Authorship Attribution, Idiolectal Style, and Online Identity:

a specialised corpus of Najdi Arabic tweets

Mashael AlAmr

Submitted in accordance with the requirements for the degree of
Doctorate of Philosophy

The University of Leeds
School of English and School of Computing

June 2022

Acknowledgements

I am truly grateful for my supervisor Dr Alison May; being her student is an honour and a privilege. I feel most fortunate to embark on this journey under her guidance. I cherish all the talks, the comments, and the opportunities she has given me throughout this experience. My sincere thanks to my supervisor Professor Eric Atwell for his constant support and guidance. This research could never have been achieved without their supervision. Also, I extend my thanks to Professor James Dickens for reviewing the Najdi Arabic features that are at the core of this study and providing me with feedback.

Also, my sincere thanks and appreciation to my home, the Kingdom of Saudi Arabia and King Saud University for offering me this opportunity and sponsoring my research.

I would like to express my heartfelt gratitude to my family for their never ending love, prayers, support, and encouragement. My mother, Mona Alhatlani, is an exceptional woman who is and continues to be my north star. My father, Mohammed Alamr, who taught me everything I know about resilience and perseverance. My crown jewels and siblings, Mishal, Mishari, and Muneera, my life pillars that I always rely on: thank you and I love you with all my heart.

My warmest thanks to all my dear extended family and friends for cheering for me since day one, and especially Sara Albassam and Ghadah Abalkhail, I will never forget all the long phone calls and the warmth they have generously given me. Fahdah Alsuwailim and Mashael Altammami, I will always cherish the words of encouragement and wisdom they shared with me. Tomader Alfuraikh, we continue to root for each other as we share this doctoral journey. Dina Alhejailan, I will never forget the support she has given me in my first year.

Abstract

This thesis investigates a synergy of approaches, namely corpus linguistics, stylistics, computational linguistics, and computer-mediated communication to address forensic authorship attribution problems in Arabic. Forensic linguists approach authorship from a position that a linguistic theory would lay the groundwork of the analysis; this theory in most of the literature is that of the idiolect. Grant argues with this by proposing that “[p]ractical authorship analysis may depend less on a strong theory of idiolect than on the simple detection of consistency and the determination of distinctiveness.” (2020:559). This thesis aims to explore how authorship as a concrete matter can communicate and reveal the theoretical notion of the idiolect. On another front, researchers in computer sciences, specifically in machine learning, tackle authorship from a standpoint of accuracy. Corpus linguistics is a common ground where both fields meet, as linguists and computer scientists use corpora to test their methods. Linguists approach it from a stylistic, qualitative point of view because it provides them with the explicability a courtroom would require when presenting their analysis report as expert witnesses. Computational scientists, on the other hand, focus on the quantitative, statistical aspect of analysis which generated, until recently, black box tools that do not necessarily show the tool’s trail of analysis.

This study investigates a specialised corpus composed of 58,005 tweets and 748,348 words by thirteen authors from the central region of Najd, Saudi Arabia with a reference corpus representing the relevant Najdi population (Larner, 2014; Heydon, 2018). Using Turell’s (2010) concept of *idiolectal style* as theoretical groundwork, this corpus-based research explores the authors’ styles. The first part takes a qualitative, stylistic angle to explore the same feature set to reveal the lexico-grammatical patterns of each author. The second part of the analysis explores the authors’ online identity, both their collective and individual themes, through the lens of *the identity approach* (Bucholtz and Hall, 2005) and Herring’s (2007) *Faceted Classification Scheme*. The third part of the analysis addresses the computational aspect of the synergy by exploring WEKA, the machine learning tool, and its potential in authorship attribution using a stylistic feature set in two varieties: Modern Standard Arabic and Najdi Arabic. The final part tests the proposed methodology in a case simulation with hypothetical ‘suspected’ authors. The qualitative findings show that the Najdi Arabic feature set has potential as markers of idiolectal style. This is further supported by the quantitative findings provided by the computational analysis, where the accuracy rates are promising for short texts. The qualitative analysis guided by the approach also reveals that the authors' identity is revealed and developed through the stances they take throughout interaction. The final empirical stage of the investigation shows that the triangulated methodology proposed in this research shows promising results and thus can be considered as a method in forensic authorship casework.

Table of Contents

List of Tables

List of Figures

Preface

﴿إِنَّ اللَّهَ يَأْمُرُكُمْ أَنْ تُؤَدُّوا الْأَمَانَاتِ إِلَى أَهْلِهَا وَإِذَا حَكَمْتُمْ بَيْنَ النَّاسِ أَنْ تَحْكُمُوا بِالْعَدْلِ إِنَّ اللَّهَ نِعِمَّا يَعِظُكُمْ بِهِ إِنَّ اللَّهَ كَانَ سَمِيعًا بَصِيرًا﴾ سورة النساء, آية 58

Surely, Allah commands you to make over the trusts (such as the affairs of the state) to those who are competent to it, and that when you judge between the people you should judge with justice. That which Allah exhorts you to do is best indeed. Allah is All-Hearing, All-Seeing. Surat An Nisa'- 58

The epigraph emphasises the importance of people achieving justice not only by practising it in court but also by paying their dues to one another (Mujahid, 2000). The emphasis in the verse reflects the challenges that face the legal system and the forensic linguist when being a participant in the system: how to "improve the delivery of justice" (Grant and MacLeod, 2020: 180)? What valid and reliable methods can we apply to do that? Forensic linguists will always aim to assist in the administration of justice.

One of the facets of open justice in the Saudi court is the hearings are conducted in open court unless the judge, on its own motion, or the motion of one of the parties, rules that it will be held in a closed session to maintain order, observe public morality or protect the privacy of the family (Article 64, Civil Procedure Law). Precedents are published and available to the public, in which the facts are included, though they will have the names of people and entities removed for their protection (laws.boe.gov.sa).

Chapter 1
LANGUAGE, AUTHORSHIP AND THE LAW

1.1 Introduction

There is an interesting relationship between language and the law. For instance, one way to refer to the field of forensic linguistics is the term *language and the law*. Language is a means to communicate and execute the law, on one hand, and evidence to be examined by the law, on the other hand. Forensic linguistics investigates this relationship from its three different positions: 'the written legal texts, the spoken legal practices and the provision of evidence for criminal and civil investigations' (Coulthard et al., 2011: 529). This thesis is concerned with forensic authorship attribution: language as evidence. For a forensic linguist, to attribute the authorship of a text is to give an opinion on the identity of the author of an anonymous or disputed text using linguistic features as markers. This task, unfortunately, is not as straightforward as one assumes it to be. The following quote by Coulthard and Johnson (2010) paints a clear picture of the complexity of authorship attribution in a forensic context:

> "[T]he ownership of a text is easy to dispute, but difficult to settle, because individual style is difficult to pin down and… the legal profession relies on certainties or, at the very least, being sure. In criminal trials, juries have to base decisions on being sure beyond 'reasonable doubt', in civil, on the 'balance of probabilities, and, to help them reach their decision, it is the expert's job to give an opinion that is neither inflating nor wavering or indecisive…"
>
> (Coulthard and Johnson, 2010: 5).

The law means to discover the truth, thereby achieving justice, but the nature of the litigation system can make the process somewhat complicated (Solan, 2020). The legal system in British and American courts accepts the testimony of an expert linguist witness, who is usually a qualified academic scholar with a background of research in the field. The Saudi court also can assign an expert witness if the judge sees the need for it, though court records do not show the assignment of linguists (https://data.gov.sa/). Each system approaches the expert testimony differently, which is discussed in more detail in 1.3, but one of the objectives that they aim to achieve is for experts to report what the expert finds and to give an expert opinion. In the American court, the demand for the truth is deducted by valid and reliable methods. Litigators on both ends aim to present the truth, where usually each party presents a version of it that is in their favour (Post, 1987; Solan, 2012; 2020). Expert witnesses are asked to be cooperative by the litigators that work with them, but there is a chance that by doing so they might be attacked by the opposite party, either by using their testimony against them or by questioning their qualifications. A cautious measure some expert witnesses do is to present a concise report that shows their findings but not so elaborately that the opposite litigating team can find loopholes (Solan, 2020). However, there is risk in taking that measure as Sanders (2007: 1558 as cited in Solan, 2020: 357) notes:

> [W]hen [experts] do fail to present adequate justification for a belief, often it is not because they fail to present the best case for a position but that they fail to tell the ‘whole truth’ about their belief and present with equal force the evidence for and against it.

The Civil Procedure Rules that govern the practice of the expert witness reflects that concern clearly (Solan, 2020). One of the challenges is for the scholars to report findings without a tone of uncertainty, because academic research is based on the cycle of proposing hypotheses, testing them, then either revising and explaining or discarding and testing new ones. The British court is more concerned with the expert in terms of their qualifications and sense of integrity; that the expert witness guarantees the findings in their testimony is bias-free. Similar to the British approach, the Saudi court also assigns an expert based on their qualifications and integrity. The expert witness is to report the findings of their analysis bias-free and without favour to any party. Thus far, it appears that academia shows more tolerance towards uncertainty than the legal system does. I revisit the legal system and variations in the admission of evidence and expert witness testimony in 1.3.

1.2 Statement of the problem

Authorship research can be achieved through a number of approaches. Statistical/stylistic ones highlight the significance of idiolect and idiolectal style (e.g. Grant and Baker, 2001; Zheng et al., 2006; Grant, 2007; Turell, 2010; Johnson and Wright, 2014) and computationally focused approaches (e.g. Juola, 2008; Koppel et al., 2009; 2010; 2012; Ebrahimpour et al., 2013; Seroussi et al., 2014; Rocha et al., 2017) highlight the significance of tools that are fast and score high accuracy rates. Some analyse evidence based on a sociolinguistic/stylistic approach (McMenamin, 2002; Coulthard and Johnson, 2007; Turell and Gavalda, 2012); others opt for an idiolect-free statistical approach (Grant, 2013; Ishihara, 2017). Using corpus linguistics as a tool to identify authorship is relatively novel and claimed to be effective (Grant and Baker, 2001; McMenamin, 2002; Coulthard, 2004; Johnson and Wright, 2007, 2014; Grant, 2007). There is a substantial amount of literature that addresses authorship attribution, but not with a focus on sociolinguistic implications (Turell, 2010). In the context of cybercrime, there is a need to investigate the construction of online identity using idiolectal style (Hardaker, 2016).

A body of literature has been published in authorship research in Saudi Arabia that focuses on computational approaches (Alruily, 2012; Althenayan and Menai, 2014; Al-Tuwairesh et al., 2015, 2018; Assiri et al., 2016) while linguistic and stylistic approaches fall short (Maslouh, 1982; Al Mufti, 2002; Awad, 2014). This calls for a need to contribute to the field of forensic linguistics in general and authorship analysis in Arabic in particular. Some of these computational authorship studies address cybercrime and cyber-terrorism in the aim of detecting it and counterattack. This is raised by the fact that social media platforms such as Twitter are heavily populated by users who sometimes abuse such mediums. Saudis are responsible for 30% of the Arabic tweets posted on this platform (Salim, 2017). Simultaneously, there are efforts to fight cybercrime and issue regulations that incriminate hate speech and offensive language published online.

Until recent times, there has been a lack of understanding in Saudi Arabia as to what constitutes digital offenders' behaviour and the best practice to deal with cybercrime in compatibility with Sharia law. The Council of Ministers issued the Anti-Cyber Crime Law, and it was approved by Royal Decree in 8/3/1428 Hijri (22nd of March, 2007) but the Act is only a step in the direction of having a law that defines the interpretation of cybercrime and produce clear and defined methods to examine such crimes (Elguindy, 2012; Alanazi and Jones, 2017). The following section discusses the law further, as well as the collection of evidence.

1.3 Admission of evidence and ethics in expert witness testimony (UK, US, and Saudi Arabia)

Understanding the workings of a legal system is an important area for a forensic linguist to learn, especially in terms of the court's policies to admit evidence and give their testimony as an expert witness. This section surveys how three legal systems operate, the United Kingdom, the United States of America, and Saudi Arabia. The first two countries have built a history of civil and criminal cases where forensic linguists provide testimony as expert witnesses for language evidence such as the dispute over the prefix Mc- as a trademark (Lentine and Shuy, 1990), the Unabomber (Fitzgerald, 2004), the Derek Bentley case (Coulthard and Johnson, 2007), the Jenny Nicholl case (Grant, 2012), the Bixby letter (Grieve et al., 2019) among others. The previous cases exhibit instances where linguistic evidence supported other physical evidence to steer the case in favour of one party over the other. Therefore, both have generated rules that govern that process. Since the sample and data of this study are Saudi, it is relevant to introduce Saudi law and address the Saudi court regulations as far as evidence admission and testimony is concerned.

The courts in the United Kingdom are concerned with the qualification of the expert witness and ensuring that the evidence or testimony presented before them is bias-free. Therefore, Civil Procedure Rules clearly dictate the following:

> 1- It is the duty of an expert to help the Court on matters within his expertise.
> 2- This duty overrides any obligation to the person from whom he has received instructions or by whom he is paid.
>
> (Civil Procedure Rule 35.3)

Recently, the UK is making a shift towards an approach that is being followed in the US, which is to pay more attention to the validity and reliability of the methodology the expert witness follows to produce their testimony.

The legal system in the United States emphasises using a methodology that is valid and replicable. The Federal Rules of Evidence first appeared in 1975 and went through some amendments after the three civil cases known as the Daubert trilogy (see Solan, 2020). The rules allow expert testimony should it ‘assist the trier of fact to understand the evidence or to determine a fact in issue,’ and if:

> 1- The testimony is based upon sufficient facts or data,
> 2- The testimony is the product of reliable principles and methods, and
> 3- The witness has applied the principles and methods reliably to the facts of the case.
>
> (The Federal Rules of Evidence, Rule 702)

Saudi Arabia applies the Sharia law derived from its main source, the Qur’an, the holy book that addresses all matters of human life whether the relationship between man and God or between man and society. It sets the laws and principles that concern ethics, jurisprudence, social relations, politics, justice, law, morality, and trade and commerce (Alfazie, 2015; Ansary, 2015; Alanazi and Jones, 2017). The second source is Sunnah, which stands for the methods and practices established by the Prophet which have been witnessed and reported by his companions and passed on by the consensus for generations (Farid, 2017; Alanazi and Jones, 2017). Other instances can be sorted by consensus of opinions, or Ijma, among Sharia scholars; this is also a principle in the Sunnah that states if all Muslims agree on a matter, then it cannot be wrong (Alanazi and Jones, 2017). In instances that both sources do not address or account for, they are used as analogical tools to find a resolution for the matter. Modern narcotics, for example, are prohibited based on the analogical reasoning that they cause the same harm as narcotic substances (Alfazie, 2015). In terms of expert testimony, the Law of Procedure before Sharia Courts, Article 124, allows appointing expert(s) when required. The court assigns the task to the expert, the amount of time given to prepare and present the report, and the time of the trial hearing. Also, the court states that the testimony given is used as guidance, not as facts.

When it comes to cybercrime, a form of offence that is recently enabled by technology and social media can be approached from the Qur'anic principle that calls for the protection of privacy. The Anti-Cybercrime law was a step in that direction, but there are several issues that need to be addressed. Atalla (2010) reports that cybercrime is a new environment that requires one to describe what is criminal behaviour in a digital context and the best practice to counteract that is compatible with Sharia law. Similarly, Elguindy (2012) says that the Anti-Cybercrime law lacks a thorough description and definition of what constitutes cybercrime and the specific methodology required when examining such crimes. Recent literature also reports that while policies are set and effective in the case of a computer incident, there is a gap when it comes to guidelines for digital forensic investigation (Alanazi et al., 2018). On one hand, the Saudi courts recognise textual data as they state:

> Electronic data is defined as data that has a number of characteristics including text, symbols, images, sounds, drawings or other electronic data.
>
> (The Law of Procedure before Sharia Courts, Article 1.11, Alanazi et al.'s translation)

On the other hand, it does not give clear instructions regarding examining the evidence and the best practice to analyse it and present its findings before the court. Interestingly, one of the reasons digital types of evidence are sometimes excluded by the court is due to the method of their seizure and analysis (Al Beshri, 2008; Alanazi et al., 2018). Moreover, the court does not allocate much weight to digital evidence (i.e., electronic data) which drives away investigators and barristers from using it. This is partly because digital evidence can be changed and altered, unlike more tangible types such as evidence of violence or blood. In the instances where digital evidence is permitted by the court, it is accepted as confession or *Iqrar* in Arabic. However, it is not regarded as clear or strong evidence or *Bayyinah* (Alanazi et al., 2018).

1.4 Aims and objectives of the study

While the fact that forensic authorship attribution is a relatively new field, there is a body of literature that examines data in English, Chinese, Portuguese among other languages. There is much to be explored as far as the Arabic language is concerned in its standard and dialectal varieties. The first objective this research aims to achieve concerns forensic authorship attribution and contributing to the field by exploring data of an Arabic dialect spoken in the central region of Saudi Arabia known as Najdi Arabic (explored in more detail in section 1.7). More specifically, this is a corpus-based study of the *idiolectal style* of 13 authors located in Riyadh, the capital of Saudi Arabia, to capture their use of the dialect-specific features in Najdi Arabic. To have a clearer understanding of how distinct each author’s *idiolectal style* is, the specialised corpus is examined in comparison to a reference corpus that represents the relevant population. Therefore, the first question this doctoral research proposes is:

1) Can the Najdi Arabic Specialised Corpus of Tweets (NASCoT) help in identifying the idiolectal style of thirteen Najdi authors when compared to the relevant population?

This research also aims to investigate authorship attribution and online identity. The NASCoT is an extensive collection of the authors' tweets to examine their style and its connections to online identity, which leads to the second question:

2) How does the idiolectal style of these authors contribute to the construction of their online identities?

Moreover, authorship attribution has been a subject of interest by linguists and computer scientists that generated a body of research in both fields. This research aims to create a synergy between these fields by proposing a triangulated method that combines corpus-based methods, stylistics, computer-mediated discourse, and machine learning tools. Waikato Environment for Knowledge Analysis (WEKA), the machine learning tool used in this research, is developed at the University of Waikato, New Zealand written in Java script (Witten et al., 2016). It is a collection of machine learning algorithms that perform data mining tasks, and its tools can achieve pre-processing, classification, clustering, and can develop new machine learning schemes. To explore the potential of WEKA in authorship research and due to the size of the NASCoT, I propose the following question:

3) What is the potential of WEKA to attribute authorship using the lexcio-grammatical feature set?

Also, there are other elements that are just as important as size to achieve the most accurate results: classifiers and parameters. Classifiers are the algorithms that recognise patterns in data and group them into pre-set categories; parameters are the values a machine algorithm uses to map between the data and the pre-set categories. Consequently, I propose these two questions:

4) Which classifier and parameter can accurately identify authorship using the NASCoT corpus?

Lastly, this research explores the potential of implementing the proposed methods in casework where language can be used as forensic evidence. The final two questions this doctoral research aims derive from that:

5) How likely is it for the methodology and the Najdi Arabic feature set proposed in this research to identify authorship in forensic casework?

The first two questions are addressed in chapters 5 and 6 respectively, the former covers corpus-based stylistic analysis and the latter focuses on CMDA and online identity. Questions 3 and 4 are covered in the computational analysis in Chapter 7, and the empirical experiment in Chapter 8 explores the final question.

1.5 A background of Najdi Arabic

Language represents a myriad of things: history, geography, and culture to name a few. This section introduces a brief and concise history and geography of Najd then transitions to the sociolinguistic implications of the Najdi Arabic variety and concludes with a survey of the sociolinguistic studies that address the dialect.

Located in the heart of the Arabian Peninsula is the central rocky plateau of *Najd* (which is Arabic for *high surface*). As illustrated in Figure 1.1, *Najd* (highlighted within the red rectangle) is cornered by three regional deserts south and east: *An Nafud*, *Ad Dahna*, *Ar Rub'a Al Kahli* and the mountainous terrain of *Jabal Shammar* to the north (Ingham, 1994). These central and northern parts of the peninsula are the homeland of Arabic language and its speakers who had infrequent contact and exposure to other languages and dialects due to its remote terrain (Ingham, 1994).

This resulted in a dialect that retains many Classical Arabic (CA) features, mainly referred to as Najdi Arabic (NA) (Watson, 2002). This dialect has been the subject of many European linguists’ attention in the 20th and 21st centuries (e.g. Johnstone, 1967; Prochazka, 1988; Ingham, 1994; Versteegh, 2014) as it preserved most of its features because of its geographic location. Nowadays the dialect is the subject of interest for local sociolinguists who investigate the shift of its features in time (e.g., Al-Rojaie, 2013; Alqahtani, 2015; Alaiyed and Abdullah, 2018) and report its speakers’ attitudes towards their dialect and other Saudi dialects (Al-Essa, 2009; Aldosaree, 2016; Alhazmi and Alfalig, 2022). Linguistically, NA nowadays is a dialect spoken by nearly 8 million inhabitants that is a distinct variety that represents its speakers of Nomadic Bedouins and sedentary community (Lewis, 2013). In terms of practice, the Najdi community are a representation of Ferguson’s (1959) perception of a diglossic community, where NA is the low variety that is mainly used in daily spoken communication and Modern Standard Arabic (MSA) is used in formal settings (e.g. academic lectures, news press, religious sermons, courtrooms, etc.). An elaborate demonstration of the linguistic features explored in this research will take place in Chapter 4. Frey et al. (2022) report that one of the conventions of social media platforms is using low varieties, or dialects, in written communications. This is observed in German (Frey, Glaznieks, and Glück, 2022) as well as English (see Einstein et al., 2014) and NA is no exception to this case being a low variety that is mostly spoken and recently became written on social media mediums (Frey, Glaznieks, and Glück, 2022). That being said, it is worth indicating that the Najdi variety is spoken by the royal family Al Saud, which assigns the dialect with some prominence and social prestige (Al-Essa, 2009; Ismail, 2012; Alhazmi and Alfalig, 2022); a notion that will be discussed further in Chapters 5 and 6.

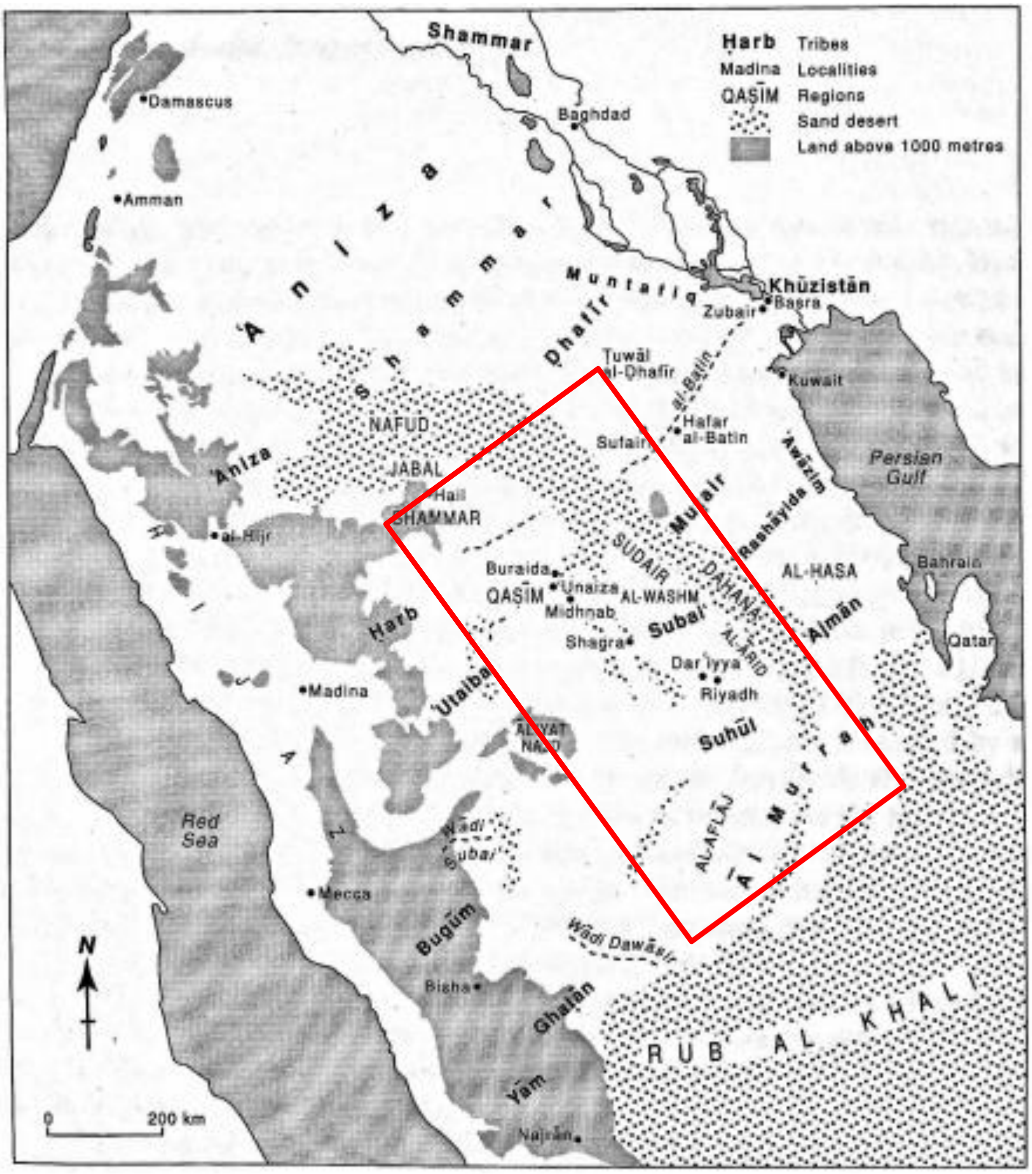


Figure 1.1: Map of Najd (Ingham, 1996)

A substantial amount of work has been produced to provide a comprehensive description of the variety spoken in the region (Abboud 1964, 1978; Badawi, 1965; Lehn, 1967; Johnstone, 1967; AlSweel, 1992; Ingham, 1996; Alothman, 2012; Binturki, 2015), which will be discussed with more detail in Chapter 3. An equally substantial amount of research has been also produced that investigates cybercrime in Arabic (e.g., Abbasi and Chen, 2005, 2008; Alfaifi and Atwell, 2016; Altakrori et al., 2018) and authorship attribution from a computational angle (e.g. Ouamour and Sayoud, 2012; Althenayan and Menai, 2014). In terms of forensic authorship analysis, there have been calls for carrying out research that examines Arabic data (Mansour, 2013; Omar and Aldawsari, 2019; Abdellah, 2019). To the best of my knowledge, there are no studies that examine NA in a forensic authorship context.

1.6 Thesis outline

Chapter 2 of this thesis starts with a discussion about the theory of idiolect and its connection to forensic authorship attribution. It reports the evolution of the theory by different scholars throughout time, and its current interpretation by forensic linguists. It also addresses the main approaches of the theory, cognitive and stylistic, and methods to measure it in empirical studies and casework using stylistics and stylometry. Then it transitions to a survey of the literature produced by linguists in authorship attribution and the methods devised to tackle it. Following that is a computational sciences' perspective on authorship attribution and the tools used to investigate it. This eventually leads the way to the common ground both fields share, which are corpus linguistics and stylometry as methodologies. At this point, the proposed synergy is introduced which would combine all four fields: stylistics, corpus linguistics, computer-mediated discourse, and computational linguistics.

Chapter 3 is a theoretical backgrounding of the identity and how it is connected to the idiolect theory and the notion of authorship. The first section is a brief survey of the identity from a third-wave sociolinguistic perspective. The following section discusses the concepts introduced in light of this thesis' research and its connections to the analytical chapters.

Chapter 4 reveals the corpus design of the Najdi Arabic Specialised Corpus of Tweets, and the reference corpus named the Najdi Arabic Corpus of Tweets. It explains the methodological logic behind each corpus in the context of forensic authorship attribution. Then it introduces the sample selected for the research with a brief background that reflects the criteria and ethical considerations that lie behind their selection. Also, it walks through the data collection process and the formatting of the data in order to be read by different computational tools. Afterwards is a discussion of the methodological approaches implemented in this research which demonstrates how each element in the synergy contributes to the research, followed by a section that addresses the stylistic features examined in this study. Finally, it reports the pilot study that was conducted at an earlier stage in this research and how its findings produced the current form of methodology.

The first analysis chapter is Chapter 5, which explores idiolect from a linguistic, stylistic perspective. It first explains how the data is represented, the corpus linguistics tools used, and the theoretical approaches that are carried out in this analysis. Concordances, collocates, and patterns are measured and indicate to implications of the idiolectal style each author has. The findings of this chapter support the computational analysis presented in Chapter 7 as far as lexico-grammatical features are concerned, which leaves room for investigating online identity construction in the next chapter.

Chapter 6 is the second aspect of analysis that addresses the authors’ online identity construction. This is investigated using the identity approach and the Faceted Classification Scheme which, combined, can help read the online identity the authors project of themselves. The conclusion shows how the qualitative analysis presented can support and complement the stylistic analyses performed in the earlier chapter, and both chapters in turn are supported by quantitative, computational analysis in the following chapter.

Chapter 7 is the final analytical chapter in the thesis and it explores computational tools in solving authorship attribution problems. Using the same tools presented earlier, the corpora and the stylistic features, this chapter investigates authorship attribution using WEKA. By testing a range of seven classifiers, the study discovers the optimum combination of classifier and parameter to obtain the most accurate results. This leads to the final stage that is to test the methodology in an empirical experiment.

Chapter 8 breaks into two parts, the first is an empirical experiment to test the methodology proposed in this research. By presenting a new set of data, this case simulation tests the methodology proposed in this research and whether it can be considered as an adequate method for a real forensic authorship problem. The findings of the case simulation pave the way for a general discussion that addresses all three studies and their connections to related literature.

The final chapter concludes this thesis by presenting the triangulated approach and a summary of the results presented earlier. It also addresses the theoretical and methodological contributions presented in this research and future recommendations for research in forensic authorship attribution.

Chapter 2
The Idiolect: Theoretical Perspectives

> "Individual variation exists, but it can properly be appraised only with reference to the social norms." (Sapir, 1927:903)

2.1 Introduction

Sapir was one of the earliest linguists who introduced and established the idea of individual variation in language, and authorship is one of the materialisations of that idea. In this epigraph, not only does he acknowledge that language can take multiple unique shapes, but he also suggests that these shapes can be recognised and attributed. According to Juola (2008: 287), "authorship attribution, broadly defined, is one of the oldest and one of the newest problems in information retrieval". It aims to identify or attribute one or more disputed texts to a single or multiple author(s), either from a closed set or an open one (Stamatatos et al., 1999; Koppel et al., 2009). In 2016, a report was published by the President's Council of Advisors on Science and Technology (PCAST) at the United States that states "[f]oundational validity for a forensic-science method requires that it be shown, based on empirical studies, to be repeatable, reproducible, and accurate, at levels that have been measured and are appropriate to the intended application." (PCAST, 2016: 47). While forensic authorship attribution was not addressed in this report per se, the statement represents what researchers in the field aim to achieve from both disciplines: linguistics and computing.

This chapter starts with the theoretical background that is the cornerstone of this research, the theory of idiolect. It presents the linguist's take on the theory, then discusses its two main approaches: cognitive and stylistic. It also talks about the empirical aspect of the theory and the methods used to measure it. The section that follows is a concise survey of studies in authorship attribution, in linguistics and computer sciences with a focus on studies that address short messages and Arabic data; this aims to demonstrate the contrast in approaches between fields. Section 2.4 presents stylometry and statistics, an area where both fields meet when dealing with authorship attribution. Figure 2.1 is a representation of the current situation of authorship research and how it is approached by each field; the figure is also an illustration representing this chapter's structure. It shows the current picture of authorship attribution, where linguistic approaches employ stylometry and statistical methods that account for elements such as the author's idiolect or style and identity. The style markers become stylometric features by giving them numeric values, which will be explained in 2.4. In the linguistic authorship research, stylometry and statistics are used to test the linguistic theory, or lack of it, and the explicability of the findings. On the other front, computational approaches also use stylometry and statistics with a slight variation compared to linguistics. Their focus when conducting research is to assess the performance of the algorithms, classifiers, and an overall model; stylometric features are the variables that enable us to accomplish this task.

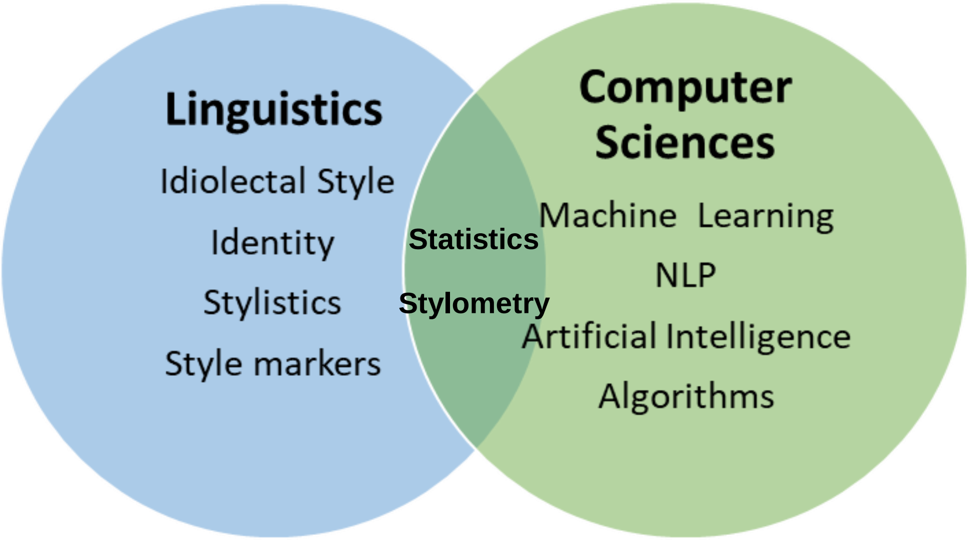


Figure 2.1: Approaches currently shared by both fields.

2.2 The idiolect

Idiolect is an integral element in forensic authorship attribution literature and in this doctoral research; hence, I attempt to define idiolect and what the term stands for. Two centuries ago, Paul (1888) called to shift our attention from the language of the community to that of the individual as an interesting element of language change. Decades later Sapir (1927) and Bloch (1948) endorsed Paul's proposal and called for differentiating group and individual levels of linguistic analysis. The term idiolect was coined by Bloch (1948) and is a blend of the Greek words 'idio' and 'lect' to better reflect the concept of *personal language variety*. Bloch (1948) defines idiolect as an individual variety that consists of a uniquely patterned set of linguistic characteristics and called for us to study this notion further. On one hand, sociolinguists Weinrich, Labov and Herzog (1968) steered clear from linguistic individualism in their theory of language and focused on social forces as agents of language change (Kuhl, 2003). On the other hand, other sociolinguists such as Evans (2012; 2016; 2018) explored idiolect and the individual variation of the royal British court in a historical context to report its role in the variation of the language and a stimulus to its change.

Bloch (1948) coined the term idiolect and Sapir (1927) recognised the connections between speech and personality where vocabulary and style are the most explicit signs that characterise an idiolect. Since then, researchers, especially sociolinguists, have explained their various interpretations of the term (Larner, 2014). Similar to Sapir's approach, Coulthard (2004) pays attention not only to the choices of words an individual makes but also the uniqueness and individuality of these choices or *idiolectal co-selection*. He explains that the process of creating a text is not as random and open as one might assume, but rather it is controlled by linguistic habits where features are co-selected thus creating a unique pattern. Hockett (1958: 321) describes it as 'the totality of speech habits of a single person at a given time'. This definition, however, brings up two challenges: the first is that one must record the entirety of someone's linguistic output in order to be able to recognise their habits, which is unrealistic; and the second is, even if this was achieved, these habits will change over time.

McMenamin's (2003) understanding of idiolect is more inclusive and accounts for cognition and extra-linguistic elements that contribute to shaping the unique linguistic habits an individual develops. Most authorship attribution methods, whether based on linguistic theories or computational methods, rely on the use of style markers (MacLeod and Grant, 2012). McMenamin describes them as 'the observable result of the habitual and usually unconscious choices an author makes in the process of writing' (2010: 488). These choices would be subject to analysis or as he puts it 'the scientific interpretation of style-markers as observed, described, and analysed in the language of groups and individuals' (2010: 488). He continues to classify these choices as a choice made from a range of options or a deviation from the norm, where the norm is what is conventionally used by the community of practice. Nonetheless, the discussion of style markers raises the debate of how to distinguish conscious and unconscious habits when working on an authorship attribution problem. Another debate is the overlap between the terms *idiolect* and *style* and what each one stands for. Some linguists emphasise that idiolect stands for written language as much as the spoken one (e.g. Coulthard, 2004; Larner, 2014), while others prefer to refer to written language as style (e.g. Sapir, 1927; Kredens, 2002). Turell's (2010) take on idiolect shows a comprehensive perspective of linguistic individuality in written and spoken contexts, hence the term *idiolectal style*:

> hav[ing] to do primarily, not with what system of language/dialect an individual has, but with a) how this system, shared by lots of people, is used in a distinctive way by a particular individual; b) the speaker/writer's production, which appears to be 'individual' and 'unique' (Coulthard, 2004) and also c) Halliday's (1989) proposal of 'options' and 'selections' from these options. (Turell, 2010: 7)

Crystal (2011) states that each individual has their own language system that generates their unique idiolect, One of the debates linguists have is the overlap between the terms *idiolect* and *style* and what each one stands for. This debate is about written and spoken language, whether they are two different categories and, if they are, then how to address each category in the context of authorship.

Moreover, researchers in time have generated two interpretations of the theory of idiolect: one is cognitive, and the other is stylistic (Grant, 2020). Cognitive theorists propose that language competence is what determines the individual's language production, with reference to Chomsky's notions of competence and performance. Their approach for forensic authorship analysis would be to measure the author's cognitive capacity such as syntactic complexity or their mental lexicon. This has been implemented in the work of computational forensic linguists who examined individual linguistic features in a range of authorship attributes such as word frequency distributions (e.g., Baayen, 2001; Grant, 2007) and stylistic structures (e.g., Chaski, 2001; Spassova and Grant, 2008). The cognitive approach assumes that stability of competence generates individual consistency in the linguistic style that endures relatively well over time and is context-free; it provides its explanation of idiolect based on this ground (Grant and MacLeod, 2020).

The stylistic approach to idiolect, on the other hand, is more concerned with understanding the linguistic variation between individuals and less concerned with the theory of idiolect. To express that shift, some linguists propose the term *the linguistic individual* (Johnstone, 1996; 2017a; Kredens, 2002); while others propose *idiolectal style* (Turell, 2010). Stylistic approaches to idiolect are more interested in variation across authors as opposed to their cognitive counterparts that focus on explaining individual consistency across different texts and genres (Grant and MacLeod, 2020). The discussion of style and variation is further developed in Chapter 3.

Another debate is concerned with the implementation of the theory in empirical ventures. Geroge (1990) claims that the notion of idiolect appears sound in theory, but it can be confusing when applied to detect whether an individual's idiolect is driven by their linguistic individuality or inspired by the communal language. Sapir anticipated such hindrance as he noted that "[i]t would be a very complicated problem to disentangle the social and individual determinants of style, but it is a theoretically possible one." (1927: 904). Garcia-Barrero et al. (2013) tackled this issue by saying it is highly unlikely that an idiolect is unique but rather similar or distant to another idiolect. They propose two fundamental notions that support their argument; one is a) language both in written and spoken forms can reveal the socio-individual and socio-collective traits, and the other is b) the idiosyncrasy that lies in the way an individual uses the language.

Moreover, the theory of idiolect is recognised in many authorship attribution studies (e.g. Grant and Baker, 2001; Grant, 2007; Feiguina and Hirst, 2007; Turell, 2010; Johnson and Wright, 2014), some of which use the term idiolectal style to indicate its measurability. In the context of forensic text comparison, idiolectal style can be represented in the linguistic decisions an individual makes while using their repertoire of a particular language as a source (Nolan, 1994; McMenamin, 2001; Coulthard and Johnson, 2007). It represents the accumulation of cultural, geographic, religious, and social experiences. However, the approach focuses on the individual level, not a group or a community of language users; that is to avoid an overlap with Labov's (1972) approaches on style, which attributes linguistic choices to social stratification in a communal fashion (Coupland, 2007). Years after George's (1990) statement, Kredens (2002; 2006), Louwerse (2004), and Coulthard (2004) continue to raise the issue that the theory of idiolect lacks empirical examination, a call that is supported also by Larner (2014) and Wright (2017). Aside from the debate over the validity of the theory of idiolect and what it entails, research in forensic authorship analysis continues to adopt it, in acknowledgement of Paul's and Sapir's call for examining the language of the individual. Also, research has met some of Bloch's goals in explaining what the term idiolect is in theory; a prominent example is Coulthard's conclusion that:

> "… the concepts of idiolect and uniqueness of utterances are robust and provide a basis for answering certain questions about authorship with a high degree of confidence." (2004: 445).

Corpus linguistics was and continues to be the tool that made the implementation and empirical investigation of the idiolect and its theoretical approaches possible. It enabled researchers to collect and analyse individual language datasets and not only reference corpora (Coulthard, 2004; Kredens and Coulthard, 2012; Biber, 2015; Wright, 2017). Both corpus-based and computational research has described the individual variation in language production using word frequency (Holmes, 1994; 1998; Baayen, 2001; Grant, 2007; Garcia-Barrero et al., 2013), syntactic structures (Chaski, 2001; Spassova and Grant, 2008; Sidorov et al., 2014; Zhang et al., 2018), and some explored other markers of authorship like characters (Belvisi et al., 2020; Marko, 2021), and n-grams (Johnson and Wright, 2014; Wright, 2017). I discuss how corpus linguistics is implemented in this research theoretically and methodologically in Chapter 4.

Furthermore, it was noted earlier in Chapter 1 that the dialects across regions in Saudi Arabia have been a subject of interest in sociolinguistic research, and the Najdi dialect is no exception (e.g. Johnstone, 1967; Abboud, 1979; Ingham, 1994; 1996; Prochazka, 1988; Miller, 2004; Madini and de Nooy, 2012; Binturki, 2015; Almutairi, 2021). These are all attempts to address the dialect as a communal language. Some sociolinguistic studies address the more generic Saudi dialect instead of focusing on a specific region, especially when investigating online communication (e.g., Al-Tuwairesh et al., 2015; 2018; Almutairi, 2021). Nonetheless, the researchers acknowledge the dialectal variation across the country that would extend from phonetic to syntactic features and account for it by marking different dialects (i.e., Najdi, Hijazi, or Southern dialects) at the data collection and analysis stages (Al-Tuwairesh et al., 2018). The focus on idiolect and individual variation and its relationship to the communal language in Saudi Arabia is yet to be explored.

After discussing the literature related to the idiolect, I introduce my critical view to approach the data in this research. Sociolinguists identify the idiolect with a focus on individuality in style and stancetaking, which will be discussed further in Chapter 3. My personal perception of the idiolect is an adaptation of Turell’s (2010) interpretation of idiolectal style. According to Turell, idiolectal style is the accumulation of variables: first is the acquisition of the variety/varieties, second is the individuality in using the constituents of said varieties, and third is the options authors make in choosing the stylistic sets to communicate their stance, which formulate their repertoire and can change over time. Additionally, I propose that idiolectal style is perceived as an act of agency, which is represented in the idiolectal selections an author makes. An author’s style is not a reflection of their age, education, or vocation, or as Johnstone (2009) puts it their ‘social identity’, it is rather that of a ‘personal identity’ that reveals their beliefs, values, and attitudes. Moreover, it is a performance that surpasses the macro-level sociological categories. The selection of a variety, its linguistic constituents, as well as the stances that express an author’s attitudes and values are all attributes of idiolectal style. This resonates with Johnstone’s (1996) linguistic individual, as these selections reconfigure a variety that is shared by a community in a way that makes it individual and attributive to a single person. The discussion of the theory of idiolect will continue in an empirical context in the analysis chapters 5 and 6.

2.3 Same toolbox, different approaches

Authorship problems are a matter of interest to ‘both groups [as they] seek to make accurate predictions about uncertainties related to textual data’ (Rocha et al., 2016: 2). To linguists, authorship attribution is a problem of classifying the texts of two or more authors who share linguistic similarities, a perception shared by computer scientists. When linguists work on an authorship attribution problem, they account for the variation of linguistic production of the author in different genres as well as their linguistic development over time. One way to achieve that is stylometry: the quantification of linguistic style, which was introduced by de Morgan (1851) and his work will be discussed in section 2.4. Computer scientists, for the past two decades, have also employed advanced machine learning techniques that use stylometric features and statistical pattern recognition (e.g., Juola, 2006; Bishop and Nasrabadi, 2006; Koppel et al., 2009; and Stamatatos, 2009).

2.3.1 Linguistic approaches

Ownership of a text has always been a matter of significance and one of the most common disputes (Olsson, 2014). One of the earliest pieces of authorship work was a letter by de Morgan responding to a biblical scholar in 1851 in which he indicated that word length can be a feature to attribute authorship (Baker and Grant, 2001; Grzybek, 2006). Another case in the late nineteenth century surrounded the authorship of a literary text attributed to either Bacon, Marlowe, or Shakespeare, investigated by Mendenhall in 1887, and in 1888, Mascol examined the gospels of the New Testament (Koppel et al., 2009). Modern era authorship attribution was established with the Federalist Papers' study by Mosteller and Wallace (1963), which supported the status of stylistics as a method in authorship research, a field that will be discussed further ahead. Since then, a substantial number of studies has been produced that explore forensic authorship attribution; Nini and Grant (2013) propose that cognitive and stylistic approaches to the idiolect should complement each other rather than alternate with each other.

There are authorship studies that paved the way for linguists to build on and expand the field, especially in the forensic context. Coulthard is one of the linguists that bridged the gap between forensic linguists and the law. He provided analysis for a number of cases such as the Bridgewater case (Coulthard, 2004) where he proposes that authorship can be unique and representative of one's idiolect. His findings in the Nicholl-Hodgson and the Jones-Campbell cases further supported his proposition that authorship and idiolect are connected, and that stylistic analysis can detect the distinctiveness of both (Grant, 2020). Coulthard's analyses in both cases helped in steering the direction of police investigation and convict the defendants of murder. Also, he emphasised the importance of context and manipulation in police investigations in the cases of the Birmingham Six (Coulthard, 1995) and the Derek Bentley appeal (Coulthard, Johnson, and Wright, 2016). Furthermore, Coulthard addressed the potential corpus linguistics has for authorship research and how its methods can make idiolect research possible. Grant (2012) implemented stylistic analysis and corpus-based methods in the case of Amanda Birks. He emphasised understanding the crucial difference between the interpretations of distinctiveness in terms of population-level or small group or pairwise levels. In the authorship analysis approach he proposes, he claims that pairwise or small group distinctiveness is more feasible to carry out in forensic casework because population-level distinctiveness can show wide linguistic variations. Also, Nini and Grant (2013) proposed a combination of cognitive and stylistic approaches in authorship analysis using the Systemic Functional Linguistics (SFL) theory of codal variation. They ran two trials, the first applying SFL theory on a small dataset of three authors' 300-word texts. The second trial was derivative of the first trial's findings which suggested incorporating Biber's (1988) multidimensional analysis framework. The researchers found that combining the theory with the multidimensional framework shows potential as a methodology for authorship analysis. Sousa Silva et al. (2011) explored stylistic markers in authorship analysis, specifically emoticons in micro-blogging text messages. They examined three authors' Twitter messages and trained Support Vector Machine (SVM) classifiers to detect and identify their individual combinations. Their findings reveal that SVM classifiers and emoticons show good potential in attributing authorship. Furthermore, Johnson and Wright (2014) tackle authorship attribution and idiolect using corpus linguistics methods by exploring a corpus of 63,000 emails and 2.5 million words by 176 Enron employees. The first study they run is a case study where they found that one of the authors produces patterns in using directives politely in a habitual manner. The second one is a statistical experiment where the researchers identify the same author's n-grams and assign anonymous emails that match the style with a success rate of 100%. They conclude that authorship can be identified even when a mass of data is reduced to key segments or 'textbites', provided that they are distinctive enough (Johnson and Wright, 2014). Another project presented at the 15th conference of the International Association of Forensic and Legal Linguistics (IAFLL) that aims to encapsulate idiolect is by Heini, Kredens and Pęzik who made a study of 100 idiolects across seven different genres such as text messages, emails, and oral interviews. The researchers claim that this project has the potential to serve as a prototype for reference databases to perform authorship analysis across genres. They collected the linguistic data for over a hundred individuals and designed the corpus as a database model where text, author, and text print tables are coded and entered. Moreover, linguistic features (i.e., words, sentences, and syntax) are annotated using NLP and all textual data can be exported into any text-based format (Heini and AlAmr, 2022). Most linguists believe that authorship has to be based on a theory of language or idiolect, but that is a debatable matter to Grant (2010) who claims that consistency and distinctiveness of stylistic features are what make a strong argument in an authorship attribution case.

