmith Tools:



2. Vergelyking: Sleutelwaardes in Deelnemer 4 se teks

Sleutelwaardes aangedui in Antconc:

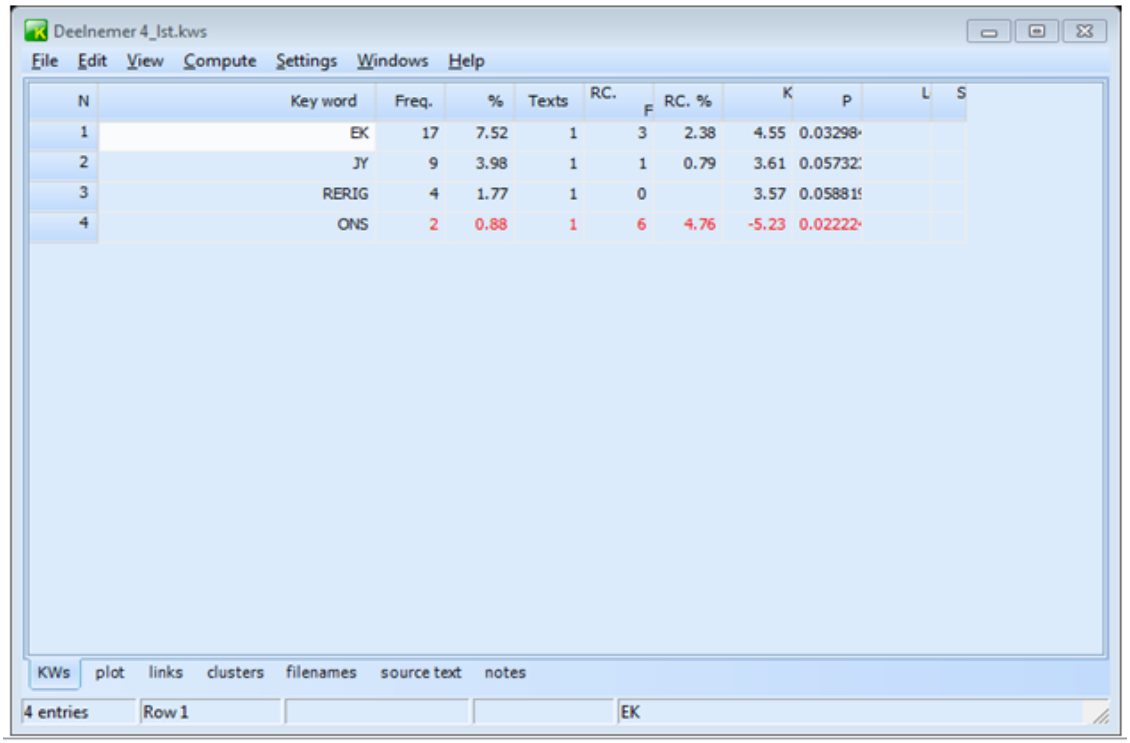


Keyness D4 - Notepad

File Edit Format View Help

```
#Types Before Cut: 121
#Types After Cut: 121
#Search Hits: 0
1        17        4.321    ek
2        4         3.545    rerig
3        9         3.529    jy
4        3         2.659    doen
5        3         2.659    gaan
6        3         2.659    weer
7        10        2.158    nie
8        2         1.772    dink
9        2         1.772    ja
10       2         1.772    julle
11       2         1.772    nogal
12       2         1.772    om
13       2         1.772    op
14       2         1.772    seker
15       2         1.772    sien
16       2         1.772    slapies
17       2         1.772    so
18       2         1.772    te
19       2         1.772    volgende
20       6         1.630    jou
21       8         1.191    dit
22       5         1.079    is
23       1         0.886    alklaar
24       1         0.886    almal
25       1         0.886    blabber
26       1         0.886    bo
27       1         0.886    boodskappe
28       1         0.886    by
29       1         0.886    cliche
30       1         0.886    daar
31       1         0.886    daaroor
32       1         0.886    dankie
33       1         0.886    dat
34       1         0.886    della
35       1         0.886    dom
36       1         0.886    doodmoeg
37       1         0.886    dreams
38       1         0.886    eintlik
39       1         0.886    eks
40       1         0.886    expire
41       1         0.886    foon
42       1         0.886    gebruik
43       1         0.886    gegaan
44       1         0.886    geinoreer
45       1         0.886    get
46       1         0.886    gewys
```

Sleutelwaardes aangedui in WordSmith Tools:



Deelnemer 4_lst.kws

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	Texts	RC. F	RC. %	K	P	L	S
1	EK	17	7.52	1	3	2.38	4.55	0.03298		
2	JY	9	3.98	1	1	0.79	3.61	0.05732		
3	RERIG	4	1.77	1	0		3.57	0.05881		
4	ONS	2	0.88	1	6	4.76	-5.23	0.02222		

KWs plot links clusters filenames source text notes

4 entries Row 1 EK

3. Vergelyking: Sleutelwaardes in Deelnemer 5 se teks

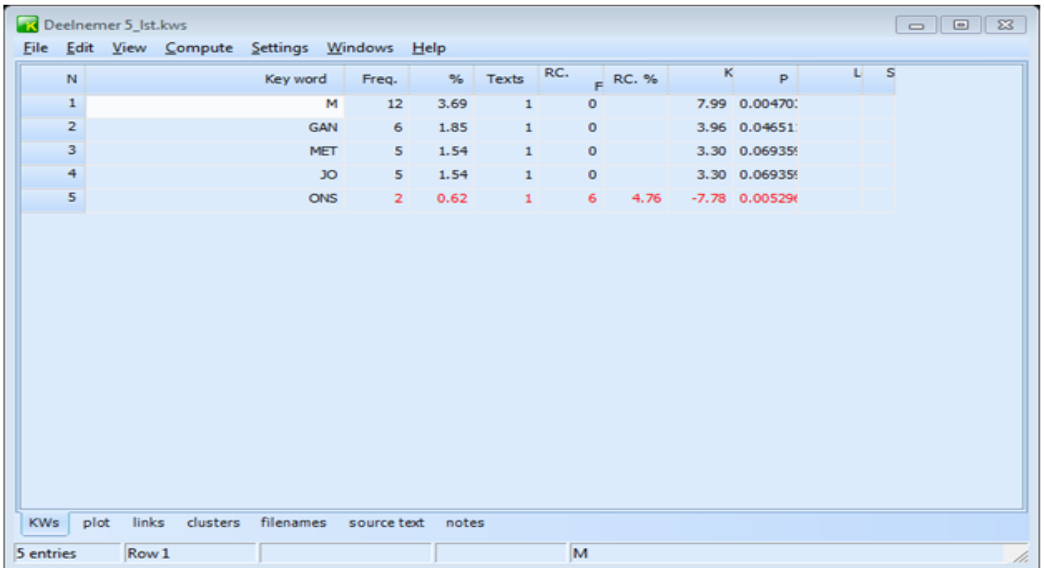
Sleutelwaardes aangedui in Antconc:

Keyness D5 - Notepad

File Edit Format View Help

```
#Types Before Cut: 181
#Types After Cut: 181
#Search Hits: 0
1        12        7.863    m
2        6         3.932    gan
3        5         3.276    jo
4        5         3.276    met
5        4         2.621    goed
6        4         2.621    werk
7        3         1.966    b
8        3         1.966    hoor
9        3         1.966    ni
10       3         1.966    so
11       17        1.883    ek
12       2         1.311    aan
13       2         1.311    afr
14       2         1.311    als
15       2         1.311    asb
16       2         1.311    bietjie
17       2         1.311    elo
18       2         1.311    ens
19       2         1.311    fb
20       2         1.311    gaan
21       2         1.311    info
22       2         1.311    ipad
23       2         1.311    k
24       2         1.311    kla
25       2         1.311    klas
26       2         1.311    man
27       2         1.311    mar
28       2         1.311    op
29       2         1.311    sit
30       2         1.311    studies
31       2         1.311    sug
32       2         1.311    sus
33       6         0.740    is
34       6         0.740    j
35       1         0.655    almal
36       1         0.655    an
37       1         0.655    baba
38       1         0.655    begin
39       1         0.655    bietji
40       1         0.655    blerrie
41       1         0.655    bly
42       1         0.655    bv
43       1         0.655    by
44       1         0.655    clicks
45       1         0.655    cover
46       1         0.655    create
```

Sleutelwaardes aangedui in WordSmith Tools:



4.2.3 *Resultate:* n-gramanalise

Verskeie kombinasies van n-gramme en profiellengtes (vektorlengtes) is in die huidige navorsing gebruik. Realisties is daar honderde kombinasies van laasgenoemde wat gebruik kan word en gevolglik sal dit tot ’n groot hoeveelheid uiteenlopende resultate lei. Vir die doeleindes van die huidige navorsing is die afbakenings in die Cavnar en Trenkle-studie oorweeg en aangepas aangesien die data wat beskikbaar is beperk is. Cavnar en Trenkle (1994) het in hul studie van n-gramme gebruik gemaak en met profiellengtes van 100, 200, 300 en 400 karakters getoets. 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 (Wolff, 2014). Daar moet in gedagte gehou word dat enige vermeerdering gebruik kan word. Vir praktiese doeleindes, aangesien die tydsbeperking nie toegelaat het vir die toetsing van elke moontlikheid nie, is daar op ’n vermeerdering van 10 n-gramme besluit. 4-gramme is as die maksimum n-gramgrootte gekies omdat langer n-gramme meestal net een keer in die opleidingsdata voorkom, of soms glad nie. Om hierdie rede is besluit dat langer n-gramme waarskynlik nie baie betekenisvol sou wees nie. 400 karakters is as die maksimum grootte van die profiellengte gekies aangesien lang profiellengtes tot onbetroubaarheid in die resultate kan lei, veral in gevalle waar die data beperk is. Wolff (2014) meen:

> ’n Lang profiel sal waarskynlik gevul word (aan die einde) met n-gramme wat net een keer voorgekom het in die opleidingsdata, maar hulle sal steeds gesorteer word in een of ander arbitrêre volgorde (dalk ewekansig). Aangesien die rang in die lys ’n groot rol speel in hoe die metode werk, raai ek dit kan ’n invloed hê. Dus raai ek is ’n te groot profiel nie sinvol nie.