Stylistics, and forensic stylistics, are the study of style in language; the word *style* refers to the human behaviour of ‘the individual variation in activities that are otherwise invariant’ (McMenamin, 2020: 540). He also describes stylistic analysis as the focus on the use of language variables that show consistency and idiosyncrasy. There is also computational stylistics, which is closely related to authorship attribution, and it refers to the use of wide range of computational tools to identify style (Rybicki et al., 2016). Language is a common commodity used by everyone for communication. However, linguistic style is formed when an individual makes individual selections from the linguistic options made available in their group(s) of speakers. Labov (2002) states that language diversity is triggered by non-linguistic factors such as the individual’s need to be distinctive and their acquisition process and social experiences. There is a ‘general tendency towards accommodation and the pressure of community norms’ (Labov, 2002: 19) that shapes the verbal behaviour of individuals to convey the group identity they associate with. Labov makes a distinction between linguistic forms that are more stable and permanent as opposed to forms that change frequently across generations; an individual develops their linguistic style through an ongoing acquisition process of personal criteria to make individual choices. The notions of distinctiveness, consistency and verbal behaviour are investigated in forensic stylistics by examining the author's style. A forensic stylistician would examine a writing sample presented by the author and, if it is large enough, would be able to detect individual variation that reveals an underlying pattern of linguistic behaviour. That pattern is the accumulation of the author's acquisition and social experiences (McMenamin, 2020).

2.3.2 Computational approaches

The main concept in computationally supported authorship attribution is to distinguish between different authors of texts by measuring a range of textual features. Before the emergence of word processors and the Internet, researchers invested in statistical methods to achieve this task (Stamatatos, 2009). Methods developed further in the late 1990s; computers offered word processors and other tools that shifted the focus away from literary texts towards other organic genres (Holmes, 1994; Stamatatos et al., 1999). This was supported by the World Wide Web that broadened the horizon of authorship analysis research by giving access to 'real world unrestricted texts' (Stamatatos et al., 1999; Larner, 2014). Computational researchers, much like linguists, explored the use of statistics and stylometry to solve authorship analysis problems reliably. Several limitations hindered the methodological development at the time. First, the data was lengthy and stylistically diverse as most authorship problems would address a whole book. Second, the limited number of candidate authors usually did not exceed two or three authors. In addition, the methods devised were intuitive and were assessed in that sense, and the lack of a benchmark made the comparison of these methods more challenging (Stamatatos, 2009). The vast size of electronic texts made available by computers and the internet made it urgent to address these limitations, which paved the way for new areas to contribute such as machine learning, information retrieval, and natural language processing (NLP). Machine learning is a subfield of artificial intelligence; it is a method of data analysis that learns from the data, identifies patterns, and provides different ways of representing the results that show these patterns. Informational retrieval also represents data and classifies large volumes of texts efficiently. Lastly, NLP analyses textual data and provides a new form of representing data.

The plethora of electronic texts also started a new era of authorship analysis in the last two decades. The emergence of technological tools showed the potential of applying authorship analysis in various areas such as intelligence and attribution of messages to terrorists or linking different messages by authorship (Abbasi and Chen, 2005), identifying harassing authors and/or verifying authenticity of suicide notes in criminal law and copyright conflicts in civil law (Chaski, 2005; Grant 2007), identifying source code of malicious software in computer forensics (Frantzeskou et al., 2006), along with the traditional literary authorship disputes of known authors (Burrows, 2002; Hoover, 2004). The dominant focus is to come up with practical applications that tackle 'real-world texts' instead of literary disputes (Stamatatos, 2009: 539). There is also a focus on assessing and comparing the different methods proposed and setting a benchmark for methodologies (Juola, 2004). Moreover, studies explored ways to increase accuracy and how to examine it, such as the size of training texts (Marton et al., 2005; Feiguina and Hirst, 2007), the number of candidate authors (Koppel et al., 2006), and the distribution of texts over candidate authors (Stamatatos, 2008). The emergence of social media platforms drove some researchers to devise methods that account for the potential of having hundreds or thousands of candidate authors compared to the usual small range in previous studies (Seroussi et al., 2011).

Another research interest triggered by social media platforms, particularly the trend of using such media for terrorism and cybercrime, is authorship analysis research on short texts. The challenge is to identify or verify authorship using a relativity small dataset e.g., a few text messages or tweets and a large pool of candidate authors. The tools used are the same in classic authorship analysis problems, as the researchers explore the range of stylometric features and different classifiers to find which configuration scores the highest rates of accuracy using as minimal dataset sizes as possible. One of the prominent studies that paved the way for many later ones was done by Zheng et al. (2006), who developed a framework to identify the authorship of online messages and trace identities. They extracted lexical, syntactic, structural, and content-specific features and used them to build classification models and tested the framework on English and Chinese online-newsgroup messages. The results show that all features contributed successfully to discriminating the authors; also, Support Vector Machine outperformed the other two classifiers: decision trees and back-propagation neural networks by accuracy rates that average 70 to 95%. More importantly, the researchers recommended the approach be implemented in multiple languages. Forensic linguistics continues to cross paths with computational approaches in authorship analysis of short texts research as well. The study by MacLeod and Grant (2012) adopted stylistic and statistical techniques proposed in Grant’s earlier work (2007; 2010) to develop a taxonomy that identifies the authors’ distinctive features.

Tweets used on the Twitter platform are a particular kind of short text that is the focus of this research. Azarbonyad et al. (2015) tackled language evolution and change in word usage over time using both tweets and emails. They propose an approach inspired by time-aware language models to investigate authorship and the temporal changes of word usage. The datasets consisting of Twitter posts and Enron emails reveal that authors do change their usage of words at different rates. In another study, Rocha et al. (2016) revisited the current computational authorship attribution methods of social media forensics. They also introduced and tested the novel supervised learning-based methods that provide explanations for their results and are effective for small-sized samples. Srinivasan and Nalini (2019) performed a study similar to Zheng et al. (2006) by creating a framework using four stylometric features (lexical, syntactic, structural, and n-grams) and inductive learning algorithms to build a feature-based classification model. After conducting experiments to measure the accuracy of two, three, and five authors, the researchers found that combining syntactic, structural, and n-gram features scores higher accuracy rates across the four different classifiers: C4.5, Random Tree, Fuzzy classifiers, and Adaptive boosting classifiers.

The field of authorship research in Arabic is dominated by computational scientists who initially were exploring the potential of their tools in identifying authorship of religious and literary texts. Sayoud (2012) composed a series of three authorship attribution experiments to discriminate between the Quran and the *Bukhari Hadith* (a book that collects the statements of the prophet Muhammed). The first experiment aimed to analyse both books in terms of word frequency, word length, and character frequency, which concluded that the books were written by different authors. The second experiment is a statistical analysis of stylometry of the texts in a segmental form and the third one is also a segmental analysis of the texts using a range of stylometric features and classifiers. Both experiments reveal that the segments extracted from each book were able to show stylistic similarity. Ouamor and Sayoud (2012) investigated authorship attribution of old Arabic texts by examining character and word n-grams. For text classification, they tested Sequential Minimal Optimization-based Support Vector Machine (SMO-SVM). They used twenty texts for training the classifiers and another ten texts for testing. The results show that character-based features, specifically trigrams and tetragrams, perform better than word-based features across all classifiers with scores up to 80%. Ouamour and Sayoud (2013) revisited the authorship of old Arabic books written by ten travellers. They used character unigrams, bi-grams, trigrams, and tetragrams and explored four classifiers: Stamatatos distance, Manhattan distance, Multi-Layer Perceptron (MLP), and Support Vector Machines (SVMs). They also used an Arabic dataset to evaluate the performance of the features and classifiers. Their findings reveal that Arabic character grams are like English; hence it is possible to exploit most of the findings discovered in English related studies and employ them in Arabic settings. Also, Manhattan distance outperformed other classifiers at an accuracy rate of 90% when using character tetragrams. Howedi and Mohd (2014) investigated the authorship of short historical Arabic texts and ten candidate authors. The researchers' dataset consisted of three short texts per author's book. They identified lexical and character features (N-grams) per author for data representation as well as exploring the authors' individual style in using punctuation marks. Also, they compared the performance of Naïve Bayes (NB) in classifying short texts compared to Support Vector Machines (SVMs). They found that punctuation marks can increase the potential of distinguishing authors' style and NB classifier performs more accurately compared to SVMs with rates up to 96% using unigrams.

Arabic authorship research also showed interest in online messages and social media platforms, specifically Twitter, as it is widely used in the Arab region. Abbasi and Chen (2005) explored the lexical, syntactic, structural, and content-specific writing style features to identify the authorship of Arabic web forum messages. The t-test analysis showed that while all Arabic features improve the accuracy for identifying authorship of online messages, the least effective were content-specific words due to the broad scope of the Arabic dataset used in the study. They also found that SVMs outperform decision tree classifiers such as C4.5. The researchers pursued this further in the following study (2008) in which they modified an existing framework to analyse the authorship of Arabic and English web forum messages. The findings confirm that all feature types improve the classification accuracy for Arabic and English except for content-specific features. Also, SVMs outperformed the decision tree classifier in all feature cases as they have proven to be better equipped to handle larger feature sets and noisier data. Twitter was the subject of a number of authorship studies in Arabic; some used the platform as a corpus to test and analyse (Alfaifi and Atwell, 2016) while others used it to solve authorship identification problems (Rabab’ah et al., 2016; Khonji and Iraqi, 2018). Another attempt to address cybercrime was made by Altakrori et al. (2018) to investigate the authorship of Arabic microblog posts on Twitter. They compared the performance of an instance-based classification tool and a profile-based approach using n-grams as features and the effect of the training set, the tweets’ length, and the number of authors on the accuracy of the classification. They also utilise an event-visualisation developed to accommodate English and Arabic for the representation of the data. The results show that while the instance-based approach can outperform its counterpart in some tests, it is complex and cannot be presented as evidence in court. On the other hand, profile-based approaches are simpler, and their results can be visualised more clearly. Other authorship attribution studies tackled author profiling in emails (Estival et al., 2007), authors’ gender identification (Alsmeerat et al., 2014; Alsmeerat et al., 2015), author identification (Otoom et al., 2014), author authentication (Alwajeeh et al., 2014; Al-Ayyoub et al., 2016), and authors’ political orientation (Abooraig et al., 2014). Menai (2012) addressed plagiarism detection by proposing, APlag, a prototype of a new detection tool mainly designed for Arabic texts. Menai implemented heuristic algorithms to compare stylistic variation across document, paragraph, and sentence levels and trace change or similarity in structure. The results reveal that APlag can detect copying exact sentences, changes at sentence level, and replacement of synonyms.

Moreover, a substantial number of authorship studies in Arabic explored text classification and categorisation techniques such as Linear Discriminant Analysis (Shaker and Corne, 2010), Support Vector Machines (Baraka et al., 2014), and some compared the performance of classification techniques (Kanaan et al., 2009; Khorsheed and Al-Thubaity, 2013; Hmeidi et al., 2015). Other studies explored text pre-processing tools to further improve the performance of classifiers (Said et al., 2009; Saad, 2010), and some focused on automatic speech tagging (Diab et al., 2004). Sadat et al. (2014) proposed a framework to classify Arabic dialects used in social media platforms using probabilistic models. After running a set of experiments using character n-gram, Markov language model and Naïve Bayes classifiers, they found that bi-gram models can identify the eighteen different Arabic dialects with a 98% accuracy rate. Althenayan and Menai (2014) explored the Naïve Bayes classifier more closely using a range of its models to attribute the authorship of Arabic books written by ten different authors. Using the same dataset presented in Abbasi and Chen (2005), the researchers measured the performance of simple Naïve Bayes (NB), multinominal Naïve Bayes (MNB), multi-variant Bernoulli Naïve Bayes (MBNB), and multi-variant Poisson Naïve Bayes (MPNB). The findings showed that MBNB scored the highest results at the rate of 97.43%. Comparing the classifiers revealed that multi-variant Bernoulli Naïve Bayes and multinominal Naïve Bayes are best suited for authorship attribution tasks. Albadarneh et al. (2015) wanted to explore authorship authentication of Arabic tweets in a large dataset. Using 53k tweets distributed over twenty authors, the researchers implemented a bag-of-words approach and computed the feature vectors of each tweet using the Term Frequency-Inverse Document Frequency (TF-IDF). After pre-processing the data, the researchers use a Naïve Bayes classifier and Hadoop, a framework to achieve parallel and distributed computing. The results show that the accuracy rate is only up to 61.6%, which they attribute to the challenging setting of the study. Hussein (2019) tested a set of Arabic function words in a corpus-based approach to verify the authorship of the famous Arabic religious book *Nahjul-Balagha*. Using two multivariate methods (Principal Component Analysis and Cluster Analysis), he found that the eleven disputed religious sermons are written by the same author. Also, both multivariate methods presented results robustly. A rare study that addresses authorship analysis in a forensic context is by Garcia-Barrero, Feria, and Turell (2012) who propose a forensic authorship attribution approach to written Modern Standard Arabic. Using a sample corpus of three Moroccan authors in two different genres (short stories and literary reviews), the researchers tokenised the texts in full. They performed their analysis in terms of type/token ratio, word length in character, punctuation, conjunctions, the combination of punctuation and conjunctions, and finally the standard deviation of sentence length in words. The results reveal that the variables examined can attribute authorship in Arabic texts successfully. Moreover, text genre proves to be an influential variable as opposed to time that made no significant impact on the authorial style of the sample.

2.3.3 Challenges that face both fields

A fundamental challenge that both fields face is to ensure a valid and reliable method to reach accurate conclusions in spite of the text’s length. Computational researchers were initially more concerned with the accuracy rates of classifiers and parameters. This interest has generated a range of black-box methods that achieve accurate results that do not necessarily provide a trace for their findings, which was later developed into white box testing techniques that provide the researcher with detailed internal logic and structure of the code (Khan and Khan, 2012). Linguists are keen on providing quantitative evidence that is founded on linguistic theory, all the while paying more attention to the qualitative aspect of the analysis and explicability of the findings. One size does not fit all; the difficulty in authorship research cannot resolved by a fixed set of features that would successfully identify author X across any text genre. There is a need for forensic linguists and computer scientists to create more synergy between the disciplines. Forensic authorship attribution is an interdisciplinary field where linguistic approaches, corpus linguistics included, cross paths with machine learning, NLP, statistics, and law. This calls attention to the many research opportunities that address these challenges. Thus far, research in forensic authorship and stylometry - whether linguistically-oriented or computational - is rarely presented as sole evidence at court but rather used as supporting opinion or to give direction in narrowing down a circle of suspects.

2.4 Stylometry and statistics

Forensic authorship research has come a long way in finding reliable methods in the forms of stylometric and stylistic approaches. Stylometry is the statistical analysis of a text by calculating a unit that represents an author’s style (Holmes, 1994). When it comes to authorship analysis, Grant (2012) makes a distinction between stylometric and stylistic approaches. Stylometric approaches aim to quantify markers by measuring frequencies of word classes or n-grams and their significance relies on the variation of their occurrence in disputed texts. Alternatively, stylistic approaches focus on the habitual choices an author makes and selecting the quantitative methods that work best with the data. Baily (1979) describes the properties of a measurable stylistic feature as a “… salient, frequent and easily quantifiable, and relatively immune from conscious control.” (as cited in Holmes, 1994: 88). This could be by word length (Mendenhall, 1887; Brinegar, 1963; Mosteller and Wallace, 1963), syllables (see Fuchs, 1952; Bruno, 1974; Brainerd, 1974), sentence-length (Yule, 1938), function words (Morton, 1978), word frequency (Zipf, 1932), and many other stylometric methods. One of the ways stylometric approaches work is to statistically measure the similarities and differences between authors’ styles using a specific set of linguistic features. Another way is to identify the style of an author using stylistic features as measuring units. Word-length is one of the earliest features measured first by Mendenhall (1887) in the authorship attribution case of Shakespeare and Bacon, who concluded that the frequency distributions of word-length in Bacon’s prose is inconsistent with the frequencies in Shakespeare’s verse; therefore, it is highly unlikely to question the authorship of some of Shakespeare’s work. Several researchers adopted Mendenhall’s approach such as Brinegar (1963) who questioned the authorship of Mark Twain in *The Quintus Curtis Snodgrass Letters,* and when he compared frequency distributions of word-lengths between the letters and his other work, he found that Twain was not the author. As mentioned earlier, Mosteller and Wallace’s (1963) study of *The Federalist Papers* marks the modern era of authorship attribution research. The author of the papers was suspected to be either Alexander Hamilton or James Madison and, following the same approach that was established by Mendenhall, they identified Madison as the author (Mosteller & Wallace, 1963; Holmes & Forsyth, 1995).

The twentieth century marked progress in authorship attribution tasks as the linguist George Zipf introduced a law of probability that states that the frequency of an event is proportionate to their rank in an inverted manner. In a linguistic context, the law states that the frequency of a word its rank on a table. For instance, a word that is ranked second in frequency would occur half as often as the word ranked first (Powers, 1998). Stylometric studies use computational tools to ensure valid and reliable results, while stylistic ones tend to rely more on the researcher's analysis. Stylometry experts have devised a range of methods to calculate and measure an author's 'stylistic fingerprint' (Holmes, 1994: 87).

In a forensic context, linguists examine an author's conformity to and divergence from the community norms in using stylistic features (McMenamin, 2002; 2017; 2020). Also, McMenamin (2020) points out several limitations that hinder the practice of forensic stylistics that affect the admissibility of linguistic evidence in court. The process of selecting the linguistic variables to measure the similarity and contrast between styles is usually described as arbitrary and subjective, as there is no specific rationale that explains the criteria of selecting style markers. Also, calculating the frequency of occurrence must be based on solid statistical analysis of the text that accounts for possible variants of the linguistic variable under examination. Studies in the past few decades have proven the success of statistical approaches in forensic authorship analysis (e.g., Grant and Baker, 2001; McMenamin, 2002; Grant, 2007; Grant and MacLeod, 2020). Another challenge is to determine the distinctiveness or uniqueness of an examined linguistic variable without a reference to the linguistic norm. This also makes it difficult to distinguish group and individual variation. Lastly, to determine the level of consciousness when making one linguistic choice or another is still not possible, which makes it difficult to assign significance to the examined linguistic variable.

Computational approaches search for 'stylometric features that capture the diversity of the language deployed therein' (Rocha et al., 2016), a gap that linguists can fill by the numerous studies such as those of Burrows' (1987, 1989, 1992) using Principal Component Analysis (PCA) which has proven to be one of the successful approaches to identify the potential markers of authorship that also discriminate between the markers of each author (Grant and Baker, 2001). Other methods other than PCA have also proven to be accurate such as Burrows' Delta, Zeta, Linear Discriminant Analysis, Support Vector Machines, deep learning. Computational approaches focus on stylometry to select linguistic features that can be calculated and analysed using machine learning algorithms, Natural Language Processing (NLP) methods.

2.5 Conclusion

Authorship attribution research has come a long way for a novel research field and its pace of progress is accelerated and accounted for by the technological breakthroughs over the past two decades. It extended beyond disputed literary texts to branch out to solve nowadays issues such as plagiarism detection, counterterrorism, cybercrime, and much more. Stylometry and statistics established a solid foundation for authorship attribution methods. The emergence of new fields like corpus linguistics and machine learning provided authorship attribution with new and diverse tools for data analysis and representation that complement its predecessors: stylometry and statistics. Research established that the combination of stylometric features, and classification models is a successful framework to tackle authorship problems (Zheng et al., 2006; Rocha et al., 2016). Researchers also generated different models using different features, n-grams sets, and classifiers in search for the optimum accuracy rates. The ability to represent data visually and clearly is essential. Furthermore, researchers explored the applicability of these frameworks in other languages beside English such as Chinese (Zheng et al., 2003; 2006), Urdu (Naqvi et al., 2017; Anwar et al., 2018), Thai (Maruktat et al., 2014; Pingjai, 2019), Spanish (Turell, 2010; Bel et al., 2012), and Portuguese (Sousa-Silva et al., 2010; 2011) in a global effort to establish authorship studies within forensic linguistics as a field in its own right.

Authorship research in Arabic is an established area in computer sciences (represented in artificial intelligence and machine learning) and NLP, exploring Classical and Modern Standard Arabic in religious and literary texts (Sayoud, 2012; Ouamour and Sayoud, 2012; 2013). Further studies explored Arabic dialects on social media platforms in response to the call for research in authorship to counterterrorism and fight cybercrime (Mansour et al., 2012). Moreover, Arabic authorship research has driven the production of corpora to further establish corpus-based approaches (e.g., Alfaifi and Atwell, 2016). The linguistic features used in the research, while indispensable in extraction and analysis processes are described and presented exclusively as stylometric features (Al-Tuwairesh et al., 2015; 2018; Khonji, 2018). Thus far, forensic authorship analysis is yet to explore the robustness of Arabic linguistic features and their stylistic and contextual implications, especially as far as the idiolect and identity are concerned.

Looking at the big picture of forensic authorship analysis, there is a need for more collaborative work not only between linguists and lawyers, as discussed in Chapter 1, but also between linguists and computer scientists. Combining the efforts of both fields and integrating their methods will work towards achieving what forensic linguistics aims to achieve: creating a valid and reliable methodology that is acknowledged and accepted by court. The work of Patrick Juola (2004; 2006; 2008; 2012; 2015; Juola et al., 2013), a computer scientist who implements linguistic approaches in his research, is a representation of what an interdisciplinary approach can accomplish. The field of authorship attribution is interdisciplinary, which is an asset as it gives the researcher room for creativity. However, generating new methods that suit the individuality of the data, whether in research or casework, must not compromise the validity and reliability of these methods.

Chapter 3
THIRD-WAVE SOCIOLINGUISTICS AND IDENTITY

> “[t]he study of variation was dominated by a definition of style as “different ways of saying the same thing” (Labov 1972, p.323)” – (Eckert, 2012: 98)

3.1 introduction

There has been a growing interest by sociolinguists to learn more about the language of the individual, how does it vary from one another, and its relationship with the society and community of practice. Sociolinguists has gone through phases, or waves, as their understanding of the identity and variation unfolds. The excerpt above quotes Eckert (2012), who describes style as per the first and second waves’ perspective and how the third wave steers away from that. Earlier sociolinguistic research would categorise individuals based on their age, sex, education, and other demographic data in a top-down approach. Instead, third-wave sociolinguists are interested in the micro-level categories that are the product of the individual and not vice versa (i.e., macro-level sociological variables). This chapter is a continuation of the theoretical discussion that started in Chapter 2, where the focus shifts from idiolect to identity. Section 3.2 is a brief survey of third-wave sociolinguists and their discussion of prior concepts of identity and its relationship with macro-level sociological categories. The following section is a discussion of the notion of identity and different perspectives on how to encapsulate it introduced by scholars such as Moore (2004), Eckert (2012), Bucholtz and Hall (2008), and Johnstone. The final part, section 3.4, introduces CMDA which is a methodological approach that encompasses the theoretical notions discussed here. The chapter concludes with a personal interpretation of identity, which is derived from the findings that will be discussed in Chapter 6.

3.2 Third wave sociolinguists

As noted earlier in the introduction, sociolinguists went through three trends, or waves, in learning about linguistic variation and how to interpret style. The earliest one, the first wave, commenced by William Labov's study in 1966, which was replicated in other countries such as Great Britain (Trudgill, 1974) and Iran (Modaressi, 1978). This wave founded the socioeconomic stratification of linguistic forms, where the standard ones are at the top and as it gets lower the non-standard forms are classified in terms regional and ethnic variation (Eckert, 2012). During the 1960s and 1970s, studies similar to Labov's all focused on placing speakers in static categories, to which their identities are affiliated. First wave variationists stigmatised non-standard forms and marked their frequent occurrence as a decrease in the socioeconomic strata. Similarly, second wave sociolinguists also perceived variation as 'incidental fallout from social space' (Eckert, 2012: 94). Moreover, their ethnographic studies provided the configurations to categorise individuals more locally compared to the macro-level sociological categories of the first wave.

Third-wave sociolinguists pay attention to stance and stancetaking strategies as they find them indicators of individual stylistic variation, which is their focus (Bucholtz and Hall, 2005, Johnstone, 2009, Eckert, 2012). Biber and Finegan (1989) examined the degree of certainty, knowledge, and attitudes in the speakers' utterances. Later, Hunston and Thomson (2000) developed their interpretation of stance and attitude into evaluation, which they define as 'the broad cover term for the expression of the speaker or writer's attitude or stance towards, viewpoint on, or feeling about the entities or propositions that he or she is talking about.' (p.5). They claim that speakers/writers use evaluation for three purposes: either to express their opinions, to manipulate the recipients' attitudes, or to mark boundaries and emphasise significant points in a discourse (Hunston and Thomson, 2000).

Johnstone (2009) finds the connection between Hunston and Thomson's *evaluation* and Goffman's (1981) *footing* in the sense that both perceptions capture the speaker/writer's attempt to manipulate the recipient through interaction. Another element that is key to third wave variationists is the connection between stance and style. Bauman (2004) and Johnstone (2009) define style is the repetitive sets of stances that emerge as part of the repertoires. In her analysis of Barbara Jordan's style, Johnstone (2009) examines the notions of the ethos of self, the personal identity, and the ethos of person, the social identity, and which is more comes to the conclusion that 'style of stance comes to index not a social identity but a personal identity…' (p.34).

3.3 Approaches to identity

Several theories define identity and the process of its construction such as speech accommodation theory (Giles et al., 1991) and social identity theory in social psychology (Tajfel and Turner, 1979; Meyerhoff and Niedzielski, 1994; Meyerhoff, 1996). Linguistic anthropology uses theories of language ideology (Silverstein, 1979; Irvine and Gal, 2000) and theories of indexicality (Silverstein, 1976; 1985; Ochs, 1992), while sociolinguistics is represented by theories of style (Eckert and Rickford, 2001; Mendoza-Denton, 2002) and models of identity (Le Page and Tabouret-Keller, 1985). But it was Goffman's (1974; 1981) interest in the self and identity that became the foundation for much of the research that followed. He explains that an individual cultivates alignments or 'footings' that they signal when engaging in an interaction whilst simultaneously acknowledging the alignments of their counterparts. The early 1980s was the start of a growing interest by linguists in identity (e.g. Gumperz, 1982) only to become the theme of most social sciences research in the 1990s (Edwards, 2009).

Researchers also took a particular interest in interaction, performance, and the processes that construct and communicate identity. Some researchers believe that identity development is an internal affair within oneself; for example, Antaki and Widdicombe (1998) claim that it is an internal mechanism that individuals develop that goes beyond self-classification and social behaviour. Their counterparts such as Butler (1990) think that identity is constructed through repetitive discursive performances. Benwell and Stokoe (2006) also argue that social actions, especially language, constitute identity saying: "identity inhabits not minds, but the public and accountable realms of discourse." (p. 49). This is further supported by advocates of the idea of the linguistic individual, such as Johnstone (1996) whose research confirms that discourse is the only channel that enables individuals to communicate their self-concept to society. Bucholtz and Hall (2004) also say that individuals employ symbolic resources, language included, to produce their identity, and language is a central and flexible resource that is not a product caused by the membership of a particular social category but rather 'interactionally negotiated.' (p. 376). But Grant and MacLeod (2020) emphasise that an identity theory that focuses on the role of interaction might fail to account for the inner workings of the personal identity, which is essential in authorship analysis casework. They argue instead to account for the possibility that an identity isn't necessarily revealed through interaction and can remain opaque.

The internet and digital media introduced a new notion: online identity. Research has shifted recently to explore how individuals perform their identities in virtual communities. To understand how online communities interact, Angouri (2016) and Grant and MacLeod (2020) find that the community of practice framework is an adequate concept that explains community norms, discourse practices, and the roles of its members. Johnstone (1996) states that 'it is more enlightening to think of factors such as gender, ethnicity, and audience as resources that speakers use to create unique voices, than determinants of how they will talk.' (p.11). Grant and MacLeod note there are two issues to observe when developing a theory of identity: first, identity is not "a set of static social categories" (2020, p. 18) and must not be perceived as such; and second, even though identity is not fixed, it can be somehow constrained.

In light of the early work, Bucholtz and Hall (2005) proposed an approach that combines the sociocultural, anthropological, and psychological dimensions. They define identity in the broad sense as "the social positioning of self and the other." (Bucholtz and Hall, 2005: 586). The framework provides a comprehensive sociolinguistic perspective of identity focusing on the structure of the language and the influence of culture and society. It comprises five principles: emergence, positionality, indexicality, relationality, and partialness, which I turn to explain in more detail (Bucholtz and Hall, 2005).

The first principle they introduce is emergence, which states that for self-perception and self-conceptualisation to take place, it would have to be through disclosure and expression. Disclosure is represented in interaction being a social and cultural phenomenon that constantly helps in building and projecting an identity through its linguistic and semiotic practices. Second, the principle of positionality is a development of Labov's (1972) macro-level social categories which identifies identity in terms of age, gender, and social class. An identity also conveys local ethnographically specific cultural positions and stances that are temporary and specific to an interaction. Furthermore, the third principle, indexicality, points to the role indexical processes play in revealing identity. Processes such as categorising, labelling, and creating presuppositions are consequently producing epistemic and evaluative positions that reflect the individual's orientations. The terms 'epistemic' and 'evaluative' stances have been used in Chapter 5, but I will develop discussion of these here. Moreover, an Individual's stance is developed using linguistic features such as vocabulary that conveys ideological connotations and reveals the author's personas.

Furthermore, Bucholtz and Hall (2005) classify indexical stance construction into two types. First is *indexical inversion*, which refers to the top-down interaction and ideological ties that are set by cultural authorities such as the media and affect the language and its speakers. The opposite process would be *stance accretion*, derived from Du Bois's (2002) work on stance, which refers to the bottom-up buildup of stances through interactional conventions into a perpetual structure of identity (Bucholtz and Hall, 2005: 596). Relationality refers to an individual's urge to connect or identify their social meaning in relation to other available identity positions and other social actors. They say that identities are not autonomous but constructed through relational processes that often overlap: adequation and distinction, authentication and denaturalisation, and authorisation and legitimisation. An individual would employ these processes to generate their sense of identity: distinction and its counterpart adequation which focus on the identity's relation of similarity or difference to a social group/category. The second process is where the identity relates to what is real/authentic and what is fake/artifice. The last process is where the identity relates structures of ideology and power while simultaneously claiming other identities as illegitimate by dismissing, censoring, or ignoring them. Bucholtz and Hall (2005) conclude with the final principle that is partialness: the construction of an identity shift as interaction develops and discourse changes in context. It is partly deliberate and partly habitual and less than conscious; it is partially an outcome of interaction and partially resulting from others' perceptions and representations; it is the result of larger ideological processes and material structures. This conceptualisation of partialness aims to describe an individual's sense of agency and how it is represented in linguistic structure. In summary, the five principles of the identity approach account for the range of online discourse that can take place and the discursive interactions reveal and represent the online identity. Nonetheless, there is a need for a paradigm or a scheme to facilitate classifying these interactions in order to analyse them adequately.

3.4 CMDA

Computer Mediated Discourse Analysis (CMDA) aims to encapsulate the finer details of this mode of communication and facilitates the analysis of the data. As the Web 2.0 platforms evolve, which are platforms that some of which emerged in the early 2000s (e.g., Facebook, YouTube, and Twitter), new discourse phenomenon is generated (Herring, 2013). Herring claims that the shift in Web 2.0 transformed computer-mediated communication (CMC) into 'the convergent media computer-mediated communication (CMCMC)' (Herring, 2013: 4). Thus, Herring (2013) proposes a three-way paradigm to classify this Web 2.0 discourse phenomena: first is to observe technological variables such as multimodality and use of media. The second aspect is social variables, which is to consider situation and culture; and finally, the third is linguistic. Nonetheless, the evolution of the web and social media does not suggest that CMC and CMDA are obsolete, as they are still useful to analyse new social media.

Furthermore, Herring (2007) proposes that there are two, non-hierarchal main categories when classifying CMD: medium factors and situation factors, which formulate the Faceted Classification Scheme. Herring explains that these main categories do not follow a particular order to account for the different contexts CMD can derive from, hence the 'non-hierarchal relationship' that aims to help 'discover the strength of the social and technical influences' of CMD (2007: 10). Each category includes subcategories with a range of realisations derived from empirical evidence observed in CMD. The medium factors describe the different modes that communication technology provides, which are demonstrated in Table 3.1 (see Herring, 2007). However, it is the situational factors that are concerned in this study's analysis (Table 3.2). Furthermore, the scheme can be used flexibly in the sense that a researcher can select and apply the subcategories and facets that apply to their data as it isn't mandatory to apply them all. In this research, I will not address the Medium factors and will move on to the Situation factors.

Table 3.1: Medium factors (Herring, 2007)

M1	Synchronicity
M2	Message transmission (1-way vs. 2-way)
M3	Persistence of transcript
M4	Seize of message buffer
M5	Channels of communication
M6	Anonymous messaging
M7	Private messaging
M8	Filtering
M9	Quoting
M10	Message format

The situation factors classify information related to the context of the communication, starting with its structure and the number of participants involved. For instance, a blog can be authored by one person addressing a group of participants, the author can be anonymous or declare their name, the participants who read the blog can be a larger group than those who actively interact by leaving comments or queries, and so on. The second category classifies the characteristics that participants show through CMD; starting with demographics, their level of proficiency in language and/or CMC, their real-life status or online persona, and whether they reveal their beliefs or ideologies, among other facets. Also, there are categories allocated to describe the purpose of the online group or interaction, whether it is for social or political purposes or to develop professional relationships. Another category is to describe the range of topics or themes of the CMD such as sports, academia, science fiction, etc. Herring is also interested in looking at the tone of that communication between the participants, which is usually set by the previous categories of purpose and topic/theme. For example, a group dedicated to professional development would have a serious, formal tone. Moreover, the activity category aims to describe the kinds of exchanges that take place in communication, which can range from information to jokes or insults, playing a game, or giving a performance. Norms refer to the rules of interaction that are acknowledged by the online community, by society in the real world, or by the language the participants use. And finally, the code category classifies the languages or varieties the participants use and the writing system of their devices.

Table 3.2: Situation factors (Herring, 2007)

S1	Participation structure	- One-to-one, one-to-many, many-to-many - Public/private - Degree of anonymity/pseudonymity - Group size; number of active participants - Amount, rate, and balance of participation
S2	Participant characteristics	- Demographics: gender, age, occupation, etc. - Proficiency: with language/computers/CMC - Experience: with address/group/topic - Role/status: in ‘real life’; of online personae - Pre-existing sociocultural knowledge and interactional norms - Attitudes, beliefs, ideologies, and motivations
S3	Purpose	- Of group, e.g., professional, social, fantasy/role-playing, aesthetic, experimental - Goal of interaction, e.g., get information, negotiate consensus, develop professional/social relationship, impress/entertain others, have fun
S4	Topic or Theme	- Of group, e.g., politics, linguistics, feminism, soap operas, sex, science fiction, South Asian culture, medieval times, pub - Of exchanges, e.g., the war in Iraq, pro-drop languages, the project budget, gay sex, vacation plans, personal information about participants, meta-discourse about CMC
S5	Tone	- Serious/playful - Formal/casual - Contentious/friendly - Cooperative/ sarcastic, etc.
S6	Activity	- E.g., debate, job announcement, information exchange, phatic exchange, problem solving, exchange of insults, joking exchange, game, theatrical performance, flirtation, virtual sex
S7	Norms	- Of organization - Of social appropriateness - Of language
S8	Code	- Language, language variety - Font/writing system

In this study, I employ the principles of Bucholtz and Hall’s (2005) identity approach to examine the collective identity themes revealed through interaction. I explore emergence, relationality, indexicality, and other principles when applicable to show the authors’ affiliation in terms of nationalism, gender, and region. The findings and some illustrative examples are discussed in Chapter 6. Furthermore, the NASCoT data is a form of CMD; thus I adopt Herring’s (2007) scheme to facilitate the analysis of the implications of the authors’ individual identities. The scheme allows for using its values liberally; therefore, I investigate some of the authors’ use of the NA features by classifying the activities (S6 in Table 3.2) they partake in. I conduct the latter analysis using a sample of the data: 25 randomly selected tweets for four authors of the NASCoT. Also, I extensively examine the activities’ trends in one of the authors and then finally discuss the findings in section 6.3.2.

3.5 Conclusion

This chapter was a theoretical discussion of an integral element of this research: identity. We learned in section 3.2 about third-wave sociolinguists who raised attention towards the individual and linguistic variation that is based on such differences. This came as a change from the prior trends that placed individual in pre-determined, static categories such as age, gender, and sex. The following section explained identity and its different interpretations in the literature. It also discussed theoretical approaches to analyse identity through style and stance. The final part discussed a methodological approach that is connected to identity and style in social media or Web 2.0 that is CMDA. After a brief introduction of this area of research, we discussed the scheme used to analyse the data in this research. The following chapter further explains the methodological approaches and tools used in data analysis.

Chapter 4
CORPUS DESIGN AND METHODOLOGICAL APPROACHES

4.1 Introduction

In this chapter I explain the design of my Najdi Arabic Specialised Corpus of Tweets (NASCoT) and the reference corpus, the Reference Corpus of Najdi Arabic Tweets (ReCNAT), and the rationale behind their structure. The section that follows discusses the data collection and preparation processes I implemented prior to analysis. Also, I discuss the selection of my sample and the criteria for that selection, a brief profile of each author, and the ethical considerations and measures that were taken regarding the data. Then I address the methodological approaches adopted in this study which are corpus linguistics, stylometry, computer-mediated discourse, and computational linguistics. In the final section, I will briefly discuss the pilot study and how its findings contributed into the development of the doctoral research.

4.2 Corpus design

A corpus has proven to be a useful tool to discover various aspects of language phenomena (Reppen, 2010; Cotterill, 2010). To answer the research questions for this study, I compiled a 'specialised corpus' (Koester, 2012) that consists of a collection of tweets by thirteen Najdi Saudis. It is specialised in the sense that it is a corpus of a particular genre: compiled of the authors' tweets and replies they publish. To classify and describe the genre of the corpus is key for an accurate analysis. Twitter is classified as a microblog where each individual publishes posts in real time and the post develops through interaction with others, which can develop into a narrative. Hardaker (2010) classified types of interaction on Twitter and their functions. Table 4.1 shows the different categories of posts or tweets users can publish. I added ‘retweet with comment’ because it is one of the aspects of interaction that I want to include in my data.

Table 4.1: Types of interaction on Twitter (Hardaker, 2010)

Interaction type	Function
Tweet	An online post made by a Twitter user
Mention	X includes Y’s username in their tweet
Retweet	X re-posts Y’s tweet so that X’s followers can see it, thereby expanding the audience originally intended.
***Retweet with comment**	X re-posts Y’s post with a comment. Sometimes used as a form of a reply.

*This feature was introduced after Hardaker’s study.

An average of 500 tweets and above per author have proven to be sufficient to identify authorship (Sadat et al., 2014). The number of tweets collected for this corpus is 58,005 with at least 1,000 per author (Table 4.2), all produced by the participants in the form of text only or images with minimal text (e.g. a caption). In corpus design, the number of words is of significance because it shows the representativeness and range of the corpus. According to Flowerdew (2004), a small-specialized corpus can be the average size of 250,000.

The Najdi Arabic Specialised corpus of Tweets comprises 748,348 words in total, so large enough for a specialised corpus, and is designed to capture the idiolectal style of the selected authors. Part of this is the variation in the sizes of their subcorpora and the average number of words per author. The design reveals the number of tweets and number of words per subcorpus, which shows the variation in the lexical density in the authors’ language. For instance, Alajlan has the highest number of tweets (12,040) and words (216,027) and the average number of words per tweet is 18, which is different compared to authors such as Altassan who has a total of 10,560 words and the word average is two per tweet. Another example is

Allahim who has 7,532 tweets in his subcorpus which is close to Aleidi's (7,952). However, he has 135,033 words in his subcorpus while she has 70,709 words only.

Table 4.2: The Najdi Arabic Specialised Corpus of Tweets (NASCoT)

#	Author	No. of tweets	No. of words	Average words	Profile
1	Faisal Alabdulkarim	2,825	60,213	21	Social activist – influencer - male
2	Mansour Alrokibah	2,412	24,727	10	Social activist – influencer - male
3	Abdulrahman Allahim	7,532	135,033	18	Lawyer – influencer - male
4	Ali Alghofaily	2,424	37,217	15	Television presenter - male
5	Abdullah Alsubayel	5,805	35,791	6	Influencer - male
6	Abdulaziz Alzamil	4,741	47,832	10	Youtuber - male
7	Taghreed Altassan	5,292	10,560	2	Writer – columnist - female
8	Wafa Alrasheed	2,200	24,236	11	Writer – columnist - female
9	Maha Alwabil	1,550	40,263	26	Writer – columnist - female
10	Arwa Almohanna	1,143	17,956	15	Cultural newspaper editor - female
11	Ghadah Aleidi	7,952	70,709	9	Humanitarian - female
12	Maha Alnuhait	2,089	27,784	13	GM of sustainability program - female
13	Amani Alajlan	12,040	216,027	18	Social activist – influencer - female
	Total	**58,005**	**748,348**		

In order to achieve forensic text comparison, there needs to be a reference corpus by which the authors' language is measured. The selection of the reference corpus is just as important as the sampling of the corpus under examination. Prior studies (Kredens, 2002; Grant, 2010, 2013; Kredens and Coulthard, 2012; Grant and MacLeod, 2018) confirm that using a relevant population as reference corpus in forensic authorship analysis is acceptable. In this case, the reference corpus is composed of tweets published in Riyadh during the same time as the sample's data collection process. The relevant population in this study is the city of Riyadh, Saudi

Arabia. Hence, a corpus composed of tweets posted in the region would be acceptable (Larner, 2014; Heydon, 2019). The Reference Corpus of Najdi Arabic Tweets (ReCNAT) is a randomly scraped collection of tweets that carry the geotag location of Riyadh, Saudi Arabia presented in Table 4.3. According to Twitter, geotagging is an optional feature that a user activates to add location information. The geolocation can show location information that is latitude and longitude as well as additional information such as neighbourhood or a specific landmark. Twitter's Help Centre webpage claims that the geolocation feature is accurate and shows precise locations that are found using API (https://help.twitter.com/en/safety-and-security/tweet-location-settings).

Table 4.3 The Reference Corpus of Najdi Arabic Tweets (ReCNAT)

Corpus	**No. of tweets**	**No. of words**
Tweets published by the population located in Riyadh, Saudi Arabia.	42, 160	448,000

4.2.1 Representativeness and sampling

A crucial factor that must be present in any corpus is that of *representativeness*. Biber (1993) classified this notion into situational and linguistic dimension. *Situational representativeness* refers to the range of genres produced by the target sample or authors. After selecting the range of genres, *linguistic representativeness* is then observed which is the range of linguistic structures (lexical and syntactic) that is found in the sample's production. As the NASCoT is a specialized corpus, the focus is exclusively on one genre that is microblogging (i.e., Twitter); hence it doesn't show situational representativeness. It does show, nonetheless, linguistic representativeness as I collected a large set of data per author to observe the ranges of their linguistic patterns.

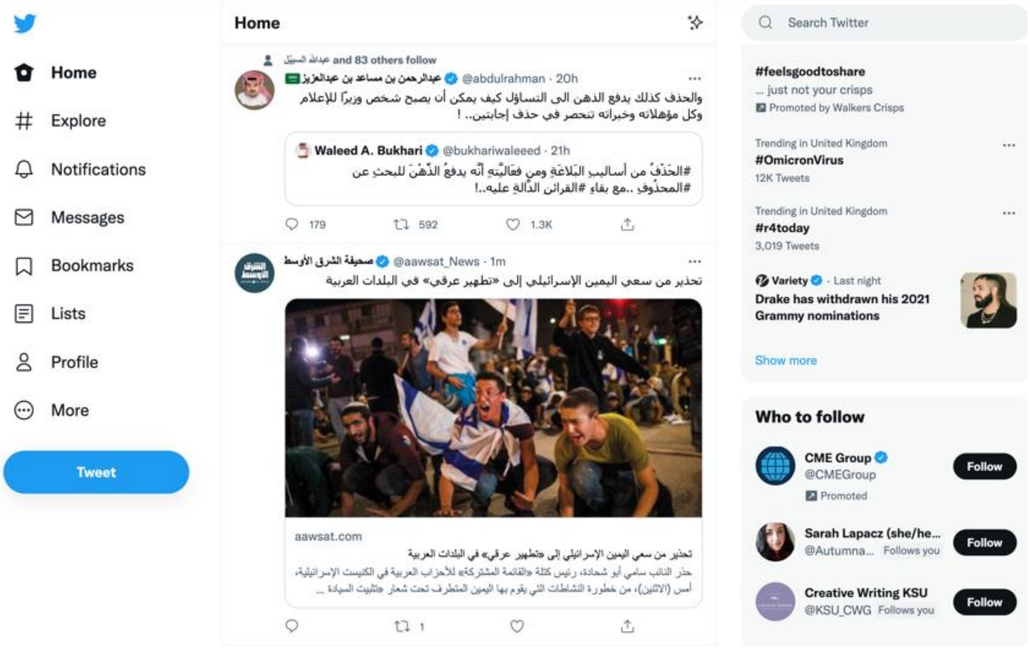
The aim of the corpus in this study is to represent the authors’ idiolectal style. Therefore, selecting a small number of authors with substantial amounts of posts published should encapsulate their style and therefore reveal their online identity performance. In forensic authorship analysis casework, the number of suspected authors is often small; this includes corpus-based studies where the corpus of the relatively small circle of suspected authors is compared to the larger relevant population (Cotterill, 2010; Larner, 2014). In this corpus, there are a number of variables and variants to be represented and considered in order to answer the research questions: linguistic and non-linguistic. The linguistic variables the corpus aims to examine are the Najdi Arabic variables represented in interrogatives, negatives, and deixis and their variants; which will be discussed in detail in section 4.3. One of the aspects to be explored in this research is the authors’ choice between Najdi and Standard Arabic variants. For instance, interrogatives are one of the linguistic variables examined and one of the interrogative variables*, why,* has three different Najdi variants /waraah/, /wišulah/, /leh/ and one Modern Standard Arabic variant /limada/. It will be interesting to see how the authors choose among these variants and if there is any consistency in their choices. That being said, it is hypothesised that each author’s subcorpus should represent an individual idiolectal style in using the linguistic variables and their variants. The other variables would be non-linguistic, which stand for the collective and individual identity implications, and they will be discussed elaborately in Chapters 5 and 6.

There are three main themes that colour the collective identity that I want to address in this research: nationalism, regionality, and gender. Nationalism stands for the authors’ expression of their patriotism or national affiliation to Saudi Arabia and whether they use dialectal features as a way of expressing it. Another variable is the regional identity that would be represented in the authors’ use of dialect-specific features. The presence of the selected features – or lack thereof – could vary across authors’ subcorpora. Gender is also a variable that could contribute to the shape of the authors’ identities and therefore their style. I wanted to explore how the authors represent their gender identity and whether they vary in the way they demonstrate it, if at all. The individual identity themes are finer and more challenging to categorise; therefore, I adopt Herring’s (2007) Faceted Classification Scheme. She proposes this scheme to classify computer-mediated discourse in terms of two variables: medium and situation. According to Herring (2007), the scheme allows the researcher to use its variables flexibly by selecting the ones that are suitable to the type of data and analysis the research requires.

4.2.2 Data collection and processing

The time span of the data collection process was November 2018 to April 2019. This presents the time allocated to data collection at the start of the research: six months. The corpus includes all textual tweets and replies published by the authors from March 1st, 2018 until September 30th, 2019: I aimed to collect a large amount of data to build up the corpus; hence a period of nineteen months represents a large continuous period in each author’s output. Also, I wanted to ensure that the data is relatively recent to capture the authors’ current idiolectal style and make sure that it is a current representation of each author’s language.

The process of data extraction was intricate and technically challenging. Twitter is an online platform that can be accessed using a web browser or by downloading the application on a smartphone. However, both the website and the application provide limited search options which can be extended using their API tools. Figure 4.1 shows a tweet posted on the homepage which is referred to as a *timeline*. The challenge was to extract the data i.e., tweets and transform them into a file that *Wordsmith* (Scott, 2020) can process. After extensive research for a program that can achieve this task, I found a website (https://www.twdocs.com), developed by a Twitter API user. This website gives access to Twitter's advanced search using one's personal account. It enables the user to extract data (tweets, number of followers, likes, and other Twitter-related features) and download them in different file formats (.TXT, .XLSX, .CSV, PDF, and others). While the website proved to be useful, it had its setbacks.



عبدالله السيّل and 83 others follow
عبدالرحمن بن مساعد بن عبدالعزيز @abdulrahman · 20h
والحذف كذلك يدفع الذهن الى التساؤل كيف يمكن أن يصبح شخص وزيرًا للإعلام
وكل مؤهلاته وخبراته تنحصر في حذف إجابتين.. !

Waleed A. Bukhari @bukhariwaleeed · 21h
#الحَذْفُ من أساليبِ البَلاغَةِ ومن فعَاليّتهِ أنّه يدفعُ الذِّهْنَ للبحثِ عن
#المحذُوفِ ..مع بقاءِ #القرائن الدّالةِ عليه..!

179 592 1.3K

صحيفة الشرق الأوسط @aawsat_News · 1m
تحذير من سعي اليمين الإسرائيلي إلى «تطهير عرقي» في البلدان العربية

aawsat.com
تحذير من سعي اليمين الإسرائيلي إلى «تطهير عرقي» في البلدات العربية
حذر النائب سامي أبو شحادة، رئيس كتلة «القائمة المشتركة» للأحزاب العربية في الكنيست الإسرائيلية،
أمس (الاثنين)، من خطورة النشاطات التي يقوم بها اليمين المتطرف تحت شعار «تثبيت السيادة ...

1

#feelsgoodtoshare
... just not your crisps
Promoted by Walkers Crisps

Trending in United Kingdom
#OmicronVirus
12K Tweets

Trending in United Kingdom
#r4today
3,019 Tweets

Variety · Last night
Drake has withdrawn his 2021 Grammy nominations

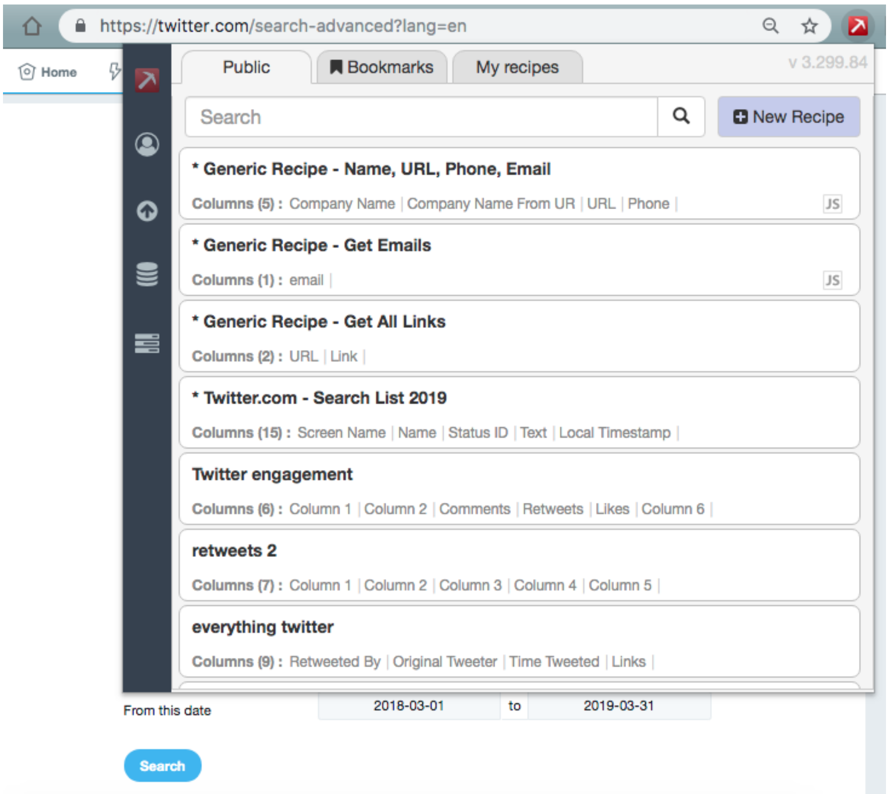
CME Group @CMEGroup
Promoted

Sarah Lapacz (she/he... @Autumna... Follows you

Creative Writing KSU @KSU_CWG Follows you

Figure 4.1: Screenshot of Twitter's timeline

One of the important issues was that it doesn't have a feature to provide a time or date stamp for the retrieved data. It retrieved the most recent tweets without any indication of their dates. It also failed to indicate the number of tweets retrieved, whether it is a reply or an original tweet, which can affect the corpus design and analysis negatively. The search for an alternative program was imperative. Data Miner (https://data-miner.io/) was far superior in its performance and accuracy in information retrieval. It is a Google Chrome extension that is available for free download. Data Miner is an extension that, once installed, can scrape any data required from a webpage and provide it in the shape of a file in any format (.XLSX and .CSV formats are most commonly used). To perform the scraping process, a script of codes in Java or Python is required. One of the advantages of Data Miner is that it provides a range of ‘recipes’ which are Java scripts that give options of types of data to scrape. Figure 4.2 below demonstrates a screenshot of the publicly available Java recipes, each of which scrapes a range of elements such as email, tweet URL, and so forth. The overall advantage of using Data Miner is its accessibility and user-friendly interface to beginners.



https://twitter.com/search-advanced?lang=en

Home

Public | Bookmarks | My recipes

v 3.299.84

Search

New Recipe

* Generic Recipe - Name, URL, Phone, Email
Columns (5) : Company Name | Company Name From UR | URL | Phone |

* Generic Recipe - Get Emails
Columns (1) : email |

* Generic Recipe - Get All Links
Columns (2) : URL | Link |

* Twitter.com - Search List 2019
Columns (15) : Screen Name | Name | Status ID | Text | Local Timestamp |

Twitter engagement
Columns (6) : Column 1 | Column 2 | Comments | Retweets | Likes | Column 6 |

retweets 2
Columns (7) : Column 1 | Column 2 | Column 3 | Column 4 | Column 5 |

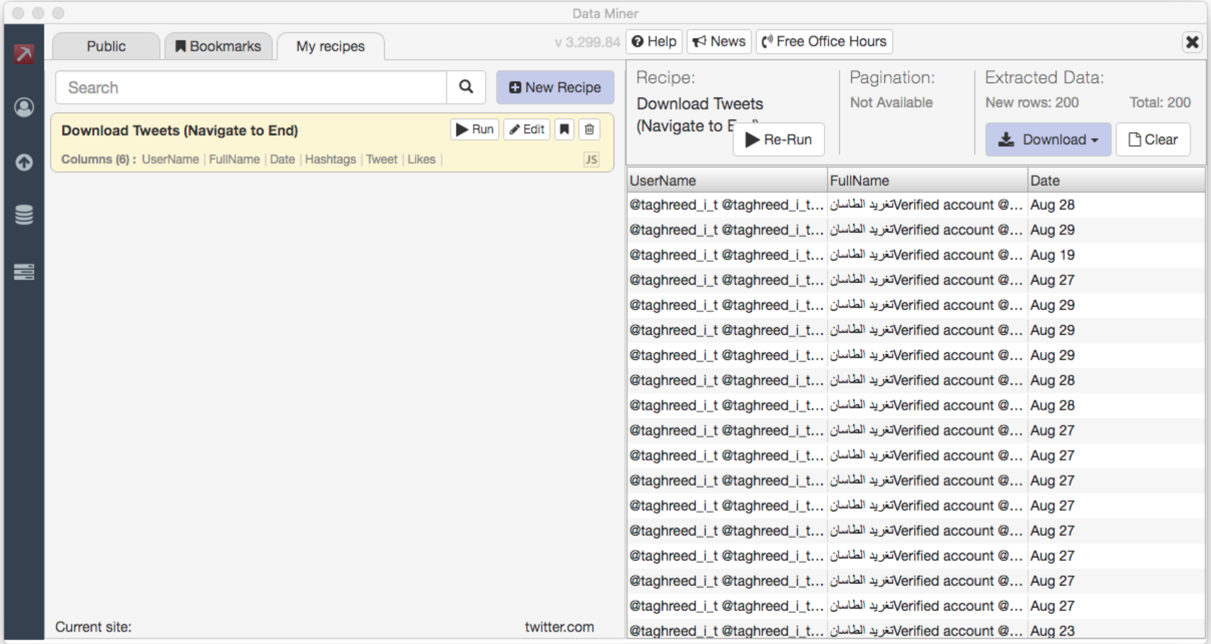
everything twitter
Columns (9) : Retweeted By | Original Tweeter | Time Tweeted | Links |

From this date 2018-03-01 to 2019-03-31

Search

Figure 4.2: Data Miner recipe tab

The recipe selected for this study aimed to retrieve six elements: username, tweet, date, hashtags, number of retweets, and path of the tweets in case it was a reply to someone else’s tweet. Each author’s data was retrieved and exported into an .XLSX file, which was then also converted into a .TXT file. Figure 4.3 shows the scraped data with an option to download as an Excel spreadsheet. As for the content of each file, the tweets are in the shape of an original tweet, a reply, or a quote retweet which is to republish someone else’s text with a comment. All images and videos were excluded to ensure a text-based corpus; however, the images’ links were part of the text as Data Miner does not provide the option to exclude them. Also, all retweets were excluded from the corpus as they do not represent the authorship attributes of the authors.



Download Tweets (Navigate to End)

Columns (6) : UserName | FullName | Date | Hashtags | Tweet | Likes |

Recipe: Download Tweets (Navigate to E…)

Pagination: Not Available

Extracted Data: New rows: 200 Total: 200

Current site: twitter.com

UserName	FullName	Date
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 28
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 29
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 19
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 27
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 29
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 29
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 29
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 28
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 28
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 27
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 27
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 27
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 27
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 27
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 27
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 27
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 27
@taghreed_i_t @taghreed_i_t…	تغريد الطاسانVerified account @…	Aug 23

Figure 4.3: Screenshot of scraped data

4.2.3 The sample

This section addresses the process of selection of the authors, their background which partly contributed to their selection and lays the context of their textual production. Lastly, the ethical considerations that were taken into this study are discussed.

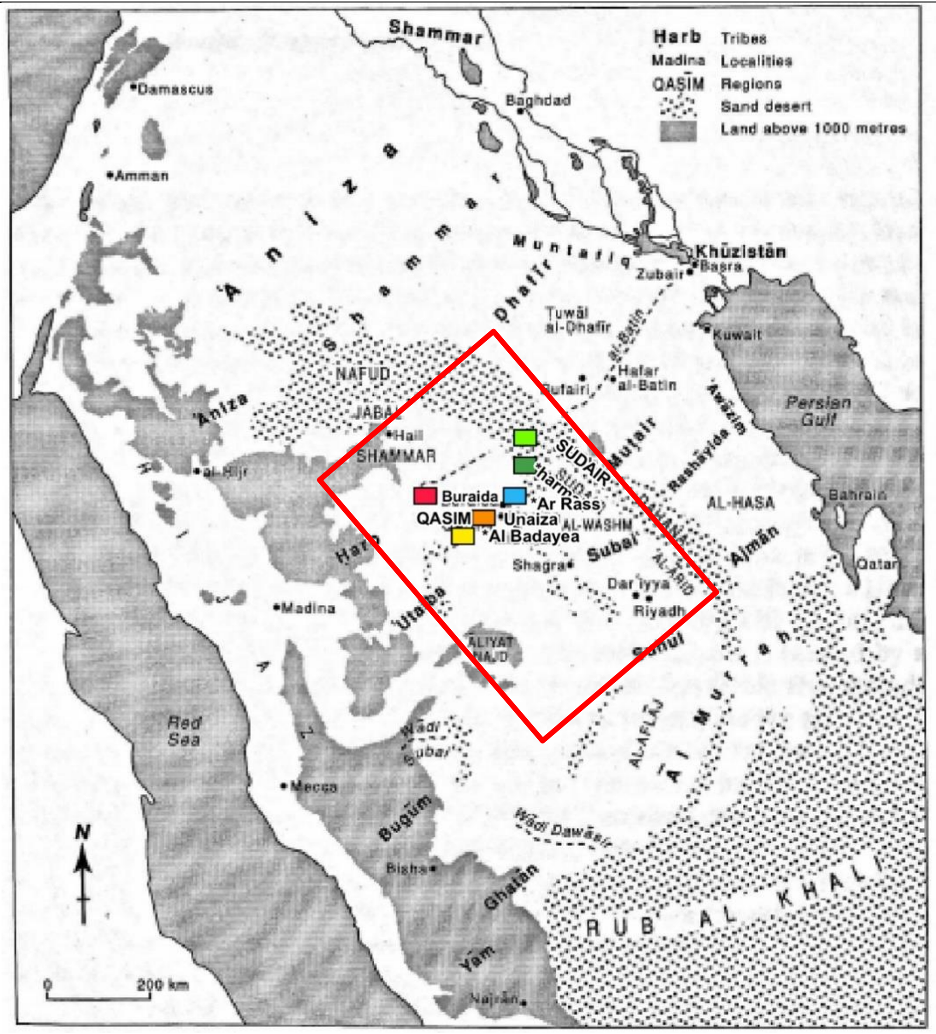
The selection of the sample went through some considerations. In terms of authorship, it is important that the authors selected for this study would use their actual names in their Twitter accounts (they will be referred to as users henceforth). Twitter commonly allows its users to create accounts using alias usernames. In addition, some of these accounts that have such names could be run by a number of authors, not just one. Therefore, all users that had an alias were excluded. To further ensure authorship validity, there was also the verification issue to consider. Twitter verifies some of its users based on their activity and number of followers. This assures that the names and identities of the individuals managing the accounts are authentic. Moreover, in order to produce a corpus of adequate size, the authors required should be prolific users or tweeters. This description fits best to social media influencers because one of the motives that drives them to be active is to maintain, if not increase, their number of followers and thereby their status as high-profile individuals. An additional and crucial aspect is the region; all participants are from Najd which can be evident or tracible from their family names and use of the dialect. I was able to trace the origin of the authors' families using historical references that document the origins of tribes and their migrations which are Aljasser (2001) and Albassam-Altamimi (1818). I was also able to locate most of the authors' birthplaces and where they currently live because of their public profiles. Also, being a Najdi native enabled me to recognise these variables during the selection process. Finally, I wanted to explore the implications of gender and how it would be represented in the data as mentioned in 4.1.1, so I included a balanced number of male and female authors.

The total authors selected for the corpus is thirteen: six males and seven females. Initially I was aiming for a total of 14 authors, seven males and seven females to have a sample that is not tokenistic and both genders are represented equally. However, the ethical considerations of my sample selection restricted having an equal number as planned. Also, due to technical difficulties using the scraping tool, I wasn't able to collect the data for the seventh male author, therefore ending up with only six males and thirteen authors in total. Table 4.4 below shows the authors selected for the study and their brief profiles. Faisal Alabdulkarim started a career in sports journalism but in the last ten years he shifted to being a social media influencer whose main activity is to raise consumer awareness and volunteer with governmental institutes to protect consumer rights and fight counterfeit products. The number of followers is of high significance in the social media influencer community, and while Snapchat is the primary platform for Alabdulakrim's activity with 2.5 million followers, he has 644,000 followers on Twitter. In 2020, he was named as *most influential* by the ministry of information for his coverage for COVID-19 news in Saudi Arabia. He was also awarded a distinction for consumer rights protection. He is a native Najdi who was born and lives in Riyadh city and his family origins are from Sudair, a town in the region of Riyadh.

Table 4.4: The NASCoT authors and profiles

#	Author	Profile	No. of Twitter followers
1	Faisal Alabdulkarim	Social activist – influencer - male	744,400
2	Mansour Alrokibah	Social activist – influencer - male	301,600
3	Abdulrahman Allahim	Lawyer – influencer - male	365,500
4	Ali Alghofaily	Television presenter - male	905,900
5	Abdullah Alsubayel	Influencer - male	1.4 million
6	Abdulaziz Alzamil	Youtuber - male	581,900
7	Taghreed Altassan	Writer – columnist - female	54,000
8	Wafa Alrasheed	Writer – columnist - female	80,600
9	Maha Alwabil	Writer – columnist - female	62,100
10	Arwa Almohanna	Cultural newspaper editor - female	4,753
11	Ghadah Aleidi	Lawyer- Humanitarian - female	181,300
12	Maha Alnuhait	GM of sustainability program - female	47,900
13	Amani Alajlan	Social activist – influencer - female	249,900

Mansour Alrokibah identifies himself as a self-made stockbroker who also pursued a role as a social media influencer starting in 2010. Similar to Alabdulkarim, Snapchat is where Alrokibah's main activity takes place followed by 2.4 million users. He refers to himself as a motivational speaker in his Snapchat and Twitter profiles, but he also posts content related to current local events. The third author is Abdulrahman Allahim, a lawyer who was born in Buraidah, Alqassim and lives in Riyadh city. Allahim uses Twitter to talk about legal issues and occasionally his casework; he has a public Snapchat account where he posts about his daily activity, but it is Twitter where he is most active and posts prolifically. Another male author is Ali Algofaily, a television presenter with a weekly show, director of a marketing and event management agency, and a social media influencer. He actively posts on Twitter, Snapchat and Instagram. The next author, Abdullah Alsubayel, mainly identifies himself as a social media influencer. He is more active on Twitter than on his Snapchat account. The last male author is Alabdulaziz Alzamil, a Youtuber with 279,000 subscribers to his channel about films and television shows. Figure 3.4 shows the geographic origins of all the NASCoT authors with reference to Aljasser's (2001) documentation of their family's geographic origin.



🟥 **Buraida:** Alajlan, A; Allahim, A; Almohanna, A; Alrasheed, W; Alrokibah, M

🟫 **Unaiza:** Aleidi, G; Alsubayel, A

🟨 **Al Badayea:** Alwabil, M

🟦 **Ar Rass:** Alghofaily, A; Altassan, T

🟩 **Sudair:** Alabdukarim, F; Alzamil, A

🟧 **Harmah:** Alnuhait, M

Figure 4.4: Geographic origin of the NASCoT authors (Ingham, 1994)

As for the female authors, there are a total of seven. Taghreed Altassan identifies herself as a newspaper columnist and a businesswoman. She has written three books that are collections of articles about her interests. Altassan has social media accounts on different platforms and was born in Riyadh where she also lives. Wafa Alrasheed is also a newspaper columnist and a businesswoman. She has a PhD in international relations and diplomacy and lives in Riyadh city. Alrasheed participates actively and exclusively on Twitter. Another columnist is Maha Alwabil, who is also involved in volunteer work. She is a co-founder and a member to a number of societies that focus on community service and personal development. She also founded the first initiative to document Saudi women's achievements. Alwabil was born in Riyadh city and runs a number of accounts on social media platforms but is most active on Twitter. The last columnist is Arwa Almohanna, who is a researcher in cultural studies as well. She also founded a salon that discusses women in a philosophical and intellectual context and her digital presence in social media is limited to Twitter. Furthermore, Maha Alnuhait is the general manager of the sustainability program at a telecom company and an impact measurement consultant. Alnuhait is mainly active on her Twitter account. The last two authors have a more prominent presence on social media. Ghadah Aleidi presents herself as a lifestyle influencer, but she is also a legal advisor, a content creator, and a podcast presenter. She is active on Snapchat, YouTube, but mostly on Twitter. Lastly is Amani Alajlan, who identifies herself as a social worker and an activist and uses the social media platforms Snapchat and Twitter to perform those roles.

Due to the nature of the collected data, ethics were highly considered in this study. There is a constant debate about the use of online data and whether to consider it free for public use or not. According to the British Association of Applied Linguistics (BAAL), information that is published on the internet is open for public use and therefore does not require any permission or consent from a certain party (BAAL, 2021). However, it is good practice to consider and abide by the terms and condition of that platform. In their privacy statement, Twitter ensures that all posted tweets are public and available for research:

> Twitter is public and Tweets are immediately viewable and searchable by anyone around the world. We give you non-public ways to communicate on Twitter too, through protected Tweets and Direct Messages. You can also use Twitter under a pseudonym if you prefer not to use your name. (https://twitter.com/en/privacy, 2018)

Therefore, no consent forms were required at the time to perform this study, and this policy remained after their last privacy policy update which was effective in 10 June 2022. There were other ethical considerations in terms of the validity of the data collected and its authenticity. To ensure that the data collected is valid and written exclusively by one identified author, all the participants selected for this pilot study use their full names and their usernames are verified. As previously mentioned, Twitter has a policy in regard to account verification based on a number of variables: most important are that the username's owner is of public interest and they are authentic (https://twitter.com). Also, I had to consider sensitivity in terms of the content of the collected data. Since the posts are public and their authors are identified, I eliminated some of the posts that could potentially be seen as controversial.

4.3 Methodological approaches

This section explains the methodological approaches implemented to investigate the data. Firstly, I briefly explain how corpus linguistics can be utilised in corpus-based or corpus-driven research and the differences between them. Furthermore, I discuss the relationship between discourse analysis and corpus linguistics and how both approaches can complement each other. The last part of this section is dedicated to computer-mediated communication and its approaches in the analysis.

4.3.1 Corpus linguistics as a method

Implementation of corpus linguistics approaches in forensic authorship analysis is a relatively new methodology that was derived from authorship research in religious and literary texts as it proved its effectiveness. This approach was generated as a response to a need to investigate authorship and plagiarism cases that are on the rise (O'Keeffe et al., 2007). Forensic authorship research identifies corpus linguistics performs as a useful instrument by which a linguist can determine the authorship of a text and exclude unlikely candidates (Coulthard, 1994; O'Keeffe et al., 2007; Turell 2010; Cotterill, 2010; Johnson and Wright 2014). A set of texts produced by an author can represent their linguistic behaviour and choices in terms of words and grammatical structures. Therefore, a forensic linguist can make use of corpus linguistic approaches in casework. For example, in a dispute between two authors claiming that they did - or did not - write a piece of text, the role of the forensic linguist would be to assess their writing individually and look for similar patterns in their corpus and in the disputed text. Other cases would involve an unknown text and a number of candidate authors.

Corpus linguistic research can either be corpus-based or corpus-driven. The primary goal in corpus-based research is:

> Attempting to describe the systematic patterns of variation and use for linguistic features and constructs that have been previously identified by linguistic theory. (Biber, 2015: 10)

It highlights and demonstrates the linguistic features and how they are used across registers. Another goal is to identify a linguistic feature that occurs either frequently or rarely in discourse from a particular variety i.e., a dialect. Such research is often based on linguistic expectations, which the analysis findings could confirm or negate. On the other hand, corpus-driven research is inductive; the generated linguistic findings define the course of the research. It describes the evidence the corpus exhibits comprehensively such that the linguistic elements are generated "systematically from the recurrent patterns and the frequency distributions that emerge from language in context" (Tognini-Bonelli, 2001: 87). An extreme version of the corpus-driven approach is that all linguistic elements or features do not have a priori status before analysis. An advantage of a corpus-driven approach is exploring linguistic elements that were not recognised before. This study is a corpus-based study in the sense that the stylometric features, discussed in 4.3, are pre-determined and are selected to represent the Najdi variety. Exploring these features in the corpus will reveal how they are used in discourse and the online identity themes their contexts of use project.

Furthermore, Biber (1993) talks about the process of corpus building which he describes as a cycle, as in Figure 4.5.

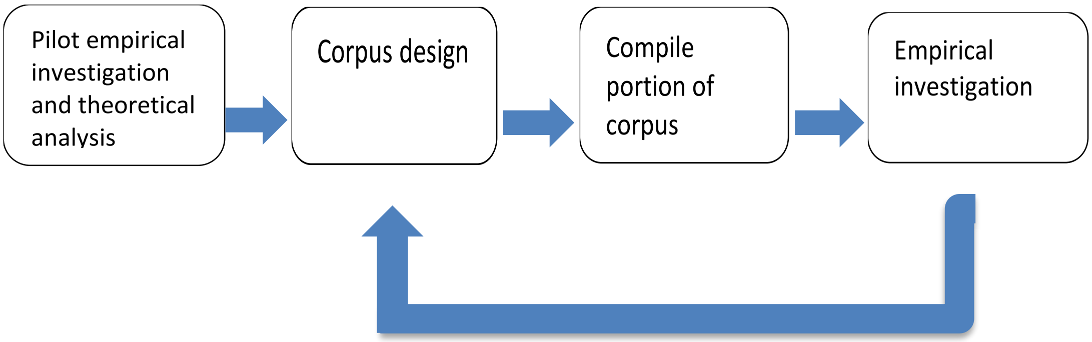

Pilot empirical investigation and theoretical analysis
Corpus design
Compile portion of corpus
Empirical investigation

Figure 4.5: The cycle of corpus design (Biber, 1993)

The design of the corpus starts with a theoretical groundwork that defines its purpose and what it aims to represent. The following step is to compile a pilot corpus; Biber suggests collecting a broad range of genres to record a variation of registers and texts. Also, to grammatically tag items to investigate further. The empirical investigation that follows will either confirm the design of the pilot corpus or require making modifications. This cycle or parts of it can take place continuously until the corpus is finalised (Biber, 1993). I adopted a similar approach to what Biber proposed by compiling a portion of the NASCoT at the pilot study stage, which will be discussed in 4.4. However, I did not design a multi-genre corpus or grammatically tag items. The empirical research confirmed that the pre-determined linguistic feature set can show stylistic variation. However, it also pointed to a number of modifications: to increase the number of authors and explore gender variation by including female authors, and to increase the size of the corpus by extending the time period it covers.

4.3.2 Methodological synergy

Baker at al. (2008) proposed a synergy between corpus linguistics and critical discourse analysis approaches. I propose a similar synergy between corpus linguistics and computer-mediated discourse analysis (CMDA), while also using techniques from stylometry and computational linguistics. I briefly explain the relationship between all these fields and how they are connected and applied as methodological approaches in this research.

In the 1980s and 90s a debate started as to how to classify this mode of communication that combines two basic modalities of language: speaking and writing. Murray (1990) and Ferrara, Brunner, and Whitmore (1991) classified this form of electronic communication as a third mode that is characterised by unique production and reception constraints. The term computer-mediated communication suggests that such communication is a single category or genre of communication characterised by emoticons, abbreviations, and non-standard spelling or how Crystal names it 'netspeak' (Crystal, 2001). Therefore, some linguists recently such as Herring (2007), Georgakopoulou (2011) and Grant and MacLeod (2020) use computer-mediated discourse to refer to a mode of communication that combines the typical features of face-to-face interaction (e.g., informality, immediacy) with features of written communication (lack of paralinguistic cues, physical distance between participants).

Baker et al.'s (2008) paper proposes a synergy between corpus linguistics (CL) and critical discourse analysis (CDA) because "CL needs to be supplemented by the close analysis of selected texts using CDA theory and methodology" (p.297). CL approaches provide a map for the researcher to read the data by indicating frequencies, keyness, and collocations that can point to a linguistic phenomenon, which can be explained by a CDA approach. In other words, CL offers objective quantification of the data that can report findings accurately while CDA provides qualitative interpretation for that data. Wodak in Baker et al. (2008) concludes that this synergy of approaches addresses data quantitatively and qualitatively equally, which mitigates the risk of overlooking any details as they note "pragmatic devices and subtle, coded strategies or concepts cannot be readily analysed through corpus linguistics means." (Wodak, 2007 as cited in Biber et al., 2008: p. 296). This research proposes a synergy between the methodological approaches of corpus linguistics and CMDA that follows the stages presented by Baker et al. (2008). In this synergy, CMDA is employed to capture and interpret the nuances in the authors' subcorpora and also to account for and explain the patterns that the concordance lines and collocates reveal.

The Faceted Classification Scheme proposed by Herring in her paper (2007) aims to account for online communication with a focus on the different medium factors (e.g., synchronicity, anonymous messaging, filtering). It also accounts for situational factors that address variables such as participant structure, tone, code, and others. Both medium and situational factors are designed to "facilitate data selection and analysis in CMD research." (Herring, 2007: 10). The scheme will be addressed elaborately in Chapter 6.

In addition, the synergy in this research is proposed also between corpus-based research and computational linguistics, namely using machine learning tools to process linguistic data. Baker et al. (2008) emphasised the significance of combining quantitative and qualitative data analysis approaches to reflect the nuances of the data objectively. WEKA is a machine learning classification tool that is trained to classify data, usually used in life sciences research (Witten et al., 2006). I introduce the machine learning tool WEKA, used in this research, and explain how it was incorporated in the research with further detail in Chapter 7. The final element in this synergy that connects all these fields together is the stylistic features, which are introduced in further detail in the following section.

4.4 Stylistic feature set

This section is derived from the literature on the notion of stylometric features and how they contribute to forensic authorship analysis presented in Chapter 2. First, I explain the translation and transliteration systems employed for presenting the data for readers of English. I also explain the logic behind the selection of the dialectal features and their standard counterparts. The final part of this section exhibits the stylometric features that I examine in this research.

4.4.1 Translation and transliteration

Considering that the data is in Arabic text, I established a system to translate all examples used in the analysis. In this study I use the Hans Wehr transliteration system for Arabic as used in the Hans Wehr dictionary of modern written Arabic (Wehr, Cowan, and Wehr, 1966). Table 4.5 demonstrates the Arabic letters, their names in Arabic, their symbol in the Hans Wehr system, and their phonetic symbol in IPA. Accordingly, all examples will be transliterated phonetically and translated into English. Example 1 below demonstrates the three-part representation: the post in Arabic text, followed by its transliteration in phonetic symbols and, finally, its translation into English.

Example 4.1: Transliteration system

مااحد تحدث عن مؤامرات وانا شخصيا تأكدت ان الموضوع
"بلاغات".

Ma aḥd tḥadath ʕan mo’amarat w ana šaḵṣian ta’kadt
’n el mawḍooʕ “balaḡat”.

Nobody said anything about conspiracies and I
personally checked that it is about “notifications”.

Table 4.5: Hans Wehr’s transliteration system

Arabic letter	Name	Transliteration	IPA	Arabic letter	Name	Transliteration	IPA

ء	hamza	’	ʔ	س	sīn	s	s
ا	alif	a	a	ش	šīn	š	ʃ
ب	ba’	b	b	ص	ṣad	ṣ	sˤ
ت	ta’	t	t	ض	ḍad	ḍ	dˤ
ث	tha’	ṯ	θ	ط	ṭa’	ṭ	tˤ
ج	jīm	j	ʔ	ظ	ẓa’	ẓ	ðˤ/zˤ
ح	ḥa’	ḥ	ħ	ع	ʕain	ʕ	ʕ
خ	ḵa’	ḵ	x	غ	ḡain	ḡ	ɣ
د	dal	d	d	ف	fa’	f	f
ذ	ḏal	ḏ	ð	ق	qaf	q	q
ر	ra’	r	r	ك	kaf	k	k
ز	za’	z	z	ل	lam	l	l
م	mīm	m	m	ن	nūn	n	n
هـ	ha’	h	h	و	waw	w	w
ي	ya’	y	j				

4.4.2 Selection of features

Since I am using a corpus-based method, this suggests that I will use a top-down approach by selecting the stylometric features. Turell (2010) discussed the idea of markedness and emphasised Jakobson’s (1956) understanding of it as a binary system in relation to the existence or absence of a mark. Turell finds Jakobson’s proposal accommodating to authorship in a forensic context in which “an individual’s ‘idiolectal style’, has to do precisely with Jakobson’s proposal that ‘the marked form conveys more precise, specific and additional information than the unmarked form’” (Turell, 2010: 219). Furthermore, Turell (2010) adopts the concept of saliency that derives from discourse analysis and corpus linguistic approaches, which refers to the words that are statistically prominent when comparing two corpora or a subcorpus that is compared against the totality of a corpus. With reference to these notions of markedness and linguistic saliency, this section demonstrates the dialect-specific stylometric features examined in this study. The selection was achieved with reference to sociolinguistic studies that report and confirm some linguistic features that are specific to the Najdi Arabic dialect (Abboud, 1964; Ingham 1994; Watson, 2002; Alothman, 2012; Badawi, 2012; Binturki, 2015). These include the plural personal pronoun, demonstratives (deixis), interrogatives, and negatives.

The Najdi Arabic (NA) personal pronouns are all derivatives of the standard ones in Modern Standard Arabic (MSA). There are nine forms, all being gender-specific except for the first person (both singular and plural forms). As opposed to the MSA paradigm, the NA pronouns do not acknowledge dual pronouns (Alothman, 2012). Table 4.6 shows the personal free pronouns in NA.

Table 4.6: Personal free pronouns in NA and MSA

Person	**Gender**	**Transliteration**	**NA forms**	**MSA forms**
1st Singular	-	‘ana	أنا	أنا
2nd Singular	Masculine	∂nta	انت	انت
	Feminine	∂nti	انتِ	انتِ
3rd Singular	Masculine	huw	هو	هو
	Feminine	hiyy	هي	هي
1st Plural	-	**ḥna**	احنا / حنا	نحن
2nd Plural	Masculine	∂ntum	أنتم	أنتم
	Feminine	∂ntum / ∂ntin	أنتم / أنتن	أنتن
3rd Plural	Masculine	humm	هم	هم
	Feminine	humm / h∂nn	هم / هن	هنّ

Being derivatives of Standard Arabic forms, both free and bound personal pronouns are used in Najdi Arabic and in most Arabic dialects. The only exception to the case is the 1st person plural pronoun /ḥna/: therefore, it will be the only form included in the examined stylometric features

Demonstratives in MSA and in NA agree with the subject in number and gender. As is the case in pronouns, NA dropped the dual forms in demonstratives as well. Moreover, the demonstratives conform to the notions of distance and space in deixis. The proximal and distal spatial deictic expressions can be found in other Arabic dialects as shown in Table 4.7.

Table 4.7: Deixis in NA and MSA

Deixis	English form	Transliteration	NA forms	MSA forms
Proximal (this / these)	This (Singular Male)	ða / haða	ذا / هذا	هذا
	This (Singular Female)	ði / hãði	ذي / هذي	هذه
	These (Plural Male)	**ðõl / haðol**	ذول / هذول	هؤلاء
	These (Plural Female)	**ðõl / haðol**	ذول / هذول	هؤلاء
Distal (that / those)	That (Singular Male)	**ðak / haðãk**	هذاك/ذاك	ذاك
	That (Singular Female)	**ðïk / hãðïk**	هذيك/ذيك	تلك
	Those (Plural Male)	**ðõlãk / haðõlãk**	ذولاك / هذولاك	أولئك
	Those (Plural Female)	**ðolïk / haðlïk**	ذوليك / هذوليك	أولئك
Temporal	Now	alḥin	الحين	الآن
	Later	bʕdain	بعدين	لاحقآ
	Today	alyom	اليوم	اليوم
	Tomorrow	bukrah/bukra/bakir	بكره / بكرا / باكر	غدًا
Spatial	Here	hina	هنا	هنا
	There	hinak	هناك	هناك
	Near	girib	قريب	قريب
	Far	biʕid	بعيد	بعيد

While most of the interrogatives are derivatives of MSA and can be found in other Arabic dialects, the interrogative forms of *what* and *why* are exclusive to NA (Ingham, 1996; Alothman, 2012). It is also useful to point out that all Najdi variants for *why* (i.e., /waraah/ /wišulah/ /leh/) are similar and can alternate with each other in most contexts. I included the forms that are examined in this research in Table 4.8, although *where* /wen/ and *how* /kef/ are not Najdi-specific. Nonetheless I incorporated them to collect comprehensive findings and to explore the contexts of where the authors would use them.

Table 4.8: Interrogatives in NA and MSA

Interrogative	Transliteration	NA forms	MSA forms
What	**/wiš/ /wišu/**	وش – وشو	ماذا

Why	**/waraah/ /wišulah/ /leh/**	ليه – وراه – وشو له	لماذا
Where	/wen/	وين	أين
How	/kef/	كيف	كيف

With five Najdi-specific forms (/wiš/, /wišu/, /waraah/, /wišulah/, and /leh/) and only two non-exclusive ones, /wen/ and /kef/, the findings of the research are not compromised but rather wholesome.

The grammatical structure of negation in NA does not vary from that of other Arabic dialects such as Kuwaiti, Jordanian, Egyptian, and so on. Nevertheless, it is distinguished with morphological markers (Abboud, 1964). Negation morphemes in NA are classified into three: anaphoric, verbal, and non-verbal morphemes. The only morphemes that are derivatives of MSA and can be found in other dialects are the anaphoric/verbal morphemes /la/ and /ma/. The non-verbal morphemes are Najdi-specific features as demonstrated in Table 4.9 below (Ingham, 1996; Binturki, 2015).

Table 4.9: Negatives in NA and MSA (Binturki, 2015)

Negatives	Translation/Transliteration	NA forms	MSA forms
Anaphoric **/la/**	Negative imperative /laḥad/ *No one*	لحد	لا أحد
Verbal **/ma/**	Negative declarative /maḥad/ *No one*	محد	ما من أحد
Pseudo-verb predicates **Negative marker** **/mub/**	1st person ***/manab/ /maneeb/*** *I'm not*	منب – مانيب	لستُ / ما أنا بـ
	2nd person ***/mantib/ /manteeb/*** *You're not*	منتب – ماتيب – مانتب	لستَ / ما أنت بـ
	3rd person ***/muhub/ /miheeb/ /mahoob/ /maheeb/*** *He / She isn't*	مهب – مهيب ـ ماهوب – ماهيب	ليس / ليست

4.5 The pilot study

This section discusses the pilot study, conducted in the first year of this research, and reports the preliminary research questions, the methodology, and the preliminary findings. The pilot study aimed to answer the following research questions:

1. Using the social media platform Twitter, can a specialized corpus of Najdi Arabic tweets show indications of individual idiolectal style?
2. To what extent can gender affect the idiolectal style of a speaker/writer and how is it represented in their online identity?

As indicated earlier in 4.2.1, the process of corpus building is cyclic and having a pilot corpus is a step in that cycle (Biber, 1993). The initial NASCoT corpus consisted of a collection of tweets by six Najdi Saudi males. The time span of the data collection started in November 2018 and until April 2019. The time span of the collected tweets was from March 1st, 2018 until March 31st, 2019. The chosen time period enabled me to collect an adequate amount of data to build up a corpus that is also relatively recent. The number of tweets collected for this corpus was 16,281 all produced by the participants in the form of a text. Table 4.10 provides the number of tweets per author within the said timespan. The reference corpus used for the pilot study was compiled by Alshutayri and Atwell (2017). It is comprised of phrases collected from tweets in the Gulf Arabic (GA) dialect. This is a dialect spoken in the Arabian Gulf region which includes Kuwait, Bahrain, Qatar, United Arab Emirates and Eastern province in Saudi Arabia.

Table 4.10: Pilot NASCoT breakdown in terms of tweets and words

Author	**No. of tweets**	**No. of words**
Faisal Alabdulkarim	1904	40,206
Mansour Alrokibah	1351	15,433
Abdulrahman Allahim	4652	70,659
Ali Alghofaily	1952	28,700

Abdullah Alsubayel	4508	25,154
Abdulaziz Alzamil	1914	23,246
Total	**16,281**	**203,398**

The reference corpus consists of 65,536 tweets with a total of 658,893 words. It will be referred to henceforth as the Gulf Arabic Corpus of Tweets (GACoT).

The findings of the pilot study revealed that while the examined style markers consistently appeared across the NASCoT subcorpora, each author used them in their own way. Findings showed that 1st person plural pronoun, distal-deixis personal demonstratives, interrogatives, and negatives are salient features of NA. The sample's subcorpora also revealed that authors either opt for salient style markers or refrain from using them; such idiolectal choices, as a result, convey their sociocultural orientation (Bucholtz and Hall, 2005). The findings of this pilot study further supported Turell's (2010) conclusion that idiolectal style does convey sociolinguistic information.

Being classified as a microblog, Twitter does allow the authors to become innovative in their use of language. The data of Alrokibah, Alzamil, and Allahim showed instances of orality and vernacular writing which indicate that social media platforms give authors the space to become creative and encourage individuality of idiolectal style. This indicates that social media as a genre can be a platform for authors to practice their individuality through their idiolectal style and as a result project their online identities stylistically. This finding agrees with Androstopolous' (2007) notions on innovation in computer-mediated communication and Page's (2015) remarks on the role of social media in the construction of an online identity.

The sample used for this pilot study was exclusively male; they exhibited salient dialectal features in their idiolectal style either on a lexical or morphological level. These features are indicators of their ethnographic identities and reveal their epistemic as well as ironic stances and how they position themselves as Najdi males. It would be interesting to explore gender as a variable as far as idiolectal style is concerned and whether male and female authors vary in their use of the features. The findings of 1st person plural pronoun /hna/ *we*, personal deictic demonstratives, interrogatives, and negatives show that interaction is key in using these features thereby enabling the emergence of these stances. This conforms with Bucholtz and Hall's (2005) principles of emergence, indexicality, and relationality. As for the reference corpus, the GACoT is a collection of several dialects spoken in the Gulf region, which might not be the best representation of the relevant population that use the Najdi variety. Therefore, another reference corpus is required.

4.6 Conclusion

This chapter breaks down the design of the NASCoT corpus and the reference corpus ReCNAT and the logic behind their design and structure. Moreover, it explains the collection and preparation of the data and introduces the authors' profiles. It also introduces the corpus, the methodological approaches, and the stylometric features that I use in the analysis. Finally, it reports the pilot study and its findings. The following chapter introduces the first part of the data analysis: corpus-based stylistic analysis of lexico-grammatical features.

Chapter 5
LINGUISTIC PATTERNS & IDIOLECTAL STYLE

5.1 Introduction

Language, dialect, and idiolect are aspects of language use that all language users share. We speak a language or two, choose standard or non-standard dialect, and have personal social experiences that produce what Coulthard (2004: 31) describes as our 'own distinct and individual version', our idiolect. Idiolect is a theoretical construct, one that was defined and explained in Chapter 2, but there is room to research what idiolect is, especially in the forensic context. In the case of Arabic speakers, there is a huge range of dialects across the Arabic speaking world from Saudi Arabia to Morocco, and the situation within Saudi Arabia is also complex. The dialects in Saudi Arabia developed partly due to their geographic location; the central region alone has four variations of Najdi Arabic: south, north, central, and Badawi Najdi (Ingham, 1994). However, the socioeconomic changes that took place in the past few decades assimilated the differences between them making the variety spoken in the capital ArRiyadh the most dominant (Aldosaree, 2016). Prestige is also the reason why the Najdi variety on a national level is perceived as a dominant variety compared to other Saudi dialects. This is because the Najdi community is a majority group in the country; the prestige is also enhanced because their variety is spoken by the royal family (Al-Essa, 2009; Aldosaree, 2016; Alaiyed, 2018; Alaiyed and Alfalig, 2022). Another interesting feature is the distinctiveness of the variety compared to other Saudi dialects. Research shows that there is no similarity between NA and any other dialect outside Saudi Arabia, unlike the other dialects that can be similar to neighbouring dialects such as Egyptian to Hijazi and Yemeni to Jizani (Alhazmi and Alfalig, 2022). The central thesis of this research is to discover whether the Najdi dialect and its key syntactic variants, distinctive interrogatives, negatives, and deictic expressions most of which are not found in MSA, are useful in distinguishing between authors who share this dialect. This attempts to answer a specific question in forensic linguistics that would enable us to identify individuals based on their unique combination of dialect features that set them apart from other users of the dialect. This chapter, therefore, asks the following question:

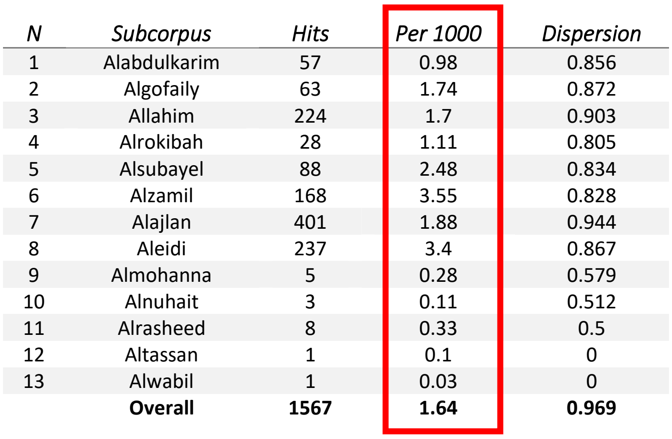
1) Can the Najdi Arabic Specialised Corpus of Tweets (NASCoT) help in identifying the idiolectal style of thirteen Najdi authors when compared to the relevant population?

The chapter starts with a section that briefly discusses the theoretical groundwork of identity and extensively goes over the identity approach (Bucholtz and Hall, 2005), which is one of the approaches implemented in the analysis. The following section explains how the data is represented and the scope of the stylistic analysis that will be carried out. The third section reports on the main patterns the NASCoT reveals and discusses those patterns in terms of the main dialect features' categories: interrogatives, negatives, and deixis. The fourth section discusses the implications of these patterns and how they contribute to the construction of the authors' online identities. I discuss both group membership and individual identity themes. Lastly, I would like to note that material in this chapter has been presented at the fourth Arabic Linguistics Forum (Alif 2020) and published as part of the proceedings at the seventeenth International Conference in Natural Language Processing (AlAmr and Atwell, 2020).