Slegs drie konfigurasies se resultate word bespreek.Tydens die eerste analise is 1-, 2-, 3- en 4-gramme gebruik en is profiellengtes van 50 tot 400 ingesluit. Uit die eerste konfigurasie blyk dit dat D4 die waarskynlikste outeur van die verdagte teks is met 19 gevalle waar D4 die kortste afstand met die opleidingsdata handhaaf. D2 is in die tweede posisie as moontlike outeur met 12 gevalle en D5 word in die derde posisie gelys met 5 gevalle.

In die tweede analise is daar slegs van 1-, 2- en 3-gramme gebruik gemaak. Profiellengtes van 50 tot 400 is steeds ingesluit. Uit hierdie konfigurasie dui die resultate daarop dat D2 die grootste moontlikheid het om die outeur van die verdagte teks te wees met 17 gevalle waar D2 die kortste

afstand met die opleidingsdata handhaaf. D4 is in die tweede posisie met 12 gevalle en D5 is weereens in die derde posisie met 7 gevalle.

Tydens die derde analise is 1- en 2-gramme gebruik. Die profiellengtes is steeds 50 tot 400. Hier dui die resultate weereens daarop dat D2 die grootste moontlikheid van outeurskap het met 22 gevalle waar D2 die kortste afstand met die opleidingsdata handhaaf. D5 is nou in die tweede posisie met 12 gevalle en D4 is derde met slegs 2 gevalle.

Uit die resultate van slegs drie verskillende konfigurasies is dit duidelik dat die moontlikheid van outeurskap tussen D2, D4 en D5 varieer. Dit is gevolglik nie moontlik om met sekerheid outeurskap van die verdagte teks te bevestig nie. Alhoewel sukses in die verlede met n-gramanalises in kleiner datastelle behaal is, is die totale korpusse wat in die studies gebruik is steeds massief teenoor die korpus in die huidige navorsing. Ishihara (2011: 54) rapporteer 80% akkuraatheid wanneer die steekproef uit 2200 woorde bestaan. Mohan e.a. (2010: 10) rapporteer 65% tot 70% akkuraatheid wanneer die groep moontlike outeurs 28 is, met 50 boodskappe elk (dit wil sê ’n korpus van 1400 woorde). Ragel e.a. (s.a.: 5) het slegs 1-gramme in hulle navorsing gebruik en maak die volgende opmerking:

> Even with a very small amount of testing data (even with one SMS to test) the algorithm produces good results. But the growth of accuracy as the number of testing SMSes grows is not linear. The accuracy growth rate goes down with the increase of testing data until it saturates around ninety per cent. But a good accuracy has been achieved with a small number of test cases (around ten SMSes stacked together).

Dit is duidelik dat suksesvolle resultate met klein hoeveelhede data nie onmoontlik is nie, maar dat die strategieë wat gebruik word vir elke scenario aangepas moet word indien akkuraatheid bo 70% verkry wil word. Verder word daar in al drie die bogenoemde studies van ’n massiewe SMS-databasis gebruik gemaak (die NUS (National University of Singapore) korpus wat uit vyftigduisend SMS-boodskappe bestaan in drie tale), waarvan die inhoud gebruik word as die opleidingsdata en toetsdata. In die huidige studie is ’n korpus van 2434 woorde gebruik vir die opleidingsdata en die toetsdata en gevolglik is die resultate nie so akkuraat nie aangesien die opleidingsdata te min was.

4.3 Generiese taalgebruik in die SMS-korpus

Ten spyte van die beperkte data wat probleme veroorsaak met die positiewe identifisering van die ware outeur van die verdagte teks, maak die resultate dit wel moontlik om vas te stel dat daar nie ’n generiese SMS-taal onder hierdie groep SMS-gebruikers bestaan nie. Met generiese taalgebruik word bedoel ’n tipe taalgebruik wat oorwegend deur elkeen van die deelnemers gebruik word. Dit beteken dat die meeste woorde (of selfs al die woorde) in elke deelnemer se SMS-boodskappe op dieselfde manier gespel sou word, indien ’n generiese taalgebruik teenwoordig was.

Olivier (2013) het in sy studie oor die konsekwentheid van taalgebruik in Afrikaanse SMS-boodskappe gebruik gemaak van tien boodskappe wat volgens die literatuur tipiese eienskappe van SMS-taal vertoon. Die SMS-boodskappe is oorspronklik uit ’n Engelse korpus geneem (Baker, 2010) en is deur Olivier in Afrikaans vertaal. Die SMS-boodskappe is na 108 deelnemers gestuur en elkeen van die deelnemers is gevra om die kort boodskappe as SMS’e oor te tik. Olivier (2013: 501–502) het gevind daar geen konsekwentheid voorkom in die manier waarop die respondente sekere woorde in SMS’e tik nie. Hy het ook bepaal dat daar variasie voorkom in die skryfstyle van individuele respondente wat byvoorbeeld een woord op twee of drie verskillende maniere tik. Olivier (2013: 501) meen dat daar steeds reëls geld vir die manier waarop woorde in SMS-taal verander of verkort word. Hy noem egter dat wanneer hierdie ‘reëls’ gebreek word, selfoongebruikers steeds mekaar se kommunikasie kan begryp. In die huidige navorsing is gepoog om dieselfde SMS-boodskappe te gebruik om vas te stel of ’n generiese SMS-taal onder die groep deelnemers in die navorsing bestaan. Die generiese teks wat die navorser aan die deelnemers gestuur het lees soos volg (Olivier, 2013: 491–492):

> Hoekom? Hoor hier ons moet vinnig koffie drink en pasteitjies eet.
>
> Wanneer kom jy terug huis toe? Vanaand?
>
> Ek wonder hoe dit met jou gaan.
>
> Dankie. Sien jou die sesde.
>
> Ek sal dit nie môre of die naweek kan maak nie.

Waar is jy? Ek gaan nou maar klas toe.

Help! Vir die derde keer werk my kaart nie by die hek nie. Ek's klaar laat.

Ag nee, jou aansoek was toe nie suksesvol nie.

'n Deel van die werk is vir my baie lekker.

Anders kan ons doen wat ons wil.

Slegs 11 uit die 13 deelnemers het op die versoek gereageer om die generiese teks (as SMS'e) aan die navorser te stuur. In die meerderheid gevalle was die data onvolledig. Nietemin is die data van die generiese SMS'e wat wel ontvang is, ook deur middel van Antconc verwerk om vas te stel of die stylaspekte (spelling van woorde, afkorting van woorde en gebruik van leestekens) van die deelnemers deurlopend dieselfde is in elkeen van die boodskappe.

Elkeen van die 11 deelnemers wat wel die generiese SMS'e, of ten minste enkele van die generiese SMS'e, aan die navorser gestuur het, se teks is skoongemaak van enige woorde wat deur SMSPortal afgesny is. Daarna is al die afsonderlike tekste tot een teks saamgevoeg sodat die variasie van elke woord binne die teks bepaal kan word. Die *Word list*-funksie van Antconc is gebruik om een woordelys van die teks te produseer.

Die woordelys het die aanname bevestig dat die groep deelnemers moontlik slegs 'n klein hoeveelheid generiese SMS-taaleienskappe deel. Enkele van die variasies op die woorde wat deur Antconc gegenereer is, word in die onderstaande tabel aangedui:

Tabel 10: Enkele voorbeelde van (idiolektiese) variasie in deelnemers se tekste

WOORD	HOEVEELHEID
Nie	20
Ni	11
Ek	16
K	5
Die	14
Di	4
Ons	14
Ns	2
Hoekom	5
Hkm	1
Hukm	1
Huko	1
Hukom	1
Dit	10
Dt	4
Gaan	10
Gan	6
Jou	12
Ju	5
Vanaand	6

Vanand	2
Vnaant	1
Vnand	1

Uit die tien woorde wat in die tabel gelys word met hul variasies, is dit duidelik dat die spel van woorde en die afkorting van woorde tussen die deelnemers verskil. Die leestekengebruik tussen die deelnemers verskil ook. Slegs een deelnemer maak van drie opeenvolgende vraagtekens gebruik (???) en slegs een deelnemer maak van dubbel uitroeptekens (!!) gebruik. Die hoeveelheid uitroeptekens en vraagtekens teenwoordig in die teks wissel ook van deelnemer tot deelnemer.

Tabel 11: Die variasie van uitroeptekens en vraagtekens in die deelnemers se tekste

Die variasie van uitroepteken -en vraagtekengebruik in die deelnemers se tekste.						
	(!)	(!!)	(!!!)	(?)	(??)	(???)
Deelnemer 1	X	x		x		
Deelnemer 2	X	x		x		
Deelnemer 4	X	x		x		
Deelnemer 5	X	x		x	x	
Deelnemer 6	X					
Deelnemer 7				x		
Deelnemer 8				x		
Deelnemer 9	X					
Deelnemer 10	X			x		
Deelnemer 11	X		x			
Deelnemer 12	X			x		
Deelnemer 13	X		x			
Deelnemer 14	X	x		x		

Volgens Olivier (2013: 501) is inkonsekwentheid in SMS-taal glad nie vreemd nie aangesien daar nie ’n gestandardiseerde vorm van SMS-taal bestaan nie. Olivier noem egter dat hy van ’n klein hoeveelheid SMS-boodskappe gebruik gemaak het (slegs 92 woorde per deelnemer) en meen daarom dat die bevindings in die studie met ’n groter korpus van SMS-taal geverifieer moet word. Dit is vanselfsprekend dat die bevindings ten opsigte van ’n generiese SMS-taal in die huidige studie ook nie veralgemeen kan word nie, maar daar kan tog met sekerheid gesê word dat daar nie ’n generiese SMS-taal onder hierdie spesifieke groep van dertien deelnemers bestaan nie.