5.2 Representation and discussion of data

In Chapter 4 I introduced the thirteen authors and the linguistic variables I plan to investigate: interrogatives, negatives, and deixis (person, time, and space) in their NA and MSA variants. Initially, I extracted the raw frequencies, rates, and dispersion of each feature per text as shown in Table 5.1. However, the *hits* or raw figures do not reflect the frequency of the feature accurately and they are superfluous. The rate i.e., *per 1000*. *Dispersion* is normally helpful in showing how a feature is distributed across a text that is rather longer than tweets (e.g., a series of letters or a play). Therefore, it too was not considered in the analysis. All forthcoming figures and tables will demonstrate the frequency rates per 1000 words of each item per author. Moreover, this research is corpus-based in the sense that the data and their findings are extracted from a corpus. This chapter is particularly focused in terms of corpus linguistics approaches because it explores the corpus data using *Wordsmith* tools (Scott, 2020). The concordance lines provide the number of occurrences of the linguistic features under investigation and more importantly their contexts.

Table 5.1: Frequency and dispersion of Najdi interrogative /wiš/ across all 13 authors



N	*Subcorpus*	*Hits*	*Per 1000*	*Dispersion*
1	Alabdulkarim	57	0.98	0.856
2	Algofaily	63	1.74	0.872
3	Allahim	224	1.7	0.903
4	Alrokibah	28	1.11	0.805
5	Alsubayel	88	2.48	0.834
6	Alzamil	168	3.55	0.828
7	Alajlan	401	1.88	0.944
8	Aleidi	237	3.4	0.867
9	Almohanna	5	0.28	0.579
10	Alnuhait	3	0.11	0.512
11	Alrasheed	8	0.33	0.5
12	Altassan	1	0.1	0
13	Alwabil	1	0.03	0
	Overall	**1567**	**1.64**	**0.969**

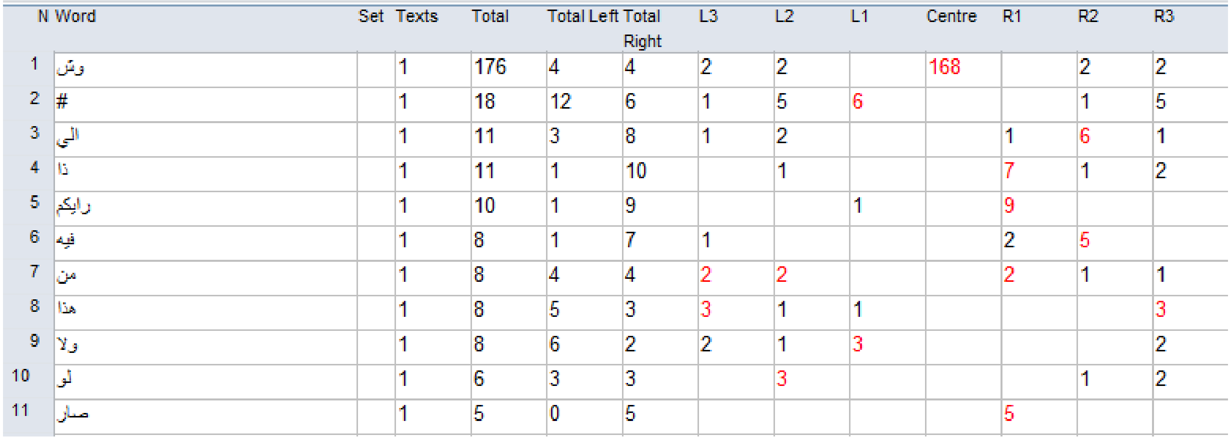
This tool is especially useful in the case of features that overlap between MSA and NA (e.g., temporal and spatial deictic expressions) where the context indicates which variety the author is using, which is usually recognised by looking at neighbouring lexical features such as verbs, interrogatives, or negatives. Also, using the collocates tool helps detect if there is a linguistic pattern or idiolectal co-selection (Coulthard, 2004) that distinguishes an author. The plot tool shows how frequently a feature occurs and its dispersion across the subcorpus, which can indicate the author's consistency in using the feature. More important is the keyword tool which enables one to extract each author's keywords compared against the wordlists of the other authors and the relevant population. The forthcoming analysis demonstrates the results that were found using these tools.

Furthermore, I discuss the findings through two angles; the first is looking at the frequency rates that shape a particular pattern in the three feature categories: interrogatives, negatives, and deixis. The other angle is evaluative language and stance that the aforementioned features communicate. Also, the implications of collective and individual identities are revealed in the authors' attitudes and emotional reactions in their posts, which can either be on a semantic or a pragmatic level. The notion of identity will be explored in Chapter 6 elaborately, but there are instances in the current analysis where it will be discussed when relevant.

5.3 Instances of idiolectal systems

Before going through the idiolectal and linguistic patterns of the NASCoT, I dedicate this section to discuss some of the authors the corpus revealed that exhibit interesting idiolectal systems when compared against the sample.

. Alzamil's subcorpus has the highest frequency rates in interrogatives, specifically the Najdi variants /wiš/ *what*, /leh/ *why* and /wen/ *where*. Similar to Aleidi, his subcorpus shows a preference to use Najdi interrogatives to initiate conversation with his audience. Moreover, his subcorpus shows that he uses Najdi interrogatives as part of his negative evaluative and sarcastic discourse. Alzamil uses the NA variant /wiš/ frequently, mainly to pose questions for his audience that are related to his work. The topmost collocates as shown in Figure 5.1 are وش رايكم/wiš raykum/ *What do you think?* and وش صار/wiš ṣar/ *What happened?*. Another frequent collocate is وش ذا/wiš ḏa/ *What is this?* which he uses to express criticism.



N	Word	Set	Texts	Total	Total Left	Total Right	L3	L2	L1	Centre	R1	R2	R3
1	وش		1	176	4	4	2	2		168		2	2
2	#		1	18	12	6	1	5	6			1	5
3	الي		1	11	3	8	1	2			1	6	1
4	ذا		1	11	1	10		1			7	1	2
5	رايكم		1	10	1	9			1		9		
6	فيه		1	8	1	7	1				2	5	
7	من		1	8	4	4	2	2			2	1	1
8	هذا		1	8	5	3	3	1	1				3
9	ولا		1	8	6	2	2	1	3				2
10	لو		1	6	3	3		3				1	2
11	صار		1	5	0	5					5		

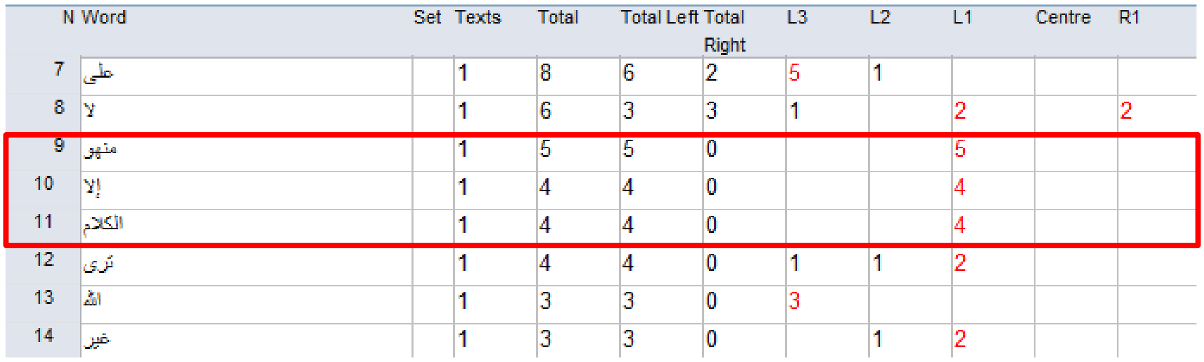
Figure 5.1: Collocates of Alzamil's use of /wiš/

Furthermore, Alzamil's subcorpus shows that he uses /leh/ *why* in a range of stances. In some instances he uses it to take a sarcastic stance such as ليه أبرر من الأساس /leh abarrir lik min al'asas/ *Why should I explain to you in the first place*, and زعلان ليه يعلق/zaʕlan leh yʕalliq/ *I am upset why he comments.* Other occurrences are more epistemic where he asks his addressee for information like ليه ما نتمرس بالانجليزية/leh ma nitmarras bil ingliziya/ *Why don't we practice our English?* and هاللبس ليه مو موجود بالمتاجر /hallibs leh mu mawjod bil matjar/ *Why is this outfit not available at the store?*.

The last NA interrogative is /wen/ *where*, which the examination of the concordance lines shows that Alzamil uses it for two functions. The first is epistemic where he asks for information such as وين أبي أفتحه؟ /wen abi aftaḥah/ *Where? I want to open it*, هالشي وين طلع قبل؟/hal šai wen ṭalaʕ gabil/ *Where did this appear before?*, and وين فيه هذي بأي مدينة؟/wen fih haḏi b'ai madina/ *Where is this in which city?*. The second is evaluative in which asks for the sake of the argument and to defend his position as in وين الحرق معليش/wen alḥarg maʕlaiš/ *Where is the spoiler?* and وين هالكلام؟ ماهو صحيح/wen halkalam mahu saḥiḥ/ *Where does it say that? That's not true*.

Alazamil's subcorpus reveals a number of implications. First, his preference to use Najdi interrogatives over standard ones highlights his regional identity prominently. He uses /wiš/ to initiate interaction with his audience about his interests, mostly being the film industry and television shows. Lastly, his use of interrogatives occasionally in some of the replies is to express his criticism or sarcasm about some people and/or ideas.

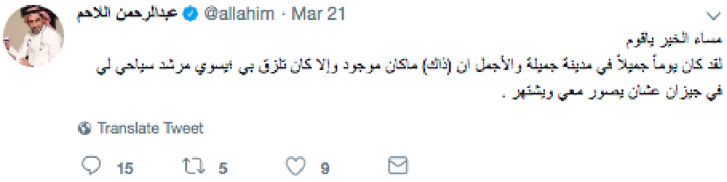
Allahim's subcorpus revealed he is the most frequent user of person deixis and the plural pronoun /ḥna/. The highest frequency rates are for the proximal masculine singular variant /d̠a/ and its feminine counterpart /d̠i/. The frequent use of these deictic expressions is part of Allahim's style in creating indexes. The first feature is the proximal masculine singular /d̠a/. The collocates in Figure 5.2 (lines 9-11 highlighted in the red rectangle) show the most frequent collocates which are the phrases منهو ذا/minho d̠a/ *who's that* and إلا ذا/illa d̠a/ *except that*, in which the expression refers to a person. Another frequent pattern that refers to objects is the expression الكلام ذا/alkalam d̠a/ *things like that*. In all of these phrases, Allahim is creating attitudinal sarcastic stances through these indexical references.



N	Word	Set	Texts	Total	Total Left	Total Right	L3	L2	L1	Centre	R1
7	على		1	8	6	2	5	1			
8	لا		1	6	3	3	1		2		2
9	منهو		1	5	5	0			5		
10	إلا		1	4	4	0			4		
11	الكلام		1	4	4	0			4		
12	ترى		1	4	4	0	1	1	2		
13	الله		1	3	3	0	3				
14	غير		1	3	3	0		1	2		

Figure 5.2: Allahim's collocates of /d̠a/

Furthermore, Allahim also uses the masculine distal /d̠ak/ to create sarcastic references. Figure 5.3 shows an original post that demonstrates his stance. Allahim uses distal deixis such as /d̠ak/ and /d̠olak/ consistently in his sarcastic discourse. Furthermore, Allahim's subcorpus also reveals that the proximal feminine /d̠i/ is used more frequently than the distal form /d̠ik/. Moreover, the collocates show that he uses it in a pattern /kuḏ ḏi/ when addressing a male and its feminine counterpart /kuḏi ḏi/.



عبدالرحمن اللاحم @allahim · Mar 21

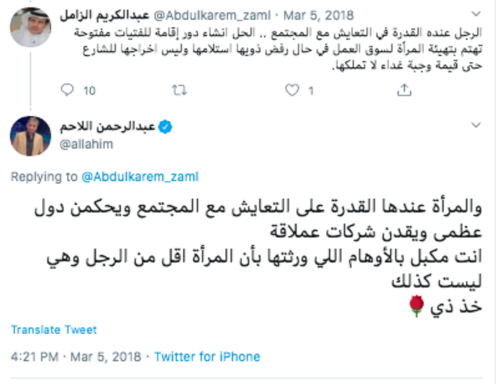
مساء الخير ياقوم
لقد كان يوماً جميلاً في مدينة جميلة والأجمل ان (ذاك) ماكان موجود وإلا كان تلزق بي ايسوي مرشد سياحي لي
في جيزان عشان يصور معي ويشتهر .

/masaa’ ilk̲air ya qawm laqad kaan yawman jameelan fee madina jameela w alajmal in (d̲ak) ma kan mawjod wila kan tlazzag bi ysawi muršid syaḥi li fee jizan ʕšan yṣawer mʕi w yištihir/

Good evening people It was a beautiful day in a beautiful city and the most beautiful part is (that he) was not there otherwise he would have followed me around acting like a tour guide in Jizan so he can take photos with me and become famous.

Figure 5.3: Screenshot of Allahim’s use of /d̲ak/

The deictic expression followed by a flower emoji, which Allahim uses as marker to end a conversation. Figure 5.4 demonstrates the collocate below.



عبدالكريم الزامل @Abdulkarem_zaml · Mar 5, 2018

الرجل عنده القدرة في التعايش مع المجتمع .. الحل انشاء دور إقامة للفتيات مفتوحة
تهتم بتهيئة المرأة لسوق العمل في حال رفض ذويها استلامها وليس اخراجها للشارع
حتى قيمة وجبة غداء لا تملكها.

عبدالرحمن اللاحم
@allahim

Replying to @Abdulkarem_zaml

والمرأة عندها القدرة على التعايش مع المجتمع ويحكمن دول
عظمى ويقدن شركات عملاقة
انت مكبل بالأوهام اللي ورثتها بأن المرأة اقل من الرجل وهي
ليست كذلك
خذ ذي 🌹

4:21 PM · Mar 5, 2018 · Twitter for iPhone

/wa almar’a ʕndha alqudra ʕla attaʕayuš maʕ almujtamaʕ w yaḥkumn duwal ʕuẓma w yaqudon šarikat ʕimlaqa

ant mukabbal bil awham illy warat̲taha b’an almar’a aqal min arrajul w hiya laysat kad̲alik

k̲ud̲ d̲i (F)/

and women are capable of adapting to society, govern great countries, and run gigantic companies.

you are imprisoned by the delusions you inherited that a woman is inferior to a man and she is not

take this (F)

Figure 5.4: Screenshot of Allahim’s collocate /kuḏ ḏi/

As noted earlier, the sarcastic indexicality also includes the use of the plural /ḏolak/, which is another frequent feature in his subcorpus. Instances such as ما يصيحون مثل ذولاك/ma ysihon mitl dolak/ *They don’t cry like those*, لا تعلم ذولاك/la tʕalmin dolak/ *Don’t tell them*, and إن فازوا علينا ذولاك ولاتعادلوا/in fazo ʕalaina dolak wila taʕadalo/ *If those* *win or tie with us* illustrates Allahim’s humour by describing himself and the addressee(s) as members of the in-group and the other (i.e, *those*/*them*) as the outgroup. These indexes he uses throughout his tweets are part of his attitudinal stance taking. He also uses the plural pronoun /ḥna/ as one of his indexical tools. In his comment on a tweet in Example 5.1, he uses /ḥna/ to refer to himself and the fans of Alnassr football team who won the local tournament. This attitudinal sarcastic stance is Allahim’s response to someone’s tweet that says:

Example 5.1: Allahim’s use of /ḥna/

هذانا حنا لابسين بشوتنا من يوم التتويج ولاتكبرنا على أحد .. نمشي بسكينة و وقار
ونلين بأيدي الناس ... اللحية مثل البشت تعطي وقار ولقب وإذا الشخص مم قد اللحية
لا يسيئ لها ولأهلها

/alliḥya mit̲l albišt tiʕṭi waqar wa laqab wa id̲a išak̲ṣ mub gad allḥiya la ysee’ laha wl’ahlha/

/had̲ana ḥna labsin bšotina min yom attatwij wa la tkabbarna ʕla aḥd .. namši bisakinah wa waqar wa nilin b’aydi innas ../

Author: a beard is like a bisht that gives who wears it grace and prestige and just because someone cannot commit to it doesn’t give them the right to offend those who can.

Allahim: here we are wearing our bishts since winning the cup and didn’t look down at anyone .. walking calmly and humbly (In his comment to a tweet)

To sum up, Allahim uses the pronoun /ḥna/ to take epistemic stances to position and identify himself in a number of political contexts. Allahim’s stances reveal how he identifies himself as a Najdi, a Saudi, an Arab, and a Muslim. These stances show macro-level information about him and how he positions himself towards this data. He uses a range of person deictic expressions to create indexical references that build up throughout his tweets. These indexical references are part of his sarcastic discourse towards the religious conservatives in the Saudi community.

5.4The patterns

Initially, The NASCoT shows a major pattern: in each linguistic category, there are authors who use some features frequently and those who rarely use them if at all as shown in Table 5.2. The other major pattern is that in every category there is a variety that is used more frequently than the other (e.g., Figure 5.5 and Figure 5.2). I will address and revisit these aspects of the pattern in every category as well as discussing the authors that stand out in these categories in terms of frequency rates.

5.4.1 Interrogatives

The first category examined is interrogatives, and Figure 5.1 shows the frequency rates of the interrogative variables in both varieties: the MSA variants are marked in green and the NA variants marked in blue. The horizontal axis shows the frequency rates for each NASCoT author in each variety. The first overall pattern I notice is that the NA variants are more frequent than the MSA variants. Most of the authors, prominent examples are Alzamil, Aleidi, Alajlan, and Alsubayel, use the Najdi interrogative variants more frequently than their MSA counterparts. The subcorpus of Alzamil shows that he uses the NA interrogative variants almost exclusively, and that is similarly the case for Aleidi, Alrokibah and Alsubayel. There are, however, exceptions to this pattern: Almohanna, Alnuhait, and Alrasheed use the MSA variants more than NA. Another interesting thing the NASCoT reveals is that Alabdulakrim uses both variants almost equally, while other authors (Altassan and Alwabil) rarely use interrogatives in either variant. I discuss the authors with the highest rates in further detail and explore their subcorpora by examining concordance lines, collocates, and patterns using *Wordsmith* (Scott, 2020).

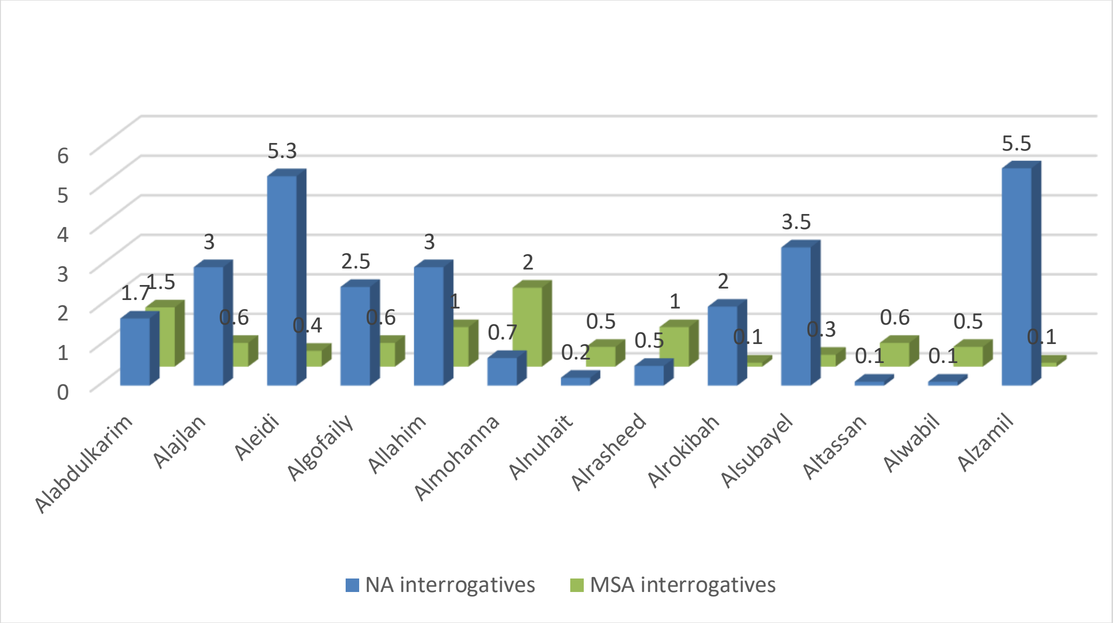


Figure 5.5: Interrogatives in NA and MSA (rates per 1000 words)

The subcorpora of Alzamil and Aleidi strike the highest rates in using interrogatives, especially the NA variants. Referring back to their introduction in Chapter 4, both authors have a strong presence in social media (Alzamil is a Youtuber and Aleidi is a lawyer with a public Snapchat account). This background information can help explain the frequent use of interrogatives as we look into their tweets. Ghadah Aleidi's subcorpus has the second highest overall frequency rates in interrogatives. Initially it reveals that her frequent use of interrogatives is to connect to her audience and the preference to do so using Najdi variants instead of the MSA ones. The most frequent forms Aleidi uses are the Najdi variant /wiš/ followed by /leh/. The concordance lines show that Aleidi uses /wiš/ to ask her audience questions relevant to the topic of her radio show like وش بتشترين/wiš btistirin/ *What will you buy?*. Examples 5.2 and 5.3 demonstrate the interrogative being used to pose questions related to current local affairs using Najdi dialect.

Example 5.2: Aleidi's use of /wiš/

بتسوقين ولالا واذا سقتي وش هي سيارتك ؟

/bitsogin wila la w iḏa sigty wiš hi sayyartk?/

Are you going to drive or not? And if you are, what is your car?

In both examples, Aleidi’s approach to communicate with the audience is by posting questions that initiate a conversation about current local events such as women driving in Example 5.2 and traffic during the holy month of Ramadhan in Example 5.3. She also uses Najdi interrogatives instead of MSA, which makes her appear approachable and less formal.

Example 5.3: Aleidi’s use of /wiš/

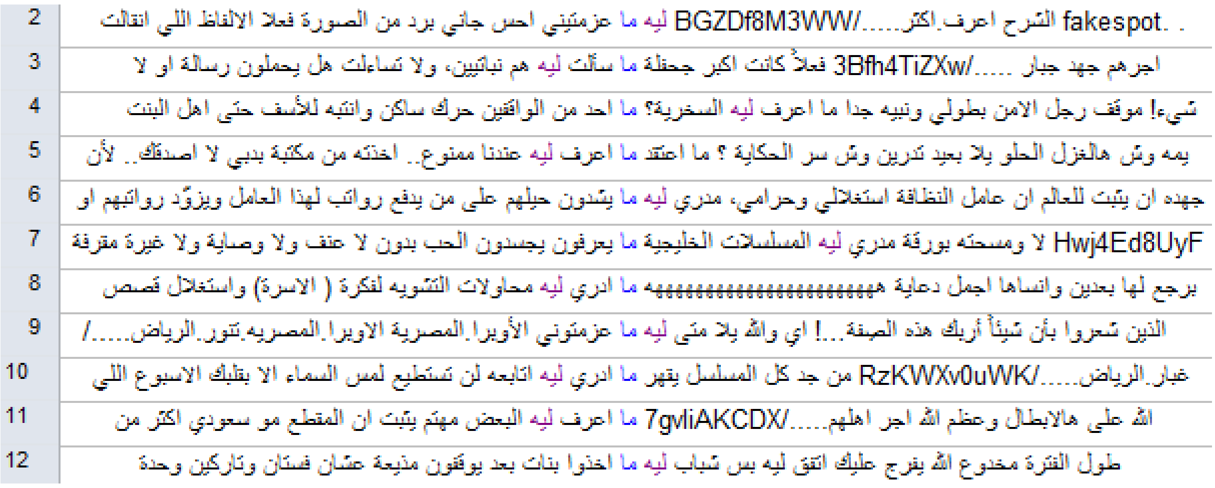
وش سر الزحمة في رمضان؟

/wiš sir azzaḥma fi ramaḍan?/

What is the secret behind Ramadhan traffic?

The selection of the informal NA variant shows how Aleidi positions herself as a Najdi woman and a member of the community. Questions such as عرفتوا وش عندنا /ʕaraftu wiš ʕndna/ *Do you know what we have*, وش داخل هذا الصندوق/wiš dakil haḏa aṣandog/ *What is inside this box,* وش تعرف عن موسم رمضان/wiš tiʕrif ʕn mawsim ramadan/ *What do you know about Ramadan season*, and وش خططكم للعيد/wiš ḵuṭaṭkm lil ʕid/ *What are your plans for Eid* indicate her selection of the informal variant to open up conversation.

Also, a frequent pattern that appears in her subcorpus is combining /leh/ with the negative /ma/ to pose questions such as *Why not?*. The other authors’ subcorpora did not exhibit this, which can be attributed to an individual stylistic choice that Aleidi makes. Figure 5.6 shows the collocates of /leh ma/ *why not* in her subcorpus in a number of contexts. The interrogative /leh/ in the Figure is highlighted in purple and the negative /ma/ is in blue.



2	. . fakespot الشرح اعرف.اكثر...../BGZDf8M3WW ليه ما عزمتيني احس جاني برد من الصورة فعلا الالفاظ اللي انقالت
3	اجرهم جهد جبار ...../3Bfh4TiZXw فعلاً كانت اكبر جحفلة ما سألت ليه هم نباتيين، ولا تساءلت هل يحملون رسالة او لا
4	شيء! موقف رجل الامن بطولي ونبيه جدا ما اعرف ليه السخرية؟ ما احد من الواقفين حرك ساكن وانتبه للأسف حتى اهل البنت
5	يمه وش هالغزل الحلو يلا بعيد تدرين وش سر الحكاية ؟ ما اعتقد ما اعرف ليه عندنا ممنوع.. اخذته من مكتبة بدبي لا اصدقك.. لأن
6	جهده ان يثبت للعالم ان عامل النظافة استغلالي وحرامي، مدري ليه ما يشدون حيلهم على من يدفع رواتب لهذا العامل ويزوّد رواتبهم او
7	Hwj4Ed8UyF لا ومسحته بورقة مدري ليه المسلسلات الخليجية ما يعرفون يجسدون الحب بدون لا عنف ولا وصاية ولا غيرة مقرفة
8	برجع لها بعدين وانساها اجمل دعاية ههههههههههههههههههههههههه ما ادري ليه محاولات التشويه لفكرة (الاسرة) واستغلال قصص
9	الذين شعروا بأن شيئاً أربك هذه الصفقة...! اي والله يلا متى ليه ما عزمتوني الأوبرا.المصرية الاوبرا.المصريه.تنور.الرياض...../
10	غبار.الرياض...../RzKWXv0uWK من جد كل المسلسل يقهر ما ادري ليه اتابعه لن تستطيع لمس السماء الا بقلبك الاسبوع اللي
11	الله على هالابطال وعظم الله اجر اهلهم...../7gvliAKCDX ما اعرف ليه البعض مهتم يثبت ان المقطع مو سعودي اكثر من
12	طول الفترة مخدوع الله يفرج عليك اتق ليه بس شباب ليه ما اخذوا بنات بعد يوقفون مذيعة عشان فستان وتاركين وحدة

Figure 5.6: Screenshot of Aleidi's collocates /leh ma/

There are other instances where Aleidi uses the NA interrogative /leh/ such as in Example 5.4 to pose a question to start a conversation, to raise awareness, and to encourage her audience to donate blood. The post shows Aleidi's positive attitude towards blood donation and it is a rhetorical device where she answers her own question and construct a socially beneficial stance about blood donation and encourage the public to participate in the donation drive.

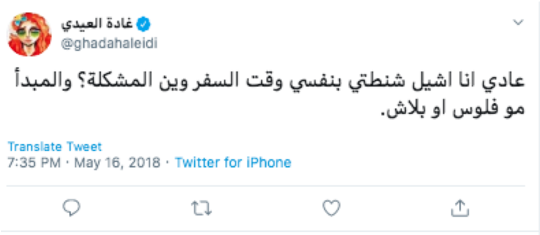
Example 5.4: Aleidi's use of /leh/

ليه تتبرع بدعمك؟ فزعتك.تفرق تبرع.بالدم.....

/leh titbarraʕ b dammik? fazʕtik tifrig .. tibarraʕ biddam/

Why do you donate blood? You can make a difference. Donate blood.

Similar to Alzamil, the last interrogative Aleidi uses is the NA variant /wen/ *where*. She uses the cluster وين المشكلة؟/wen almuškila/ *where is the problem?* to take an evaluative stance and defend her position towards an addressee or a topic. In Figure 5.7, Aleidi replies to a post saying: *I carry my own suitcase when travelling, where is the problem?* To express her criticism of people who think it is inappropriate to carry their own suitcases at airports.



غادة العيدي
@ghadahaleidi

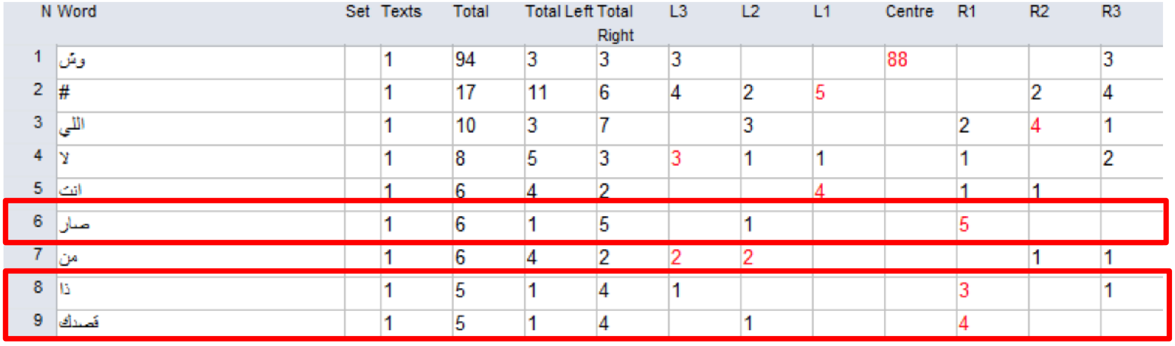
عادي انا اشيل شنطتي بنفسي وقت السفر وين المشكلة؟ والمبدأ
مو فلوس او بلاش.

7:35 PM · May 16, 2018 · Twitter for iPhone

Figure 5.7: Screenshot of Aleidi’s use of /wen/

Aleidi’s subcorpus reveals that she uses a limited range of interrogatives but with a high frequency. She uses the NA variants /wiš/, /leh/ and /wen/ out of the total nine interrogative variants in MSA and NA. It also shows her preference of NA forms, which reflects her sense of regional identity as a Najdi woman. However, she uses them occasionally as part of her stylistic stancetaking when interacting with her audience. She uses them as well to express her sarcastic criticism of what she perceives as negative social behaviour like using the phrase /wen almuškila/ *Where is the problem?*. Nonetheless, Aleidi, uses the Najdi interrogatives almost exclusively to pose questions for her followers and create interaction. The selection of the NA feature reflects Aleidi's self-perception as a Najdi woman and a member of the community. Moreover, her approach in highlighting her gender identity is by asking questions that target women such as /wiš btištirin/ *What will you buy?* and /wiš hi sayyartik/ *What is your car?*.

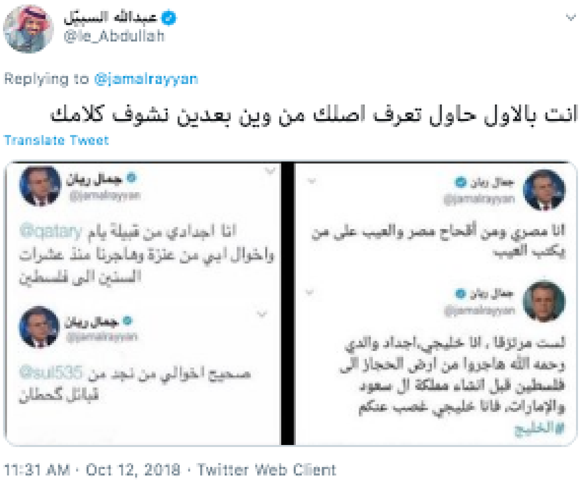
Moreover, Alsubayel’s subcorpus shares similarities with Alrokibah. It shows his preference for Najdi Arabic forms and rare occurrences of Modern Standard Arabic. Alsubayel’s infrequent use of MSA forms is strictly within a religious context. Such occurrences show his religious ethnic identity which is shared by most Muslims who use the formal variety in such contexts (Alenezi et al., 2018). The subcorpora show that both Alrokibah and Alsubayel have the NA variant /wiš/ as their most frequent interrogative. While each author uses a different set of collocates, eventually both use it in a manner that projects their regional affiliation. Figure 5.8 shows Alsubayel’s frequent collocates (highlighted in red rectangles in lines 6, 8 and 9) such as وش صار/wiš ṣar/ *What happened*, وش قصدك/wiš gaṣdk/ *What do you mean*, and وش ذا/wiš ḏa/ *What is this*. All of these phrases can be described as part of Alsubayel's humorous and sarcastic discourse.



N	Word	Set	Texts	Total	Total Left	Total Right	L3	L2	L1	Centre	R1	R2	R3
1	وش		1	94	3	3	3			88			3
2	#		1	17	11	6	4	2	5			2	4
3	اللي		1	10	3	7		3			2	4	1
4	لا		1	8	5	3	3	1	1		1		2
5	انت		1	6	4	2			4		1	1	
6	صار		1	6	1	5		1			5		
7	من		1	6	4	2	2	2				1	1
8	ذا		1	5	1	4	1				3		1
9	قصدك		1	5	1	4		1			4		

Figure 5.8: Alsubayel's pattern of use of /wiš/

Alternatively, Figure 5.9 demonstrates Alsubyael's style in sarcasm. This post also reflects his ethnic values and the notion of family lineage, which is important and valuable to the Saudi community and the Najdi specifically (Akers, 2001; Maisel, 2015; Alenezi et al., 2018). In this example, Alsubayel is asking about the addressee's ancestry and attached photos of older posts where he claims different origins in each one. The choice of using the Najdi variety to respond solidifies these identities and makes the stance prominent.



عبدالله السبيّل
@le_Abdullah

Replying to @jamalrayyan

انت بالاول حاول تعرف اصلك من وين بعدين نشوف كلامك

Translate Tweet

جمال ريان
@jamalrayyan

انا مصري ومن ألحاح مصر والعيب على من يكتب العيب

جمال ريان
@jamalrayyan

لست مرتزقا , انا خليجي,اجداد والدي رحمه الله هاجروا من ارض الحجاز الى فلسطين قبل انشاء مملكة ال سعود والإمارات، فانا خليجي غصب عنكم #الخليج

جمال ريان
@jamalrayyan

@qatary انا اجدادي من قبيلة يام واخوال ابي من عنزة وهاجرنا منذ عشرات السنين الى فلسطين

جمال ريان
@jamalrayyan

@sul535 صحيح اخوالي من نجد من قبائل كحطان

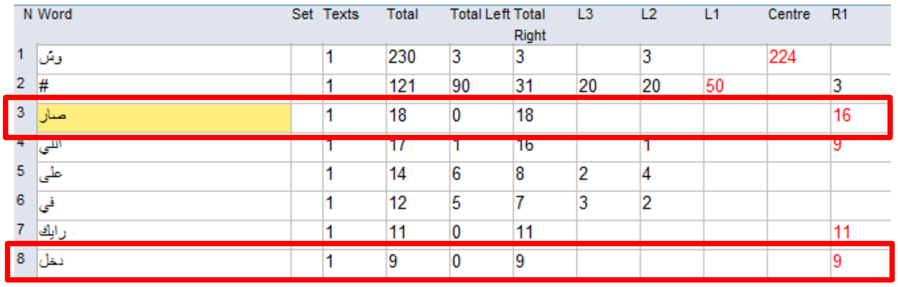
11:31 AM · Oct 12, 2018 · Twitter Web Client

/ent bil awal ḥawil tiʕrif aṣlik min wen bʕdain nšof kalamik/

First try and find out where are you from then we will consider what you say.

Figure 5.9: Screenshot of Alsubayel's use of /wen/

Alsubayel’s subcorpus shows that his preference in using Najdi interrogatives is a form of communication with his audience, which in time made him influential. This preference also shows his individual style in humour. The numerous occurrences found in his subcorpora build an image of Alsubayel as a Najdi male who is seeking to become relevant to many members of the local community through his humour and sarcasm. This is his influencing style. Allahim’s subcorpus shows one of the highest frequency rates in interrogatives. Also, it appears that the Najdi variants occur more than the MSA ones, the topmost being /wiš/ and /wen/. Figure 5.10 shows the most frequent collocates of the interrogative which are وش صار/wiš ṣar/ *What happened?* and وش دخل/wiš daḵal/ *How is that relevant?* (highlighted in the red rectangles in lines 3 and 8).



N	Word	Set	Texts	Total	Total Left	Total Right	L3	L2	L1	Centre	R1
1	وش		1	230	3	3		3		224	
2	#		1	121	90	31	20	20	50		3
3	صار		1	18	0	18					16
4	اللي		1	17	1	16		1			9
5	على		1	14	6	8	2	4			
6	في		1	12	5	7	3	2			
7	رايك		1	11	0	11					11
8	دخل		1	9	0	9					9

Figure 5.10: Allahim’s collocates of /wiš/

The concordance lines also show that Allahim uses the phrase /wiš ṣar/ *what happened* to create sarcastic interrogatives such as وش صار على ولي أمري أدرى بأمري؟/wiš ṣar ʕla wali amri adra b’amri/ *What happened to ‘my guardian knows best’?* which mocks the social media campaign that was against empowering Saudi women. The second collocate is /wiš daḵal/ *How is that relevant?*, which is part of Allahim’s sarcastic discourse mainly directed at the addressee such as وش دخل المهنة بالموضوع؟/wiš daḵal almihna bil mawdoʕ/ *What does the career have to do with it?* and وش دخل المطاوعة و رجال الدين؟/wiš daḵal almitawʕa wa rijal addin/ *What does it have to do with men of religion?*.

The second highest interrogative in Allahim’s subcorpus is the NA variant /wen/ *where.* Example 5.5 shows Allahim’s post in support of giving women freedom of choice.

Example 5.5: Allahim’s use of /wen/

لاحظوا ؛ تراي ادافع عن حق اخواتنا في ارتداء العباءة على الرأس ....
عشان الوضع يكون واضح

#وين التعميم يالطيفه الدليهان

/laḥẓo; tarai adafiʕ ʕan ḥag aḵwatna fi ertida’ alʕaba’a ʕla alra’s … ʕašan ilwaẓʕ ykon waẓḥ #wen ittaʕmim ya liṭifa addilaihan/

Note that I am defending our sister’s right to wear the abaya on the head ... so that the situation is clear #WhereIsTheMemoLatifaAddulihan

He has a pattern of creating ironic evaluative posts by using elements such as Najdi-specific words then switching to MSA variants. In Example 5.6, Allahim uses the NA variant /warah/ *why* repetitively as a stylistic choice to express his critical stance about some of the religiously conservative groups. It is also part of his accumulative construct of his ideological orientation (Bucholtz and Hall, 2005). Allahim is a moderate and a patriot, who expresses his criticism towards the religious conservative members of the community as he describes them as traces of the Alsahwa movement that holds the country’s development back. His construction of stances is built over a series of posts, combining MSA and NA variants.

Example 5.6: Allahim’s use of /warah/

هذا وراه مستهيش؟

انا قلت (بغيت اقول) واستعذت بالله من الشيطان و تبت   طيب
وراه زعلان علي؟

/haḏa warah mistihiš? ana gilt (bağait agol) w istaʕaḏt billah min išayṭan w tibt ṭayb warah zaʕlan ʕalai?/

Why is he upset? I said (I was about to say) and I repented So why is he mad at me?

Example 5.7 shows Allahim’s style of criticism regarding the regulations of women’s prison release procedures in Saudi Arabia. He uses the NA variant

/wišulah/ to raise a rhetorical question expressing his position on the topic. The NA form /wišulah/ does not occur in the reference corpus ReCNAT and rarely occurs in the NASCoT authors' subcorpora, except for Allahim and Alrokibah. Both authors are originally from the subregion of Algasseem where this form is commonly used, which reveals their regional orientation (McMenamin, 2002).

Example 5.7: Allahim's use of /wišulah/

أب يرفض خروج ابنته من السجن
وشوله اصلا يوخذ رأيه

/'ab yarfuḍ k̲urooj ibnatuh min al sijn wišulah aṣlan yok̲ad̲ rayah/

A father who refuses that his daughter leaves prison. Why is his opinion considered in the first place?

Another evaluative post Allahim publishes that extends to his ideological discourse is in Example 5.8. He is posing rhetorical questions criticizing the religious extremists who urge young men to commit suicide attacks, such as in *why does he talk to them about death?* and *why doesn't he urge them to participate in the (life) projects the country launched?* These questions are examples of Allahim's indexical processes using interrogatives to take a critical stance towards Islamic extremists. He criticises their discourse that calls on youth to take part in suicidal attacks instead of encouraging them to be active in improving their country.

Example 5.8: Allahim's use of /limad̲a/

شبابٌ يافعين مقبلين على الحياة؛ لماذا يحدثهم عن الموت؟ لماذا لا يرغبهم في الحياة والاستمتاع بها وبجمالها ، لماذا لا يحثهم على الانخراط في مشاريع(الحياة)التي اطلقتها الدولة؟

متى يتوقفون عن تداول هذه القصص التي لا فائدة منها سوى الكآبة والتعاسة؟

/šabab yafʕin muqbilin ʕla alḥayah; limaḍa yuḥadithhum ʕan ilmawt? limaḍa la yuraḡibhum fi alḥayah w alistimtaʕ biha wa bijamaliha, limaḍa la yaḥuthhum ʕla alink̲iraṭ fi mašariʕ (alḥayah) allati aṭlaqat'ha addawla?

mata yatawaqqfon ʕan tadawul had̲ih ilqiṣaṣ allati la fa'ida minha siwa alka'aba wa altaʕsa?/

Young men whose lives are still ahead of them, why does he talk to them about death? Why doesn't he encourage them to live and enjoy life in all its beauty, why doesn't he urge them to participate in the (life) projects the country launched?

When will they stop telling these stories that have no use other than depression and despair?

th

ir

interrogatives more frequently than others, the concordance lines, collocates and patterns also revealed the contexts in which these interrogatives are used thereby enabling us to understand the authors' style. The examination unfolds further with the following category as we learn more about the authors' use of negatives.

5.4.2 Negatives

The first thing to notice when looking at Figure 5.11 is that there is a preference to use MSA variants (green) in negatives compared to the Najdi variants (blue). I should point that the Figure shows the amalgamated negatives, because the MSA/NA negative variant /la/ *no* was distorting the view of other variants. The NASCoT shows that Alrasheed's subcorpus has the overall highest frequency rates in negatives, especially in the MSA variants. Other authors such as Alabdulkarim, Almohanna, and Altassan also show the same pattern of using the MSA variants more than the NA. Again, there is an exception to this pattern who is Alzamil, using the NA variants more frequently than the MSA. The other interesting pattern the NASCoT reveals is not only the variation in frequency between the authors but also in the stylistic purposes for which they use them, which we explore by looking into their subcorpora closely.

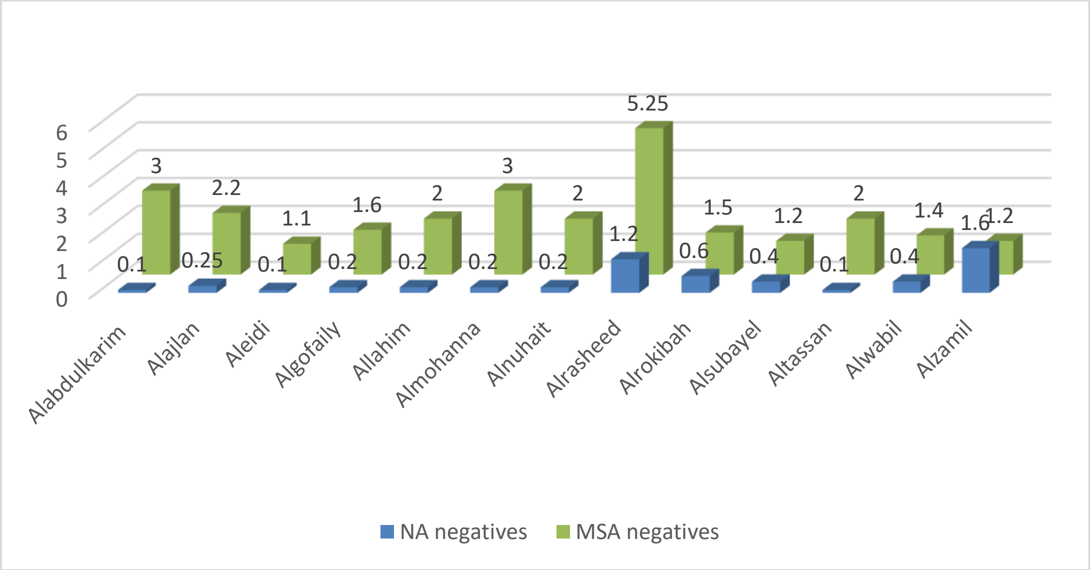
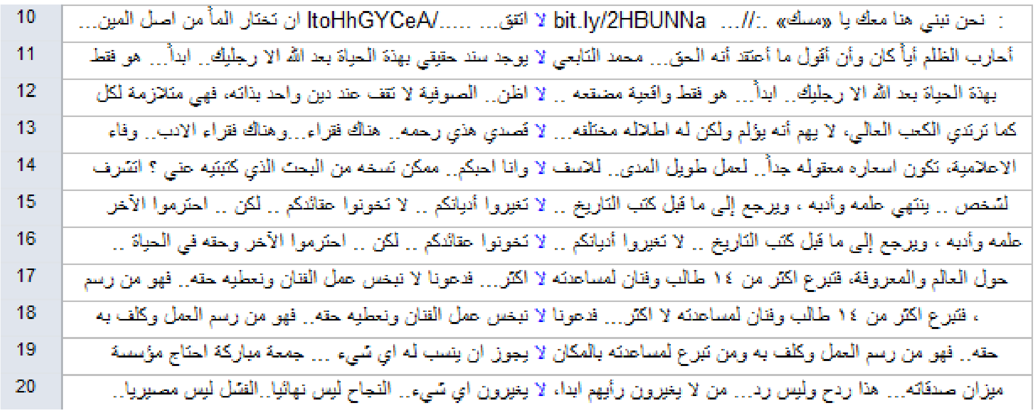


Figure 5.11: Negatives in NA and MSA (rates per 1000 words)

Alrasheed's subcorpus stands out with the highest frequency rates in using negative forms, specifically the MSA variant /la/, though figure 5.11 only shows the amalgamated negatives. Figure 5.12 shows concordance lines that represent her use of /la/. These stances convey ethnic and gender identity information.



10	: نحن نبني هنا معك يا «مسك» ..//: bit.ly/2HBUNNa لا اتفق ... ItoHhGYCeA/..... ان تختار الماأ من اصل المين...
11	أحارب الظلم أياً كان وأن أقول ما أعتقد أنه الحق... محمد التابعي لا يوجد سند حقيقي بهذة الحياة بعد الله الا رجليك.. ابداً... هو فقط
12	بهذة الحياة بعد الله الا رجليك.. ابداً... هو فقط واقعية مضقعه .. لا اظن.. الصوفية لا تقف عند دين واحد بذاته، فهي متلازمة لكل
13	كما ترتدي الكعب العالي، لا يهم أنه يؤلم ولكن له اطلاله مختلفه... لا قصدي هذي رحمه.. هناك قراء...وهناك قراء الادب.. وفاء
14	الاعلامية، تكون اسعاره معقوله جداً.. لعمل طويل المدى.. للاسف لا وانا احبكم.. ممكن نسخه من البحث الذي كتبتيه عني ؟ اتشرف
15	لشخص .. ينتهي علمه وأدبه ، ويرجع إلى ما قبل كتب التاريخ .. لا تغيروا أدیانكم .. لا تخونوا عقائدكم .. لكن .. احترموا الآخر
16	علمه وأدبه ، ويرجع إلى ما قبل كتب التاريخ .. لا تغيروا أديانكم .. لا تخونوا عقائدكم .. لكن .. احترموا الآخر وحقه في الحياة ..
17	حول العالم والمعروفة، فتبرع اكثر من ١٤ طالب وفنان لمساعدته لا اكثر... فدعونا لا تبخس عمل الفنان ونعطيه حقه.. فهو من رسم
18	، فتبرع اكثر من ١٤ طالب وفنان لمساعدته لا اكثر... فدعونا لا تبخس عمل الفنان ونعطيه حقه.. فهو من رسم العمل وكلف به
19	حقه.. فهو من رسم العمل وكلف به ومن تبرع لمساعدته بالمكان لا يجوز ان ينسب له اي شيء ... جمعة مباركة احتاج مؤسسة
20	ميزان صدقاته... هذا ردح وليس رد... من لا يغيرون رأيهم ابدا، لا يغيرون اي شيء.. النجاح ليس نهائيا..الفشل ليس مصيريا..

Figure 5.12: Concordance lines of Alrasheed's use of /la/

The second most frequent variant is /lam/, which Alrasheed uses as part of her idiolectal style to highlight the opposite of what she wants to communicate. This style of stating the negation of what she finds true is demonstrated in Example 5.9. She negates the common Arabic proverb that states silence is a sign of consent by saying it is a sign of boredom. This proverb has feminine connotations as it is usually used when a man proposes to a woman and she does not reply and thus her silence is understood as a sign of consent. Alrasheed uses the negative here to object to this common notion and bring another interpretation to the silence.

Example 5.9: Alrasheed's use of /lam/

السكوت لم يعد علامة الرضا... فقط اصبح علامة الملل..

/assukot lam yaʕud ʕalamat irrida … faqaṭ aṣbaḥ ʕalama lil malal../
Silence is no longer a sign of consent, but a sign of boredom..

Looking at her subcorpus, she uses negation as evaluative tool to express the experiences she gained and share her views. The pattern of negatives in Alrasheed's subcorpus reveals her beliefs and values about independence and strength - especially for women -, religion, and tolerance and acceptance of the other.

Alsubayel also uses /la/ frequently as part of his stylistic stancetaking. As noted earlier, /la/ is one of the forms where there is an overlap between Modern Standard Arabic and Najdi Arabic. My examination of the concordance lines and clusters show that Alsubayel uses /la/ in standard form for religious texts such as لا إله إلا الله/la ilah illa Allah/ *There is no god but Allah*. Other instances, as shown in the clusters, are used in the Najdi variety when replying to participants such as لا صح/la ṣaḥ/ *No that is right* or لا لا خطأ/la la k̲aṭa'/ *No no that is wrong*. As noted earlier in 5.2, some of the variants introduced here overlap across varieties and the negative /la/ is one of them. The only way to identify which variety is being used is through context, which was provided in Alsubayel's interaction in the previous examples. He uses the standard form /lam/ frequently but strictly in religious contexts; for instance, لم تيأس من الصبر/lam tai'as min alsabr/ *Not to give up patience*, لم تمل من الدعاء/lam tamil min adduʕa'/ *Never tired of praying*, and لم يكن الرضا سهلًا/lam yakun arrida sahlan/ *Being content was never easy*. These occurrences reveal Alsubayel's beliefs in terms of religion. On another note, he uses the Najdi /mahib/ to take an attitudinal stance criticizing his addressee's language in Example 5.10. The reply post shows his understanding of what is polite, thereby positioning his participant as impolite.

Example 5.10: Alsubayel's use of /mahib/

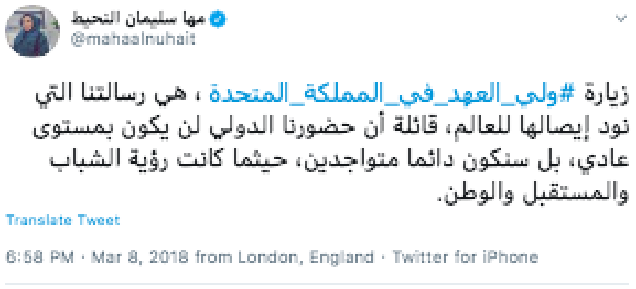
لا انا عبدالله السبيل والمسألة مهيب انك تعرف تذب او لا
بس عيب انك تطلق الفاظ غير لائقة لمجموعة اشخاص عشانهم خالفوك بالرأي

/la ana ʕabdalla assibayel w almas'ala mahib innik tiʕrif tithib wila la bas ʕaib innik tiṭliq alfaẓ g̅air la'iqa limajmoʕat ašk̲aṣ ʕašanhum k̲alafok birra'i/

No I'm Abdullah Alsubayel and it is not a matter of knowing how to banter. It's just inappropriate to call people bad names just because they disagree with you.

Alsubayel's use of negatives projects his ethnic identity in terms of religious affiliation when he uses /la/ and /lam/ in the standard variety. In terms of collective identity, his regional identity appears in the evaluative posts he publishes when interacting with his followers. On an individual level, these posts show his sense of humour and occasionally his sense of politeness.

Looking at the negative forms in Altassan's subcorpus, it shows her preference for using standard forms /la/, /lam/, and /lan/. There are different occurrences of stylistic stances about saying the word *no*, such as متى نقول لا؟/mata naqol la/ *When do we say no?*, كلمة لا قد تغير مجرى حياتك/kalimat la qad tuḡayir majra ḥayatik/ *The word no can change your life*, كلمة لا هي الكلمة المناسبة/kalimat la hiya alkalima almunasiba/ *The word no is the right word*, and لا يجب أن نخاف من لا/la yajib 'an naḵaf min la/ *We should not fear saying no*. Her use of the bound pronoun in (نقول) *we say* or (نخاف) *we fear* are rhetoric devices that do not necessarily communicate any collective identity values. They do ,however, reflect Altassan's individual views about boundaries as well as show us her stylistic choice to express them. Alnuhait's subcorpus also has high frequency rates of using negatives but strictly the MSA variants /la/, /lam/, and /lan/. Figure 5.13 shows a post by Alnuhait using /lan/ to project a positive view of the Saudi youth as *beyond average* and they are *the bright future of the country*.



مها سليمان النحيط
@mahaalnuhait

زيارة #ولي_العهد_في_المملكة_المتحدة ، هي رسالتنا التي
نود إيصالها للعالم، قائلة أن حضورنا الدولي لن يكون بمستوى
عادي، بل سنكون دائما متواجدين، حيثما كانت رؤية الشباب
والمستقبل والوطن.

Translate Tweet

6:58 PM · Mar 8, 2018 from London, England · Twitter for iPhone

/ziyarat wali alʕahd fi almamlaka almuttaḥida hiya risalatna allati nawad 'iṣalha lilʕalam qa'ila 'an ḥuḍorna addawli lan yakon bimustawa ʕadi, bal sanakon da'iman mutawajidin, ḥaiṯma kanat ru'yat ašabab wa almustaqbal wa alwaṭan/

The visit of the crown prince to the United Kingdom is our message to the world that our international representation will not be average, we will always be present wherever the vision of our youth, future, and country.

Figure 5.13: Screenshot of Alnuhait's use of /lan/

Altassan uses the MSA variant /lam/ for the same function in instances like من منا لم ينتظر ذات يوم/man minna lam yantaẓir d̲at yom/ *Who among us has not waited*, or الظروف الصعبة لم تصبح أسهل/alẓurof aṣṣaʕba lam tuṣbiḥ ashal/ *Hard times do not become easier*. There are instances in the concordance lines I have studied where she constructs epistemic and attitudinal stances that reflect her sense of patriotism; for instance, حضورنا الدولي مهم لم يأتِ من فراغ/ḥuḍorna addawli muhim lam ya’ti min faraḡ/ *Our international attendance is not without a purpose*. Another example is رادع باسم القانون لمن لم تردعه رجولته/radʕ bi ism alqanon liman lam tardaʕuh rujolatuh/ *Deterrence by law for men who do not deter themselves*. Looking at the last variant /lan/, Altassan also uses it for stylistic purposes to distinguish one group from another. This shows in instances such as لن يشعر بالمعاناة من لم يمر بها/lan yašʕur bilmuʕanah man lam yamur biha/ *One cannot feel the agony if one did not go through it*. There are other occurrences of /lan/ that are part of Altassan’s national discourse addressing support of other countries like لن نسمح أن يؤخرنا شيء عن مواقفنا الداعمة/lan nasmaḥ ‘an yu’k̲irna šay’ ʕan mawaqifna addaʕima/ *Nothing will keep us away from showing support*. Other examples address attacking terrorism لن نقضي عليه إلا بالبناء/lan naqḍi ʕalaih illa bil bina’/ *We will not destroy it unless we continue to build* and celebrating empowering women in فرح لن يشعر به إلا من عانى/faraḥ lan yašʕur bih illa man ʕana/ *A joy no one would feel unless they suffered*.

In terms of negative forms, Altassan’s subcorpus shows a number of findings. Similar to all NASCoT authors, the frequency rates show her preference in using standard forms and did not exhibit the use of Najdi. The occurrences appear to highlight her individual identity more and the perceptions she gained through personal experiences. Nonetheless, there are also few instances where her national and gender identity appear in epistemic and attitudinal stances that address national achievements and particularly for empowering Saudi women. Alnuhait’s use of negatives shows little diversity in forms. She uses the Modern Standard Arabic forms /la/, /lam/, and /lan/ as all the authors of NASCoT. Nonetheless, she uses negatives to post positive highlights about Saudi women and youth.

5.4.3 Deixis

As noted in Chapter 4, the deictic expressions investigated in this study denote person, time and space both in NA and MSA. I divided the variants into three sets to show the differences between the use of personal deictic expressions compared to temporal/spatial ones. Demonstrating the data in two separate figures would better project the idiolectal style of the authors. Figure 5.14 shows all deictic expressions of the NASCoT, where the NA deixis that mark time and space are marked in blue, their MSA counterparts are marked in green, and the NA person deixis variants are marked in turquoise. Also, it is relevant to note that some temporal and spatial deictic expressions have the same orthographic shape; therefore the overlap between both varieties is likely to occur (i.e., *today* /alyom/, *here* /huna/, *there* /hunak/, *near* /qarib/ and *far* /baʕid/). I included the variants nonetheless to investigate the authors' idiolectal selections in this category.

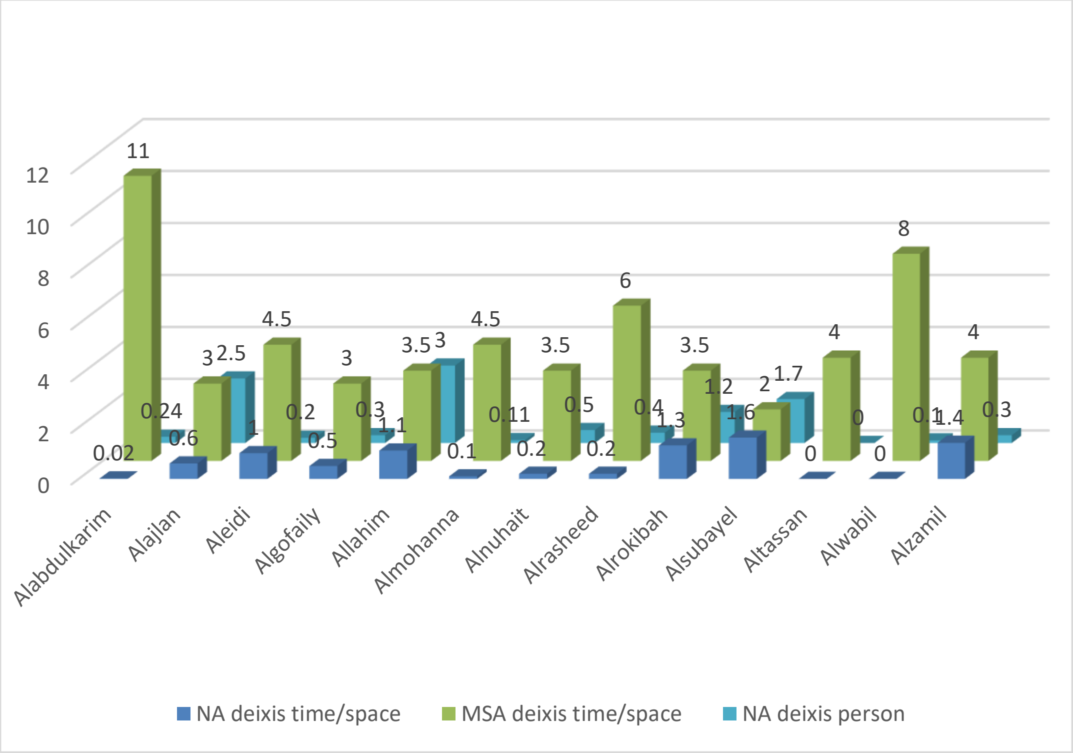
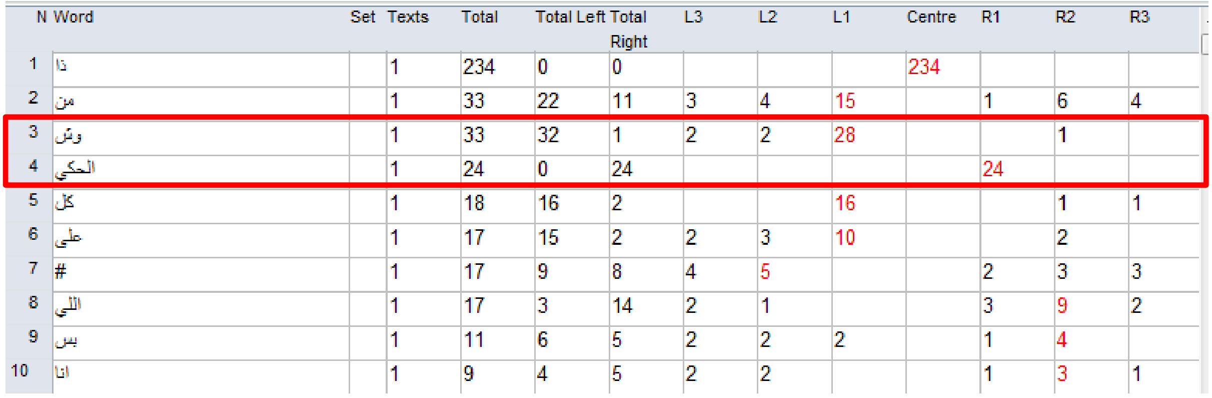


Figure 5.14: Deixis in NA and MSA (rates per 1000 words)

What is interesting is that, similar to negatives, the MSA variants of temporal and spatial deixis are higher in frequency than the NA ones, especially in the subcorpora of Alabdulkarim and Alwabil. Personal deictic expressions also show the authors use them infrequently, but they show more prominently in the subcorpora of Allahim, Alajlan and Alsubayel.

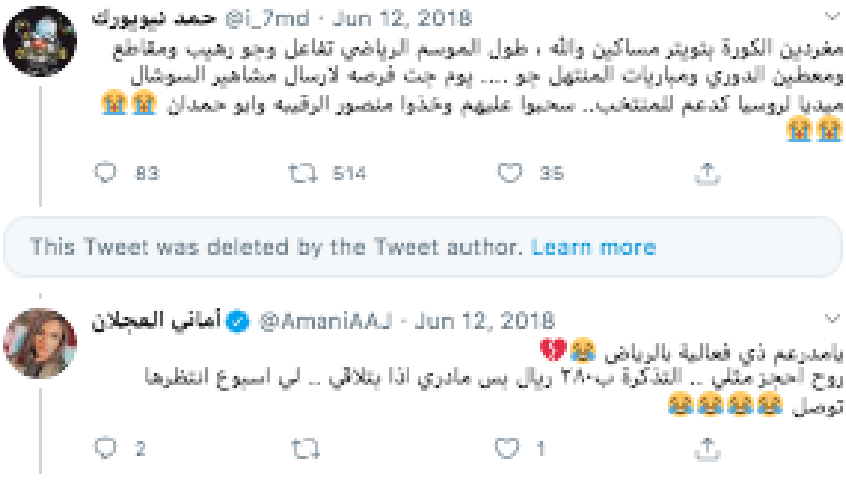
Alajlan's subcorpus is the second highest in frequency rates using person deixis. The most frequent forms were /ḏa/ and /ḏi/ followed by /haḏol/ and /ḏik/. The first form is the masculine singular /ḏa/ and Figure 5.15 shows its most frequent collocates. There are five phrases that repeatedly occur in the subcorpus, four of which she uses to create sarcastic attitudinal stances and they are: من ذا/min ḏa/ *Who is that*, وش ذا/wiš ḏa/ *What is that*, الحكي ذا/alḥaki ḏa/ *That kind of talk*, كل ذا/kil ḏa/ *All of that*. While the last phrase على ذا/ʕla ḏa/ *Over that* Alajlan uses to refer to objects. She uses collocates to create phrases such as من ذا الحكي/min ḏa alḥaki/ *That kind of talk*, وش ذا الجيل/wiš ḏa aljil/ *What sort of generation is that*, or وش ذا العقلية/wiš ḏa alʕaqliya/ *What sort of mindset is that* to express her criticism in a sarcastic manner.



N	Word	Set	Texts	Total	Total Left	Total Right	L3	L2	L1	Centre	R1	R2	R3
1	ذا		1	234	0	0				234			
2	من		1	33	22	11	3	4	15		1	6	4
3	وش		1	33	32	1	2	2	28			1	
4	الحكي		1	24	0	24					24		
5	كل		1	18	16	2			16			1	1
6	على		1	17	15	2	2	3	10			2	
7	#		1	17	9	8	4	5			2	3	3
8	اللي		1	17	3	14	2	1			3	9	2
9	بس		1	11	6	5	2	2	2		1	4	
10	انا		1	9	4	5	2	2			1	3	1

Figure 5.15: Alajlan's collocates of /ḏa/

Alajlan uses the feminine singular form /ḏi/ also as part of her sarcastic and humorous stancetaking as demonstrated in Figure 5.16. Other occurrences show that unlike the masculine form, she uses the feminine /ḏi/ as part of her sarcasm and humour and not in criticism.



حمد نيويورك @i_7md · Jun 12, 2018

مغردين الكورة بتويتر مساكين والله ، طول الموسم الرياضي تفاعل وجو رهيب ومقاطع ومعطين الدوري ومباريات المنتخب جو .... يوم جت فرصه لارسال مشاهير السوشال ميديا لروسيا كدعم للمنتخب.. سحبوا عليهم وخذوا منصور الرقيبه وابو حمدان 😭😭
😭😭

This Tweet was deleted by the Tweet author. Learn more

أماني العجلان @AmaniAAJ · Jun 12, 2018

يامدرعم ذي فعالية بالرياض 😂💔
روح احجز مثلي .. التذكرة ب٢٨٠ ريال بس مادري اذا بتلاقي .. لي اسبوع انتظرها توصل 😂😂😂😂

/ya mdarʕim d̠i faʕaliya birryaḍ

roh eḥjiz mit̠li.. attad̠kira bi 280 riyal bas madri id̠a bitlagi.. li esboʕ antiẓrha toṣal/

Not so fast this is an event in Riyadh

go and get your ticket like me.. it costs 280 Riyals but I don't know if you will find any.. I have been waiting for them for a week.

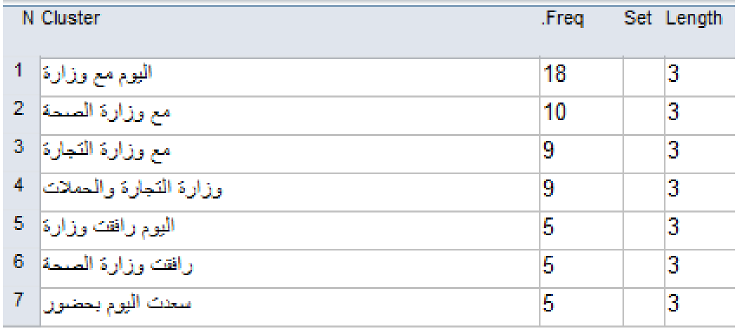
Figure 5.16: Screenshot of Alajlan's use of /d̠i/

Moreover, she uses the plural form /had̠ol/ to create critical attitudinal stances in phrases such as هذول ربع الرشاوي/had̠ol rabʕ arrašawi/ *Those who take bribes*, هذول حقون الطاقة/had̠ol ḥagon aṭṭaqa/ *Those who believe in energy*, and هذول عينة المثقفين/had̠ol ʕainat almut̠aqafin/ *That type of intellectuals*. These exhibited stances aim to criticise people who take bribes or claim to heal using energy or claim to be intellectual.

In Alajlan's regard, *the other* whether they are those who take bribes or believe in energy or some of the fake intellectuals are portrayed as artifice or illegitimate. In return, this shows her self-perception and position as someone who is authentic and has ethics. The last frequently used form in Alajlan's subcorpus is /d̠ik/. Her collocates reveal that she uses it to refer to time such as ذيك الأيام/d̠ik alayyam/ *Those days*, ذيك الساعة/d̠ik assaʕa/ *That time*, and ذيك الفترة/d̠ik alfatra/ *That period of time*. To sum up, Alajlan's subcorpus shows her use of person deixis is to create sarcastic and attitudinal stances based on how she perceives the other, or their action in terms of ethics and authenticity. Nonetheless, the distal deictic expression such as /ḏik/ is used to create references to objects.

Furthermore, the subcorpus of Alsubayel shows a relatively frequent occurrence of three Najdi forms, the topmost frequent being the singular masculine /ḏa/. The concordance lines exhibit phrases that are part of his humorous discourse like /wiš ḏa/ *What is this*, /wiš ṣar ʕla ḏa aliḵtiraʕ/ *What happened to that invention*, /ṣaḥiḥah ḏa almaʕloma/ *Is that information true*, and / wiš ḏa attaʕziz/ *What kind of support is this*. Most of the occurrences are in the interrogative, which also shows Alsubyael's idiolectal style. Alsubayel also uses the NA plural pronoun /ḥna/ to create multiple references that are also epistemic stances, one of which is an employee in the government sector in /ḥna ma nzalat ʕalawatna/ *We did not have our raise.* Another reference is a member of the Najdi community such as /birriyaḍ ḥna alḥin/ *We are in Riyadh now*, /ḥna nsamih asansair/ *We call it an elevator*, and /ḥna kiṯirin bi twitter/ *We are so many on Twitter*. Alsubayel's subcorpus shows that his use of person deixis and plural pronouns prominently project his regional identity and how he uses it in his stylistic and humorous stances to interact with his followers. On the level of collective identity, the pronoun /ḥna/ reveals how his self-positioning as a Najdi and how he includes himself with the others he addresses in his posts. Another instance that can be part of the author's individual and collective identity construct is /ḥna nitʕallam min alostaḏ ʕabdilʕaziz/ *We learn from Mr. Abdulaziz*, in which the pronoun shows his self-position as an employee talking about his work colleague.

Alabdulkarim is one of the authors with the highest frequency rates in temporal and spatial deixis, who uses the MSA variants of /alaan/ *now*, /alyom/ *today*, /huna/ *here* and /hunak/ *there* the most. The most frequent word is /alyom/ *today* and Figure 5.17 shows clusters of phrases like /alyom maʕ wizara/ *Today with* *the ministry*, /alyom rafaqt wizara/ *I accompanied the ministry today*, and /saʕidt alyom biḥuḍor/ *I was happy to attend today*. These epistemic stances that address Alabdulkarim's work with different ministries on a daily basis.

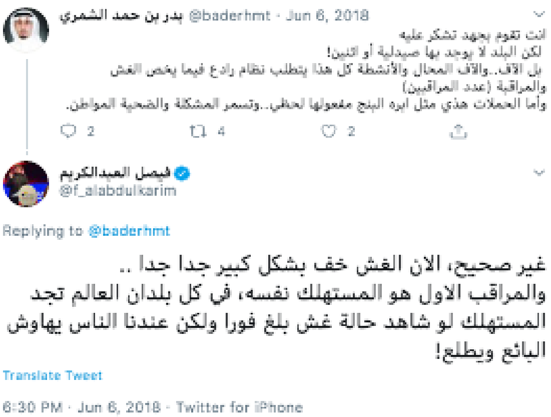


N	Cluster	.Freq	Set	Length
1	اليوم مع وزارة	18		3
2	مع وزارة الصحة	10		3
3	مع وزارة التجارة	9		3
4	وزارة التجارة والحملات	9		3
5	اليوم رافقت وزارة	5		3
6	رافقت وزارة الصحة	5		3
7	سعدت اليوم بحضور	5		3

Figure 5.17: Alabdulkarim's clusters of /alyom/

Another standard form Alabdulkarim uses is /hunak/, which overlaps in Najdi but the concordance lines in his subcorpus shows he uses it in MSA. His posts include phrases such as /hunak jihat ḥukumiya/ *There are governmental bodies*, /hunak qanon yuḥasubh/ *There is a law that will hold him liable*, and /hunak man yuṣadiq tilk alḥisabat alwahmiya/ *There are those who believe such fake accounts*, which also construct epistemic stances that relate to his volunteer work in protecting consumer rights. He uses the MSA time deictic /al'aan/ *now* either to report changes in consumer rights and corruption or current activities with the ministries he collaborates with. Through his posts like Figure 5.18, he is taking the temporary role of an authority figure who is actively involved in monitoring consumer rights. The epistemic stances he takes highlight his national identity.

The last form Alabdulkarim frequently uses is also the standard form /huna/. There are examples exhibit some of the occurrences such as تحدثت عنه هنا/taḥadaṯt ʕanha huna/ *I talked about it here*, هنا خبر القبض عليهم/huna ḵabar alqabḍ ʕalaihm/ *Here is the news about catching them*, and جزء من المداهمة هنا والبقية في سنابي/juz' min almudahama huna wa albaqiya fi snaby/ *Part of the raid is here and the rest is on my Snapchat story*.



بدر بن حمد الشمري @baderhmt · Jun 6, 2018

انت تقوم بجهد تشكر عليه
لكن البلد لا يوجد بها صيدلية أو اثنين!
بل الاف..والاف المحال والأنشطة كل هذا يتطلب نظام رادع فيما يخص الغش
والمراقبة (عدد المراقبين)
وأما الحملات هذي مثل ابرة البنج مفعولها لحظي..وتسمر المشكلة والضحية المواطن.

فيصل العبدالكريم
@f_alabdulkarim

Replying to @baderhmt

غير صحيح، الان الغش خف بشكل كبير جدا جدا ..
والمراقب الاول هو المستهلك نفسه، في كل بلدان العالم تجد
المستهلك لو شاهد حالة غش بلغ فورا ولكن عندنا الناس يهاوش
البائع ويطلع!

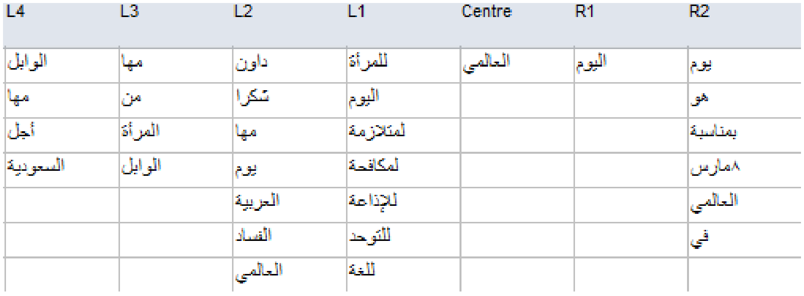
6:30 PM · Jun 6, 2018 · Twitter for iPhone

/ḡair ṣaḥiḥ, al'aan alḡiš ḵaf bišakil kabir jiddan jiddan.. wa almuraqib alawwal huwa almustahlik nafsuh, fi kil buldan alʕalam tajid almustahlik lao šahad ḥalat ḡiš ballaḡ fawran wa lakin ʕndina ennas yhawiš alba'ʕ wa yitlaʕ/

Not true, fraud has decreased to a large extent.. and the top auditor is the consumer themselves, in every country around the world you find the consumer if they witness a case of fraud they report immediately but here people just yell at the vendor and leave!

Figure 5.18: Screenshot of Alabdulkarim's use of /al'aan/

Alwabil's subcorpus is the second highest in temporal and spatial deictic forms. Figure 5.19 shows her pattern of using the proximal /alyom/. She uses temporal deixis to mark events in time as in اليوم العالمي للمرأة/alyom alʕalami lilmar'a/ *Today is International Day for Women*, اليوم العالمي لمتلازمة الداون/alyom alʕalami limutalazimat addaown/ *Today is international day for Down Syndrome*, and اليوم العالمي للتوحد/alyom alʕalami liltawaḥud/ *Today is international day for Autism.*



L4	L3	L2	L1	Centre	R1	R2
الوابل	مها	داون	للمرأة	العالمي	اليوم	يوم
مها	من	شكرا	اليوم			هو
أجل	المرأة	مها	لمتلازمة			بمناسبة
السعودية	الوابل	يوم	لمكافحة			٨مارس
		العربية	للإذاعة			العالمي
		الفساد	للتوحد			في
		العالمي	للغة			

Figure 5.19: Alwabil's pattern of use of /alyom/

The subcorpus reveals a frequent collocate /su'al alyom/ also shown in the Figure. Alwabil uses the collocates to post questions for her followers such as سؤال اليوم شاركونا أجمل قرار اتخذتوه/su'al alyom šarkona ajmal qarar itaḵadṯuh/ *Today's question is what is the best decision you have made?*. Another temporal form in Alwabil's subcorpus is the proximal /al'aan/. My study of concordance lines shows a range of epistemic stances as she shares her activities such as يحدث الآن ملتقى علوم الإدارة/yaḥdu̱t al'aan multaqa ʕulom alidara/ *Happening now is the administrative sciences forum*, and يحدث الآن في النادي الأدبي/yaḥdu̱t al'aan fi annadi aladabi/ *Happening now at the literary club*. Alwabil also uses spatial proximal deixis /huna/ to mark her current location. For instance, there are occurrences of /huna/ to refer to Alwabil's physical location as in هنا عنيزة/huna ʕunaizah/ *Here is Unaizah* and هنا القصيم/huna alqaṣim/ *Here is Alqassim*. In other occurrences, she refers to Twitter as a space such as هنا في تويتر/huna fi twitter/ *Here in Twitter* or to talk about arguments that take place in the platform and people who take it seriously يضع كل جهوده هنا/yaḍaʕ kul juhadahu huna/ *Places all their effort here*, which also show her negative views of some of the tweeps. The last form is the temporal distal /ḡadan/, which is also Modern Standard Arabic. Her epistemic stancetaking shows that she uses it to refer to events taking place in the future as in غداً اليوم الأول/ḡadan alyom alawal/ *Tomorrow is the first day*, غداً موعدي معكم/ḡadan mawʕdi maʕkum/ *My appointment with you is tomorrow*, and غداً الساعة 5 مساءً/ḡadan assaʕa 5 masa'an/ *Tomorrow at 5 in the evening*.

Looking at the temporal and spatial deixis in Alabdulkarim's subcorpus, he uses them as tools to create epistemic stances that position him as a figure of authority. His preference to choose the MSA variants over the NA ones conforms with these stances because they represent formality and authority. In terms of temporal and spatial deixis, Alwabil's subcorpus shows consistency in her preference to use MSA variants over Najdi ones. She uses both proximal and distal temporal deixis to address the events and activities she either partakes in or is interested about. The only spatial deixis in her subcorpus is /huna/, which she uses to denote her location or space. Occasionally Alwabil's use of /huna/ to criticise some people's behaviour on Twitter, reflects her self-perception compared to them.

One of the major patterns observed in Figure 5.5 is the authors' preference to use Najdi forms in the interrogatives and deixis categories over MSA forms (Figure 5.14). However, Figure 5.11 shows preference to use MSA forms in the negatives. Table 5.2 shows the idiolectal choices each author makes and their frequency rates with bright green showing highest frequency (1 ≤), light green showing (0.5 – 1), yellow showing rates (0.2 - 0.5), and red showing rates (≥ 0.2). The NA forms are marked in bold to distinguish them from the MSA forms. In diglossic communities such as Saudi Arabia, each variety has a cultural value; MSA is connected to discursive activities of institutions of authority and represents education and formality (Ismail, 2012). Alternatively, NA symbolises locality, informality, and indexes regional and social connotations (Silverstein, 2003; Ismail, 2012). This could explain the partialness the NASCoT authors show when using MSA variants in negatives and temporal/spatial deictic expressions.

Table 5.2: The lexico-grammatical patterns for the NASCoT authors

	Alabdulkarim	Alajlan	Aleidi	Algofaily	Allahim	Almohanna	Alnuhait	Alrasheed	Alrokibah	Alsubayel	Altassan	Alwabil	Alzamil
/mada/	▩	🟥	🟥	🟥	🟨	▩	🟨	🟥	🟥	🟥	🟨	🟨	🟥
/wis/	▩	🟩	🟩	🟩	🟩	🟨	🟥	🟨	🟩	🟩	🟥	🟥	🟩
/wisu/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟥
/limada/	🟨	🟨	🟥	🟨	🟨	🟩	🟥	🟩	🟥	🟥	🟥	🟨	🟥
/leh/	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟩
/warah/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟨
/wisulah/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/aina/	🟨	🟥	🟥	🟥	🟨	🟨	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/wen/	▩	▩	▩	▩	▩	🟨	🟥	🟥	🟨	🟨	🟥	🟥	🟩
/la/	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
/lam/	🟩	🟩	▩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	▩
/lan/	🟩	▩	🟨	🟨	▩	🟨	🟩	🟩	🟨	🟥	🟩	🟨	▩
/mub/	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟩	🟩	🟥	🟥	🟥	🟩
/manib/	🟥	🟨	🟥	🟥	▩	🟥	🟥	🟥	🟥	🟨	🟥	🟨	🟥
/muhub/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	▩	🟥	🟥	🟥	🟥

/mahub/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/mahib/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/hna/	🟥	🟥	🟥	🟥	🟨	🟥	🟨	🟥	▨	🟨	🟥	🟥	🟨
/da/	🟥	🟩	🟥	🟥	▨	🟥	🟥	🟥	🟥	▨	🟥	🟥	▨
/dak/	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟨	🟥	🟥	🟥	🟥	🟨
/hadak/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/di/	🟥	▨	🟥	🟥	▨	🟥	🟥	🟥	🟨	🟨	🟥	🟥	🟥
/dik/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/hadik/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/dola/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/dolak/	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/al’aan/	🟩	▨	🟩	🟩	🟩	▨	🟥	🟥	▨	🟨	🟨	▨	▨
/alhin/	🟥	▨	🟨	🟥	🟨	🟥	🟥	🟥	▨	▨	🟥	🟥	🟨
/lahiqan/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/badain/	🟥	🟥	🟨	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/alyom/	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
/gadan/	🟥	🟥	🟨	🟥	🟥	▨	🟥	🟩	🟥	🟥	🟥	🟨	🟥
/bukra/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟥
/bakir/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/hina/	🟩	▨	🟩	▨	▨	▨	🟨	▨	🟨	🟨	🟩	▨	🟩
/hinak/	🟩	🟩	🟨	🟨	▨	🟩	▨	🟩	▨	▨	🟩	🟨	🟨
/girib/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	▨
/bi’id/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥

🟥 ≥ 0.2 🟨 0.2 – 0.5 ▨ 0.5 – 1 🟩 1 ≤

There are authors who use their Twitter accounts primarily to engage with their audience and interact with them as opposed to others who use the platform mainly to express themselves and share their activities. This trend shows a tendency to use interrogatives in the Najdi forms (i.e., Aleidi and Alzamil). Furthermore, the linguistic patterns highlight the authors’ individual styles in interacting with participants, expressing their positive or negative views, and creating sarcastic comments. Furthermore, while the authors’ subcorpora include phrases and words that are commonly used and understood by the local population of the region (Burbano-Elizondo, 2006; Alenezi et al. 2018), when combining the linguistic features of each author into a pattern, it does set them apart from each other and the population. There is also a pattern of authors who use their Twitter accounts to directly engage with their audience. For instance, the subcorpora of Aleidi and Alzamil show consistency in their frequent use of interrogatives to initiate interaction with their audience. Other authors' activity is to post statements that represent their evaluative perspective about others and things such as Alrasheed's frequent use of negatives to express her opinions. Also, Allahim consistently uses person deixis in sarcastic posts and Alabdulkarim always uses temporal and spatial deixis when posting about his work activities.

5.5 Conclusion

As this chapter approaches its end, I revisit the research question this chapter aimed to investigate:

1) Can the Najdi Arabic Specialised Corpus of Tweets (NASCoT) help in identifying idiolectal style of thirteen Najdi authors when compared to a relevant population?

The analysis of the NASCoT reveals frequency patterns that the authors habitually practise when using their Twitter accounts. It enables us to see the individual stylistic preferences of each author when using interrogatives, negatives, and deixis and also show their preference for Najdi and MSA forms. The linguistic variables and their variants show us the kinds of activities the authors practise through their Twitter accounts. In the case of Aleidi, Alzamil, and Alsubayel, their purpose is to engage with the audience, which is reflected in their frequent use of interrogatives. Another category that shows high rates of frequency is the temporal/spatial deictic expressions. Again, the authors who use this category frequently (i.e., Alabdulkarim, Alwabil, Alrasheed) are using them as a way to engage with their followers and talk about their activities. For instance, Alrasheed and Alwabil frequently use the temporal deixis today /alyom/ to talk about their newspaper columns. The overall infrequent use of negatives also has indications, one of them being the authors' stylistic preferences to use interrogatives to express disagreement in a sarcastic manner (e.g., Alzamil). The personal deictic expressions also appear to be a stylistic preference for some authors such as Allahim and Alajlan to use when making sarcastic comments. The following analysis chapter further explores the linguistic patterns the authors have and the implications of their online identities.

Chapter 6
THEMES IN IDENTITY: COLLECTIVE AND INDIVIDUAL

6.1 Introduction

Idiolect and identity are similar constructs in the sense that they are unique representations of an individual's accumulation of sociocultural experiences. However, in the context of research, we have seen that an idiolect is more tangible and traceable compared to the more imperceptible concept that is the identity. In Chapter 2, we learned that an idiolect is represented in the way an individual uses a language/variety system and that representation is adjusted according to their sociocultural experiences. As for identity, it is not fixed or defined but rather continuously unfolding through interaction as it is 'interactionally negotiated' (Bucholtz and Hall, 2004: 376), a notion that is claimed by Grant and MacLeod (2018; 2020). Moreover, an individual's identity is not only shaped by their affiliation or membership of a group but also by the validation and acknowledgement of its community (Davies, 2005). In the previous chapter, we explored the lexico-grammatical patterns the authors' subcorpora revealed using corpus linguistics tools. This chapter addresses the identity implications that lie behind the linguistic features in answer to the following question:

2) How does the idiolectal style of these authors contribute to the construction of their online identities?

I continue to use forensic stylistic and statistical approaches to understand how the 13 authors' idiolectal styles create stances, which contribute to the construction of their online identity. The aim of this chapter is to investigate the relationship between the authors' idiolectal style and identity themes, collective and individual, by examining the same set of Najdi Arabic features explored in Chapters 5 and 7. I venture this investigation using two frameworks, the identity approach by Bucholtz and Hall (2005), which is a framework that explores online identity, and the Faceted Classification Scheme by Herring (2007) which facilitates the analysis of Computer Mediated Discourse (CMD). The following section explains both frameworks and how they are adopted in this study. The analysis section 6.3 begins with the collective identity themes, namely nationalism, gender, and region. The individual identity themes are explored through the authors' activities (Herring, 2007) which can reveal the interests by which their audiences identify them and engage with them.

6.2 The themes

This section addresses the identity themes that linguistic patterns revealed in Chapter 5. The first part covers collective identity themes, namely nationalism, gender, and region. The second part focuses on the individual variations between authors exhibited in the range of activities they perform. In spite of the classification of collective and individual identity themes in the analysis, this does not suggest that an individual's identity would be projected or interpreted in such a categorising manner. Identities always coexist; therefore, it is difficult to discuss one without addressing the social variables of ethnicity, regional background, and gender (Block, 2006).

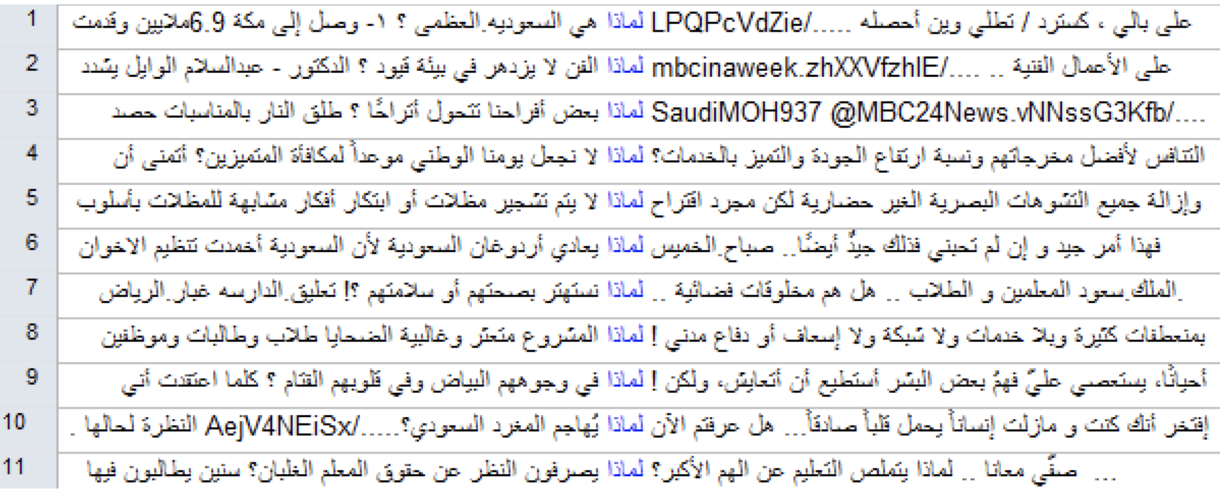
6.2.1 Collective identity

The collective identity can be a myriad of aspects that an individual identifies with or is exposed to starting from ethnicity and extending to social class. Nevo (1998) highlights the Islamic, the Arab, and the national (includes tribe, extended family, or region) as the three main elements that shape the collective identity of most Muslim Arabs of the Middle East, which consequently is applicable to Saudis. There were three main aspects of collective identity themes that appeared consistently and prominently across the NASCoT. The first one is nationalism, which extends to include patriotism as well. The subcorpora of the female authors reveal a strong projection of gendered identity in the content of their posts. And finally, being a corpus that concerns a regional dialect, it is essential to explore and address the regional identity of the authors, particularly those who reflected it clearly. In my analysis, I adopt Nevo's (1998) interpretation of nationalism and region and refer to Ismail's (2012) interpretation of gender.

Nationalism

A nationalist identity usually refers to a common language that symbolises national identity for individuals (Garri, 2016; Alsohaibani, 2016). Saudi national identity is encouraged by embracing the national symbols namely the flag, the anthem, the national holidays (e.g., the Saudi National Day), and Islam as a national value and the cornerstone of the society and state (Nevo, 1998). I extend the references here to include themes of patriotism and the individual's support and positive, or negative, views of the country and the monarchy.

The NASCoT authors' tweets frequently show affiliation to Saudi Arabia by talking about local affairs and their evaluative language and stance can either express positive or negative views. Alajlan's subcorpus shows instances where she uses /alyom/ in MSA to refer to the present time such as اليوم أصبح لدينا فرق وطنية طبية /alyom aṣbaḥ ladaina furuq waṭaniya ṭibya/ *Today we have national medical teams*, الوضع اليوم أفضل/alwaḍʕ alyom afḍal/ *The situation today is better*, and حكمة اليوم وكل يوم/ḥikmat alyom wa kil yom/ *Today's and everyday's word of wisdom*. These epistemic stances describe the current and positive progress the country is making; they are also instances of emergence where Alajlan expresses what she believes. Similarly, Alnuhait's subcorpus shows occurrences of the MSA variant /alyom/ also to take epistemic stances talking about Saudi Arabia such as ما نمر به اليوم تاريخ لا ينسى/ma namur bih alyom tarik̲ la yunsa/ *We are witnessing history in the making*, and شباب السعودية اليوم أكثر احترافًا/šabab assoʕudiya alyom alakt̲ar iḥtirafan/ *The Saudi youth of today are most professional*. Both authors use the MSA variant /alyom/ to talk about changes in Saudi Arabia positively, which can be described as indexical inversion. Moreover, when Algofaily uses the MSA variant /limad̲a/ *why* he resorts to a more formal discourse that addresses national topics. The concordance lines in Figure 6.1 show that in most instances Algofaily creates attitudinal stances to criticise local issues like concordance line 3 لماذا بعض أفراحنا تتحول أتراحًا/limad̲a baʕḍ afraḥna tatḥawal atraḥan/ *Why do some people turn celebrations into travesty?*, line 8 لماذا المشروع متعثر/limad̲a almašroʕ mutaʕat̲ir/ *Why is the project in default?*, and concordance line 11 لماذا يصرفون النظر عن حقوق المعلم/limad̲a yaṣrifon annaẓar ʕan ḥuqoq almuʕallim/ *Why do they overlook teachers' rights?*. Nonetheless, there are other instances where he uses the MSA variant to express positive stances such as line 1 لماذا هي السعودية العظمى/limad̲a hiya assoʕudiya alʕuḍma/ *Why is it great Saudi Arabia?*.