4.4 Opsomming van die resultate.

Die analises wat in hoofstuk 4 op die data uitgevoer is, het ten doel om die navorsingsvrae, soos onder 1.5 uiteengesit, te beantwoord. Die analises en daaropvolgende resultate maak dit moontlik om die antwoorde op laasgenoemde navorsingsvrae soos volg op te som:

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

Uit die analise op 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 geanaliseer. Die *Word list*-funksie genereer ’n lys van al die woorde wat in ’n bepaalde teks verskyn. Die woorde word op grond van spelling geïdentifiseer wat beteken dat elke unieke spelling van ’n woord as ’n aparte woord in die lys sal verskyn. Uit die *Word list* wat gegenereer is, is dit duidelik dat daar geen generiese SMS-taal onder die groep deelnemers bestaan nie (op grond van die variasies wat op die woorde in die teks voorkom). Indien daar wel ’n generiese SMS-taal onder die groep teenwoordig was, sou dit beteken dat daar slegs enkele spellingvariasie in die tekste sou voorkom, aangesien elke deelnemer dan die meerderheid woorde dieselfde 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 analises is dit duidelik dat daar deelnemers is wie se skryfstyle met dié van die outeur van die verdagte teks ooreenstem en verskil. Aangesien hierdie deelnemers van mekaar onderskei kan word, is dit moontlik om aan te voer dat die idiolektiese eienskappe van die deelnemers verskil. Die gevolgtrekking word daarom

gemaak dat die deelnemers almal oor ’n idiolektiese taalgebruik beskik. 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. Hierdie probleem word deur die resultate in beide die stilistiese en stilometriese analises geïllustreer.

In die stilistiese analise wat deur middel van die kleurgids uitgevoer is, is gevind dat daar een deelnemer (D2) is wat die hoogste persentasieooreenkoms met Teks X deel (64.7%). Drie ander deelnemers (D1, D4 en D5) deel die tweede hoogste persentasieooreenkoms met die verdagte teks (58.8%). Alhoewel D2 korrek as die moontlike outeur van Teks X geïdentifiseer is, is die persentasieooreenkoms steeds te laag om met sekerheid aan te voer dat D2 die outeur van Teks X is.

Verskeie stilometriese analises is ook uitgevoer en die resultate van elkeen van hierdie analises is onbeslis. Dit beteken dat hierdie analises geen resultate gelewer het wat slegs een outeur as die outeur van Teks X identifiseer nie.

Tydens die eerste stilometriese analise is die Pearson Chi-kwadraattoets gebruik om vas te stel of die data wat deur die *Word list*-funksie in Antconc verkry is, die nulhipotese (dat daar geen verhouding tussen die veranderlikes is nie) ondersteun of verwerp.

Die Chi-kwadraattoets is op 2 verskillende stelle teks uitgevoer, naamlik: die 11 mees frekwente funksiewoorde, die 10 mees frekwente algemene woorde.

Die resultate van die Chi-kwadraattoets op die **11 mees frekwente funksiewoorde** het aangedui dat twee van hierdie deelnemers (D2 en D11) ’n ooreenkomspersentasie bo 50% met die verdagte teks toon. Die persentasies is egter steeds te laag om definitiewe gevolgtrekkings te maak. Die resultate van die Chi-kwadraattoets op die **10 mees algemene woorde** het aangedui dat geen deelnemers ’n persentasieooreenkoms bo 50% met Teks X het nie. Die twee deelnemrs met die hoogste persentasies is D2 (46.83%) en D8 (47.85%). Die Yates korreksie is ook gebruik. Alhoewel die Yates korreksie tot baie hoë p-waardes in die analise van die 11

meesfrekwente funksiewoorde gelei het (D6 en D10 = 100%; D1 = 97.96%; D7 = 95.5%), is die resultate nie bruikbaar nie aangesien die beperkte data en die groepering van die data tydens die Yates korreksie duidelik tot onmoontlike resultate gelei het. Dit is uiteraard onmoontlik vir twee deelnemers om een honderd persent ooreenkomste met die verdagte teks te toon.

Die volgende stilometriese analise is deur middel van die *Keyword list*-funksie in Antconc uitgevoer. Die *Keyword-list* funksie word gebruik om te bepaal hoe ongewoon frekwent sekere woorde in ’n bepaalde korpus voorkom in vergelyking met dieselfde woorde in ’n verwysingskorpus deur die sleutelwaardes van die woorde vas te stel (vergelyk 4.2.2). Tydens hierdie analise is daar drie verskillende toetse uitgevoer.

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. Weereens is dit duidelik dat die analise nie een outeur as die moontlike outeur van Teks X kon identifiseer nie. Uit die analise wat deur die *Keyword list*-funksie uitgevoer is blyk dit met ander woorde dat ten minste drie tot vier deelnemers tot ’n mate idiolektiese ooreenkomste met die outeur van Teks X deel.

Die laaste stilometriese analise wat op die data uitgevoer is, was ’n n-gramanalise. 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 (vergelyk 4.2.3). Slegs drie konfigurasies se resultate word in ag geneem aangesien die aard van die studie beperk

is en daar moontlik honderde konfigurasies getoets kan word met kombinasies van laasgenoemde aantal n-gramme en profiellengtes.

Die eerste analise (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 met 19 gevalle waar D4 die kortste afstand met die opleidingsdata handhaaf. D2 en D5 is ook onderskeidelik as moontlike outeurs gelys.
Tydens die tweede analise (bestaande uit: 1-, 2- en 3-gramme en profiellengtes van 50 tot 400) is D2 as die waarskynlikste outeur van Teks X geïdentifiseer met 17 gevalle waar D2 die kortste afstand met die opleidingsdata handhaaf. D4 en D5 is ook as moontlik outeurs gelys.
Die derde analise (bestaande uit: 1- en 2-gramme en profiellengtes van 50 tot 400) het bevind dat D2 weereens die waarskynlikste outeur van Teks X is met 22 gevalle waar D2 die kortste afstand met die opleidingsdata handhaaf. D5 en D4 is weereens gelys, maar het posisie verander.

Al die bogenoemde resultate bevestig die vroeër stelling dat dit baie moeilik is om aan te voer dat daar in die navorsing ’n duidelike idiolektiese onderskeid tussen die deelnemers getref kan word. Dit is waar dat idiolektiese eienskappe teenwoordig is, aangesien net sekere deelnemers in elke analise as die moontlike outeurs van die verdagte teks geïdentifiseer word. D2 word ook, in verskeie van die analises, korrek as een van die moontlike outeurs of die enigste moontlike outeur van die verdagte teks geïdentifiseer. Omdat die beperkte data dit laat blyk dat D2 se idiolek met ander deelnemers se idiolek ooreenstem, kan geen beslissende gevolgtrekkings egter gemaak word nie.

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 analises 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 onbeslissende resultate is die beperkte hoeveelheid data wat tot die navorser se beskikking is. Soos duidelik blyk uit die bespreking van idiolek (vergelyk 2.6) speel die hoeveelheid data ’n belangrike rol in die identifisering van

idiolek. 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 verdagte outeur.

In die huidige navorsing is die data net genoeg om aan te dui dat

(1) daar nie ’n generiese SMS-taal onder die groep deelnemers bestaan nie

en

(2) dat idiolektiese eienskappe gevolglik teenwoordig is.

Die data is egter te beperk om vas te stel tot watter mate die idiolek van die deelnemers van mekaar verskil. Dit is moontlik dat D11, byvoorbeeld, 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 persentasieooreenkoms met die verdagte teks deel in werklikheid geen idiolektiese eienskappe met mekaar deel nie.

Die bespreking in hoofstuk 4 maak dit duidelik dat die navorsing suskesvol was om die navorsingsvrae in die huidige studie te beantwoord. Dit is egter nodig om die resultate verder binne die groter konteks van forensiese linguistiek te ondersoek. In hoofstuk 5 word belangrike gevolgtrekkings gemaak op grond van die resultate wat in vier ondersoekareas verkry is.

Hoofstuk 5: Gevolgtrekkings

Uit die huidige navorsing, asook navorsing in 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. Dit is ook duidelik te sien dat daar potensiaal is vir die navorsingsveld van outeuridentifikasie om in Suid-Afrika te groei.

5.1 Doelstelling en navorsingsvrae.

Soos duidelik blyk uit die opsomming in 4.4, is die doelstelling van die huidige navorsing nie bereik nie. Dit is, op grond van die resultate wat uit beide die stilistiese en stilometriese analises verkry is, nie moontlik om met sekerheid die ware outeur van die verdagte SMS-teks te identifiseer nie. Dit is nietemin moontlik om vas te stel dat:

1. Daar nie ’n generiese SMS-taal onder die groep deelnemers bestaan nie.

2. Daar, tot ’n mate, idiolektiese eienskappe onder die groep deelnemers geïdentifiseer kan word.

Dit is, op grond van selfs die beperkte data in die huidige navorsing, ’n geldige argument om aan te voer dat idiolek bestaan. Uit die navorsing wat ek gedoen het, is dit duidelik dat daar tussen verskeie outeurs onderskei kan word op grond van hulle idiolektiese taalgebruik. Dit is egter belangrik om daarop te let dat ’n ‘suiwer’ idiolek nie bestaan nie en dat een persoon se idiolek beïnvloed word deur interaksie met ander individue. Dit is ook belangrik om in ag te neem dat die akkuraatheid van die bepaling van idiolek toeneem na mate die hoeveelheid teks wat gebruik word om die idiolek te identifiseer, toeneem. Dit beteken dat die idiolek van ’n individu met groter sekerheid beskryf kan word indien daar genoeg teks is om aannames ten opsigte van iemand se idiolek te verifieer. Minder teks of data word gevolglik geassossieer met ’n laer akkuraatheid ten opsigte van die beskrywing van een persoon se idiolek. As gevolg van die beperkte data in die huidige navorsing beteken dit dat die stelling gemaak kan word dat idiolek teenwoordig is, maar akkurate beskrywings van die verskillende idiolekte is nie moontlik 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 analises dit wel moontlik

gemaak het om die aantal waarskynlike outeurs van 13 tot 11 te verminder. Deur die resultate in elkeen van die analises 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 as een van die moontlike outeurs van Teks X geïdentifiseer is. Hierdie verskynsel kan moontlik die forensiese linguis oortuig om veral D2 as die hoofverdagte te oorweeg.

Die resultate dui daarop dat daar potensiaal is ten opsigte van suksevolle 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.

5.2 Bruikbaarheid van resultate in die hof.

Dit is belangrik om die bruikbaarheid van die resultate in die studie te oorweeg aangesien dit juis in hofsake is waar die resultate verkry uit forensies-linguistiese analises getoets moet word.

Die resultate wat in die huidige navorsing verkry is, sal **nie as geldige bewyse** in die hof gebruik kan word nie. Dit beteken dat daar nie op grond van slegs hierdie resultate ’n skuldigbevinding of onskuldigbevinding kan wees nie. In die huidige navorsing lê die probleem hoofsaaklik by die lae persentasie van sekerheid ten opsigte van ooreenkomspersentasies tussen die verdagte teks en die verdagtes. Dit beteken dat die navorser nie met ’n hoë persentasie van sekerheid kan bewys dat één verdagte moontlik die ware outeur van die verdagte teks is nie.

Die resultate sal egter gebruik kan word as omstandigheidsgetuienis. Dit is moontlik dat die resultate wat verkry is, gebruik kan word om bloot die groep moontlike outeurs van die verdagte teks te verminder. Soos reeds genoem, was dit moontlik om die groep verdagte outeurs in die huidige studie van 13 tot 6 te verminder. Verder kan die forensiese linguis aanvoer dat die verdagte, wat in die meeste toetse geïdentifiseer word as ’n moontlike outeur, eerste ondervra of ondersoek moet word.

Dit is belangrik om daarop te let dat alhoewel SMS-boodskappe in Suid-Afrika ook as dokumente (elektroniese dokumente) beskou word (Streicher, 2010: 2) en as toelaatbare bewyse in beide kriminele en siviele sake gebruik kan word, outeuridentifikasie nog nie in Suid-Afrika op sulke dokumente (SMS-tekste) toegepas is en as bewyse in die howe gebruik is nie.

As daarna gestrewe wil word dat die resultate van outeuridentifikasieanalises in kort elektroniese tekste in die toekoms as geldige bewyse in Suid-Afrikaanse howe gebruik kan word, is dit uiters belangrik dat die resultate aan streng vereistes moet voldoen. Daar sal verseker moet word dat die resultate hoë persentasies van sekerheid toon ten opsigte van die outeur van ’n bepaalde teks.

Die enigste manier om sulke sekerheid te bereik, is om verdere studies in moeilike forensies-linguistiese situasies aan te pak in ’n poging om die metodologie en instrumente aan te pas en verder te ontwikkel totdat dit moontlik is om sukses te behaal met outeuridentifikasieanalises in beperkte, kort, elektroniese tekste. Die belangrikheid van resultate soos dié wat in die huidige navorsing verkry is, wat as omstandigheidsgetuienis in ‘n hof kan dien, moet egter ook nie onderskat word nie.

5.3 Die vestiging van forensiese linguistiek as selfstandige vakgebied in Suid-Afrika.

Uit die literatuuroorsig is dit duidelik dat outeuridentifikasiestudies reeds in verskeie scenario’s met sukses gebruik is (Mikros, s.a.; Mohan e.a., 2010; Ishihara, 2011; McLeod en Grant, 2012; Michell, 2013). Die gebruik van outeuridentifikasie om die outeurs van elektroniese tekste soos SMS-boodskappe, e-posse en ander aanlyn tekste te identifiseer, sal ook in Suid-Afrika van nut wees aangesien dieselfde kriminele aktiwiteite regoor die wêreld plaasvind. Dit is om hierdie rede dat volledige kursusse in forensiese linguistiek by universiteite in Suid-Afrika aangebied moet word. Ideaal gesproke sou ’n gedoseerde meestersgraad in forensiese linguistiek, wat modules soos onder andere bewysreg, statistiek, rekenaarwetenskap, linguistiek en ook ’n basiese kursus in forensiese wetenskap insluit, voordelig wees vir die opleiding van professionele forensiese linguiste. So ‘n program sou uiteraard ook navorsing in dié veld in Suid-Afrika stimuleer.

Forensiese linguistiek word reeds suksesvol in verskeie lande gebruik en word reeds by verskeie universiteite regoor die wêreld aangebied (Broeders, 2001: 64,69; Blackwell, 2012: 1–4). Alhoewel daar navorsers is wat reeds binne die veld van forensiese linguistiek in Suid-Afrika

werk, is daar nie enige merkbare groei in die navorsingsveld nie. Daar is wel ’n merkbare groei in die hoeveelheid individue wat toegang het tot selfone en ander vorme van tegnologie soos die internet en e-poskommunikasie. Dit beteken dat daar die potensiaal is vir ’n groot groep individue om misdade deur middel van hierdie tegnologieë te pleeg. Forensiese linguistiek, in al sy vorme (outeuridentifikasie, sprekeridentifikasie, plagiaatidentifisering, ensovoorts) sal ook in Suid-Afrika moet ontwikkel om te verseker dat ons, soos verskeie ander lande, ook die vermoë het om ’n wye verskeidenheid misdade te ondersoek en so ook deel te word van die internasionale gemeenskap van forensiese linguiste.

5.4 Probleme en beperkings.

5.4.1 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 uitgevoer word (Barry en Luna, 2012; Stamatatos e.a., 2001; Chaski, 2001). In die huidige navorsing is die beperkte hoeveelheid data ook problematies. Dit het daartoe gelei dat daar nie in die analises 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-gramanalises, SVM en *log-likelihood ratios*) aan te pas in ’n poging om groter sukses in die analisering van beperkte data te behaal. Dit is ook nodig 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-gramanalises). Alhoewel daar slegs beperkte data tot die navorser se beskikking was, was dit nietemin moontlik om tot ’n mate idiolek onder die huidige groep deelnemers te identifiseer.

5.4.2 Vaardighede en kennis van die forensiese linguis.

Dit is belangrik om daarop te let dat die forensiese linguis nie ’n forensies-linguistiese analise alleen kan uitvoer nie. Forensiese linguiste is in baie gevalle afhanklik van kenners op ander vakgebiede, soos byvoorbeeld statistiek en rekenaarwetenskap, aangesien ’n forensiese linguis self nie noodwendig ’n kenner in laasgenoemde vakgebiede is nie. In die huidige navorsing was

dit nodig om die hulp van ’n rekenaarwetenskaplike in te roep aangesien die navorser nie self genoeg kennis dra van die vakgebied nie.

Bibliografie

Abeywickrema, P. 2008. *Now SMS deemed admissible evidence*. Besoek: 30/09/2013. http://www.island.lk/2008/09/03/features2.html

allAfrica. 2012. *Africa: Predatory cyber crime in South Africa - current risks and realities*. Besoek: 19/03/2013. http://allafrica.com/stories/201208010184.html

Angouri, J. 2010. Quantitative, qualitative or both? Combinig methods in linguistic research. In: Litosseliti (ed.). *Research methods in linguistics*. London: Continuum International Publishing Group. pp. 29-45.

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. Besoek: 05/08/2014. http://www.antlab.sci.waseda.ac.jp/research/iwlel_2004_anthony_antconc.pdf

Anthony, L. 2005. *Antconc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom*. In: 2005 IEEE International Professional Communication Conference Proceedings. Besoek: 11/07/2014. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1494244

Anthony, L. 2014. *Antconc: Readme*. Besoek: 16/07/2014. http://www.antlab.sci.waseda.ac.jp/software/antconc341/AntConc_readme.pdf

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press Ltd.

Barber, A. 2004. *Idiolects*. In: E.N. Zalta (ed.), The Stanford Encyclopedia of Philosophy: Winter 2004 Edition. Besoek: 23/04/2013. http://plato.stanford.edu/archives/win2004/entries/idiolects/

Barry, K en Luna, K. 2012. *Stylometry for online forums*. Stanford University. Besoek: 18/07/2013. http://cs229.stanford.edu/proj2012/BarryLuna-StylometryforOnlineForums.pdf

Blackwell, S. 2012. *History of forensic linguistics*. Wiley online library. Besoek: 16/01/2013. http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0508/full

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

Broeders, A.P.A. 2001. *Forensic speech and audio analysis forensic linguistics. A review.* 13th Interpol Forensic Science Symposium. 16-19 Oktober 2001. Frankryk: Lyon. Besoek: 16/01/2013. http://www.taracentar.hr/attachments/interpol_forensic.pdf

Brown-Jackson, M. 2013. *How linguistic analysis helped unmask Robert Galbraith as J.K Rowling*. Geekosystem. Besoek: 18/07/2013. http://www.geekosystem.com/linguistic-tool-rowling/

Buck, C. 2012. *Cybercrimes (via cell phones) up in 2011*. Phys.org. Besoek: 08/10/2014. http://phys.org/news/2012-05-cybercrimes-cell.html

Caliskan Islam, A, Greenstadt, R en Afroz, S. 2013. *Stylometry and online underground markets*, Video (aanlyn), besoek: 14/10/2013. http://www.youtube.com/watch?v=xL9aam3ZUlk&list=PLQICoZ0SpVvtgqeGYg4Fait9S1z7mxNEm&index=1

Carney, T.R. 2012. 'n Forensies-semantiese beskouing van die woordgebruik 'onkoste' in die hofsaak Commissioner for South African Revenue Service vs. Labat Africa Limited. *South African Linguistics and Applied Language Studies*, 30(4): 487-496.