1	على بالي ، كسترد / تطلي وين أحصله LPQPcVdZie/..... لماذا هي السعوديه_العظمى ؟ ١- وصل إلى مكة 6.9ملايين وقدمت
2	على الأعمال الفنية .. mbcinaweek.zhXXVfzhIE/.... لماذا الفن لا يزدهر في بيئة قيود ؟ الدكتور - عبدالسلام الوايل يشدد
3	SaudiMOH937 @MBC24News.vNNssG3Kfb/.... لماذا بعض أفراحنا تتحول أتراحًا ؟ طلق النار بالمناسبات حصد
4	التنافس لأفضل مخرجاتهم ونسبة ارتفاع الجودة والتميز بالخدمات؟ لماذا لا نجعل يومنا الوطني موعداً لمكافأة المتميزين؟ أتمنى أن
5	وإزالة جميع التشوهات البصرية الغير حضارية لكن مجرد اقتراح لماذا لا يتم تشجير مظلات أو ابتكار أفكار مشابهة للمظلات بأسلوب
6	فهذا أمر جيد و إن لم تحبني فذلك جيدٌ أيضًا.. صباح_الخميس لماذا يعادي أردوغان السعودية لأن السعودية أخمدت تنظيم الاخوان
7	_الملك_سعود المعلمين و الطلاب .. هل هم مخلوقات فضائية .. لماذا تستهتر بصحتهم أو سلامتهم ؟! تعليق_الدارسه غبار_الرياض
8	بمتعطفات كثيرة وبلا خدمات ولا شبكة ولا إسعاف أو دفاع مدني ! لماذا المشروع متعثر وغالبية الضحايا طلاب وطالبات وموظفين
9	أحيانًا، يستعصي عليّ فهمُ بعض البشر أستطيع أن أتعايش، ولكن ! لماذا في وجوههم البياض وفي قلوبهم القتام ؟ كلما اعتقدت أني
10	إفتخر أنك كنت و مازلت إنساناً يحمل قلباً صادقاً.. هل عرفتم الآن لماذا يُهاجم المغرد السعودي؟.....AejV4NEiSx/ النظرة لحالها .
11	... صفّي معانا .. لماذا يتملص التعليم عن الهم الأكبر؟ لماذا يصرفون النظر عن حقوق المعلم الغلبان؟ سنين يطالبون فيها

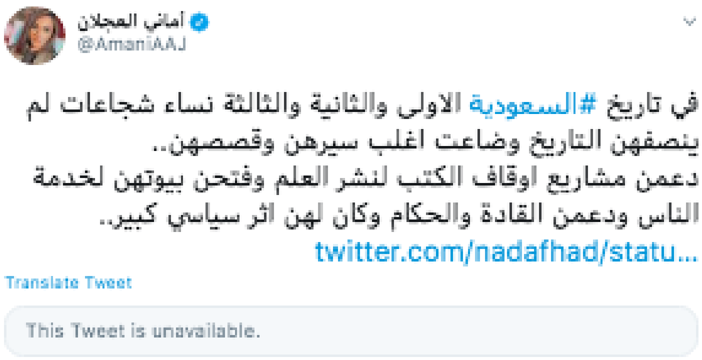
Figure 6.1: Concordance lines of Algofaily's use of /limad̲a/

In the indexicality principle, Bucholtz and Hall talk about 'the use of linguistic structures and systems that are ideologically associated with specific persons and groups' (2005: 594). The examples extracted from the NASCoT show us each author's position in terms of nationalism and how they choose to express that position. MSA appears to be the variety Alajlan, Alnuhait, and Algofaily use when talking about Saudi Arabia, especially when talking about positive events or emotions they have towards the country's past and current feats. The stylistic variation across the authors is that Alajlan and Alnuhait incorporate the MSA variant /alyom/ to express their positive views which seems to reveal where the authors' focus, that is the present time. Algofaily, on the other hand, mostly tends to use MSA in the context of criticism. The concordance lines show that there is a pattern where Algofaily uses the MSA interrogative /limaḏa/ to express his negative position about local issues like teacher's rights and failing construction projects. On the one hand, the authors exhibit indexical inversion by addressing positive progress and issues for the sake of improvement, and they are using the formal variety to express it. This is part of the encouraged national discourse, which also shows how they position themselves as members of the in-group. On the other hand, the authors' selection when communicating these ideas can also be perceived as stance accretion. The variants the authors use (i.e., /alyom/ for Alajlan and Alnuhait and /limaḏa/ for Algofaily), when talking about these themes contribute to their national discourse.

Gender

Lakoff's (1975) *Language and Woman's Place* was one of the foundational works that attribute language variation between men and women to power. She suggested that language features used by men and women are power cues, where women use hedges, tag questions, and rising intonation for declaratives to mark their subordinate status to men. Tannen (1990) later proposed that language variation between men and women can be attributed to the gendered subcultures that a community assigns to boys and girls (Hall, 2020). More recently gender theorists looking at language have talked about gender as performance, drawing on Butler (1988, 2004). Duranti (2003) called for studies that explore ‘the role of language in establishing gender, ethnic, and class identities’ (p. 332). The contexts of the posts that convey gender tend to observe government, religion, and society (Aloufi, 2017). There are a number of instances across the female authors that project their gender by posting about issues and facts that address and concern Saudi women. Interestingly, the male authors’ subcorpora do not show any explicit demonstration of gender in their posts.

The instances related to gender are examples of the positionality principle in the sense that they reveal the authors’ macro-level social category. Starting with Alajlan’s subcorpus, Figure 6.2 shows an original post published by the author. The epistemic stance /nisa’ šujaʕat lam unṣifhun attariḵ/ *Brave women overlooked by history* shows her national and gender identities by writing about the role of women in Saudi history that has been obscured and kept from the public.



أماني العجلان
@AmaniAAJ

في تاريخ #السعودية الاولى والثانية والثالثة نساء شجاعات لم ينصفهن التاريخ وضاعت اغلب سيرهن وقصصهن..
دعمن مشاريع اوقاف الكتب لنشر العلم وفتحن بيوتهن لخدمة الناس ودعمن القادة والحكام وكان لهن اثر سياسي كبير..
twitter.com/nadafhad/statu...

This Tweet is unavailable.

/fi tariḵ assoʕudiya al’ola wa aṯania wa aṯaliṯa nisa’ šujaʕat **lam** unṣifhun attariḵ wa daʕat aḡlab syarhun wa qiṣaṣhun.. daʕamn mšariʕ awqaf alkutub linašir alʕilm wa fataḥn byotihun lḵidmat annas wa daʕamn alqadah wa alḥukam wa kan lahun aṯar syasi kabir../

In the history of the first, second, and third Saudi states are brave women who were ***not*** *mentioned by history and most of their stories are lost.. they supported book publishing to support education, opened their houses to serve people and encouraged leaders and rulers and had a strong political influence..*

Figure 6.2: Screenshot of Alajlan’s use of /lam/

In Example 6.1, she positions herself as a moderate person and a member of the Saudi community who witness the ideological shift that is taking place by saying /mu gadra tḥisin bil muʕana/ *you cannot understand the suffering*, /yarmon 40 sana b saʕa/ *you want them to throw away 40 years in one hour* and more importantly / aḥawil agarrib wijhat al naẓar/ *I try to bridge the gaps*. The Saudi community particularly in Najd have been highly conservative both religiously and culturally for the past forty years, especially when it comes to women (Ingham, 1996; Nevo, 1998; Ismail, 2012), but that is undergoing a lot of change in the past five years (Alkarni, 2018). She also positions herself as a member of the moderate in-group, a community that is growing to become the majority currently, being critical of the behaviour of the supporters of the Alsahwa movement whose *extremists used to call us infidels*. Her position is revealed when she calls Alsahwa movement as extremist and refers to herself as one of the moderates, represented in *us*, that the Sahwa extremists would call infidels.

Example 6.1: Alajlan's use of /wiš/

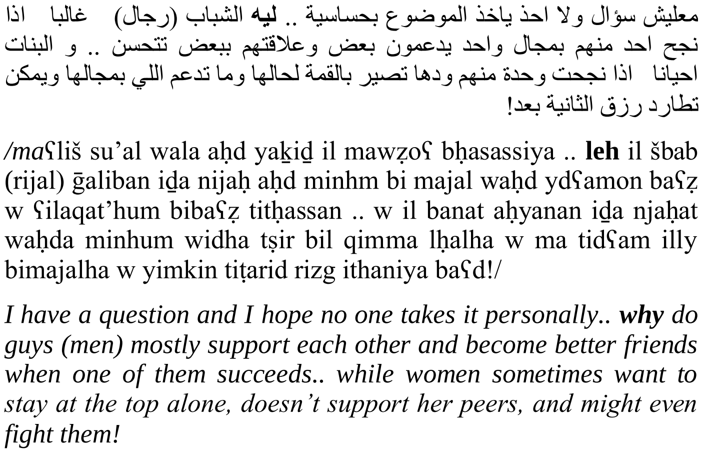
لا مو غيرة، انت مو قادرة تحسين بالمعاناة اللي يعيشوها وهم مو متأقلمين على الوضع الجديد.. اقعد مع كثير واسمع ذا الكلام كثير واحاول اقرب وجهات النظر .. انت تبينهم يرمون ٤٠ سنة بساعة .. مايصير يختي.. اذا كل وحده بتقولين لها انت مستشرفه **وش** فرقتي عن الصحويين اللي كانوا يكفرونا؟.

/la mu ḡira, inti mu gadra tḥisin bil muʕana illi yʕišuha w hum mu mit'aqlimin ʕla al waḍʕ al jideed .. agʕid maʕ kiṯir w asmʕ ḏa al kalam kiṯir w aḥawil agarrib wijhat al naẓar .. inti tbinhum yrmon 40 sana b saʕa .. ma yiṣir ikẖti .. iḏa kil waḥda bitgolin laha inti mistašrifa **wiš** faragti ʕan al ṣaḥawiin illi kanu ykafronina?/

No it is not jealousy, you cannot understand the agony they go through as they cannot adjust to the new system.. I try to bridge the gaps every time I hear such things.. you expect them to throw away 40 years in one hour.. that's not fair.. ***what*** *difference is this from back when the Sahwa extremists used to call us infidels?*

Moreover, Example 6.2 shows a different context and identity aspect of Aleidi. She takes an attitudinal stance in this original post to criticise gender differences at work. This post reveals two positions; the first one is how Aleidi perceives herself as a career woman and talks about the different behaviours between men and women when they become successful at work. Also, using Najdi rather than MSA reveals another position as she chooses to talk about it using the Najdi variety to be more relevant to the Saudi professional community and job market.

Example 6.2: Aleidi’s use of /leh/



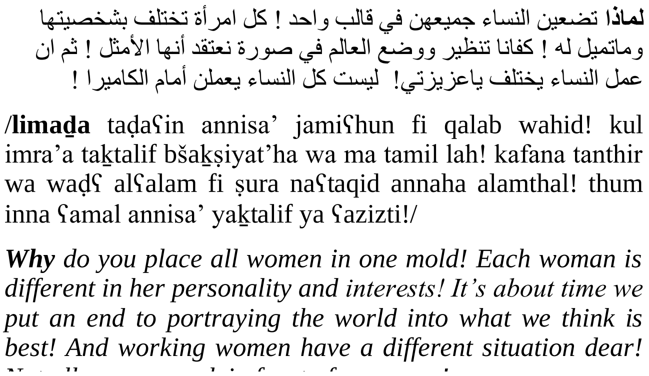
معليش سؤال ولا احذ ياخذ الموضوع بحساسية .. **ليه** الشباب (رجال) غالبا اذا نجح احد منهم بمجال واحد يدعمون بعض وعلاقتهم ببعض تتحسن .. و البنات احيانا اذا نجحت وحدة منهم ودها تصير بالقمة لحالها وما تدعم اللي بمجالها ويمكن تطارد رزق الثانية بعد!

/*ma*ʕliš su’al wala aḥd yak̲id̲ il mawẓoʕ bḥasassiya .. **leh** il šbab (rijal) ḡaliban id̲a nijaḥ aḥd minhm bi majal waḥd ydʕamon baʕẓ w ʕilaqat’hum bibaʕẓ titḥassan .. w il banat aḥyanan id̲a njaḥat waḥda minhum widha tṣir bil qimma lḥalha w ma tidʕam illy bimajalha w yimkin tiṭarid rizg ithaniya baʕd!/

I have a question and I hope no one takes it personally.. ***why*** *do guys (men) mostly support each other and become better friends when one of them succeeds.. while women sometimes want to stay at the top alone, doesn’t support her peers, and might even fight them!*

Example 6.3. shows an instance for another female author, Almohanna. This post is an instance where she uses the MSA variant /limad̲a/ *why* to talk about stereotyping women. She posted a reply to the addressee criticising her stereotypical portrayal of the intellectual woman, which reveals how she perceives herself as a member of that group.

Example 6.3: Almohanna’s use of /limad̲a/



لماذا تضعين النساء جميعهن في قالب واحد ! كل امرأة تختلف بشخصيتها وماتميل له ! كفانا تنظير ووضع العالم في صورة نعتقد أنها الأمثل ! ثم ان عمل النساء يختلف ياعزيزتي! ليست كل النساء يعملن أمام الكاميرا !

/**limad̲a** taḍaʕin annisa’ jamiʕhun fi qalab wahid! kul imra’a tak̲talif bšak̲ṣiyat’ha wa ma tamil lah! kafana tanthir wa waḍʕ alʕalam fi ṣura naʕtaqid annaha alamthal! thum inna ʕamal annisa’ yak̲talif ya ʕazizti!/

Why *do you place all women in one mold! Each woman is different in her personality and interests! It’s about time we put an end to portraying the world into what we think is best! And working women have a different situation dear!*

Furthermore, Almohanna’s performs her gender identity again in Example 6.4 in another attitudinal stance /lam nara šai’an mima yaddaʕin/ *We have not seen what they claim*. She is expressing her criticism of women who attack other women instead of supporting and empowering them.

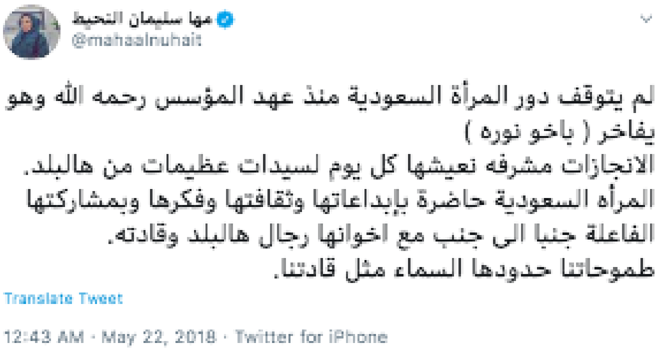
Example 6.4: Almohanna's use of /lam/

وعي المرأة بحقوقها يشكل رعباً في مجتمع لم يألف على أن يكون لصوت المرأة
حضور، الوعي دوماً مربك للآخر ، مايأسف أن الهجوم يأتي من نساء يدعين أنهن
خدمن المرأة في مكانة ما ، و**لم** نرى شيء مما يدعين!!

/waʕi almar'a biḥqoqha yušakil ruʕban fi mujtamaʕ lam ya'laf ʕla 'an yakon liṣawt almar'a ḥuḍor, alwaʕi dawman murbik lilaḵar , ma ya'saf 'an alhujom ya'ti min nissa' yaddaʕin annahun ḵadamn almar'a fi makanatin ma , wa **lam** nara šai'an mima yaddaʕin!!/

A woman being aware of her rights is causing terror in a community that is not used to a woman having a voice, awareness always causes confusion to the other, what is unfortunate is this attack comes from women who claim they serve women at some point, and we have **not** *seen what they claim!!*

Both stances illustrate Almohanna's use of MSA variants to express her views about women and how she relates herself to one group of women while criticising the other. The examples show us Almohanna's self-positioning, which is as an intellectual woman who advocates for empowerment and self-expression. The posts also show her perception of what she portrays as an out-group, women who antagonise other women. Alhunait's tweets have a number of instances that project her position towards her gender identity, Figure 6.3 is an example. She takes an epistemic stance stating that Saudi women have always had a role in the development of the country. The stance also expresses Alnuhait's positive attitude about that fact.



مها سليمان النحيط
@mahaalnuhait

لم يتوقف دور المرأة السعودية منذ عهد المؤسس رحمه الله وهو
يفاخر (باخو نوره)
الانجازات مشرفه تعيشها كل يوم لسيدات عظيمات من هالبلد.
المرأه السعودية حاضرة بإبداعاتها وثقافتها وفكرها وبمشاركتها
الفاعلة جنبا الى جنب مع اخوانها رجال هالبلد وقادته.
طموحاتنا حدودها السماء مثل قادتنا.

12:43 AM · May 22, 2018 · Twitter for iPhone

/lam yatwaqqaf dawr almar'a asso3udiya munthu 3ahd almo'assis rahinah Allah wa huwa yufakir (b'aku nora) alinjazat musarrifa na3isha kil yom lisayidat 3azimat min halbalad

The role of the Saudi woman has not stopped since the time of the founder who was proud to be called (Nora's brother) Every day we witness feats achieved by the great women of this country

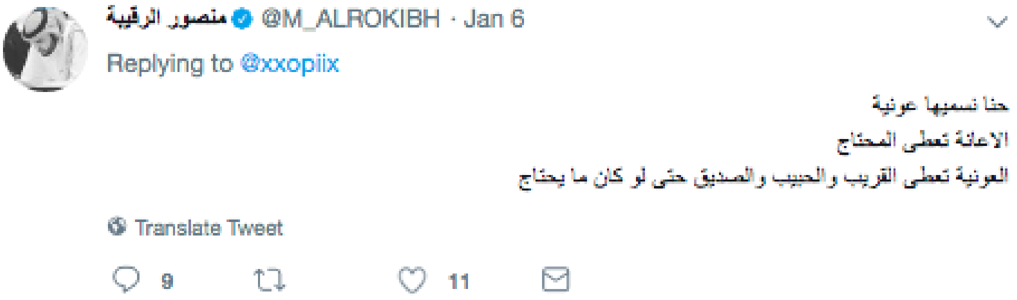
Figure 6.3: Screenshot of Alnuhait's use of /lam/

The gender instances in the NASCoT show that the female authors, unlike in Lakoff's proposition, are not expressing their gender by using hedges or tag questions, but by publishing content that concerns them as women. Moreover, the corpus shows that female authors use interrogatives and negatives, some opting for the NA variety like Aleidi but most in the MSA variety like Almohanna and Alnuhait using the formal, standard variants. Referring to the authors' profiles in Chapter 4, Almohanna and Alnuhait identify themselves in a formal, authoritative capacity and using the MSA is an extension to their online identity. Alajlan's subcorpus shows that she uses variants from both varieties, Ismail (2012) claims that Saudi women would opt for vernacular or dialectal variants even in formal contexts as opposed to men who prefer to use MSA; a claim that is supported by other sociolinguistic studies of Arab communities (see Bakir, 1986; Holes, 1987; Abu-Melhim, 1992; Sadiqi, 2003; Miller, 2004). She interprets this cross-gender linguistic variation as a result of a 'gender-segregated social order' and social norms that restrict women's access to public mediums of communication (p. 275). Such interpretation might have been valid at earlier times and in the wider world, but social media platforms, and Twitter is a prominent example, have made an impact in the way women express themselves (Alothman, 2012; Almahmoud, 2015). The female authors in the NASCoT show frequent use of MSA variants; for instance, Almohanna and Alnuhait's subcorpora show a preference for using MSA variants over the NA variety. However, gender variation appears in the corpus in terms of the contexts, indexes, and positioning in the female authors' tweets. Ismail (2012) interprets this as 'gender asymmetry' that is 'sustained by socio-cultural practices and it seems to be reproduced in linguistic behaviour through dynamics of Arabic glossia.’ (p. 274). I adopt this interpretation to explain the female authors’ choices, both in terms of context and linguistic features. In this study, women using MSA are rejecting gender asymmetry and assimilating with the male linguistic behaviour to assert their authority.

Region

Regional identity denotes the meaningful places with which people identify themselves and the variety they speak (Johnstone, 2004). There is a clear reference to regionality in Saudi Arabia (Montagu, 2015); its various territories (the prominent regions are Najd in the centre, Hijaz in the west, and Al-Hasa in the east) retain local and regional characteristics (Ingham, 1996; Nevo, 1998). The authors of the NASCoT exhibit clear instances of regional identity that are discussed in this subsection.

The subcorpus of Alrokibah shows that he consistently uses the pronoun /ḥna/ *we* exclusively to denote the Najdi community. There are a number of occurrences such as حنا في السعودية/ḥna fi assoʕudiya/ *We at Saudi Arabia*, حنا نعامل الناس/ḥna niʕmil annas/ *We treat people*, and حنا ننتظر المستقبل/ḥna nintizir almustaqbal/ *We are waiting for the future*, all of which denote the Saudi community and the Najdi specifically. Moreover, Figure 6.4 exhibits a post where Alrokibah explains the difference between the Arabic word /iʔana/, which translates to *charity*, and the Najdi word عونية/ʔouniah/ which means the financial help one offers to his relative or friend when getting married. The use of the pronoun /ḥna/ as he explains positions him as a Najdi and discloses his self-orientation as a member of the community.



منصور الرقيبة @M_ALROKIBH · Jan 6

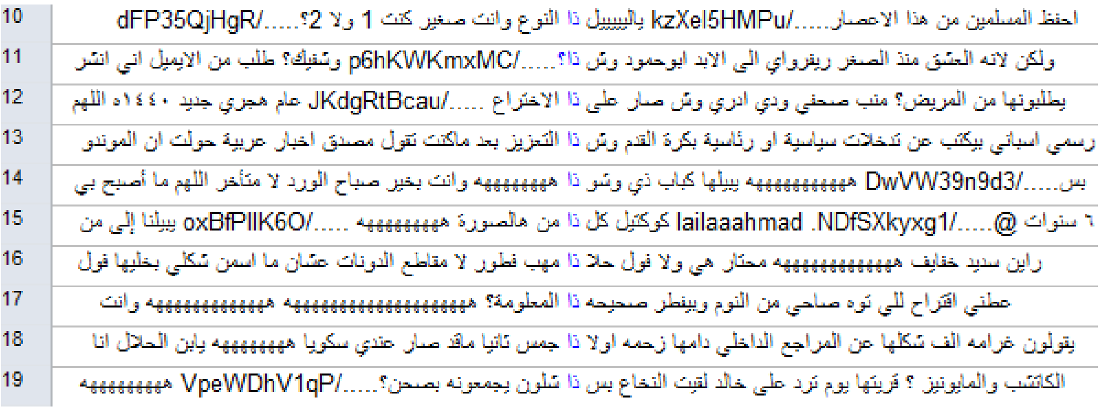
Replying to @xxopiix

حنا نسميها عونية
الاعانة تعطى المحتاج
العونية تعطى القريب والحبيب والصديق حتى لو كان ما يحتاج

Figure 6.4: Screenshot of Alrokibah’s use of /ḥna/

Similarly, Alnuhait uses the Najdi pronoun /ḥna/ *we* to refer to Saudis in expressions such as حنا أهلك/ḥna ahalk/ *We belong to you* and حنا خدام الحرم/ḥna kiddam alḥaram/ *We are the custodians of the holy mosque*. These epistemic stances show how Alnuhait perceives Saudis and reveals her regional and national identity in including herself by using /ḥna/ rather than the MSA pronoun /naḥnu/.

Furthermore, the subcorpus of Alsubayel shows relatively frequent occurrence of three Najdi variants, the topmost frequent is the singular masculine deictic /ḏa/ *that*. The concordance lines in Figure 6.5 exhibit phrases that are part of his humorous discourse like وش ذا/wiš ḏa/ *What is this* in concordance line 11, وش صار على ذا الاختراع/wiš ṣar ʕla ḏa alikṯiraʕ/ *What happened to that invention* in concordance line 12, صحيحه ذا المعلومة؟/ṣaḥiḥah ḏa almaʕloma/ *Is that information true* in concordance line 17, and وش ذا التعزيز/ wiš ḏa attaʕziz/ *What kind of support is this* in concordance line 13. Most of the occurrences are in the interrogative, which also shows Alsubyael’s stylistic stancetaking.

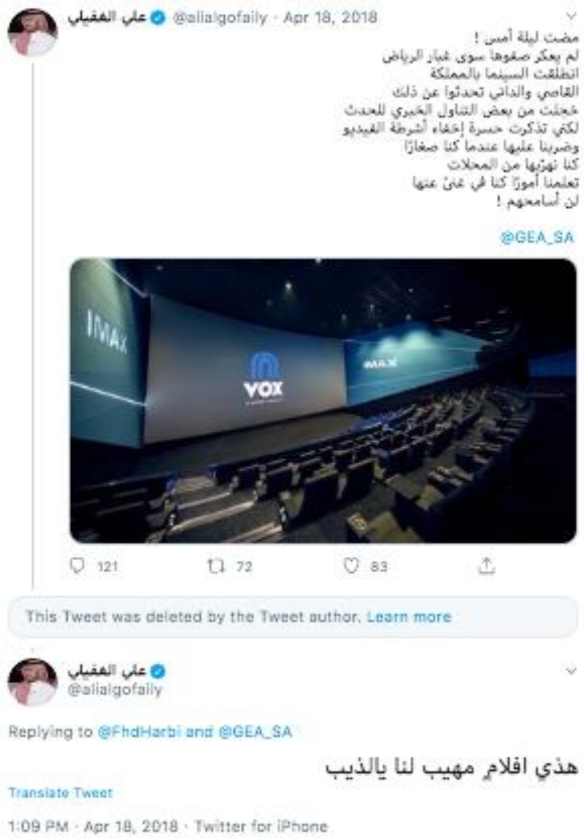


10	احفظ المسلمين من هذا الاعصار...../kzXeI5HMPu بالبيبيل ذا النوع وانت صغير كنت 1 ولا 2؟...../dFP35QjHgR
11	ولكن لانه العشق منذ الصغر ريفرواي الى الابد ابوحمود وش ذا؟...../p6hKWKmxMC وشفيك؟ طلب من الايميل اني انشر
12	يطلبونها من المريض؟ منب صحفي ودي ادري وش صار على ذا الاختراع ...../JKdgRtBcau عام هجري جديد ١٤٤٠هـ اللهم
13	رسمي اسباني يكتب عن تدخلات سياسية او رئاسية بكرة القدم وش ذا التعزيز بعد ماكتت تقول مصدق اخبار عربية حولت ان الموندو
14	بس...../DwVW39n9d3 هههههههههههه يبيلها كباب ذي وشو ذا هههههههه وانت بخير صباح الورد لا متأخر اللهم ما أصبح بي
15	٦ سنوات @...../NDfSXkyxg1 lailaaahmad كوكتيل كل ذا من هالصورة هههههههههه ...../oxBfPIlK6O يبيلنا إلي من
16	راين سديد خقايف هههههههههههههههه محتار هي ولا فول حلا ذا مهب فطور لا مقاطع الدونات عشان ما اسمن شكلي بخليها فول
17	عطني اقتراح للي توه صاحي من النوم وبيفطر صحيحه ذا المعلومة؟ هههههههههههههههههههههه هههههههههههههه وانت
18	يقولون غرامه الف شكلها من المراجع الداخلي دامها زحمه اولا ذا جمس ثانيا ماقد صار عندي سكوبا ههههههههه يابن الحلال انا
19	الكاتشب والمايونيز ؟ قريتها يوم ترد على خالد لقيت النخاع بس ذا تشلون يجمعونه بصحن؟...../VpeWDhV1qP هههههههههه

Figure 6.5: Concordance lines of Alsubayel's use of /d̠a/

Alsubayel also uses the NA plural pronoun /ḥna/ rather than the MSA pronoun /naḥnu/ to create multiple references that are also epistemic stances. For instance, he would use the pronoun to position himself as an employee in the government sector in حنا ما نزلت علاواتنا/ḥna ma nzalat ʕalawatna/ *We did not have our raise.* Another reference is as a member of the Najdi community such as بالرياض حنا الحين/birriyaḍ ḥna alḥin/ *We are in Riyadh now*, حنا نسميه أصنصير/ḥna nsamih asansair/ *We call it an elevator*, and حنا كثيرين بتويتر/ḥna kit̠irin bi twitter/ *We are so many on Twitter*. Butler (1990) explains that repetition of discourse behaviour builds up the identity, and Alsubyel's repetition of some of the NA variants constructs his regional identity and how he affiliates himself as a Najdi.

In Figure 6.6, Algofaily used the MSA /lam/ *not* in the original post to take attitudinal stances لم يعكر/lam yuʕakir/ *Did not disturb* and /lan/ *never* for لن نسامحهم/lan nusamiḥhm/ *We will not forgive them* but then shifted to Najdi in his reply in a humorous stance in هذي أفلام مهيب لنا يالذيب/had̠i aflam mihib lina yal d̠ib/ *These movies are not for us my friend*. The author used the standard variety in the original post to discuss the cultural shift the Saudi community is making with the debut of the cinema in Riyadh. The phrases /lam yuʕakir/ and /lan nusamiḥhm/ refer to the conservatives who were against having film theatres.



علي القفيلي @alialgofaily · Apr 18, 2018

مضت ليلة أمس !
لم يعكر صفوها سوى غبار الرياض
انطلقت السينما بالمملكة
القاصي والداني تحدثوا عن ذلك
خجلت من بعض التناول الخبري للحدث
لكني تذكرت حسرة إخفاء أشرطة الفيديو
وضربنا عليها عندما كنا صغارًا
كنا نهرّبها من المحلات
تعلمنا أمورًا كنا في غنىً عنها
لن أسامحهم !

@GEA_SA

This Tweet was deleted by the Tweet author. Learn more

علي القفيلي
@alialgofaily

Replying to @FhdHarbi and @GEA_SA

هذي افلام مهيب لنا يالذيب

Translate Tweet

1:09 PM · Apr 18, 2018 · Twitter for iPhone

Figure 6.6: Screenshot of Algoafily’s use of /mihib/

Regional identity is more robust in the case of Alrokaibah and Alsubayel through the frequent use of NA variants, especially the pronoun /ḥna/ *we*. The contexts in which this pronoun is used always indexes regional affiliation whether it is to Saudi Arabia, Najd, or the capital of Riyadh. Some authors such as Alabdulkarim and Alnuhait who show consistency in using MSA variants exhibit instances when they switch to the NA variety to use the pronoun /ḥna/ *we* when talking about themselves or Saudi people to create regional affiliations and to index their local identity. Other authors, however, such as Almohanna, Alwabil and Altassan express their regional identity using MSA variants. This can be interpreted as a preference to make the regional affiliation more inclusive and about the country as a whole, instead of a specific region or city.

6.2.2 Individual identity

While aspects of individual identity are more fluid and elusive to be categorised, classified and identified (Antaki and Widdcombe, 1998; Grant and MacLeod, 2020), the NASCoT reveals some patterns in the authors’ subcorpora that convey implications of individuality represented in their expression. In this section, I address the notable patterns by which the authors reveal their views and orientations using evaluative language and stance (Hunston and Thompson, 2000). I adopt Herring’s (2007) Faceted Classification Scheme, explained in Chapter 3, to enable data classification and analysis with a particular focus on the *activity* facet. By randomly selecting 25 tweets from four authors, two males and two females, I thought it would be interesting to examine how the authors use the NA features to talk about the range of activities they’re interested in (e.g., exchange of jokes, information, etc.). The selection of the tweets is random; however, the selection of the sample is not. I examine male and female authors to investigate whether gender variation has any weight in the activities the authors carry out. Also, I opted for authors who show a higher tendency to use NA variables, to investigate how they use them in communicating their activities. Table 6.1 shows the raw figures for each author: the female authors are Alajlan and Aleidi, and the male authors are Alrokibah and Alsubayel, and the classification of the activity as per Herring’s scheme. I noticed that the most frequent activities in the subcorpora can be classified as debate, information exchange, joke exchange, or insult exchange. Moreover, I came across a few tweets where the authors are posting games to entertain their followers, so I included the activity in the table. One of the interesting things to be observed is that the sample can show us is how each author has a different pattern in their activities. The first author in the sample, Alajlan, shows insult exchange at the top of the activities, joke exchange comes second, and debate last. Observing the instances in Alajlan’s subcorpus in Chapter 5 and the previous section reveals to us that she engages in discourse that involves women, feminists and their opposers, as well as people who advocate for the return of Alsahwa movement.

Table 6.1: The NASCoT sample authors' activity

Author	**Activity**				
	Debate	**Information exchange**	**Joke exchange**	**Insult exchange**	**Game**
Alajlan	5	-	7	11	-
Aleidi	2	13	3	5	2
Alrokibah	8	6	10	1	-
Alsubayel	-	7	12	3	2

Example 6.5 shows an instance of her defensive posts against Saudi feminists who accuse her of being against women's empowerment in Saudi Arabia and question her loyalty.

Example 6.5: Alajlan's use of /wen/

وين الانفصام؟! فعلا كان هذا الحال موجود قبل ٢٠١٦ ومازال هناك اثر لليوم..
لكن الفرق اني ما ابيع نفسي لجهات خارجية عشان احارب بلدي، وكمان كويس
انكم طلعتوا التغريدات اللي تثبت اني عمري ماكنت ضد حصول المراة على
حقوقها كما ادعوا الحقوقيات

/**wen** ilinfiṣam?! fiʕlan kan haḏa alḥaal mawjod gabil 2016 w ma zaal hunak athar lilyom .. lakin ilfarg inni ma abiʕ nafsi lijihat ḵarjiya ʕašan aḥarib baladi, w kaman kuwais innikum ṭallaʕto ittaḡridat illi tithbit inni ʕumri ma kint ḍid huṣol ilmar'a ʕla huqoqha kama iddaʕo ilhuqoqiyat/

***Where** is the discrepancy? That really was the situation before 2016 and the effect is still there till today .. but the difference is that I will not sell myself to foreign entities to fight against my country, and it is a good thing you dug up the tweets that prove I was never against women rights as the feminists claim.*

Alajlan chooses the NA variant /wen/ *where* in /wen alinfisam/ *where is the discrepancy?* and continues to use the dialect throughout the tweet to identify herself as a member of the region and of the community. She also shows criticism to the people who sell themselves for external entities which implicates her values in terms of nationalism and loyalty. Aleidi's sample exceptionally covers all activities selected for this analysis, although it appears that information exchange is at the top.

Also, both Alrokibah and Alsubayel show a trend in joke exchange compared to other activities; the numbers indicate that the male authors of the sample partake in joke exchange more than their female counterparts. Looking at the tweets, some of which are shown in Table 6.2 with the NA variants highlighted in orange, we can see that all of the authors have a sense of sarcasm when joking but the male authors practice it more frequently.

Table 6.2: Examples of the NASCoT authors' activity

Author	Tweet	Activity
Aleidi	معي ضيفة جميلة بسناب شات **وش** تبونها تتكلم عنه؟ I'm hosting a lovely guest on Snapchat, **what** do you want her to talk about?	Asking for information
Aleidi	شاركونا **وش** اكثر موقف مستحيل تنسونه بأيام الدراسة؟ Tell us **what** the most memorable situation from school days is?	Asking for information
Alajlan	اذا كان هذا اعلان لشركة كبيرة ومحترمة.. **وش** تتوقع تكون النتيجة؟ ! If this is an advertisement by a big respectable company then **what** result do you expect?	Joke exchange
Alajlan	**مانيب** حاذفتها.. بشوف من فاهم قصدي ومن الغبي .. لقيت شي اتسلى فيه I'm **not** deleting it, I want to see who get me and who is stupid, I find it entertaining.	Exchange of insult
Alrokibah	**وش** الحل مع المغردين اللي يحس انه من الفخامه والرقي والتحرر والانفتاح (سب الدين) **What** to do with tweeps who find offending religion as a sign of sophistication, liberation, and open-mindedness?	Initiating a debate
Alrokibah	عسى حظنا **موب** بالكوره Let's hope our luck is **not** in football.	Joke exchange
Alsubayel	الدوام **ليه** مافيه فسحه؟ **Why** is there no lunch break in the workspace?	Joke exchange
Alsubayel	اللي بيعرف **وش** الشي المشترك بين الاشخاص بالصور له 1000 ريال وبعطيكم مهلة الى السبت العصر Whoever knows **what** is the common thing between the people in the picture gets 1000SAR and you have until Saturday afternoon.	Game

Exploring Alrokibah's sample shows frequent temporal and spatial deixis, mostly in the Najdi variety. The topmost frequent temporal deixis is the proximal /alyom/, whose most frequent collocate is the preposition /fi/. The collocates show Alrokibah's pattern of using them to talk about his daily activities and promote his Snapchat account as an influencer such as في سنابك اليوم/fi snabik alyom/ *In your*

Snapchat today, في تقريري اليوم/fi taqriri alyom/ *In my report today*, and تقريري اليوم في سنابتشات/taqriri alyom fi snapchat/ *My report today in Snapchat*. Alrokibah also uses the MSA variant /al’aan/ to create similar posts in Modern Standard Arabic. The subcorpus exhibits instances such as أرسل لهم في تويتر/arsil lihum al’aan fi twitter/ *Send them now via Twitter*, نستقبل المكالمات الآن/nastaqbil almukalamat al’aan/ *We receive phone calls now*, التغطية الآن مباشرة/attaḡtṭya al’aan mubašara/ *The live coverage is now*, and شاهد عرفة الآن/šahid ʕarafa al’aan/ *Watch Arafat now*. Examining the instances of information exchange in Alrokibah’s sample reveals that he provides information rather than asks for it by sharing information about his Snapchat activities.

I expand the trends Aleidi exhibits in her use of /wiš/ for information exchange to investigate whether she uses it exclusively for that purpose, and if not, to discover the other activities. Figure 6.7 shows a classification of all the occurrences of the NA variant in terms of activity. As discussed in Chapters 3 and 5, Aleidi’s subcorpus has the second highest frequency rates in interrogatives. The total number of occurrences is 237 times, where 128 occurrences are classified as information exchange, specifically asking for information. The first example in Table 6.2 shows us Aleidi asking her followers for a suggestion; in the second one she is asking them to share their memories from school. The information exchange exceeds the other categories, which is an indicator of how Aleidi uses her platform as an agent that initiates interaction. The interrogative was also used in jokes and sarcastic comments 36 times, 33 times to start a debate, 21 times to ask questions as games, and only five occurrences where she used it for insults. The jokes exchange reveals her behaviour in developing the interaction, which is different from Alajlan’s approach for instance. She also uses interrogatives as part of the games she plays with the audience. These figures are a quantitative representation of Aleidi’s discursive behaviour as far as the interrogative /wiš/ is concerned, which is a unique pattern compared to the other authors in Table 6.1.

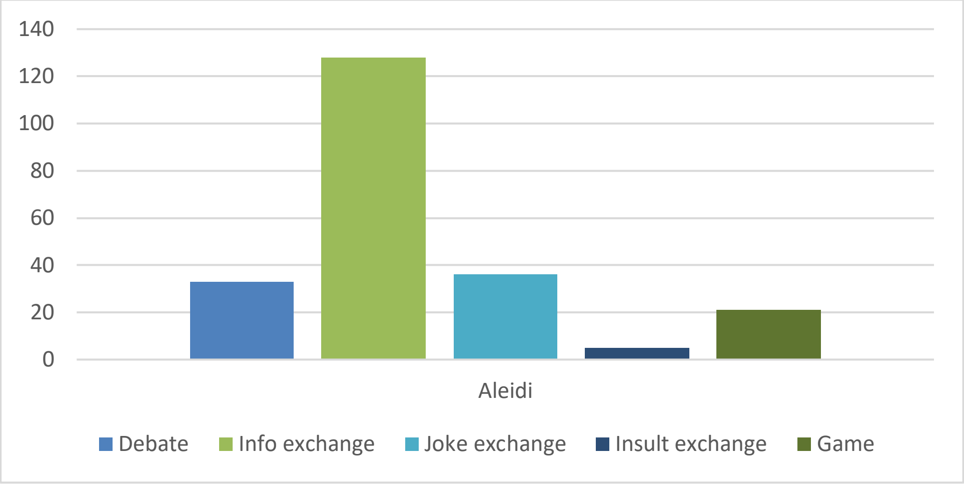


Figure 6.7: Aleidi’s activities using the NA interrogative /wiš/ (raw figures)

Furthermore, the NASCoT shows that the NA temporal deixis /alyom/ is a keyword for the authors when posting about their work, which can be classified as exchange of information with their audience. Alnuhait’s subcorpus had instances of /alyom/ in contexts related to her work such as أتشرف اليوم بالمشاركة كمتحدثة/atšarraf alyom bilmušaraka kamutaḥadiṯa/ *I am honoured to be speaking today*, سعدنا اليوم بزيارة وزارة الخارجية/saʕidna ayom biziyarat wizarat alḵarijiya/ *We are delighted to visit the Ministry of Foreign Affairs today*, and سعدت اليوم بحضور أمسية جميلة/saʕidt alyom buḥḍor omsiya jamila/ *I am delighted to attend a beautiful event today*. Alrasheed’s subcorpus has one of the highest frequency rates in temporal and spatial deixis and the most frequent collocate is مقالي اليوم/maqali alyom/ *My article today*.

Alrasheed also uses the NA distal /ḡadan/ to create epistemic stances about the publishing of her newspaper column. My study of concordance lines of the deixis shows occurrences such as غدًا مقالي للأحد/ḡadan maqali lilaḥad/ *My article for Sunday tomorrow* and غدًا في عكاظ/ḡadan fi ʕukaẓ/ *Tomorrow in Okaz*. Alwabil is another author who posts about her, whose subcorpus shows a number of occurrences of the proximal /al’aan/. The concordance lines show a range of posts sharing her activities such as يحدث الآن ملتقى علوم الإدارة/yaḥduṯ al’aan multaqa ʕulom alidara/ *Happening now is the administrative sciences forum*, and يحدث الآن في النادي الأدبي/yaḥduṯ al'aan fi annadi aladabi/ *Happening now at the literary club*. Alzamil also uses the proximal temporal /alyom/ to post about his work, the subcorpus revealing two topmost collocates. The first collocate is /alyom fi/ *Today in/at*, which concordance lines show instances of, such as لقائي اليوم في برنامج ترند/liqa'i alyom fi barnamaj trend/ *My interview today on Trend*, في حلقة اليوم/fi halqat alyom/ *In today's episode*, and اليوم في رد مول/alyom fi red sea mal/ *Today at Red Sea mall*. The second frequent collocate is /alyom ʕan/ *today about* whose occurrences are similar to the previous collocates. Alzamil uses these to specifically refer to activities related to his YouTube channel posts. Examples like أتكلم اليوم عن مسلسل/atkallam alyom ʕan almusalsal/ *Today I talk about the series*, مراجعة اليوم عن فيلم براد بيت/murajaʕat alyom ʕan film Brad Pitt/ *Today's review is about Brad Pitt's film*, and تكلمت اليوم عن الحلقة/tikallamt alyom ʕan alhalqah/ *I talked today about the episode*.

What is interesting is that when the authors share information about their daily activities, they tend to use the MSA variety. In the case of Alnuhait, Alabdulkarim, Alwabil, and Alrasheed, using such a variety complements the formal nature of the activities they perform. For instance, Alwabil and Alrasheed are writers at local newspapers which publish in MSA; therefore, their journalistic identity crosses over into their social media identity when talking about their columns with the same variety they use in that genre. What is more interesting is that social media influencers such as Alrokibah, who is usually informal, shifts to the formal variety MSA when talking about the daily activities. However, one could argue that it is different in the case of Alzamil whose subcorpus shows preference to use NA variants across all categories. But since the time deixis /alyom/ overlaps in both varieties, it is difficult to claim which variant Alzamil's subcorpora exhibits.

The themes observed in the NASCoT reveal the identity implications of each author. Collective identity themes appear more prominently, namely nationalism, gender, and region. The authors' preference to use MSA and NA variants in particular contexts and sometimes switch between glosses within a single tweet can implicate on the one hand their regional affiliation to the region of Najd and more commonly to Saudi Arabia. On the other hand, their occasional preference to use MSA can be interpreted as a way to override such regional affiliation and to be more inclusive of all Saudi followers. It appears that regional identity is often revealed through discourse, which conforms to the approach's principles of emergence and positionality. The authors' replies or comments disclose how they affiliate themselves regionally and signals their awareness of their position as members of the Najdi community. The corpus shows a pattern of instances where the authors post a tweet in MSA but reply or comment using the NA variants, which can be interpreted in terms of the principles of emergence and positionality as well as linguistic accommodation. That doesn't seem to be the case for nationalism, which is prominently communicated across all authors' subcorpora in both varieties. Alajlan's national identity is reflected in the epistemic and attitudinal stances that support her country and in the way she criticises those who argue with her. Her gender and regional identities reveal her self-perception and how she positions and identifies herself as a Najdi woman who is concerned with empowerment. As for gender, many studies emphasise that women utilise linguistic devices that establish relationships with their audience. While Lakoff (1975) and Labov's (1990) work reports that women opt for using the standard variety, Ismail (2012) claims differently in the case of Saudi women who use the dialectal variety more frequently. In the case of the NASCoT, female authors used both MSA and NA variants liberally, but gender implications are revealed and conveyed in the issues they raise in their tweets. Evaluative language and stance appear to be active variables in the authors' revelation of the individual aspects of their identities. The NASCoT shows numerous instances where the authors express their positive and negative perceptions and views about local events, issues in the community, or comment on someone in their conversation.

6.3 Conclusion

This chapter concludes the analysis part of this research, which approaches authorship with a focus on online identity by proposing the following question:

1) How does the idiolectal style of these authors contribute to the construction of their online identities?

I employed the identity approach (Bucholtz and Hall, 2005) to analyse the authors' discourse and its implications for the construction of their collective identities. I also adopted the Faceted Classification Scheme (Herring, 2007) to classify a sample of the NASCoT authors' tweets and the activities they carry out to explore implications for their individual identity. The findings show that the principles of the identity approach can reveal macro-level themes such as nationalism, gender, and region. The authors show individual variation in choosing the variants in either variety, which has some revelations in terms of relationality, positionality and indexicality. For instance, Altassan's post in MSA رادع باسم القانون لمن لم تردعه رجولته/radʕ bi ism alqanon liman lam tardaʕuh rujolatuh/ *Deterrence by law for men who do not deter themselves* is an example of authentication and authorisation because her stance shows support of civil laws that protect women from domestic violence. Another example is Alajlan's subcorpus that exhibits phrases such as هذول ربع الرشاوي/haḏol rabʕ arrašawi/ *Those who take bribes* and هذول حقون الطاقة/haḏol ḥagon aṭṭaqa/ *Those who believe in energy*, which are also instances of authentication. The second part of the analysis tackled implications of individual identity detected through the authors' activity, with an investigation into the trends for one of the authors. The findings show that the authors’ subcorpora reveal patterns of the kinds of discourse activities they perform and the variants they select in that performance. Finally, investigating into the activity trends of Aleidi enabled us to examine closely her distinctive pattern, particularly how she uses the NA variant /wiš/ more often to ask for information than to share them. The discussion of this chapter’s findings will further unfold in Chapter 7.

Chapter 7
COMPUTATIONAL ANALYSIS: EXPLORING WEKA

7.1 Introduction

Authorship research has been an interest shared by linguistics and computational sciences, as noted in Abbasi and Chen (2005). There have been recent works where both fields align to develop methods. This data analysis chapter aims to further support the qualitative, descriptive analyses in the previous ones by offering objective and quantitative findings I shift the focus to computational approaches and explore how machine learning tools can help tackle authorship attribution problems and to also explore the configuration that yields most accurate findings, which concerns the first research question in this chapter:

3) What is the potential of WEKA to attribute authorship using the lexcio-grammatical feature set?

Using the stylometric feature set that includes Najdi Arabic and Modern Standard Arabic discussed in Chapter 4, I venture to experiment with a machine learning tool, WEKA, and test a range of classifiers and parameters. Classifiers are algorithms that help the machine learning tool to categorise any given data; and the term parameter in machine learning refers to tuning the settings of word frequency to enhance the performance of the classifier. This venture is proposed in the other two research questions:

4) Which classifier and parameter can accurately identify authorship using the NASCoT corpus?

These questions address the computational approach in the methodological synergy of this research. The chapter begins with a brief discussion of the methods used in authorship attribution research from a computational perspective. Then in 7.3, I introduce WEKA, its background and how it operates. I also explain the methodology including the pre-processing of the data and configuration of the classifiers and parameters in 7.4. The chapter concludes with the findings and their connection to the literature. I would like to note that material in this chapter has been presented at the seventeenth International Conference in Natural Language Processing and published as part of the proceedings (AlAmr and Atwell, 2020).