Carney, T.R. 2013. 'Het doet emmer toverie': 'n Forensiese ondersoek na die (on)waarskynlikheid van nekromansie in Die Hexe. *Tydskrif vir Letterkunde*, 50(3): 24-38.

Carney, T.R. 2014. Being (im)polite: a forensic linguistic approach to interpreting a hate speech case. *Language Matters*, 45(3).

Cavnar, W.B. en Trenkle, J.M. 1994. *N-gram-based text categorization*. Besoek: 15/09/2014. http://odur.let.rug.nl/vannoord/TextCat/textcat.pdf

Chaski, C.E. 2001. *Empirical evaluations of language-based author identification techniques*. Besoek: 01/08/2014. http://www.iula.upf.edu/materials/050520spassova.pdf.

Chaski, C.E. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International journal of digital evidence*, 4(1).

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Besoek: 10/10/2013. http://babel.ucsc.edu/~hank/aspects_ch3.pdf

Clement, R en Sharp, D. 2003. Ngram and Bayesian classification of documents for topic and authorship. *Literary and Linguistic Computing*, 18(4): 423-447. Besoek: 11/03/2014. http://llc.oxfordjournals.org/content/18/4/423.full.pdf+html

Corney, M.W. 2003. 'Analysing e-mail text authorship for forensic purposes', Masters thesis, Queensland University of Technology, besoek: 12/08/2014. http://eprints.qut.edu.au/16069/1/Malcolm_Corney_Thesis.pdf

Costello, S.E. 2013. *Establishing that text messages are admissible*. Besoek: 30/09/2013. http://apps.americanbar.org/litigation/litigationnews/top_stories/040113-text-message-admissible.html

Coulthard, M. 2004. Author identification, idiolect and linguistic uniqueness. *Applied Linguistics*, 25(4): 431-447.

Coulthard, M. en Johnson, A. 2007. *An introduction to Forensic Linguistics: Language in Evidence*. VSA en Kanada: Routledge.

Coulthard, M. 2010. *Experts and opinions: In my opinion.* In: The Routledge Handbook of Forensic Linguistics. New York: Routledge.

Crankshaw, R. 2012. The validity of the Linguistic Fingerprint in forensic investigation. *Diffusion* 5(2). University of Central Lancashire. Besoek: 22/04/2013. http://atp.uclan.ac.uk/buddypress/diffusion/?p=1228

Croker, R.A. 2009. An introduction to qualitative research. In: Heigham, J en Croker, R.A. (eds.) *Qualitative research in applied linguistics.* Verenigde Koningkryk: Palgrave Macmillan. pp. 3-24.

Crystal, D. 1987. *The Cambridge encyclopedia of language*. Cambridge: University Press.

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

Davis, M. 2013. *Taking text messages to court*. Besoek: 30/09/2013. http://www.markdavisllc.com/Entries/2013/1/19_Taking_text_messages_to_court.html.

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

Easton, V.J. en McColl, J.H. n.d. Statistics Glossary (v1.1). Besoek: 18/07/2014. http://www.stats.gla.ac.uk/steps/glossary/categorical_data.html

'Gangs use of the internet and cell phones', *I look both ways*, 14 Junie 2010. Besoek: 08/10/2014 http://ilookbothways.com/2010/06/14/gangs-use-of-the-internet-and-cell-phones/

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. Birmingham: Aston University.

Grant, T. 2010. *Txt 4n6: Idiolect free authorship analysis?* In: The Routledge Handbook of Forensic Linguistics. New York: Routledge

Grieve, J.W. 2005. *Quantitative authorship attribution: a history and an evaluation of techniques*. Meestersverhandeling. Simon Fraser universiteit. Besoek: 18/07/2014. summit.sfu.ca/system/files/iritems1/8840/etd1721.pdf

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

Heigham, J en Croker, R.A. 2009. *Qualitative Research in Applied Linguistics*. Verenigde Koningkryk: Palgrave Macmillan.

Hershensohn, J. 2005. *I.T. forensics: the collection and presentation of digital evidence*. Besoek: 08/10/2013. http://icsa.cs.up.ac.za/issa/2005/Proceedings/Full/076_Article.pdf

Hockey, S. s.a. *The history of humanities computing*. Besoek: 13/06/2014. http://www.digitalhumanities.org/companion/view?docId=blackwell /9781405103213/9781405103213.xml&doc.view= content&chunk.id=ss1-2-1&toc.depth=1&brand=9781405103213_brand&anchor. id=0%23ss1-2-1_ b15#ss1-2-1_b37

Holmes, D.I. 1994. Authorship Attribution. *Computers and the Humanities* 28: 87-106

Holmes, D.I. en Tweedie, F.J. 1995. *Forensic stylometry: a review of the Cusum controversy*. Besoek: 16/042014. http://promethee.philo.ulg.ac.be/RISSHpdf/Annee1995/ Articles/DHolmesetc.pdf

Holmes, D.I. 1998. The evolution of stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3): 111-117

How2stats. 2011. *Yates' correction*. Besoek: 13/02/2015. https:www.how2stats.net/2011/09/yates-correction.html

Hubbard, E.H. 1994. Errors in court: A forensic application of error analysis. *South African Journal of Linguistics*, 12(20): 3-16

Hubbard, E.H. 1995. Linguistic fingerprinting: A case study in forensic stylometrics. *South African Journal of Linguistics*, 13(26): 55-72.

Hubbard, E.H. 2009. 'Stylometric and error analysis in the context of a style shift in abusive e-mail texts', 9th Biennial Conference on Forensic Linguistics/Language and Law, VU University, Amsterdam.

Hunt, H. en Zabel, M.P. 2012. *United States: Text messages as trial evidence - authentication*. Besoek: 08/10/2013. http://www.mondaq.com/unitedstates/x/200778/court+procedure/Text+Messages+As+Trail+Evidence+Authentication

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. pp. 47-56

Ivankova, N.V. en Creswell, J.W. 2009. Mixed methods. In: Heigham, J en Croker, R.A. (eds.) *Qualitative research in applied linguistics.* Verenigde Koningkryk: Palgrave Macmillan. pp. 135-161.

Jackson, A.R.W en Jackson, J.M. 2004. *Forensic Science*. Verenigde Koningkryk: Pearson Education Limited.

Johnson, A. en Coulthard, M. 2010. *Current debates in forensic linguistics*. In: The Routledge Handbook of Forensic Linguistics. New York: Routledge.

Johnson, K. en Johnson, H (eds.). 1999. *Macro/microlinguistics*. In: Encyclopedic Dictionary of Applied Linguistics. Blackwell Publishing.

Juola, P. 2006. Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3): 233-334

Kacmarcik, G. en Gamon, M. 2006. *Obfuscating document stylometry to preserve author anonymity*. In: Proceedings of the COLING/ACL on Main conference poster sessions. pp. 444-451

Khmelev, D en Tweedie, F.J. 2001. Using Markov chains for identification of writers. *Literary and Linguistic Computing*, 16(3): 299-307. Besoek: 09/05/2014. http://llc.oxfordjournals.org/content/16/3/299.full.pdf+html

Klopper, R. 2009. The case for cyber forensic linguistics. *Alternation*, 16(1): 261-294

Koppel, M en Schler, J. 2004. *Authorship verification as a one-class classification problem*. In: Proceedings of the Twenty-First International Conference on Machine Learning (ICML). pp. 489-495

Koppel, M., Schler, J. en Argamon, S. 2009. *Computational methods in authorship attribution*. Besoek: 18/04/2013. http://u.cs.biu.ac.il/~koppel/papers/authorship-JASIST-final.pdf

Koppel, M., Schler, J., Argamon, S. en Winter, Y. 2012. The “fundamental problem” of Authorship Attribution. *English Studies*, 93(3): 284-291.

Kotzé, E.F. 2007. Die vangnet van die woord: forensies-linguistiese getuienis in ’n lastersaak. *South African Linguistics and Applied Language Studies*, 25(3): 385-399

Kotzé, E.F. 2010. Author identification from opposing perspectives in forensic linguistics. *South African Linguistics and Applied Language Studies*, 28(2): 185-197

Krige, R. 2012. *The admissibility of electronically generated evidence in a court of law*. [Powerpoint] Aanbieding by: CyberCon Africa 2012. Besoek: 12/07/2013. http://www.google.co.za/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=5&ved=0CD4QfjAE&url=http%3A%2F%2Fwww.lexisnexis.co.za%2Fpdf%2Fcybercon-1-2-The-Admissibility-of-Electronically-Generated-Evidence-in-a-Court-of-Law-Roux-Krige.ppt&ei=qd_fUdbwGbSQ7Ab0slGoBA&usg=AFQjCNGzHhqOA-WQPPCvUuzgoUrm4s3Dkw&bvm=bv.48705608,d.d2k

Lazaraton, A. 2009. Discourse analysis. In: Heigham, J en Croker, R.A. (eds.). *Qualitative research in applied linguistics.* Verenigde Koningkryk: Palgrave Macmillan. pp. 242-259.

Leonard, R.A. 2006. Forensic Linguistics: *Applying the scientific principles of language analysis to issues of the law*. Besoek: 11/07/2014. http://www.robertleonardassociates.com/PDF/ForensicLinguistics_Applying-Scientific-Principles.pdf

Litosseliti, L. 2010. *Research methods in Linguistics*. Verenigde Koningkryk: A&C Black

Lombard, E. en Carney, T.R. 2011. Die wenslikheid van Afrikaans as vaktaal vir regstudente. *Potchefstroomse Elektroniese Regsjoernaal*, 14 (1): 164-187.

Luyckx, K en Daelemans, W. 2011. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26(1): 35-55.

McLeod, N. en Grant,T. 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 University.

McMenamin, G.R. 2002. *Forensic Linguistics: Advances in Forensic Stylistics*. CRC Press LLC. Besoek: 09/08/2013. http://www.sciencelib.net/2983/forensic-linguistics-advances-in-forensic-stylistics-g-mcmenamin-crc-2.html

McMenanim, G.R. 2010. *Theory and practice of forensic stylistics.* In: The Routledge Handbook of Forensic Linguistics. New York: Routledge.

MedicineNet.com. 2014. Definition of forensic. Besoek: 09/10/2014. http://www.medicinenet.com/script/main/mobileart.asp?articlekey=10604

Mendenhall, T.C. 1887. The characteristic curves of composition. *Science*, 9(214): 237-249 Besoek: 03/06/2014. http://www.jstor.org/stable/1764604

Mendenhall, T.C. 1901. A mechanical solution to a literary problem. *Popular Science Monthly*, 9(60): 97-110. Besoek: 05/03/2014. http://en.wikisource.org/wiki/Popular_Science_Monthly/Volume_60/December_1901/A_Mechanical_Solution_of_a_Literary_Problem

Mikros, G.K. nd. *Authorship Attribution in Greek blogs*. Besoek: 07/05/2013. http://users.uoa.gr/~gmikros/Pdf/AA%20and%20GI%20in%20Greek%20blogs_Qualico12.pdf

Michell, C.S. 2013. 'Investigating the use of forensic stylistic and stylometric techniques in the analyses of authorship on a publicly accessible social networking site (Facebook)'. Meestersverhandeling. Universiteit van Suid-Afrika.

Mitchell, E. 2008. The case for forensic linguistics. *BBC NEWS*, 8 September. Besoek: 17/09/2013. http://news.bbc.co.uk/2/hi/science/nature/7600769.stm

Mobile Pronto. 2010. *The history of SMS text messaging*. Besoek: 29/01/2013. http://www.mobilepronto.org/en-us/the-history-of-sms-html

Moeketsi, R.H. 1997. 'Of African languages and forensic linguistics: the South African multicultural criminal courtroom'.Dlit et Phil. Universiteit van Suid-Afrika.

Moeketsi, R.H. 1999. *Discourse in a multilingual and multicutural courtroom: a court interpreter's guide*. Pretoria: JL van Schaik.

Mohan, A, Baggili, I.M en Rogers, M.K. 2010. Authorship attribution of SMS messages using an N-grams approach. Besoek: 09/06/2014. http://www.cerias.purdue.edu/assets/pdf/bibtex_archive/2010-11-report.pdf

National Research Council of the National Academics. 2009. *Strengthening forensic science in the United States: a path forward*. Besoek: 01/08/2014. http://www.nap.edu/catalog/12589.html

Olivier, J. 2013. *Die mate van konsekwentheid in SMS-Afrikaans*. Litnet Akademies 10(2). Besoek: 01/08/2014. http://www.litnet.co.za/Article/die-mate-van-konsekwentheid-in-sms-afrikaans

Olsson, J. s.a. *Forensic Linguistics*. Proefhoofstuk. Besoek: 17/01/2013. http://www.eolss.net/Sample-Chapters/CO4/E6-91-13.pdf

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

PewInternet. 2012. *Cell phone activity 2012*. Besoek: 15/03/2013. http://pewinternet.org/Reports/2012/Cell-Activities/Additional-Demographic-Analysis/Demographics.aspx

Philbrick, F.A. 1949. *Language and the law: The semantics of forensic English*. New York: The Macmillan Company. Besoek: 10/10/2014. https://archive.org/details/languagelawseman00phil

Ponelis, F. 2009. *Die taal wat ons praat*. Besoek: 30/01/2013. http://www.volksblad.com/ByNuus/Die-taal-wat-ons-praat-20091218-2

Prinsloo, D.J. en Prinsloo, D. 2011. *WordSmit Tools*. Besoek: 12/09/2014. https://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4493/prinsloo.pdf?sequence=1

Ragel, R, Herath, P en Senanayake, U. n.d. *Authorship detection of SMS messages using unigrams*. Besoek: 02/04/2014. http://arxiv.org/ftp/arxiv/papers/1403/1403.1314.pdf.

Rao, J.R en Rohatgi, P. 2000. *Can pseudonymity really guarantee privacy?* In: Proceedings of the 9th USENIX Security Symposium.

Rawlinson, C. 2011. *Infographic: Cellphone usage+SA mobile stats*. Besoek: 25/01/2013. http://www.chrisrawlinson.com/tag/south-african-cellphone-stats/

Reddy, V en Potgieter, C. 2006. 'Real men stand up for the truth': discursive meanings in the Jacob Zuma rape trial. *Southern African Linguistics and Applied Language Studies*, 24(4): 511-521.

Rusko, M. en Garabík, R. n.d. *Corpus of spoken Slovak language*. Besoek: 23/01/2013. http://korpus.juls.savba.sk/attachments/publications/garabik-spoken_slovak_corpus.pdf

Sanderson, P. 2007. Linguistic analysis of competing trademarks. *Language Matters*, 38(1): 132-149.

Schulstad, I, Boga, M, Jordan, C en Pally, K. 2012. *Evaluation of a stylometry system on various length portions of books*. Besoek: 14/10/2013. http://csis.pace.edu/~ctappert/srd2012/d5.pdf

Seegogga se Bloggie. 2010. *SMS-TAAL*. Besoek: 30/01/2013. http://seegogga.wordpress.com/2010/10/19/sms-taal-2/

Sierra, G., López, F., Méndez, C., Solórzano, J. en Méndez, E. 2013. *Exploring stylometric measures for authorship attribution*. [Powerpoint]. Aanbieding by die: 11th Biennial Conference on Forensic Linguistics. National Autonomous University of Mexico.

Solan, L.M. 2010. *The forensic linguist: The expert linguist meets the adversarial system.* In: The Routledge Handbook of Forensic Linguistics. New York: Routledge

Somers, H.s.a. *Stylometry and Authorship*. [Powerpoint]. University of Manchester: School of Computer Science. Besoek: 17/04/2013. http://personalpages.manchester.ac.uk/staff/harold.somers/LELA30922/Authorship.ppt.

Stamatatos, E. s.a. *A survey of modern authorship attribution methods*. Besoek: 16/04/2013. http://www.icsd.aegean.gr/lecturers/stamatatos/papers/survey.pdf

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

'State experts differ on voice analysis in George Zimmerman case', *Click Orlando*, 14 Mei 2013. Besoek: 08/10/2014. http://www.clickorlando.com/news/state-audio-expert-ids- trayvon-martins-voice-in-911-calls/20139744

Statistics How To. 2015. *Yates correction: What is it used for in statistics*. Besoek: 13/02/2015. http://www.statisticshowto.com/what-is-the-yates-correction/

Stevenson, K. 2008. *Courts confront admissibility of text and instant messages*. Besoek: 30/09/2013. http://www.buchalter.com/bt/index.php?option=com_content&task=view&id=248&ltemid=1

Streicher, P. 2010. *The role of SMS in the law*. Besoek: 30/09/2013. http://www.prlog.org/11140507-the-role-of-sms-in-the-law.html

Svartvik, J. 1968. *The Evans statements: A case for forensic linguistics*. Part 1 of 2. Besoek: 13/06/2014. http://www.thetext.co.uk/Evans%20Statements%20Part%201.pdf.

Svartvik, J. 1968. *The Evans statements: A case for forensic linguistics*. Part 2 of 2. Besoek: 13/06/2014. http://www.thetext.co.uk/Evans%20Statements%20Part%202.pdf.

Taylor, D.C. 1998. Addressing the insane language of law. *Tydskrif vir Hedendaagse Romeins-Hollandse Reg*, 61(1): 668-677.

Thetela, P.H. 2002. Sex discourses and gender constructions in Southern Sotho: a case study of police interviews of rape/sexual assault victims. *Southern African linguistics and applied language studies*, 20(3): 177-189.

Turell, M.T. 2008. Resensie van: *Introduction to Forensic Linguistics: Language in Evidence*, deur Malcolm Coulthard and Alison Johnson. ISBN:978-0-415-32023. Besoek: 13/04/2014. http://www.atlantisjournal.org/ARCHIVE/30.2/2008Turell.pdf

Unicef. 2012. *South African mobile generation: Sudy on South African young people on mobiles*. Besoek: 25/01/2013. http://www.unicef.org/southafrica/SAF_resources_mobilegeneration.pdf

Van der Vyver, L. (samest.). 2010. *101 SMS afkortings*. Op: Blessies Babbel. Besoek: 29/01/2013. http://blessieland.webs.com/101smsafkortings.htm

Watney, M. 2009. Admissibility of electronic evidence in criminal proceedings: An outline of the South African legal position. *Journal of Information, Law & Technology*, 2009 (1): 1-13. Besoek: 26/08/2013. http://go.warwick.ac.uk/jilt/2009_1/watney

Wei, L.(ed). 2011. *From pedagogical practice to critical enquiry: an introduction to applied linguistics.* In: The Routledge Applied Linguistics Reader. New York: Routledge

Wikipedia. 2013. *Emoticon*. Besoek: 30/01/2013. http://en.wikipedia.org/wiki/Emoticon

Wikipedia. 2013. *Smiley*. Besoek: 30/01/2013. http://en.wikipedia.org/wiki/Smiley

Wilkinson, M. 2012. *The best freeware corpus analysis program for translators?* Besoek: 05/08/2014. http://www.bokorlang.com/journal/60corpus.htm

Zax, D. 2014. *How did computers uncover J.K. Rowling's Pseudonym?* Besoek: 27/02/2014. http://www.smithsonianmag.com/science-nature/how-did-computers-uncover-jk-rowlings-pseudonym-180949824/

Zechner, N. n.d. *The past, present and future of text classification*. Besoek: 17/09/2013. http://www8.cs.umu.se/~zechner/classum.pdf

Bylaag 1: Die volledige SMS-tekste van elke deelnemer.