7.2 Computational methods in authorship attribution

Computational studies in authorship analysis, specifically authorship identification, aim to find the optimum classifiers, parameters, and n-grams that achieve the task with the highest accuracy rates. Numerous studies confirm that accurate authorship identification results can be achieved using small-sized data (Rico-Sulayes, 2011; Brocardo et al., 2014; Saha et al. 2018). Moreover, several studies found that Linear Support Vector Machine (SVM) demonstrate accurate classifying results compared to others (Brocado et al., 2014). Alternatively, other classifiers such as Decision Tree J-48 and Multinominal Naïve Bayes perform more accurately with numeric data (Maruktat et al., 2014). Al-Harbi et al. (2008) explored their Arabic Text Classification Tool (ATC Tool) to classify the texts of seven different Saudi newspaper corpora and two classifiers (SVM and C5.0). They found that the performance of the algorithm C5.0 has the accuracy average of 78.42% while the SVM algorithm scored 68.65%.

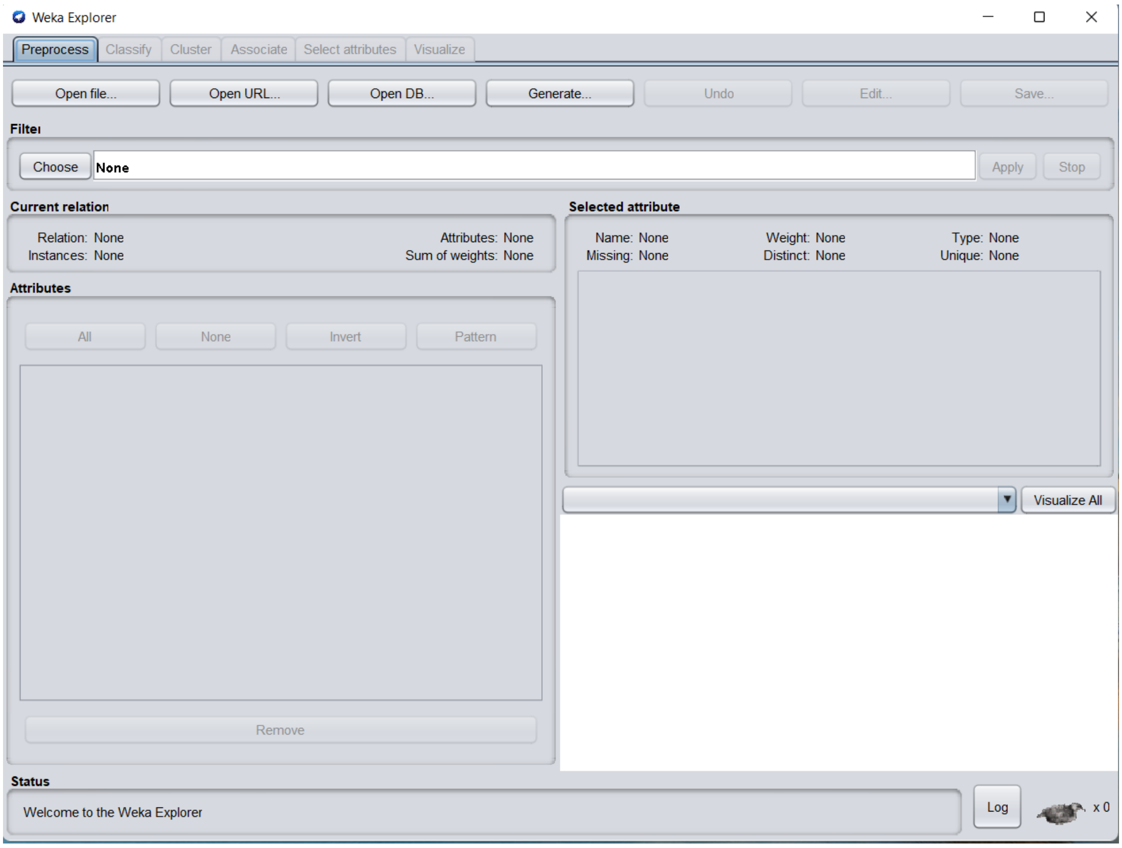
In terms of n-grams, character grams proved to perform well in short texts such as WhatsApp and Twitter but with some limitations to identify authors’ texts without qualitative analysis (Shrestha et al., 2017; Banga et al., 2018). As for word grams, some studies conclude that unigrams even in the shape of an emoticon can show good results in identifying authorship (Fissette, 2010). Bigrams proved to be successful in identifying authorship in literary texts (Feiguina and Hirst, 2007) and in Arabic dialect identification (Sadat et al., 2014).

In addition, a body of literature has been published in the authorship identification field in Saudi Arabia that investigates computational approaches (Alruily, 2012; Althenayan and Menai, 2014; Al-Tuwairesh et al., 2015, 2018; Assiri et al., 2016) which rely on stylometric features (e.g., syntactic, lexical, or morphological variables). However, these studies do not account for a linguistic theory or a theory of idiolect. This calls for a need to contribute to the field of forensic linguistics in general and forensic authorship analysis in Arabic in particular. Social media platforms such as Twitter are heavily populated by users who sometimes abuse such media. Saudis are responsible for 30% of the tweets posted (Salim, 2017). Simultaneously, there are efforts to fight cybercrime and issue regulations that incriminate people who use and publish hate speech and offensive language online.

To sum up, employing computational tools in authorship attribution problems is not new. Since electronic texts and word processors were made available in the 1990s, research has investigated various machine learning tools and different configurations of classifiers and parameters to conduct classification tasks and authorship analysis problems. Corpora and electronic texts were the subject of these studies in English, Chinese, Arabic in its standard and dialectal varieties.

7.3 About WEKA

As mentioned earlier in Chapter 1, Waikato Environment for Knowledge Analysis (WEKA) is a machine learning tool developed at the University of Waikato, New Zealand written in Java script (Witten et al., 2016). It is a collection of machine learning algorithms that perform data mining tasks, and its tools can achieve pre-processing, classification, clustering, and can develop new machine learning schemes. One of the qualities of WEKA is that it is user friendly in the case of those with limited computational or programming skills because of the way it is designed. As shown in Figure 7.1, the home screen is easy to read and selecting and operating the data files can be straightforward. One of the challenges in WEKA, however, is that it only reads its own file format, that is ARFF files. The ARFF format strictly allows data to be in a single line and no punctuation marks or symbols are allowed except for periods. While WEKA does provide the option of converting .CSV files into ARFF, that feature is only effective if the files are in the aforementioned configuration.



Filter: None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

Welcome to the Weka Explorer

Figure 7.1: WEKA home screen

7.4 Methodology

The approach implemented at this stage is computationally focused. Nonetheless, it contributes to the synergy between corpus, stylometric and computational methods, hence the triangulation of approaches that lay the groundwork of this study. I use the

NASCoT as my data, which was discussed in Chapter 4. I explain in 7.4.1 the pre-processing and preparation of the data as well as the adjustments to the size of the files that were made to explore WEKA's potential.

One of the qualities of WEKA is that it accounts for numeric and nominal values, but since the data is in standard Arabic and some of it is dialectal, I was concerned that nominal values might compromise the data. Instead, I assigned numeric values for each variant and used a bag of words representation by assigning a fixed integer that identifies each linguistic variant that would occur in any of the training data files. I continue to explain the role of stylometric features in the following sections.

7.4.1 Balanced vs unbalanced data

To explore the NASCoT using WEKA, I had to convert the .CSV files into ARFF file format. That means that all data has to be in a single-line and one needs to eliminate all punctuation marks and symbols except for periods. Moreover, the coding of the Arabic text had to be converted to UTF-8 so that the software can read it. In order to train WEKA, I reassembled the corpus into thirteen separate ARFF files. I created two data training sets as shown in Table 7.1; the first is an unbalanced data set (TS1) which includes the full corpus. The aim of this unbalanced set is to explore how WEKA performs in spite of variation between the authors' subcorpora sizes, which can be the case in an authorship attribution problem. The subcorpora of some authors are substantially larger in size compared to others. Therefore, I created a balanced data set (TS2) which includes an equal number of tweets per author. The balanced set aims to explore whether training WEKA with equal-sized files affects accuracy rates or not. Both balanced and unbalanced data sets include 80% of the authors' full data, and a header describing the types of linguistic variables and their variants under investigation. The remaining 20% of the authors' subcorpora was combined into one ARFF file for testing.

Table 7.1: Number of tweets per dataset (TS1 and TS2)

Unbalanced data set = TS1	Balanced data set = TS2
44192	25247

After preparing the datasets, I move to explore the ranges of classifiers and parameters to pursue the ones that perform most accurately.

7.4.2 Classifiers and parameters

Some classifiers perform more accurately when dealing with verbal data such as Linear Support Vector Machine (SVM), and other classifiers work best with numeric data like decision tree J-48. I wanted to examine different classifiers to see which performs most accurately, hence I ran a test using seven classifiers: Linear SVM, Multinominal Naïve Bayes, Decision tree J-48, KNN Depth=3, KNN Depth=5, Random Forest Estimator=5, and Random Forest Estimator=15. Table 7.2 shows the performance of seven different classifiers in three categories: unigrams, bigrams, and trigrams.

Table 7.2: Accuracy rates per classifier

Parameter	Unigram	Bigrams	Trigrams
Linear SVM	0.59	0.6	0.6
M Naïve Bayes	0.47	0.48	0.48
J-48	0.4	0.4	0.4
KNN Depth=3	0.25	0.25	0.25
KNN Depth=5	0.25	0.25	0.25
Random FE=5	0.4	0.42	0.42
Random FE=15	0.46	0.47	0.47

The results of the classifiers test shows that Linear SVM scores the highest accuracy rates; therefore it was implemented in the next stage: the parameters. I ran three tests to explore a range of parameters that can ensure the highest accuracy rates. In the first range, the minimum value is words that appear once and words that appear in 60% of the data files (min_df=1 – max_df=int (60/100)). The second one eliminates words that appear twice or less and words that occur in 80% of the data files (min_df=1 – max_df=int (80/100)). The last parameter test eliminates words that appear once and words that appear in 95% of the data files (min_df=1 – max_df=int (95/100)). Table 7.3 shows the accuracy rates of different parameters in both data sets.

Table 7.3: Accuracy rates per parameter using Linear SVM

Parameter	**Unigram**		**Bigrams**		**Trigrams**	
	TS1	**TS2**	**TS1**	**TS2**	**TS1**	**TS2**
1–60/100	0.59	0.58	0.6	0.6	0.59	0.6
2–80/100	0.59	0.58	0.6	0.45	0.6	0.29
0.001–95/100	0.49	0.58	0.49	0.59	0.49	0.59

In the first parameter test, both data sets scored the highest accuracy rates. The scores were most accurate across the three n-gram categories (0.59-0.6 respectively). The first data set TS1 scored a consistent and better performance compared to TS2 in the second parameter. The accuracy rates of the balanced data TS2 in the second parameter test were inconsistent. On the other hand, TS2 scored consistently higher results in the third parameter test compared to the unbalanced data set TS1. Nonetheless, both training data sets scored consistent results in unigrams, bigrams, and trigrams. Furthermore, it appears that the balanced data set scores the highest overall results in unigrams, while the unbalanced data set scores the highest overall results in bigrams. However, the optimum parameter is the first test using bigrams.

7.5 Findings and discussion

This study aims to explore the potential of WEKA in forensic authorship analysis research. Moreover, it aims to explore how incorporating computational methods into the synergy of approaches can enhance and support the corpus-based stylistic findings discussed in the previous chapter. To answer the first question, I found that the unbalanced data set performed better than the smaller, balanced one.

Interestingly, the large size of the data provides WEKA with sufficient training which enabled the tool to recognise the stylistic variables and their variants more accurately. The more the tool is trained to recognise them, the better its performance in assigning them to the correct authors at the testing stage. As for which classifier performs best, results show that Linear SVM has the most accurate performance. This conforms to the findings of Fissette (2010) and Braocardo et al. (2014). As for Al-Harbi et al. (2008), their findings show that C5.0 outperforms SVM, but since this algorithm wasn't explored in this study, it can be claimed that the SVM scores reported in their study support our findings. Lastly, the results show that bigrams can accurately identify authorship, which confirms the findings of Feiguina and Hirst (2007) and Sadat et al. (2014).

7.6 Conclusion

This chapter discussed the first part of the methodological synergy proposed in this doctoral research, which produced some discoveries. The first one is the potential of the machine learning tool WEKA in attributing authorship in a forensic context. The final revelation is the range of classifiers and parameters that perform most accurately. Moreover, the literature in computational authorship research supports these findings. Also, in a conversation with forensic linguists who use computational approaches, it appears that we can have the confidence in accuracy rates of 0.59 – 0.6 with short texts. Nevertheless, for the purposes of cross-checking required in opinions for evidential purposes, qualitative and white-box results are also beneficial as part of the methodological approach. It is to those approaches that I now turn, the following chapter addresses a different part of the synergy that explores linguistic patterns revealed by corpus linguistics and stylistic methods.

Chapter 8
TESTING CONTINUITIES IN IDIOLECTAL PRACTICE

8.1 Introduction

Up to this point, this thesis has explored research questions that aim to advance our understanding of idiolectal style and identity from a theoretical and an empirical point of view. However, one of the main objectives of forensic authorship research is to explore the potential of its methods in forensic casework. The literature of forensic authorship attribution, discussed in Chapter 2 and throughout this thesis, usually tackles the important matter that is the applicability of methodologies and research findings in authorship casework and any subsequent court proceedings. The previous data analysis chapters demonstrated the potential of the triangulated methodology from different perspectives: machine learning, corpus-stylistics, and stylistics-CMDA. This chapter begins with an empirical investigation that leads to a discussion of the three approaches used in this research and their implications for casework. The following section introduces the scenario of the case simulation followed by a demonstration of the data analysis and a discussion of the findings. The general discussion of the thesis’s findings is carried out with respect to the theory of idiolect implemented in this research as well as its implications in terms of forensic authorship research and casework.

8.2 The scenario

A common scenario in the literature of authorship attribution is that the dispute over the authorship of a questioned or unknown text would be between two candidate authors (Coulthard and Johnson, 2016; Grant, 2020). It was noted in Chapter 2 that authorship attribution problems have been the subject of cases such as in the murder of Jenny Nicholl; the questioned text messages were written and sent from her phone, but linguistic evidence reported by Coulthard suggested (see Grant 2020) that it was her lover, David Hodgson, who wrote those messages and he was, in fact, convicted of her murder in 2008. Another case is Danielle Jones who disappeared in June 2001 and the police suspected that it was her uncle, Stuart Campbell, who sent the last two messages from her phone hours after her disappearance. Coulthard offered a linguistic analysis that stated that it was unlikely that Danielle was the author of those messages, which partly contributed to the conviction of Campbell in December 2002 for her murder (Grant, 2020). These cases show the importance of stylistic features in revealing the identity of a suspected author. This case simulation study investigates the following questions:

5) How likely is it for the methodology and the Najdi Arabic feature set proposed earlier in this research to identify authorship in forensic casework?

In the forthcoming sections I explain the design of the empirical test and the selection of the authors chosen for this experiment and revisit the stylistic features that distinguish their idiolectal style, specifically the features in the Najdi variety. I introduce the new set of tweets, explain the process and time of their collection, and demonstrate the data analysis and findings.

8.2.1 The experiment’s design

Grant and MacLeod (2020) propose that stylistic consistency in authorship casework, supported by a theory of idiolect, can be reliable ‘where any points of consistency can… be demonstrated to be distinctive against a relevant comparison corp[us]’ (p. 16). With reference to the theory of idiolect discussed earlier in Chapters 1 and 2, I explore the idiolectal style of two authors in terms of the system of the dialect they are using and their individual production of that dialect, following Turell’s (2010) definition of idiolectal style. In this case simulation, I intend to focus on the Najdi variety exclusively to explore the potential of its variants in solving an authorship problem. Since the NASCoT contains the subcorpora of 13 authors, I employ them in this case simulation as known authors with known texts or tweets. I then select two of the 13 for this empirical test. These are Abdulrahman Allahim and Abdulaziz Alzamil; I chose these two men to have a homogenous sample of male authors who show preference to use the NA variety. Table 8.1 shows a sample of previous tweets by Allahim and Table 8.2 shows a sample of previous tweets by Alzamil, both extracted from their NASCoT subcorpora.

Table 8.1: Sample of 5 tweets posted by Abdulrahman Allahim.

Tweet
1
تعبنا من هذه الشركة .. لا أحد يجيب و**حنا** نفسياتنا اليوم لا تخفى على أحد We are tired of this company.. no one answers and **our** moods today aren’t hidden from anyone
2
ذولاك مو كنهم بالغوا بالفرحة والاحتفال وكذا ؟ Aren’t **those** exaggerating in their celebrations and so on?
3
إن صدقتَ ؛ فإن تأويل هذه الرؤيا بأن النصر سيعود لمجده القديم (شعار الجزيرة) وسيعود لجلد **ذولاك** و سيحقق ١٤ بطولة متواصلة .. والله أعلم If what you’re telling me is true, then the interpretation of this dream is that Alnasser will return to its old glory (with their Peninsula logo) and will beat **those** with 14 consecutive championships .. and Allah only knows
4
ذولاك هوامير وانا على باب الله ثم اخاف يتنمر علي **ذولاك** **Those** are big fish and I’m a simple man so I’m afraid **they** will bully me
5
صباحكم خير و بركة ومطر الحين **وراهم** مايخلون دوام المحاكم يبدأ ٦ صباحاً ؟ Have a good, blessed, rainy morning now **why** don’t they start court at 6 in the morning?

I mark the NA stylistic features that can be classified as within-author consistent as per the findings in Chapters 5 and 7 for both authors. The subcorpora for both authors provide an ample amount of data to identify their stylistic consistencies, each being represented by the sample tweets. Additionally, it is important at this point to see if there is any linguistic relevance between the author’s stylistic choices and the questioned tweets which will be introduced in Table 8.4.

Table 8.2: Sample of 5 tweets posted by Abdulaziz Alzamil.

Tweet
1
عندي فكرتين متردد بهم لليوتيوب وابي تساعدوني بالإختيار:١ -حلقة أسئلة أجاوب فيها على كل شي تبونه ٢ - حلقة قصيرة نناقش تيزر أو تشويقية كل حلقة قبل نزولها و**وش** نتوقع ممكن يصير .شرايكم؟ I have a couple of ideas for YouTube but I’m hesitant and need your help: 1- an

episode where I answer all your questions 2- a short episode to talk about teasers before the release of each episode and **what** do we expect to happen .what do you think?
2
روضه وكذا اجل **وش** بيسوون بالثانوي ذولا Kindergarten and that’s how they act then **what** will they do in high school
3
مب عن تشجيع، بس ابي نقابل الاتحاد احس افضل. It’s **not** about rooting, I just want us to meet Alittihad I think it’s better
4
اشاعات كبيرة حول ظهور **قريب** لشخصية "والتر وايت" في مسلسل Big rumors about the appearance of Walter White in a series **soon**
5
ماعمري فهمت الي يطمر حلقات بالمسلسل، بالذات الي معروف انها **مب** شي فرعي بالقصة. هذا ماتدري يشوف للمتعة ولا يلخص لإختبار. I’ll never understand the one who skips episodes in a series, especially when it’s known that it’s **not** irrelevant to the storyline. You never know if they’re watching for fun or summarising for a test

Comparing the authors’ subcorpora in Table 8.3, we see the three topmost features in each stylistic category: interrogatives, negatives, and deixis (person, time, and space) along with the percentages of their frequency rates. At this stage, I identify what Grant (2012) describes as between-author distinctiveness, which is to identify each author’s idiolectal style through pairwise demonstration. The idiolectal style of both authors is similar in terms of interrogatives and negatives, but the differences reveal themselves in the ranking of the features. For instance, both authors use /wiš/ *what* and /wen/ *where*, but Allahim shows preferences to use the Najdi variant /warah/ for *why* while Alzamil uses a different NA variant /leh/. Similarly, the MSA negative variants of *not*, /lam/ and /lan/, are used by both authors, but it is the NA variants of *not* /manib/ and /mub/ that set the authors’ styles apart. The variation between the authors shows in the deictic expressions, specifically person, where each uses different expressions and at different rates except for the NA variant /ḏa/ *that*. As for time and space deixis, the ranking of the features is relatively similar, but the frequency rates are different. The MSA temporal deixis /alyom/ *today*, for example, is the topmost item for both authors, but it occurs almost three times more frequently in Alzamil’s subcorpus (3.66) compared to Allahim (1.34). On the other hand, /al'aan/ *now* occurs at the rate (0.55) in Alzamil's subcorpus while it is almost doubly frequent in Allahim's (1).

Once I had chosen my two authors, I created a NA feature list for each one as in Table 8.3.

Table 8.3: The NA stylistic features of the authors (per 1000 rates)

Author	Interrogatives	Negatives	Deixis (person)	Deixis (time/space)
Allahim	/wis/ 1.7	/lam/ 1.41	/di/ 1	/alyom/ 1.34
	/wen/ 0.7	/lan/ 0.66	/da/ 0.7	/hinak/ 1
	/warah/ 0.18	/manib/ 0.6	/dolak/ 0.52	/al'aan/ 1
Alzamil	/wis/ 3.55	/mub/ 1.46	/da/ 0.66	/alyom/ 3.66
	/leh/ 0.8	/lam/ 0.63	/hna/ 0.23	/girib/ 0.75
	/wen/ 0.7	/lan/ 0.6	/dak/ 0.2	/al'aan/ 0.55

For the questioned tweets, I collected new data posted throughout January and February 2022. Given the small amount of data, I collected them manually by copying each tweet and its permanent URL into an Excel sheet. Looking at the questioned tweets (Table 8.4), a preliminary observation is that all the published tweets are in Najdi Arabic except for three tweets (no. 2, 3, and 17) that are in Modern Standard Arabic. This increases the chances of examining the authors' idiolectal style in terms of Najdi features.

Table 8.4: The twenty tweets under examination

#	Tweet
1	احتفالات بالعام الجديد و مطر وخير و الناس سعيدة ... يالله لك الحمد New year celebrations, rain, blessings, and everyone is happy …. Thank God
2	صباح الخير للجميع ان شاء الله كل أيامكم في هذه السنة الجديدة تشبه هذا **اليوم** وجمال جو هذا **اليوم** Good morning to all I pray to God that all your days in this new year are as beautiful as **today** and the beauty of the weather
3	**اليوم** عشت ليلة من اجمل الليالي في #ليلة_تريو_الرياض ليلة من الخيال لقد تغيرنا كثيراً وأصبحت حياتنا تفيض سعادة وبهجة بفضل هذا الزعيم العظيم الذي خلق المعجزة .كل الشكر لمنظمي هذه الليلة ومن صنع هذه الأجواء الجميلة و على رأسهم ابوناصر **Today** I experienced one of the most beautiful nights at Riyadh Trio Night it was beyond imagination we have changed so much and our lives exude with happiness and joy thanks to this great leader who made this miracle .All thanks goes to the organisers of this beautiful night headed by abo Nassir
4	كوده ينفع بك وانا اطبل لك I hope flattering you pays off
5	أنا ودي افهم ؛ اللي يسوون **ذي** الاعلانات المبدعة الخرافية .. **وين** كانوا ؟ وشلون طلعت هالطاقات الإبداعية فجأة ؟ I just want to know ; those who make **these** creative amazing advertisements .. **where** were they? And how such talented people come out of the blue?
6	بس لا تصير وراي وتبلش بي وصراخ واحبك يابو نورة وهالقصص.. ماعليش انت اخو و حبيب بس الحفلة **ذي** تبي روقان و تراي ابي ابلغ عنك .. لاتزعل Just don't sit behind me shouting out at Abo Noura and screaming .. you are my brother and I love you but you need to be calm for **this** concert or I will report you .. don't be upset
7	أبي افله بالقصيم الاسبوع **ذا** كشتات و وناسة .(هذه معلومة وليست رأي) I'm going to Alqassim **this** week for camping and some fun ..(this is an information not an

	opinion)
8	مع بداية اجازة ومع هذه الأجواء ؛ كثير منا عنده مشاريع جمعات أو مخيمات جماعية أو لقاءات عائلية .أعتقد من الحكمة تأجيلها لأن كورونا شرسة وتنتشر بسرعة ، وكل شيء ملحوق عليه ؛ الحمدلله الارض لا تتوقف عن الحركة و فصول السنة تتوالى.أنا كنت مخطط أخذ مخيم بالقصيم وهوّنت وابي الزم بيتي. With the holidays starting and the lovely weather ; many of us have plans for group camps or family gatherings . I think it is wise to put them off for now because covid is hitting hard and spreading fast , there is always time to catch up ; the earth keeps moving and seasons change thank God I was planning to go camping in Alqassim and now I plan to stay home.
9	بغيت أكتب تغريدة أنه من البر بالأب أنه يقنع بإقامة الصلاة بالمنزل ويشاركه ابناءه تحقيقاً لأجر الجماعة ، حتى تتكشف هذه الموجة العنيفة من كورونا.. بس خفت ينبرش علي ابوشطوب وجماعته من الغلاة و خفت منهم ومن تنمرهم علي و هوّنت I was going to tweet about one of the ways to show kindness to your father is persuading him to pray at home with his sons to get the Ajr of group prayer , until this aggressive covid wave is over .. but I worried that the fanatics would bully me and scare me so I backed down
10	أخي المتابع أختي المتابعة:السلام.. الخ تراي بتويتر اغرد بشخصيتي الطبيعية؛ شخصيتي بالشارع وبالمحكمة وعند الخباز وعند اصدقائي . شخصيتي الطبيعية إني ساخر ولا **نيب** دايم جاد ، ومو لازم اذا صرت محامي إني ما اكتب تغريدات ساخرة ، ومو لازم اكتب حكم ومقولات سقراط والبس بشت عشان أصير محامي. My brother follower and my sister follower: peace .. etc. I’m being myself on Twitter; my personality on the street, in court, at the baker, and with my friends my natural character is to be sarcastic and **not** always serious, and just because I’m a lawyer doesn’t mean I can’t write sarcastic tweets ; and
11	البنت **ذي** عظيمة على فكرة .. مثل **ذي** العقول لازم تستثمر ، لازم يكون لها دور في التوعية في المجتمع . هذي البنت لازم تكون في الصفوف الأولى. **This** girl is great by the way . mindsets like **these** should be invested , they should contribute to raise awareness in the community . this girl should be in the front lines
12	والله يا عندي حكمة قانونية عظيمة وعميقة ، بس خايف انزله و يزرفه **ذاك** .إذا نام نزلته I have a piece of great and profound legal wisdom , but I fear that I post it and **that** guy steals it. I’ll post it when he sleeps
13	ماشاء الله مبروك والله تحشون دراهم بالخيول يامحمد و **حنا** راحت اعمارنا بالمحاكم ؛ جب الصفة رد الصفة ، و قصور بالتسبيب ما قصور بالتسبيب! Congratulations you sure make a lot of money in the horse races Mohammed while us **we** spend our lives at courtrooms ; bring the capacity take back the capacity , and lack of causation and what not !
14	تدرون **وش** مشروعي القادم؟ اطلع لمزارع الشاي بسيرلانكا اللي تعبي لكل ماركات الشاي،و احط لي علامة تجارية(شاي الصوت الجبلي) و اعبي منه كمية قليلة واشغل #مشاهير_الفلس واخلي شح بالسوق ثم انزل كميات قليلة بحراج(سوق سوداء)وارفع سعره ثم انزله بالسوق وابيع العلامة لأحد الأثرياء العصاميين Do you know **what** is my next project? I will go to the tea farms in Sri Lanka that supply all tea brands, and I will make my own commercial brand called (mountainous voice tea) and make a small supply then hire influencers and create shortage in the market then sell it at a high price in the black market and I will sell the brand to one of the self-made wealthy men
15	وين تباع العصا **ذي** ؟ لاني خلاص قررت أصير قائد أوركسترا Where is **this** cane sold? Because I finally decided to become an orchestra leader
16	امس تعشيت ألذ جريش،و الغريب أنه من مطعم!بالعادة هذي الأكلات صعبة تنضبط إذا انتجت بشكل تجاري عن طريق المطاعم لأنهم ينتجون كميات كبيرة صعب تضبط معايير الجودة في إعدادها. لكن هالمطعم صراحة ضبطها طبعاً المطعم(اللي **مانيب** قليل اسمه عشان ماتكون دعاية بلاش)بتوصية من أعظم صديق ذويق أكل I had the best Jirish for dinner last night, and the odd fact is that it’s from a restaurant! Usually this kind of food is hard to master when producing at a commercial rate by restaurants. But this restaurant in particular pulled it off of course this place (which I will **not** say it’s name so it’s not free advertising) was recommended by the greatest friend who is a foodie
17	لازلنا نرى ضرورة أن تكون مهنةالمحاماة تمثل من خلال هيئة مستقلة عن وزير العدل. معظم المشاكل التي يواجهها المحامي أثناء ممارسته لمهنته ؛إما مصادمة مع قاض أو مواجهة خلل الكتروني أو بشري من المنظومة العدلية،تعرقل عما المحامي؛ وكلها ترجع لوزير العدل أو لرئيس المجلس اللي هو وزير العدل! We still find it necessary that barristers should be represented by an entity independent of the minister of Justice. Most of the issues a barrister encounters during his profession ;either in a conflict with a judge or electronic or cleric malfunction in the judicial establishment, that hinders the barrister’s work; and these issues are reported to the minister of Justice or the chairman who is the minister himself!
18	على أي قناة تبي تذاع مباراتنا مع **ذولاك** ؟ Which channel airs our game with **those** guys?
19	**وش** صار عليه البنت اللي شايلتن الاسد كنه شايلتن عيّل ؟ **What** happened to the girl that was holding a lion like a child?
20	في هذا الفيديو ؛ ناقشت الفكرة الغبية (**حنا** اللي علمناكم) In this video ; I discussed the stupid notion of (**we** are the ones who educated you)

Furthermore, in Grant’s (2012) discussion of distinctiveness he pointed out its different aspects in the context of forensic authorship casework. He explains that there is population level distinctiveness, which is the distinction of an author’s use of a stylistic feature against the relevant population or community of practice. Alternatively, there is the distinction of a smaller scale either by examining a small group of authors or pairwise which is to compare two authors. The pairwise analysis unfolds in the following section.

8.2.2 Data analysis and findings

The NASCoT represents partial information about the authors’ idiolect because it is a time-restricted sample of their total output to date. Nonetheless, it reports some patterns of stylistic consistency and, therefore, it is capable of attributing authorship (Coulthard, 2004; Johnson and Wright, 2014; Grant, 2020). As demonstrated in Chapter 6, the NASCoT reveals variation across authors in using certain feature variants in NA and MSA. The stylistic analysis findings in Chapter 5 also reveal that each author has a set of feature variants that they use frequently. Table 8.3 shows the difference in frequency rates between the subjects of this case study as per their subcorpora in the NASCoT in interrogatives, negatives, personal deixis, and temporal/spatial deixis. Stylistic variation has precedence in being used as a method in authorship analysis (Coulthard, 2004; Grieve et al., 2019; Grant, 2020). Therefore, it can be hypothesised that Allahim and Alzamil have their own stylistic feature preferences that they could use consistently in future language output. I compare the stylistic features in the questioned tweets against those of the ‘suspected’ authors to detect a pattern of stylistic consistency against their subcorpora. Therefore, I present the questioned tweets in Table 8.5 and mark Allahim’s stylistic features from the feature set in blue and Alzamil’s stylistic features in orange.

Table 8.5: Detection of the authors’ stylistic features in the questioned tweets.

#	Tweet
1	احتفالات بالعام الجديد و مطر وخير و الناس سعيدة ... يالله لك الحمد New year celebrations, rain, blessings, and everyone is happy …. Thank God
2	صباح الخير للجميع ان شاء الله كل أيامكم في هذه السنة الجديدة تشبه هذا **اليوم** وجمال جو هذا **اليوم** Good morning to all I pray to God that all your days in this new year are as beautiful as **today** and the beauty of the weather
3	**اليوم** عشت ليلة من اجمل الليالي في #ليلة_تريو_الرياض ليلة من الخيال لقد تغيرنا كثيراً وأصبحت حياتنا تفيض سعادة وبهجة بفضل هذا الزعيم العظيم الذي خلق المعجزة .كل الشكر لمنظمي هذه الليلة ومن صنع هذه الأجواء الجميلة و على رأسهم ابوناصر **Today** I experienced one of the most beautiful nights at Riyadh Trio Night it was beyond imagination we have changed so much and our lives exude with happiness and joy thanks to this great leader who made this miracle .All thanks goes to the organisers of this beautiful night headed by abo Nassir
4	كوده ينفع بك وانا اطبل لك I hope flattering you pays off
5	أنا ودي افهم ؛ اللي يسوون **ذي** الاعلانات المبدعة الخرافية .. **وين** كانوا ؟ وشلون طلعت هالطاقات الإبداعية فجأة ؟ I just want to know ; those who make **these** creative amazing advertisements .. **where** were they? And how such talented people come out of the blue?
6	بس لا تصير وراي وتبلش بي وصراخ واحبك يابو نورة وهالقصص.. ماعليش انت اخو و حبيب بس الحفلة **ذي** تبي روقان و تراي ابي ابلغ عنك .. لاتزعل Just don’t sit behind me shouting out at abo Noura and screaming .. you are my brother and I love you but you need to be calm for this concert or I will report you .. don’t be upset
7	أبي افله بالقصيم الاسبوع **ذا** كشتات و وناسة .(هذه معلومة وليست رأي) I’m going to Alqassim **this** week for camping and some fun ..(this is an information not an opinion)
8	مع بداية اجازة ومع هذه الأجواء ؛ كثير منا عنده مشاريع جمعات أو مخيمات جماعية أو لقاءات عائلية .أعتقد من الحكمة تأجيلها لأن كورونا شرسة وتنتشر بسرعة ، وكل شيء ملحوق عليه ؛ الحمدلله الارض لا تتوقف عن الحركة و فصول السنة تتوالى.أنا كنت مخطط أخذ مخيم بالقصيم وهوّنت وابي الزم بيتي. With the holidays starting and the lovely weather; many of us have plans for group camps or family gatherings. I think it is wise to put them off for now because covid is hitting hard and spreading fast, there is always time to catch up; the earth keeps moving and seasons change thank God I was planning to go camping in Alqassim and now I plan to stay home.
9	بغيت أكتب تغريدة أنه من البر بالأب أنه يقنع بإقامة الصلاة بالمنزل ويشاركه ابناءه تحقيقاً لأجر الجماعة ، حتى تنكشف هذه الموجة العنيفة من كورونا.. بس خفت ينبرش علي ابوشطوب وجماعته من الغلاة و خفت منهم ومن تنمرهم علي و هوّنت I was going to tweet about one of the ways to show kindness to your father is persuading him to pray at home with his sons to get the Ajr of group prayer, until this aggressive covid wave is over .. but I worried that the fanatics would bully me and scare me so I backed down
10	أخي المتابع أختي المتابعة:السلام.. الخ تراي بتويتر اغرد بشخصيتي الطبيعية؛ شخصيتي بالشارع وبالمحكمة وعند الخباز وعند اصدقائي . شخصيتي الطبيعية إني ساخر و**لا نيب** دايم جاد ، ومو لازم اذا صرت محامي إني ما اكتب تغريدات ساخرة ، ومو لازم اكتب حكم ومقولات سقراط والبس بشت عشان أصير محامي. My brother follower and my sister follower: peace .. etc. I’m being myself on Twitter; my personality on the street, in court, at the baker, and with my friends my natural character is to be sarcastic and **not** always intense, and just because I’m a lawyer doesn’t mean I can’t write sarcastic tweets ; and
11	البنت **ذي** عظيمة على فكرة .. مثل **ذي** العقول لازم تستثمر ، لازم يكون لها دور في التوعية في المجتمع . هذي البنت لازم تكون في الصفوف الأولى. **This** girl is great by the way. mindsets like **these** should be invested, they should contribute to raise awareness in the community. this girl should be in the front lines
12	والله يا عندي حكمة قانونية عظيمة وعميقة ، بس خايف انزله و يزرفه **ذاك** .إذا نام نزلته I have a piece of great and profound legal wisdom , but I fear that I post it and **that** guy steals it. I’ll post it when he sleeps
13	ماشاء الله مبروك والله تحشون دراهم بالخيول يامحمد و **حنا** راحت اعمارنا بالمحاكم ؛ جب الصفة رد الصفة ، و قصور بالتسبيب ما قصور بالتسبيب! Congratulations you sure make a lot of money in the horse races Mohammed while **we** spend our lives at courtrooms; bring the capacity take back the capacity, and lack of causation and what not !
14	تدرون **وش** مشروعي القادم؟ اطلع لمزارع الشاي بسيرلانكا اللي تعبي لكل ماركات الشاي،و احط لي علامة تجارية(شاي الصوت الجبلي) و اعبي منه كمية قليلة واشغل #مشاهير_الفلس واخلي شح بالسوق ثم انزل كميات قليلة بحراج(سوق سوداء)وارفع سعره ثم انزله بالسوق وابيع العلامة لأحد الاثرياء العصاميين Do you know **what** is my next project? I will go to the tea farms in Sri Lanka that supply all tea brands, and I will make my own commercial brand called (mountainous voice tea) and make a

	small supply then hire influencers and create shortage in the market then sell it at a high price in the black market and I will sell the brand to one of the self-made wealthy men
15	**وين** تباع العصا **ذي** ؟ لاني خلاص قررت أصير قائد أوركسترا **Where** is **that** cane sold? Because I finally decided to become an orchestra leader
16	امس تعشيت ألذ جريش،و الغريب أنه من مطعم!بالعادة هذي الأكلات صعبة تنضبط إذا انتجت بشكل تجاري عن طريق المطاعم لأنهم ينتجون كميات كبيرة صعب تضبط معايير الجودة في إعدادها. لكن هالمطعم صراحة ضبطها طبعاً المطعم(اللي **مانيب** قليل اسمه عشان ماتكون دعاية بلاش)بتوصية من أعظم صديق ذويق أكل I had the best Jirish for dinner last night, and the odd fact is that it’s from a restaurant! Usually this kind of food is hard to master when producing at a commercial rate by restaurants. But this restaurant in particular pulled it off of course this place (which I will **not** say its name so it’s not free advertising) was recommended by the greatest friend who is a foodie
17	لازلنا نرى ضرورة أن تكون مهنةالمحاماة تمثل من خلال هيئة مستقلة عن وزير العدل. معظم المشاكل التي يواجهها المحامي أثناء ممارسته لمهنته ؛إما مصادمة مع قاض أو مواجهة خلل الكتروني أو بشري من المنظومة العدلية،تعرقل عما المحامي؛ وكلها ترجع لوزير العدل أو لرئيس المجلس اللي هو وزير العدل! We still find it necessary that barristers should be represented by an entity independent of the minister of Justice. Most of the issues a barrister encounters during his profession; either in a conflict with a judge or electronic or cleric malfunction in the judicial establishment, that hinders the barrister’s work; and these issues are reported to the minister of Justice or the chairman who is the minister himself!
18	على أي قناة تبي تذاع مباراتنا مع **ذولاك** ؟ Which channel airs our game with **those** guys?
19	**وش** صار عليه البنت اللي شايلتن الاسد كنه شايلتن عيّل ؟ **What** happened to the girl that was holding a lion like a child?
20	في هذا الفيديو ؛ ناقشت الفكرة الغبية (**حنا** اللي علمناكم) In this video ; I discussed the stupid notion of (**we** are the ones who educated you)

To measure the frequency of the features statistically, I’m using a binary system and marking all existing stylistic features as one and the absence of features as zero. In the cases where both authors use the same feature, I mark it as one for both authors. I created a table for each author, placed each stylistic feature in a separate column, and the numbers from one to twenty in each row represent each of the questioned tweets. Each occurrence of each feature is marked as one, and to avoid any confusion, I did not account for multiple occurrences of the features. For instance, the NA person deixis /ḏi/ *this* occurred twice in the questioned tweet number 11, but I’m marking these occurrences for the value of one where it applies. Table 8.6 shows the examination of Allahim’s identified stylistic features in the questioned tweets.

Table 8.6: Coding Allahim’s features in the questioned tweets.

Tweet	/wis/	/wen/	/warah/	/lam/	/lan/	/manib/	/di/	/da/	/dolak/	/alyom/	/hinak/	/alaan/
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	0	0	0	0	0	0	0	1	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	1	0	0	0	0	0

6	0	0	0	0	0	0	1	0	0	0	0	0
7	0	0	0	0	0	0	0	1	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	1	0	0	0	0	0	0
11	0	0	0	0	0	0	1	0	0	0	0	0
12	0	0	0	0	0	0	0	1	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0
14	1	0	0	0	0	0	0	0	0	0	0	0
15	0	1	0	0	0	0	1	0	0	0	0	0
16	0	0	0	0	0	1	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	1	0	0	0
19	1	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0

In the table, we can see that Allahim’s distinctive? features occur in 14 out of the 20 questioned tweets. The interrogatives /wiš/ and /wen/ are found in four different occurrences while /warah/ is absent. The negatives category is interesting because the MSA negatives /lam/ and /lan/ are not observed whereas /manib/, a feature that is distinctive of Allahim’s subcorpus, is identified twice. In Chapter 5 we learned that the NA variants of person deixis are distinctive features for Allahim, and they are identified in seven different tweets. The singular feminine variant /ḏi/ is marked in four different tweets and the singular masculine /ḏa/ is marked in two. Also, the plural variant /ḏolak/ is marked once in tweet number 18. Lastly is the temporal/spatial deixis, where the MSA variant /alyom/ is marked in two tweets. The final two features in Allahim’s list, the NA variant for spatial deixis /hinak/ and the MSA variant for temporal deixis /al’aan/ are both marked as zero in the questioned tweets. Thus, the total of the stylistic features observed in the questioned tweets are seven features out of the 44-feature set.

The next step is to examine Alzamil’s stylistic feature list in the questioned tweets. Table 8.7 shows that his features are marked in 10 out of the 10 tweets, only half, compared with Allahim’s 70%. The NA variants for interrogatives /wiš/ and /wen/ overlap between both authors, and they are marked in two occurrences for each feature. However, /leh/, which is distinctive of Azamil is marked as zero. Similarly, the NA negative variant /mub/ that is a distinctive feature of Alzamil is also absent in the questioned tweets. As reported earlier, the NA person deixis variant /ḏa/ is marked twice, which also overlaps with Allahim. The other person deixis NA variants that are distinctive of Alzamil are /hna/ and the singular distal /ḏa/, both of which occur twice respectively.

Table 8.7: Coding Alzamil's features in the questioned tweets

Tweet	**/wis/**	**/wen/**	**/leh/**	**/lam/**	**/lan/**	**/mub/**	**/da/**	**/hna/**	**/dak/**	**/alyom/**	**/alaan/**	**/girib/**
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	0	0	0	0	0	0	0	1	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	1	0	0	0	0	0
13	0	0	0	0	0	0	0	1	0	0	0	0
14	1	0	0	0	0	0	0	0	0	0	0	0
15	0	1	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0
19	1	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	1	0	0	0	0

/ḏak/, which is absent. Furthermore, the MSA variant for time deixis /alyom/ is identified twice, while the proximal variants /al'aan/ and /qarib/ have zero occurrences. The total of Alzamil's stylistic features identified in the questioned tweets is five out of twelve. Table 8.8 shows the results of the binary analysis (with the authors highlighted in bold in the final two columns) compared to the stylistic findings discovered in Chapter 5.

Table 8.8: Binary analysis of questioned tweets against NASCoT authors

	Alabdulkarim	Alajlan	Aleidi	Algofaily	Allahim	Almohanna	Alnuhait	Alrasheed	Alrokibah	Alsubayel	Altassan	Alwabil	Alzamil	**Author 1**	**Author 2**
/mada/	▨	🟥	🟥	🟥	🟨	▨	🟨	🟥	🟥	🟥	🟨	🟨	🟥	🟥	🟥
/wiš/	▨	🟩	🟩	🟩	🟩	🟨	🟥	🟨	🟩	🟩	🟥	🟥	🟩	🟩	🟩

/wišu/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟥	🟥	🟥
/limada/	🟨	🟨	🟥	🟨	🟨	🟩	🟥	🟩	🟥	🟥	🟥	🟨	🟥	🟥	🟥
/leh/	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟩	🟥	🟥
/warah/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟥	🟥
/wišulah	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/aina/	🟨	🟥	🟥	🟥	🟨	🟨	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/wen/	▨	▨	▨	▨	▨	🟨	🟥	🟥	🟨	🟨	🟥	🟥	🟩	🟩	🟩
/la/	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥
/lam/	🟩	🟩	▨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	▨	🟥	🟥
/lan/	🟩	▨	🟨	🟨	▨	🟨	🟩	🟩	🟨	🟥	🟩	🟨	▨	🟥	🟥
/mub/	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟩	🟩	🟥	🟥	🟥	🟩	🟥	🟥
/manib/	🟥	🟥	🟥	🟥	▨	🟥	🟥	🟥	🟥	🟨	🟥	🟨	🟥	🟩	🟥
/muhub/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	▨	🟥	🟥	🟥	🟥	🟥	🟥
/mahub/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/mahib/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/hna/	🟥	🟥	🟥	🟥	🟨	🟥	🟨	🟥	▨	🟨	🟥	🟥	🟨	🟥	🟩
/ḏa/	🟥	🟩	🟥	🟥	▨	🟥	🟥	🟥	🟥	▨	🟥	🟥	▨	🟩	🟩
/ḏak/	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟨	🟥	🟥	🟥	🟥	🟨	🟥	🟥
/haḏak/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/ḏi/	🟥	▨	🟥	🟥	▨	🟥	🟥	🟥	🟨	🟨	🟥	🟥	🟥	🟩	🟥
/ḏik/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/haḏik/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/ḏola/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/ḏolak/	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥
/al’aan/	🟩	▨	🟩	🟩	🟩	▨	🟥	🟥	▨	🟨	🟨	▨	▨	🟥	🟥
/alhin/	🟥	▨	🟨	🟥	🟨	🟥	🟥	🟥	▨	▨	🟥	🟥	🟨	🟥	🟥
/lahiqan/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/badain/	🟥	🟥	🟨	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/alyom/	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
/gadan/	🟥	🟥	🟨	🟥	🟥	▨	🟥	🟩	🟥	🟥	🟥	🟨	🟥	🟥	🟥
/bukra/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟥	🟥	🟥
/bakir/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
/huna/	🟩	▨	🟩	▨	▨	▨	🟨	▨	🟨	🟨	🟩	▨	🟩	🟥	🟥
/hunak/	🟩	🟩	🟨	🟨	▨	🟩	▨	🟩	▨	▨	🟩	🟨	🟨	🟥	🟥
/qarib/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	▨	🟥	🟥
/biʕid/	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥

🟥 ≥ 0.2 🟨 0.2 – 0.5 ▨ 0.5 – 1 🟩 1 ≤

The pairwise binary analysis reveals that the questioned tweets contain seven of the twelve stylistic features that belong to Allahim while five of the twelve features that mark Alzamil. Moreover, Allahim’s stylistic features occurred in 14 out of the 20 questioned tweets, while on the other hand Alzamil’s features were found in only ten. There are some limitations that line the findings of this analysis: one being the size of the ‘questioned’ tweets is too small to attribute authorship more accurately. Another limitation is the similarity in the idiolectal style between the ‘suspect’ authors as far as the feature set is concerned. Having a larger data size or a more extensive feature list would have a higher potential for more accurate findings.