Teks X

Haha dis nogals cool!! En jy kry nou lekker geld! Ek moet nog vju dai soundtrack download j moet my remind :)

Heyy!! Wie is die uitgewers van ons hanbdoek en waar isit uitgegee? Oh en ini voetnotas se mens para of par? Haha

heeey tjommmie!! ai jammer ek reply nou eers!! hoop jyt n baie lekker daggie gehad gister? ekt soortvan al klaar planne gehad vir vandag, maar e

ons het n groot taak wat dinsdag moet in, maar as ons dit afgehandel kan kry voor dan sal ons graag wil kom. Sal jou Sondag se! Lekker dag?

Heeyy.. Ons was biki laat gewees ma heti studiegidse gekry nie, syt net gepraat van die boek wat ons moet kry ma ek weet nie wat dit isi.. En v

Deelnemer 1

ja nee dt is bja min werk. ek hoop neti di 2de toets gan te erg wees ni! dit moes beslis n tweede jaars vak gwees het..ma ek dink VBB ini tweede

wel ek weet presies hu jy vul!! en ds weird dat dt gbeu in verhoudings...want dan sodra mens wee saam is dan isit awesome..ma ek weni ds vreemd.

jis mar die man is besig om su lewe agter mekaar te kry klink dit vir my!! ek begin volgende week maandag met somerskool in twee regsvakke. kan

drink nou sulke cappuchino wat ek gekoop het nounet.. ds veronderstel om white chocolate koffie te wees.. tu verwag ek iets soos fego s'n, en tu

Hi Gizela!! haha geen probleem nie.. die getroude lewe vat seker nogals baie tyd! mar ja ds doodreg! ekt sommer gedink jy kan vi my paar goed wh

liefieqi ds 335-340 en dan 372. haha ds soooo min omw. ma eks nie doodseker oor 333 en 334 nie. maybe dit ook ma die studiegids se nie ons moet

Deelnemer 2

Heey! ekt ongelukkig klaar planne maar indien dit verander sal ek jou def laat weet! baie dankiee :)

Hey dude! Dit lyk asof die game in ekstra tyd ingaan, wil julle nie solank na ons kom nie? Jap, ons ry oor so 5min!! is op pad.

Ja, klare se in sy artikel mens moet verbeelding gebruik.. So dan verwys ek na antaki wat daaroor praat, en dat ek een kies as oplossing. En ja

sal jy sommer die boek kampus toe bring en my laat weet as daar is, as jy kan! anders sal ek dit more aand kry. lekker slapies xx

buddy hoop dit gaan goed! mis jou man! is so besig maar sodra ons bietjie tyd het reel ons ietsie kay!! hoop jyt n leke week!??

Is reg, ons kyk die sokker klaar, so as dt nog aanhou naby 8 sal ek jou se. Parkview se eds neh!

Hey! Ons het nogals baie werk voor ons ipw skryf so ekt nog nie daaraan gedink nie.. die week is bietjie mal! so baie werk!!

Deelnemer 4

Ek wag bo in jou kantoor tot liona klaar is. Waars jy? Kry jy my boodskappe? Ons kry jou by die kar

Ek ek sien alklaar uit! Maar eks nogal doodmoeg. Dit was so lekker om met jou te praat!! Lekker slapies en sweet dreams! Baie lief vir jou della!!

Haha wel dankie! Ek voel nou goed oor myself! Ek gaan net die verskoning gebruik en se "jy verstaan my nie". Cliche. Ek blabber rerig nou net.

Ja ek het eintlik nog werk maar ek wil dit nie nou doen nie. Ek dink ek sal dit later doen. Ek is nogal moeg! Lekker slapies! Sien jou more!

Kan jy nie? Seker nie. Het jy n boek saamgevat? Oh want julle moet rerig gaan. Julle moet nou die kaartjies koop en rerig gaan. Dit is baie baie lekker

Dit was rerig lekker gewees!! Dis si weird dat jy weer terug gegaan het. Wanneer jy weer kom doen ons weer iets! :) stuur groete vir almal daar!

Jammer ek was op whatsapp maar ek was nie op my foon nie so ek get online gewys maar ek jou nie geinoreer nie. Ja ek is seker maar net dom. Ai

Dit expire eers volgende jaar. Haha ek dink jy sal dit volgende keer kry! Jy weet nou hoe dit is en wat om te verwag. Moenie verder stres daaroor nie!

Deelnemer 5

Groot asb se vir my j gaan OPV toe?? Ek is n slegstudent en gan klas skip. k sit en werk an afr, sug, ek kry net ni klaar nie! Kan j m updated hou PLZ?

Elo skat! Hoor jyt m gsoek! Sorry man, jlo en sepedi het kla gemaak so ekt nie klas vandag nie! Ek sit juis nou en werk aan afr ;) sterkte met jo taak!

M iPad..k weet dis net n ipad, mar m hele lewe is daarop.. Al m info vir die exams ens. Al die lesings wat ek record het en notas ens. En m nuwe cover

Almal.. Ek moes vanoggend begin.. Ek samel nou info in van al m vriende.. Sug. Ek sou goed gedoen het. Waarskynlik n prestasiebeurs ontvang. Damn..

Nee als gaan heel goed dankie! Die studies kry m biki onder.. Dis ongelooflik baie werk.. Gan Baie goed met B! En met ons, haha! Marinus se baba kom di

Hello sus!! Jy is blerrie oulik man!! Hoor jy was bietjie in n tjor knor... BAIE bly jys oky! Hgd nog jong! Met jo, eni studies en als! Mis jo stuknd

Ello sus!! Ek wil baie graag, maar ek kannie :(ek is bietjie vas met m werk as n tutor by di universiteit en oor naweke is al wane ek kan uitkom.Ek sa

Oki ek hoef ni meer grove to t gan ni ekt reg gekom b ons clicks, so j kan gan wanr j wil, mar as j wil he ek moet samet jo gan is dit ook doodreg!! Xx

Ek sal einklik eerder b m ma se huis wil kuier.. Het n paar goedjies wat ek wil doen en so aan, bv m invites kla maak, n event op fb create,sulke goed.

Elo! Gan bietji op FB of whatsapp asb? Ekt vi jo n invite gestuur en moet dringend weet of j kom? Sleg vir m ego as ek mense moet vra of hul kom hoor!

Deelnemer 6

Hello Katy. Ek wil jou net weer herinner dat Parkerstraat 17 se hek nie behoorlik wil werk nie. Dankie. Su-Mia.

Ek het van 1 tot 4 klas op Maandae. So as julle in die erf in kan kom sal ons die sleutel in die tuinboksie in die trap se muur los.

Elandr? het al 12:30 klas so ons ry al half twaalf kampus toe.

Sms maar vandag mamsie. Die data is klaar, en ek wil nie al my airtime daarop spandeer nie, netnou is daar krisis. Mwah mwah!

Die radio waarsku oor vloede reg oor Gauteng, veral die Wes-Rand. Monument is blykbaar doorstroom.

Jy is reg hy kritiseer Komminisme omdat dit teen sy agenda gaan. Ons kry mekaar 12:30 voor die bib.

Ja.Ek het al my werk verloor van wat ek deur die dag gedoen het want die rekenaar het aan en af gegaan... So toe ek van klas af kom power ek maar deur.

En dis koud. En ek het net my jean ingepak wat warm is. Kon nie kouse vanoggend vind in my ma se kaste nie.

En jou pa is van nee hy sal nie kan oor dit kom terwyl hy hier om al haar goed is nie,maar hy sal ook nie trek nie. En hy gaan homself dood treur.

Deelnemer 7

Ek sien ek het die huiswerk boek in my sak gesit. Ek sal dit more weer in Stehan se tas sit. Skryf sommer vir my in Stehan se spelwoordeboek?

Ek gaan nou eers na my-niggie-het-woensdag-voor-die-magistraat-getrou-want-sy-is-so-cool-ete toe saam die heeele familie en ek weet nie hoe lan

Op pad na hulle huis toe met my eie stapels merkwerk. Sal koffie maak en in gees by jou wees.

Ek is nou amper klaar. Maar ek dink dit is een van my swakste take ooit. Myne is vreeslik oppervlakkig en kort. Ek gee nie meer om nie.

Hallo, wil net seker maak oor vanmiddag. ML se sy gaan saam n juffrou anton van wou toe. Moet ek direk daarnatoe gaan? En wat gebeur met die seuns?

Sal sien oor vanaand. Waar gaan julle? Anders is jy welkom om more oggend erens by my te kom boeke uitneem.

Kom kry more oggend gou n boek by my as jy tyd het. Dit gaan oor fashion. You maait laaik it.

Deelnemer 8

Sal jy asb baie oulik wees en by kfc vir H n oreo krusher en n zinger burger kry en vir my n dairy milk krusher? Eks snotterig en wil nie ry nie

Hmmm okay dan sal ek nie kan saamgan nie.

Okay, vraag: wil julle by kolonnade iets gan doen? Of is dit n nee-nee?

Haha inderdaad. Sal maar besprekings moet begin maak om mekaar te sien. Gan jy donderdag ENG toe?

Hulle lyk ook basies dieselfde behalwe vir die silwer hakkie.

Hey eks bietjie vroeg. My kar is hier by die winkeltjie gestop, so laat weet net as jy hier is.

Deelnemer 9

Le en dink hoe ek nie meer kan wag om by jou te doeks more aand nie !*:* ())

Ek voel pap en my neus loop soos crazy ! (X_X bestuur mooi !*:* pwt nounou ek is lief vir jou! <3<3

Ek het 88% gekry vir my sielkunde semester toets

Dis deel van die lewe. Ek belowe jou ek het al 2 keer deurgegaan waar deur jy nou gaan. Ek belowe jy gaan nog gevoelens kry vir n ander meisie

Baie dankie! Ek kom nie klas toe. Die verkeer is erg

Hello tanni. Ons het more n familie ete in pretoria so ek sal ongelukig nie kan uithelp more nie. Jammer ek laat weet nou eers

Eks nie dan hier nie ek en joe se ma gaan winkels toe, ek sal dit by die hek los vir julle

Wees 11 uur op hoofkampus as jy wil saam Kempton toe ry

Deelnemer 10

"Verdeelbord" daar is nou 'n mooi Afrikaanse woord! Ek is beindruk.

Is daar 'n spesifieke rede hkm die vergadering geskuif is?

Agge shame... Eintlik het ek bedoel daar was n swetterjoel kinders gister.

Dankie vir almal se boodskappe en ondersteuning! Die matrieks reken hulle behoort almal deur te gaan.

Haai, Praat jy van die kommandobestuur vergadering? Ons het dit geskuif om meer mense te akkommodeer. Indien daar sake is wat dringend bespreek

Hi, kan ek asb volgende Dinsdag (22 April) die middag of volgende Donderdag (24 April) die oggend my hare kom sny? Groete

As julle kan kry by Veearts: Duplocillin (S4) en Phenylbutazone (Bute). Die " blou olie" by Afgri: Wound Expell (Coopers) Ons het reeds: Dect

Great sien j dan

Deelnemer 11

'n Deel vd werk is vir my baie lekker. Anders kan ons doen wat ons wil!

Hallo liewe maat! Dis gouache! Geniet die pragtige swazila

Ek is só bly jy is gelukkig! Dis hoe dit moet wees!

Ekt nie geleer nie, dus het dit dôners sleg gegaan! Maar d

Ja! Jy kan kom dan kan ons gaan eet! Eks nog nie aangetrek nie, maar kom solank!!!

Ons het ook Freud en sy seksmotiewe behandel in Engels toe ons Sons and Lovers gedoen het! Waansinnig as jy my wil vra.

Ek kan die maan nie sien nie!!! Daar is nie 'n maan aan my

Dis lieflik! Het nou eers gesien dit reen, as ek by my ven

Deelnemer 12

Aners kan ons doen wat ons wil.

Eks by ed's nou.. Mar ons gan nou-nou brazenheads toe apparently. Flip het n ding da

Maaaaarz! Waars jy? Baie geluk met jou verjaarsdag. Ek hoop jyt n great jaar. Geniet dit mar wees versigtig. Lief jou meer as woorde xx

Marz more is donderdag.. Ekt vegeet ekt vir mense gese ek sal samet hulle ed's toe gan.. Los die pizza en kom ed's toe?

More vriendjie.. Eks half 11 da vimy tutorial mar ek glo nie my klas gan lank anhou nie. Sal jou msg as ek kla is?

Hey bul.. Sal jy net seker maak jyt die kaart by jou as jy my kom haal? Dan kan ons gou goedjies gan koop terwyl ons vir minette wag

Ja petools dis eintlik leke warm hie deur die dag.. Dis wyn is leke, die sigarette is leke en die mense is leke.. En dis so mooi hie petools..

Deelnemer 13

Sjoe daai foto se sommer "eks single!" Hehehe Ag dankie vir jou man! As jy nie saam gekom het nie sou dit nie so fun gewees het nie! Dit voel vi

Happy bday tay!!! Ek hoop jy gaan n verskriklike lekker daggie he! Jy is nou offisieel n hoerskool meisie! ;) ek hoop jy word vrek baie bederf e

Nee oupa. Die proef is vir die tydperk van 6 maande. 3 maande waar die uni vir ons n lys gee met skool name waaruit ons kan kies en 3 wat ons ka

Die plek waar ons gaan bly het twee enkelbedjiea in die kamer en een sleepercouch in die sitkamer. Nou wil eugene nie meer saamgaan nie maar wil

My ma se ons kan daai tas sommer net gewoonweg afvee met water en seep en dan as hy mooi droog is hom smeer met daai leer goed. Ons het nog van

Ek vra vanaand sommer uit interessantheid vir my ma wat sal sy doen as ek nou swanger raak. Sal sy my uitskop en verwyt of ondersteunend wees to

Ek wil graag as ek 24 is swanger wees. Ek dink dis die regte ouderdom vir my en ons verhoudimg en getroude lewe... lekker slapies my skat. Jy i

Deelnemer 14

Hehe ja natuurlik en jy gani weet waar ek dt kry of by wie ni.

Ok. Latweet my net om seker te mak ons is bydi huis. Wnt ons gan mre stem.

Al wt ek mre ht om te dun is wasgud en skottelgud.

Wanr gan ons wee laeveld tu? As j gan latweet my seblief?

Kan ek by ju km kyer vnand? Ek ht nix anrs om te dun ni.

Hulat bgn klas mre? My rooster is weg, stur asb vimy june. En wate vakke ht ons als mre? Sien mre leker and

Hi bokkie, hoe lat land jy mre? Wi gan ju by di lughawe kry? Latweet my as ek mut. Lief ju

My tani kom mre vn nelsp af pta tu. So ek gan mt ha kyer en sy slap by my oor so ek gani sam jule kn ytgani. Jamer!

Di dag mut nt verby km! Kani wag om ju mre te sien ni.

Ek kani wag vr my susi se baby om te km ni! Dis nog nt 3 slapies. Yay!!

Bylaag 2: Volledige tabel met vergelykings en die persentasie ooreenkoms tussen die tekste (Stilistiese analise).

KENMERKE VAN TEKS X	**1**	**2**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1. Meer dubbel uitroeptekens teenoor enkel uitroeptekens: (!!) =3 (!) =2	Gebruik ewe veel van altwee: (!!) = 3 (!) = 3	x (!!) =3 (!) =10	x (!!) = 3 (!) = 10	x (!!) = 4 (!) = 10	(!) =1	Geen uitroeptekens	Geen uitroeptekens	(!) = 5	(!!) = 2	*(!!!) = 2 (!) = 10	(!) = 1	*(!!!) = 1 (!) =4	(!!) = 1 (!) = 3
2. Gebruik nie diakritiese tekens nie: “se” (i.p.v “sê”) “n” (i.p.v ”‘n”)	x (se, n)	x (se, n, reel)	x (se, more, cliche)	x (se, n, he)	Geen woorde wat diakritiese tekens benodig	x (se, n, more, erens)	x (n)	X (le, n)			x (n, more)	x (n, he, hoerskool, se)	Geen woorde wat diakriti tekens benodi
3. Gebruik in 1 geval nie hoofletters by selfstandigenaam woorde nie: “dinsdag”	x 5 gevalle	x 4 gevalle	x 3 gevalle	x 7 gevalle		x 1 geval	x 6 gevalle	X 3 gevalle	x 1 geval	x 1 geval	x 9 gevalle	x 1 geval	x 3 geval
4. Bevat 5 sinne wat nie met hoofletters begin nie.	x 14 gevalle	x 10 gevalle	x 1 geval										
5. Gebruik verskillede vorme van ‘dit’ + ‘is’ (en ‘nie’): ‘dis’, ‘isit’, ‘dit isi’	x Gebruik ‘isit’												
6. Gebruik die dubbel ellips [..] i.p.v die vol ellips […]	x	x	Geen ellips	x		Geen ellips	Geen ellips	Geen ellips		Geen ellips	x		Geen e
7. Verkort graag	x	x	x	x			x	x		x	Verkorting		x

woorde deur samesmelting - veral met 'het': syt / isit / jyt / isi. (12 gevalle - 4 met "het")	4 gevalle - geen met "het"	3 gevalle - 3 met "het"	3 gevalle - geen met "het"	8 gevalle - 4 met "het"			2 gevalle - geen met "het"	2 gevalle - geen met "het"		2 gevalle - 1 met "het"	s het geen ooreenkomste.		2 geval geen m "het"
8. Verkort in 2 gevalle dubbel konsonante: maar/ma	x 5 gevalle			x Verkort 'maar' na 'mar'			x 2 gevalle				x Verkort 'maar' na 'mar'		x 11 geva
9. Gebruik die woord 'lekker'		x 1 geval (lekker) en 1 geval (leke)	x 5 gevalle (lekker)							x 1 geval (lekker)	(leke)	x 1 geval (lekker)	(leker)
10. Gebruik 'haha'	x		x	x 1 geval			x						
11. Gebruik wisselvorme van die groet 'hey': 'heyy' / 'heeey' / 'heeyy'		x Gebruik 'heey' en 'hey'					Gebruik een keer 'hey'						
12.Lagtekens (2 gevalle)		x 1 geval	x 1 geval	x 2 gevalle				x 2 gevalle				x 1 geval	
13. Gebruik Engelse woorde	x 4 gevalle	x 3 gevalle	x 4 gevalle	x 14 gevalle	x 3 gevalle	x 3 gevall (een volledige Englese sin)					x 3 gevalle	x 5 gevalle	x 2 geval
14. Gebruik 'daai' (gespel 'dai') i.p.v 'daardie'	Geen gebruik van 'daardie'	Geen gebruik van 'daardie'	Geen gebruik van 'daardie'	Geen gebruik van 'daardie'		Geen gebruik van 'daardie'	Geen gebruik van 'daardie'			Geen gebruik van 'daardie'	Geen gebruik van 'daardie'	x 3 gevalle (daai)	
15. Vorm graag				Geen					Geen		Geen		

verkortings met 'nie' - 'heti', 'isi'				gebruik van die negatief					gebruik van die negatief		gebruik van die negatief		
16. Een ongewone verkorting: 'vju' (vir jou)													
17. Onafgekorte, dubbelvokaalwoorde kom ook voor.	x (saam)	x (o.a: "maar" en "daar")	x (o.a: "maar" en "wanneer")	x (o.a: "gaan" en "klaar")	x (o.a: "daar" en "gaan")	x (o.a: "gaan" en "saam")	x ("laat")	x (o.a: "gaan" en "saam")	x ("daar")	x (o.a: "gaan" en "maar")		x (o.a: "saamgaan")	
TOTALE OOREENKOMSTE:	10/17= 58.8%	11/17 = 64.7%	10/17 = 58.8%	10/17 = 58.8%	2/17 = 11.8%	4/17 = 23.5%	6/17 = 35.3%	5/17= 29.4%	2/17 = 11.7%	4/17 = 23.5%	5/17 = 29.4%	7/17 = 41.2%	4/17 = 23.5%