The following stage is to examine the authors’ discourse activities by referring back to the indexical processes they make and the range of activities they engage in. In Chapters 5 and 6 we learned that authors employ lexico-grammatical features to take stances and communicate their positions. The findings reveal that Allahim uses the NA person deixis (e.g., /ḏi/, /ḏa/, and /ḏolak/) in sarcastic and sometimes insulting contexts and the NA interrogatives (e.g., /wiš/ and /warah/) to initiate debates and express criticism. He also uses the NA negative /manib/ to express sarcasm and the time deictic expression /alyom/ to share information about his work and daily activities. As for Alzamil the most frequent category he uses is interrogatives, especially the NA variant /wiš/, and it is mostly used to ask his followers for information. the second most frequent category is time/space deixis (e.g., /alyom/ and /qarib/), which he uses to talk about his work on YouTube and news about movies and tv shows. The other categories that are less frequent are person deixis (e.g., /ḏa/), which he would use for criticism, and negatives, mostly the NA variant /mub/, which he uses also for criticism and debates. Table 8.3 shows that the most frequent variants both authors use are /wiš/ and /alyom/, so I want to investigate the range of activities where they use these variants across their subcorpora. The findings in Table 8.9 show that one of the similarities the authors share is exchange of information and they use /alyom/ more frequently than /wiš/ to do so. They also share the frequency of using /wiš/ when debating more than /alyom/ and they do not participate or initiate games except for one time by Alzamil using /alyom/. Furthermore, we can see that jokes are more frequent in Allahim's subcorpus than Alzamil, so is exchange of insults. The findings also show that it is more likely for Allahim to use /wiš/ in his jokes and insults than /alyom/.

Table 8.9: The authors' activities in the NASCoT subcorpora

Author	Variant	Debate	Info exchange	Joke exchange	Insult exchange	Game
Allahim	**/wiš/**	60	73	73	19	-
	/alyom/	13	130	16	3	-
Alzamil	**/wiš/**	24	126	14	4	-
	/alyom/	11	155	5	1	1

In terms of frequency, there are some variations between the authors. For instance, debates are more frequent in Allahim's subcorpus, so are jokes and insults. On the other hand, information exchange is more frequent in Alzamil's subcorpus as well as games, even though it is one occurrence. The authors engage in debates and information exchange at a relatively similar frequency, but they perform differently when it comes to jokes and insults. Allahim refrains from partaking in games almost completely, and Alzamil rarely. I compare the numbers in Table 8.9 against the activity analysis of the questioned tweets in Table 8.10. I mark the activity identified in each tweet in green. In tweets number 14, 15 and 19 the texts appear in the form of an interrogative, but the context is sarcastic; therefore, they were marked as the exchange of a joke. The total count of the activities is 1 debate, 9 exchanges of information, 10 exchanges of jokes, and zero for insult exchange and game. When we look back to Table 8.9, we can observe that exchange of information is the most frequent activity in Alzamil's subcorpus. Alternatively, Allahim shows frequent occurrences of debates, exchange of jokes in addition to exchange of information. Moreover, tweets number 6, 7, 10, 12, 15, and 15 are identified as jokes because of the use of person deixis expressions /ḏi/, /ḏa/, /ḏak/ and /ḏi/ respectively. Another person deixis is /d̠olak/ in the exchange of information in tweet number 18, which also a distinctive index Allahim uses when he mocks an out-group whether in sports or politics.

Table 8.10: Analysis of the questioned tweets activities

Tweet number	Activity				
	Debate	Info exchange	Joke exchange	Insult exchange	Game
1	🟥	🟩	🟥	🟥	🟥
2	🟥	🟩	🟥	🟥	🟥
3	🟥	🟩	🟥	🟥	🟥
4	🟥	🟥	🟩	🟥	🟥
5	🟥	🟩	🟥	🟥	🟥
6	🟥	🟥	🟩	🟥	🟥
7	🟥	🟥	🟩	🟥	🟥
8	🟥	🟩	🟥	🟥	🟥
9	🟥	🟥	🟩	🟥	🟥
10	🟥	🟥	🟩	🟥	🟥
11	🟥	🟩	🟥	🟥	🟥
12	🟥	🟥	🟩	🟥	🟥
13	🟥	🟥	🟩	🟥	🟥
14	🟥	🟥	🟩	🟥	🟥
15	🟥	🟥	🟩	🟥	🟥
16	🟥	🟩	🟥	🟥	🟥
17	🟩	🟥	🟥	🟥	🟥
18	🟥	🟩	🟥	🟥	🟥
19	🟥	🟥	🟩	🟥	🟥
20	🟥	🟩	🟥	🟥	🟥

The use of personal deictic expressions, especially the variants /d̠i/ and /d̠olak/, in a joke/sarcasm context has been identified as part of Allahim's indexical processes in Chapter 5. Moreover, the negative /manib/ is a distinctive feature of Allahim, which is also used in tweet number 10 in a sarcastic context and tweet number 16 in an exchange of information. Furthermore, tweet number 14 is marked as a joke exchange with the interrogative /wiš/ occurs in the text. Table 8.8 shows that it is more likely for Allahim to use /wiš/ in a joke than Alzamil. Alternatively, the time deixis /alyom/ is observed in tweets number 2 and 3 in the context of information exchange, which can be classified as part of Alzamil's behaviour. Also, the use of the person pronoun /hna/ in tweet number 13, which is identified as a joke, is another variant that Alzamil is more likely to use as per the findings in Table 8.3.

In Chapter 2 I reported that Coulthard's (2004) notion of idiolect states "every native speaker has their own distinct and individual version of the language they speak and write, their own idiolect…" (p. 432). The findings both stylistically and statistically suggest that it is more likely that the author of the questioned tweets is the first author Abdulrahman Allahim than the second author Abdulaziz Alzamil. The stylistic preferences represented in the NA features: the negative /manib/ and the person deictic expressions /d̲i/ and /d̲olak/ appear as distinctive idiolectal stylistic features of Allahim compared to the total authors of the NASCoT. This confirms the findings exhibited in Chapters 5 and 7, that it is more likely to match the stylistic choices Allahim makes compared to Alzamil as per the pairwise analysis. In terms of activities, the analysis of the activities identified in the questioned tweets suggests that they are more compatible with Allahim's range of activities than those of Alzamil. Also, the indexical processes detected in the questioned tweets resonate with the processes identified in Allahim's subcorpus.

8.2.3 Discussion and conclusion

The research question this empirical experiment aims to answer are:

1- How likely is it for the methodology and the Najdi Arabic feature set proposed earlier in this research to identify authorship in a forensic casework?

The methodology proposed in this research shows potential in identifying authorship. The stylistic analysis composed of identifying the features, detecting them in the questioned tweets, and measuring their occurrences statistically is a method that is "repeatable, reproducible, and accurate" PCAST (2016, cited in Ainsworth and Juola, 2019, p. 1171). Moreover, the findings show that it is likely that the author of the questioned tweets is Allahim; the stylistic preferences and range of activities observed in the questioned tweets are more likely to be produced by Allahim than Alzamil. These findings are reached by stylistic analysis and the fact that since this is a simulation, the author of the ‘questioned’ tweets is identified beforehand. There are a few limitations in this case simulation; firstly the fact that it is simulated scenario can appear less authentic compared with real casework. Also, the absence of some of the stylistic features in some of the questioned tweets made the attribution problem more challenging. Finally, this case study aims to validate the methodology to be used in forensic casework that addresses authorship attribution, which concludes that it is applicable and can be replicated in a forensic casework context, with some limitations. Furthermore, this simulation is an empirical experiment that leads to the general discussion of the triangulated method applied in the data analysis and the implications of its findings in terms of idiolect and forensic authorship analysis.

8.3 General discussion

The experiment presented above is the final layer of analyses for this doctoral research. I stated in Chapter 2 that idiolect and authorship attribution are intertwined. This section is a development of this statement, as I discuss the implications of the findings in terms of these two key aspects: idiolect and forensic authorship attribution, both as fields of research and casework.

8.3.1 Turell’s idiolectal style

I revisit Turell’s (2010) interpretation of idiolectal style, stated earlier in Chapter 2, which is the pillar that supports my analysis of the NASCoT authors’ idiolectal styles throughout this research. She portrays it as:

> Hav[ing] to do primarily, not with what system of language/dialect an individual has, but with a) how this system, shared by lots of people, is used in a distinctive way by a particular individual; b) the speaker/writer’s production, which appears to be ‘individual’ and ‘unique’ (Coulthard, 2004) and also c) Halliday’s (1989) proposal of ‘options’ and ‘selections’ from these options. (Turell, 2010: 7)

Turell first describes idiolect as the individual’s distinctive use of the language/dialect system shared by a community. The examined linguistic features of interrogatives, negatives and deictic expressions in the NA variety are shared by the Najdi community in the central region of Saudi Arabia. The MSA version of the features are shared by the much wider community of the Arab region and some Islamic countries. Nonetheless, the NASCoT findings in Chapter 5 revealed two major patterns; the first one is the authors’ overall preference to use interrogatives in the NA variety over the MSA as illustrated in Figure 5.5. Oppositely, the second pattern is the authors’ overall preference to use MSA negative forms instead of the NA (Figure 5.11) and the same applies in the case of temporal and spatial deixis (Figure 5.14). Although some of the stylistic features occur infrequently, nonetheless, they have been able to report the authors’ variation in using MSA and NA varieties. The variation is not only visible across the NASCoT subcorpora but also when compared against the reference corpus ReCNAT presented in Chapter 4. The patterns of preferences combined correspond with Turell’s notion of using the system in a personalised manner. Nevertheless, these major patterns that somewhat distinguish the NASCoT authors form their community contain finer details that define each author as we look closer.

In the quote Turell (2010) also talks about the unique and individual production each author produces that makes it idiolectal, which is similar to Coulthard’s (2004) notions of stylistic consistency and distinctiveness. Grant’s (2012) description of stylistic distinctiveness at population level and at small group or pairwise level echoes these notions as well. Chapters 5 and 7 demonstrate that some of the NASCoT authors use NA and MSA variants in manners that set them apart and make them individual. Authors such as Aleidi and Alzamil stand out in their use of NA interrogatives, which also aligns with the findings in Chapter 6 that shows that Aleidi’s most frequent activity is exchange of information using the NA interrogative /wiš/. Another idiolectal style pattern that is found in Chapter 5 is Alrasheed, who consistently selects MSA negatives /la/ and /lam/ whenever she talks about her personal perceptions of relationships, whether positively or negatively. Allahim consistently uses the NA person deictic variants /d̠i/, /d̠ak/, and /d̠olak/ as indexical tools in his sarcastic discourse. Also, Alzamil shows exclusive preference for using the NA variants across all categories, except for negatives where he uses both varieties almost equally. Furthermore, discursive activity reveals that some of the authors show consistency in shifting from the MSA variety in the original posts to NA in their replies. There are instances where Alabdulkarim, Alajlan, Algofaily, and Alrasheed publish a tweet in the MSA variety then switch to NA as interaction unfolds. In the case of Alabdulkarim, the shift reflects relationality, because in the original posts he would identify himself as someone with authority and legitimacy, but when he replies he shifts to a position of limited authority. On the other hand, the instances of Alajlan, Algofaily, and Alrasheed exhibit speech accommodation because the replies are in the Najdi variety as well, which also show how they position themselves as members of the in-group.

Alternatively, the subcorpora of some authors did not reflect such distinctive patterns, which is a possible limitation reported by the literature (Johnson and Wright, 2017; Grant, 2020). The subcorpora of the authors Altassan, Almohanna, Alnuhait, and Alwabil did not exhibit the feature set as frequently or prominently as the other authors, especially in the Najdi variety. Nonetheless, their subcorpora were able to reveal their preference to use MSA variants in interrogatives, negatives, and deixis, and the frequency of their occurrences enables us to recognise a small part of their idiolectal style in relation to the use of the feature set. In the case of Altassan, we learn from her subcorpus that there is a consistency in her use the MSA variants. The most frequent category is time/space deixis followed by negatives then interrogatives and she did not exhibit any occurrences for person deixis.

Almohanna’s subcorpus varies in some ways, firstly with the frequency rates. The most frequent category in her subcorpus is time/space deixis, but unlike Altassan, the rates are higher and the variants she uses are more diverse. Negatives are also the second most frequent category in Almohanna’s subcorpus. However, her frequency rates are much less compared to Altassan’s and the variants are more diverse. As for the interrogatives, she has higher frequency rates compared to Altassan, Alnuhait and Alwabil. Finally, the person deixis shows only one variant /ḏak/ *that*, which is more than what Altassan’s subcorpus shows and less diverse than Alnuhait and Alwabil. Therefore, it can be claimed that the absence of some of the stylistic features in either variety is an indication of the author’s idiolectal style, which would be an interesting take on Coulthard’s (2004) concept of markedness. For instance, the empirical experiment presented in 8.2 shows that while the authors share some of the stylistic choices, Alzamil’s subcorpus shows that he does not use person deixis as frequently as Allahim, especially the variant /ḏolak/ *those*. This finding supported the conclusion that it is unlikely that Alzamil is the author of the tweets as they exhibited frequent use of /ḏolak/. It can also be claimed that in the case of the authors whose subcorpora appear somewhat similar due to their limited size, the finer details when examining the variants closely do reveal idiolectal variation.

Lastly, Turell’s definition addresses the notions of the options and selections that an author makes that comprise their idiolectal style, which is similar to Coulthard’s concept of idiolectal co-selection discussed in Chapters 1 and 2. Chapters 5 and 6 reveal that the NA interrogative variant /wiš/ *what* and the MSA deixis variant /alyom/ *today* occur in all the NASCoT subcorpora, which can compromise their position as idiolectal markers of authorship. Corpus analysis also reveals that /wiš/ is the only interrogative variant, whether in NA or MSA, that occurs across all authors’ subcorpora, unlike the negative and deixis variants. Nevertheless, it can be argued that the NA interrogative /wiš/ can disclose a range of activities that can be informative, as the analysis in Chapter 6 demonstrates. The chapter reveals that the linguistic variants the authors use reflect the identity themes such as nationalism, regionalism, gender, and types of activities.

8.3.2 Forensic authorship attribution

The synergy in this research has implications in terms of forensic authorship research and casework. The literature shows that computational tools are successful in solving authorship problems with a high degree of confidence and accuracy (e.g., Abbasi and Chen, 2005; Zheng et al., 2009). In this study, WEKA (Witten et al., 2016) classified the data and assigned each author with their subcorpora at a promising rate of accuracy. While it does not provide a trace that explains which of the variants is most salient or marks authorship, its findings are further confirmed and supported by the linguistic patterns and stylistic preferences detected by *Wordsmith* (Scott, 2020) in Chapter 5. In other words, the machine learning tool is statistically validating the quantitative findings revealed by corpus tools and qualitative stylistic analysis or vice versa.

This methodology proposes that corpus linguistics is used as a tool and an approach to identify the authors' linguistic patterns objectively and analyse them stylistically (Baker et al., 2008). *Wordsmith* (Scott, 2020) provided that roadmap as concordances lines, collocates, and keyword lists informed us of their individual variation. Furthermore, this study implemented a top-down approach using a selection of linguistic variables that are dialectally salient and their variants. Baker et al. (2008:296) state that "the corpus-based analysis tends to focus on what has been explicitly written, rather than what could have been but was not, or what is implied, inferred, insinuated, or latently hinted at." This claim resonates with Turell's quote in Chapter 4 in reference to Jackobson (1956) that the marked features are more revealing and informative of authorship than the unmarked ones. Grant (2020) argues differently by saying that the absence of a stylistic feature can also help in identifying authorship. The corpus analysis revealed that there are variants that appeared in all the NASCoT authors, as shown in Table 5.2: the MSA negative variants /la/, /lam/, and /lan/, the MSA time/space deixis /al'aan/, /alyom/, /huna/, and /hunak/, as well as the NA interrogative /wiš/. Alternatively, the NA person deixis variants occur inconsistently by comparison. For instance, Figure 5.11 shows the subcorpora of Allahim, Alajlan and Alsubayel score the highest frequency rates in person deixis while opposite to them are Altassan, Almohanna and Alwabil who rarely use them if at all. These statistical values reflect the kind of discourse activities the authors perform. Moreover, corpus tools revealed the linguistic patterns that represent the idiolectal style of each author. Grant (2013) and Wright (2017) argue that attributing authorship does not rely on a strong language theory exclusively but on a case of stylistic consistency and distinctiveness. The qualitative analysis in the corpus stylistic Chapters 5 and 6 highlighted most of the authors' stylistic patterns that appear consistent when looking at the variants' dispersion across their subcorpora. Consequently, each author's pattern appears distinctive against the sample, which was confirmed by the quantitative findings in Chapter 7.

In terms of stylistics, the analysis explored the notion of linguistic saliency and authorship markedness by investigating the Najdi-specific features and whether they have potential as markers of authorship. Table 5.2 in Chapter 5 demonstrates the stylistic choices each of the authors make and which of the features marks their authorship. To claim which variants are most revealing of authorship is arguable and ambitious. However, the findings in the case simulation suggest that the NA variants show potential as markers of idiolectal style. The analysis shows that interrogatives and person deixis in the Najdi dialect do not necessarily occur frequently but they can be informative of the authors' linguistic behaviour. Furthermore, Larner (2014) proposed an argument that the idiolect definitions provided by Hockett (1958), Labov (1972) and Coulthard (2004) address the fundamental issue that is whether an idiolect is a reliable sign of authorship in spite of time. The way the NASCoT authors use the linguistic variables and the variety they choose does show a linguistic habit or choice and, when accumulating these choices together, we can partially identify each author's idiolectal style. The reason why I claim it is a linguistic habit/choice is the time period the data covers is 19 months, which can change over time.

The final part is CMDA, which was investigated in Chapter 6 to identify the online individual identity themes. In the context of online interaction, the researcher has access to collect and analyse an author's behaviour using their digital footprint (Herring, 2004; MacLeod and Wright, 2020). A substantial amount of the analysis in this research was revealed by the interaction threads the authors create with their addressees. Their original posts, replies and comments on their addressees' replies enabled us to see the different positions each author takes as the conversation unfolds. Furthermore, the literature proposes the idea that individuality is created and represented by language as Johnstone explains: "[t]hrough talk and other aspects of behaviour, individuals display their individuality. In other words, people express their individuality with everything they do…" (2000: 407). The analysis supports this notion as we were able to observe each author's online identity represented in their linguistic choices, whether the varieties or the variants. The preference to use MSA or NA is connected to how the authors identify themselves to the public. Moreover, each author uses a range of variants (e.g., interrogatives, negatives, etc.) to create individual indexical processes to communicate their stances.

8.4 Conclusion

This chapter showcased a simulation of a forensic authorship casework involving 'suspected' authors and twenty questioned tweets. The aim of this case simulation is to test the methodology proposed in this doctoral research and explore its potential in authorship casework. The findings are promising, thus suggesting that the methodology can be tailored and reapplied to text queries in Najdi Arabic and potentially to other dialectal varieties of Arabic. Furthermore, the second part of the chapter carried a general discussion of the findings discovered in this thesis. The first aspect is in terms of the theory of idiolect that Turell (2010) proposed that is idiolectal style. The findings show that NA variants have a potential in attributing authorship using computational and stylistic approaches. The findings also show that the occurrences or absence of any of the variants in the feature set contributes to the identification of the author's idiolectal style. The second aspect is the implications of the findings with respect to forensic authorship research.

Chapter 9
CONCLUSION AND FUTURE RECOMMENDATIONS

9.1 Bridging the gap between linguistics and computer science

This chapter concludes this doctoral research by presenting a summary of the results, discussing the limitations that constrained this research and their extent, and finally highlighting the contributions it offers and the recommendations for further endeavours.

The aim of this thesis is twofold: the first is specific and the second is more global. As indicated in Chapters 1 and 2, Arabic and its dialects are new vistas to be explored in the context of forensic linguistics and in authorship research. Although Arabic is the sixth most spoken language in the world, there is little if any research that tackles forensic authorship analysis in Arabic and a few calls that bring this gap to attention (Mansour et al., 2012). Therefore, the first aim, which is to explore authorship in MSA and the spoken dialect of Najdi Arabic, reflects the status of the language worldwide and the status of the dialect in Saudi Arabia, as discussed in Chapter 5. Also, the research investigates the concept of idiolect and idiolectal style, another area that is rarely visited in Arabic authorship studies. Using a predetermined set of the dialect-specific features and their MSA counterparts, this research ventured to investigate the range of stylistic features that can be markers of authorship as well as cues of individual linguistic variation.

In terms of forensic authorship research, this research proposes a methodological synergy that combines stylistics, corpus linguistics, and machine learning. This is to address one of the challenges that the research encounters when used in court to present evidence and opinion: the reliability and explicability of the analysis presented (Ainsworth and Juola, 2019). There are studies that set good examples of combining corpus tools with stylistic analysis (e.g., Grant and Baker, 2001; Coulthard, 2004; Grant, 2007; Kredens and Coulthard, 2012; Johnson and Wright, 2014). And this research adds to that work by investigating a specialised corpus of tweets, a set of predetermined features in an Arabic dialect, and the idiolectal style of the sample of 13 authors through the lens of corpus linguistics.

The research that addresses authorship and idiolect tends to discuss the debate over stylometric or stylistic approaches. Wright (2014) paints a clear picture of the situation as stylistic findings are perceived as "too subjective, intuitive, and unreliable, as well as being impossible to generalize beyond the scope of the particular case in question." (p. 248). In Chapter 5, we were able to observe that incorporating corpus linguistic tools into stylistic analysis can present us more objective, statistical results that are guided by an idiolectal theory. Therefore, it can be claimed that the subjective nature of stylistic analysis can be mitigated by using corpus linguistic tools. Wright also says that stylometric and computational findings are objective and reliable but often it is "impossible to explain *why* a particular algorithm or set of features have worked in identifying authorship" (Wright, 2014: 249). This resonates with the WEKA findings (Witten et al., 2016) produced in Chapter 7, where the set of experiments revealed the optimum classifiers and parameters but did not show which features in the set that are most likely to attribute authorship. The experiments also revealed that the set of stylistic features can attribute authorship; nonetheless, WEKA does not report which features that are attributive of the authors' style. The set of features has been used to test which of the word grams: unigrams, bigrams, or trigrams would work best with the range of parameters.

9.2 Summary of results

Chapter 5 presented the corpus and stylistic aspects of the analysis, where corpus linguistic tools and stylistic analysis revealed the lexico-grammatical features each author uses. In this study, I used *Wordsmith* V.8 (Scott, 2020) to investigate the same set of stylistic features that consists of interrogatives, negatives, and deictic expressions. The corpus tools that *Wordsmith* provides, including concordance lines, collocates, pattern, and plot, reveal how each author uses each of the linguistic variants, or refrains from using them. The stylistic analysis of the contexts where these variants occur added more depth to the quantitative findings of the corpus tools and reported patterns of shifts in register for several authors, each driven by a different motive. The findings yielded some striking general patterns: the authors' overall preference to use the Najdi interrogatives, the topmost frequent variant being /wiš/. Another pattern was their overall preference to use negatives in the MSA variety, where the most frequent variants are /la/, /lam/, and /lan/. The same preference is observed in the time/space deixis, where most authors use the MSA variants /alyom/ for time, and /huna/, and /hunak/ for space. When looking closely into these general patterns we can see that there are individual variations between the authors, which were illustrated clearly in Table 5.2. The findings confirm to us that each of the 13 authors has their own idiolectal style in relation to use of the feature set, which is not only represented in the range of the variants they use but also in their frequency rates. The data in Table 5.2 is a quantitative representation of the authors' individual patterns and it is the stylistic analysis that gives a qualitative depth to the frequency rates. The stylistic analysis of the authors' interaction enabled us to observe their shift between varieties which is triggered by shifts in stance or opinion as examples will show below. One can conclude that the general patterns of preferences to use interrogatives or negatives in one variety or another are not arbitrary. These corpus stylistic revelations can have implications in terms of attributing authorship.

We learned from the stylistic analysis that authors shift from one variety to another within a single interaction, and that shift can be driven by motives such as power or stance. For instance, Alabdulkarim would shift from MSA to NA to cue his shift from an authoritative individual to one of limited authority. Another example is Algofaily who publishes an original post in MSA then shifts to NA in his replies to position himself as a member of the in-group.

Chapter 6 also explored stylistic approaches combined with CMDA to capture the online identity performance of the authors. In this study, I use the identity approach (Bucholtz and Hall, 2005) to investigate the collective identity positions the authors convey in their communication. Nationalism and regionalism were sporadically present across all authors' subcorpora represented in their stancetaking. Moreover, a trend was revealed about women's group identity awareness and construction and how the female authors communicate it in their posts, which was not present in the male authors' subcorpora. The instances observed in the female authors' subcorpora addressed gender in the content rather than using specific linguistic features, such as talking about the current empowerment of women in Saudi Arabia or their role in its history. This gender trend, nonetheless, does report some individual variation in the way these female authors talk about women's place in society. For instance, authors such as Aleidi and Alajlan show a preference to use the NA variety while Alrasheed, Alnuhait, and Almohanna would consistently opt for MSA. Also, the findings support the notion that Bucholtz and Hall (2005) propose in their approach which is that identity unfolds and develops through interaction. The authors' interactions, whether in original posts, replies, or comments, took their stances or revealed their perceptions on numerous topics as well as implications related to their identities. For instance, we have seen examples of the female authors, especially Alajlan, and her fluctuating views about feminism. Her tweets show that she is in full support of women's empowerment, but when engaging with her audience she takes a critical stance towards feminism. Another interesting example is Allahim who built his discourse over a period of time with a pattern of sarcastic stances, which are usually cued by Najdi person deixis variants such as /kud di/ *take this*, /dak/ *that*, and /dolak/ *those*.

Furthermore, I used the Faceted Classification Scheme (Herring, 2007) to investigate the authors' activities and encapsulate their individual identity themes. I looked at the authors' range of exchanges of information, jokes, and insults as well as debate and games. Similar to the stylistic features, the sample selected to examine the activities showed individual patterns. One can propose that, while producing their own idiolectal style, the authors produce a pattern of activities in CMD that is unique and individual. I revisit Turell's (2010) definition of idiolectal style once more, as:

> hav[ing] to do primarily, not with what system of language/dialect an individual has, but with a) how this system, shared by lots of people, is used in a distinctive way by a particular individual; b) the speaker/writer's production, which appears to be 'individual' and 'unique' (Coulthard, 2004) and also c) Halliday's (1989) proposal of 'options' and 'selections' from these options. (Turell, 2010: 7)

The findings of this research propose an extra dimension to Turell's concept of idiolectal style. Stances, whether epistemic or evaluative, are an additional aspect the authors reveal in their individual and unique linguistic selections. We learned from this research that it is not only the language that can be used in an individual manner but also the stances and the activities in which an author partakes. These findings also assimilate to the identity approach (Bucholtz and Hall, 2005) which boils down to the idea that online identity is built and developed through interaction, something that has been exhibited throughout data analysis.

Chapter 7 presented the computational aspect of the methodological synergy and explored the concept of idiolectal style using a machine learning tool. It proposed an authorship attribution problem for WEKA to solve by assigning the texts to the 13 authors, where 80% of the authors' text files were used to train the tool. The remaining 20% were combined into one text file for the tool to classify the text and assign each part to its author. The study employed the set of predetermined stylistic features and consisted of two stages: the first was to identify which classifier performs best and the second was to discover which parameter yields most accurate results. In the first stage, the test aimed to discover which of the seven classifiers assigns the texts to the correct authors most accurately, which was Linear SVM. The findings are supported by Al-Harbi et al. (2008), Fissette (2010) and Brocardo et al., (2014). The second stage was to find which of the three prospective parameters works best with the classifier Linear SVM in three categories of word grams: unigrams, bigrams, and trigrams. The findings show that the second parameter (min_df=1 – max_df=int (80/100)) and word bigrams outperforms the other parameters and unigrams and trigrams, which is further supported by previous studies (Feiguina and Hirst, 2007; Sadat et al., 2014).

Chapter 8 put the proposed methodology to the test by designing an empirical experiment of an authorship attribution dispute. The aim is to test the performance of the methodology proposed by this research. The scenario is a dispute over the authorship of a set of twenty hypothetically threatening tweets possibly produced by one of the two authors, using the same feature set to perform a stylistic, statistical analysis and to use corpus tools to investigate concordance lines and collocates. This is followed by an examination of the authors' activities and a comparison of them against the activities in the questioned tweets. The methodology shows promising potential in solving authorship attribution problems in a reliable, reproducible fashion.

9.3 Limitations

There are some limitations that constrained the study and results achieved in this research. The literature confirms that one of the challenges in corpus linguistic research is to compile a representative corpus. Forensic linguists and sociolinguists interested in individual variation agree that the mission of providing a full description of one's idiolect is impossible. In the case of the NASCoT design, it is safe to claim that the corpus represents the authors' idiolectal style in the specific genre of microblogs (i.e., Twitter) and during the period of data collection, since I collected all of their output in a 19-month period (apart from a few tweets that I discarded for ethical reasons). Therefore, the limitation lies in the fact that this corpus only represents the authors in the microblog genre and their style could vary in other forms of communication. The same claim can be made in terms of time; the NASCoT is a snapshot of the authors' current idiolectal style, which can change over a period of time. At the same time, I am restricting my analysis of style simply to the set of 44 syntactic features and other aspects of style are not part of the analysis. Therefore, I cannot claim that the feature set characterises the authors' styles, as this is only one aspect of their wider stylistic habits. Another important limitation to point out about style is how it is interpreted and examined in this research. To encapsulate the style of 13 different individuals in one research project would be an ambitious project. Therefore, I decided to explore a small set of 44 syntactic dialectal features to represent style instead of addressing it more broadly by investigating sentences length, verb choice, emojis, etc.

Another limitation has been addressed in Chapters 5 and 7 which is regarding some of the stylistic features examined. Najdi Arabic is one of the few dialects that is a derivative of Classical Arabic and retains some of its attributes. This results in some overlap between its features with the Modern Standard ones.

The NA variants take ranges of orthographic shapes and that was accounted for in the data analysis stage by identifying every possible shape and examining them in *Wordsmith*. For instance, the NA variant for *tomorrow* can be written as بكرا ending with the vowel alif (ا); the other possible form is بكره ending with the semi-vowel haa' (ه). Therefore I had to account for both shapes in the analysis by searching for them as separate queries. However, the orthographic shape of the overlapping features is identical in both varieties, which makes it impossible to make a distinction when examining them in *Wordsmith* without considering their contexts. For example, the spatial deixis هنا *here* is orthographically identical in NA and MSA, which makes it impossible for the corpus tool to tell them apart or to detect which variety the authors are using without human supervision.

In addition, previous computational studies report the challenge of tackling short text data in authorship research that is often referred to as brevity of text (e.g., MacLeod and Grant, 2012; Rocha et al., 2016; Khonji and Iraqi, 2018; 2020). This is also one of the limitations observed in this research that is the brief nature of the tweets, which as a result affected the accuracy rates of the computational analysis.

Finally, the data used in this research is a substantial number of tweets posted by a group of individuals whom I use as informants to test a methodology. Nonetheless, the data cannot be classified as forensic data nor the individuals as suspects.

9.4 Contributions and future directions

This thesis addresses the theory of idiolect and its methodological approaches to solve authorship attribution problems, and on that note, it offers a range of contributions. This section discusses the contributions made by this research and recommendations for future studies.

The literature related to authorship research both in computational and linguistic fields tends to have separate toolboxes to solve the problem. In this research, I'm proposing a methodological synergy that combines these tools by incorporating corpus linguistics, stylistics, machine learning, and CMD. This successful attempt to bridge the gap between qualitative and quantitative approaches contributes to the forensic authorship field. Also, one cannot overlook the research gap in forensic linguistics research in Arabic and especially in authorship attribution. Therefore, this research is a contribution in that regard. This is also a new venture in which idiolectal style is explored in an Arabic variety, another research area that is rarely addressed in Arabic. The final contribution this research offers is exploring an Arabic corpus in a stylistic context. Arabic has been widely explored in corpus linguistics research that is rooted in computational approaches and that aims to explore tools. The approach pursued in this research is supported by qualitative, stylistic analysis that accounts for context and how the discourse unfolds.

On a final note, this research can be a steppingstone into future directions, one of which is to establish forensic authorship research in Arabic dialects. As noted earlier in Chapter 2, authorship analysis, in its different subcategories - attribution, identification, and profiling - is a familiar area of research in Arabic computational studies. However, it is rarely addressed in a forensic linguistic context and this research is one of the firsts to do so following the lead of García-Barrero et al. (2013). It will also be important to explore the different facets of idiolectal style in other genres of communication (e.g., emails, text messages, etc.) in Arabic dialects. In terms of computational tools, it would be interesting to venture into other machine learning tools using the same synergy. Lastly, investigating authentic forensic data using the proposed methodology can further authenticate its reliability. In a wider context, there is much to be investigated when it comes to forensic authorship research in Arabic and its dialects, and to achieve that using authentic forensic data is a step in that direction.

List of References

ABBADI, S. M. B. (2018) Arbitration in Saudi Arabia: The Reform of Law and Practice, PhD thesis, Pen State Law.

ABBASI, A. & CHEN, H. (2005) ‘Applying authorship analysis to extremist-group web forum messages’, *IEEE Intelligent Systems,* 20, pp. 67-75.

ABBASI, A. & CHEN, H. (2008) ‘Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace’, *ACM Transactions on Information Systems (TOIS),* 26, pp. 1-29.

ABBASI, A., CHEN, H. & NUNAMAKER, J. F. (2008) ‘Stylometric identification in electronic markets: Scalability and robustness’, *Journal of Management Information Systems,* 25, pp. 49-78.

ABBOUD, P. F. (1964) *The syntax of Najdi Arabic*, The University of Texas at Austin.

ABBOUD, P.F. (1978) ‘The vowel of the imperfect prefix in Najdi Arabic’, *Linguistic and Literary Studies*, pp.129-138.

ABDELLAH, A. (2019) ‘What does a Forensic Linguist Really Do? A Close Reading of three Cases of Authorship’/ طبيعة عمل عالم اللغة الجنائي قراءة في ثلاث قضايا في تحقيق نسبة النص’, *Arab Journal of Forensic Sciences & Forensic Medicine*, 1(9), pp.1322-1322.

ABDUL-MAGEED, M., ZHANG, C., ELMADANY, A. & UNGAR, L. (2020) ‘Toward micro-dialect identification in diaglossic and code-switched environments’, *arXiv preprint arXiv:2010.04900*.

AHEARN, L. M. (2001) ‘Language and agency’, *Annual review of anthropology,* 30, pp. 109-137.

AINSWORTH, J. & JUOLA, P. (2018) ‘Who wrote this: Modern forensic authorship analysis as a model for valid forensic science’, *Washington University Law Review*, 96, p.1159.

AL-BASSAM AL-TAMIMI, M. (1818) Addurar Almafakhir Fi Akhbar Alarab Alawakhir الدرر المفاخر في أخبار العرب الأواخر (قبائل العرب). 2nd edition. ALAJMI, S. G. (ed.).

AL-ESSA, A. (2009) When Najd meets Hijaz: dialect contact in Jeddah. In *Arabic Dialectology,* pp. 201-222.

AL-HARBI, S., ALMUHAREB, A., AL-THUBAITY, A., KHORSHEED, M. & AL-RAJEH, A. (2008) Automatic Arabic text classification.

ALJASSIR, H. (2001) Jamharat Ansab Alussar Almutahidhira جمهرة أنساب الأسر المتحضرة. Al-Yamamah, Riyadh.

AL-MEHDAR, M. & AL-MEHDAR, H. (2019) *Litigation and enforcement in Saudi Arabia: overview | Practical Law*. [online] Practical Law. Available at: <https://uk.practicallaw.thomsonreuters.com/w-020-5670?transitionType=Default&contextData=(sc.Default)&firstPage=true#co_anchor_a259488> [Accessed 25 March 2022].

AL-ROJAIE, Y. (2013) ‘Regional dialect leveling in Najdi Arabic: The case of the deaffrication of [k] in the Qaṣīmī dialect’, *Language Variation and Change*, 25(1), pp.43-63.

AL-THUBAITY, A. & ALMUJAIWEL, S. (2018) ‘A quantitative inquiry into the keywords between primary and reference Arabic corpora’, *Journal of Quantitative Linguistics,* 25**,** pp. 122-141.

AL-TWAIRESH, N., AL-KHALIFA, H. & AL-SALMAN, A. (2015) ‘Towards analyzing Saudi tweets’, *First International Conference on Arabic Computational Linguistics (ACLing)*, IEEE, pp.114-117.

AL-TWAIRESH, N., AL-MATHAM, R., MADI, N., ALMUGREN, N., AL-ALJMI, A.-H., ALSHALAN, S., ALSHALAN, R., ALRUMAYYAN, N., AL-MANEA, S. & BAWAZEER, S. (2018) ‘Suar: Towards building a corpus for the Saudi dialect’, *Procedia computer science,* 142**,** pp. 72-82.

ALAIYED, M. and ABDULLAH, S. (2018) Diglossic code-switching between standard Arabic and Najdi Arabic in religious Discourse (Doctoral dissertation, Durham University).

ALANAZI, F., JONES, A. and MENON, C. (2018) ‘Sharia Law and Digital Forensics in Saudi Arabia’, *The Journal of Digital Forensics, Security and Law: JDFSL*, 13(3), pp.4-19.

ALBADARNEH, J., TALAFHA, B., AL-AYYOUB, M., ZAQAIBEH, B., AL-SMADI, M., JARARWEH, Y. & BENKHELIFA, E. (2015) ‘Using big data analytics for authorship authentication of arabic tweets’, In *2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC),* IEEE, pp. 448-452.

ALDOASREE, O.M. (2016) *Language attitudes toward Saudi dialects*. California State University, Long Beach.

ALENAZY, H. M. (2017) The Construction of Gender in Saudi Arabia. Proceedings of International Academic Conferences, 2017. International Institute of Social and Economic Sciences.

ALHAZMI, L.M. and ALFALIG, H.A. ‘Saudis’ Attitudes Towards Their Dialects: A Keyword Technique’, *Sciences*, 23(1), pp.114-21

ALFAIFI, A. & ATWELL, E. 2016 ‘Comparative evaluation of tools for Arabic corpora search and analysis’, *International Journal of Speech Technology,* 19**,** 347-357.

ALKARNI, S. (2018) *Twitter response to Vision 2030: a case study on current perceptions of normative disorder within Saudi social media.* Université d'Ottawa/University of Ottawa.

ALMAHMOUD, J. (2015) *Framing on Twitter: How Saudi Arabians intertextually frame the women2drive campaign,* Georgetown University.

ALMUT, K. (2010) Building small specialised corpora. In *The Routledge handbook of corpus linguistics.* Routledge.

ALMUTAIRI, S. (2021) ‘Disagreement strategies and (Im) politeness in Saudis’ Twitter Communication’, *Jounral of Languages, Text, and Society,* 5, pp.1-40.

ALOSAIMY, A. & ATWELL, E. (2017) ‘Sunnah Arabic Corpus: Design and Methodology’, In *Proceedings of the 5th International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2017),* Leeds.

ALOTHMAN, E. (2012) *Digital Vernaculars: An Investigation of Najdi Arabic in Multilingual Synchronous Computer-Mediated Communication,* The University of Manchester (United Kingdom).

ALQAHTANI, K. (2015) A sociolinguistic study of the Tihami Qahtani dialect in Asir, Southern Arabia, PhD thesis, University of Essex.

ALRUILY, M. (2014) 'Issues of dialectal saudi twitter corpus', *International Arab Journal of Information Technology*, 17(3), pp.367-374.

ALSAAIDI, H. Z. (2020) *Nation Branding and The Case of Saudi Vision 2030 and The Use of Twitter*, Rochester Institute of Technology.

ALSHUTAYRI, A. & ATWELL, E. (2017) 'Exploring Twitter as a source of an Arabic dialect corpus', *International Journal of Computational Linguistics (IJCL),* 8**,** pp. 37-44.

ALTAKRORI, M.H., IQBAL, F., FUNG, B.C., DING, S.H. and TUBAISHAT, A., (2018) 'Arabic authorship attribution: An extensive study on twitter posts', In ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(1), pp.1-51.

ALTHENEYAN, A. S. & MENAI, M. E. B. (2014) 'Naïve Bayes classifiers for authorship attribution of Arabic texts', *Journal of King Saud University-Computer and Information Sciences,* 26**,** pp. 473-484.

ANGOURI, J. 'Studying identity', *Research methods in intercultural communication: A practical guide***,** pp. 37-52.

ANTAKI, C. & WIDDICOMBE, S. (1998) *Identities in talk*, Sage.

ANWAR, W., BAJWA, I. S., CHOUDHARY, M. A. & RAMZAN, S. (2018) 'An empirical study on forensic analysis of urdu text using LDA-based authorship attribution', *IEEE Access,* 7**,** pp. 3224-3234.

ARGAMON, S. (2018) 'Computational forensic authorship analysis: Promises and pitfalls', *Language and Law/Linguagem e Direito,* 5**,** pp. 7-37.

ARGAMON, S., KOPPEL, M., FINE, J. & SHIMONI, A. R. (2003) 'Gender, genre, and writing style in formal written texts', *Text & Talk,* 23**,** pp. 321-346.

ARONSSON, K. (1998) 'Identity-in-interaction and social choreography', *Research on language and social interaction,* 31**,** pp.75-89.

ASHMORE, R. D., DEAUX, K. & MCLAUGHLIN-VOLPE, T. (2004) 'An organizing framework for collective identity: articulation and significance of multidimensionality', *Psychological bulletin,* 130**,** 80.

AZARBONYAD, H., DEHGHANI, M., MARX, M. & KAMPS, J. (2015) 'Time-aware authorship attribution for short text streams', *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 727-730.

BAAYEN, R.H., (2001) *Word frequency distributions (Vol. 18)*. Springer Science & Business Media.

BAKER, P. (2010) *Sociolinguistics and corpus linguistics,* Edinburgh University Press.

BANGA, R., BHARDWAJ, A., PENG, S.-L. & SHRIVASTAVA, G. (2018) 'Authorship attribution for online social media', *Social Network Analytics for Contemporary Business Organizations.* IGI Global.

BASSIOUNEY, R. & WALTERS, K. (2020) *The Routledge Handbook of Arabic and Identity,* Routledge.

BELVISI, N.M.S., MUHAMMAD, N. and ALONSO-FERNANDEZ, F. (2020) 'Forensic authorship analysis of microblogging texts using n-grams and stylometric features', In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, pp. 1-6.

BIBER, D. (2006) *University language : a corpus-based study of spoken and written registers,* Amsterdam: J. Benjamins.

BIBER, D. (2015) 'Corpus-based and corpus-driven analyses of language variation and use', In *The Oxford handbook of linguistic analysis.*

BIBER, D. & EGBERT, J. (2016) ‘Register variation on the searchable web: A multi-dimensional analysis’, *Journal of English Linguistics,* 44**,** pp. 95-137.

BINTURKI, T. A. S. (2015) *The acquisition of negation in Najdi Arabic.* University of Kansas.

BISCHOFF, S., DECKERS, N., SCHLIEBS, M., THIES, B., HAGEN, M., STAMATATOS, E., STEIN, B. & POTTHAST, M. (2020) The importance of suppressing domain style in authorship analysis. *arXiv preprint arXiv:2005.14714*.

BISHOP, C.M. and NASRABADI, N.M. (2006) Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

BOZKURT, I. N., BAGHOGLU, O. & UYAR, E. (2007) ‘Authorship attribution’, In *The 22nd International Symposium on Computer and Information Sciences, 2007*. IEEE, pp. 1-5.

BROCARDO, M. L., TRAORE, I. & WOUNGANG, I. (2015) ‘Authorship verification of e-mail and tweet messages applied for continuous authentication’, *Journal of Computer and System Sciences,* 81**,** pp. 1429-1440.

BRUBAKER, R. & COOPER, F. (2000) ‘Beyond" identity"’, *Theory and society,* 29**,** pp. 1-47.

BUCHOLTZ, M. & HALL, K. (2005) ‘Identity and interaction: A sociocultural linguistic approach’, *Discourse studies,* 7**,** pp. 585-614.

BUCHOLTZ, M. & HALL, K. (2008) Finding identity: Theory and data. *Multilingua*, 27 (1-2), pp. 151-163.

BUTLER, J., (1998) *Sex and gender in Simone de Beauvoir's Second Sex*. Yale University Press.

BUTLER, J., (2004) *Undoing gender*. Routledge.

CALLE-MARTÍN, J. & MIRANDA-GARCÍA, A. (2012) ‘Stylometry and authorship attribution: introduction to the special issue’, *English Studies,* 93**,** pp. 251-258.

CANBAY, P., SEZER, E. A. & SEVER, H. (2020) ‘Deep Combination of Stylometry Features for Authorship Analysis’, *International Journal of Information Security Science,* 9**,** pp.154-163.

CHASKI, C. (2007) ‘The keyboard dilemma and authorship identification’, In *IFIP International Conference on Digital Forensics, 2007*. Springer, pp. 133-146.

CHAUDHRY, I. (2014) ‘Arab Revolutions: Breaking fear|# hashtags for change: Can Twitter generate social progress in Saudi Arabia’, *International Journal of Communication,* 8**,** pp. 943-961.

COLE, S. A. (2009) ‘Forensics without uniqueness, conclusions without individualization: the new epistemology of forensic identification’, *Law, probability and risk,* 8**,** pp. 233-255.

COLLINS, L. (2019) *Corpus Linguistics for Online Communication*. London.

COTTERILL, J. (2010) ‘How to use corpus linguistics in forensic linguistics’, *The Routledge handbook of corpus linguistics.* Routledge.

COULTHARD, M. (1994) ‘On the use of corpora in the analysis of forensic texts’, *The International Journal of Speech, Language and the Law,* 1**,** pp. 27-43.

COULTHARD, M. (2004) ‘Author identification, idiolect, and linguistic uniqueness’, *Applied linguistics,* 25**,** pp. 431-447.

COULTHARD, M., GRANT, T. and KREDENS, K. (2011) ‘Forensic Linguistics’, In Ruth Wodak, Barbara Johnstone and Paul Kerswill (eds.). *The SAGE Handbook of Sociolinguistics*. London: Sage, pp. 531–544.

COULTHARD, M., JOHNSON, A. and WRIGHT, D. (2016) *An introduction to forensic linguistics: Language in evidence*. Routledge.

COULTHARD, M., MAY, A. & SOUSA-SILVA, R. (2020) *The Routledge Handbook of Forensic Linguistics*, Routledge.

COUPLAND, N. (2007) *Style: Language variation and identity*. Cambridge University Press.

CRYSTAL, D. (2011) *Internet linguistics: A student guide*. Routledge.

DAVIES, B. (2005) 'Communities of practice: Legitimacy not choice', *Journal of Sociolinguistics,* 9**,** pp. 557-581.

DE FINA, A. (2013a) 'Positioning level 3: Connecting local identity displays to macro social processes', *Narrative Inquiry,* 23**,** pp. 40-61.

DE FINA, A. (2013b) 'Top-down and bottom-up strategies of identity construction in ethnic media', *Applied linguistics,* 34**,** pp. 554-573.

DE VEL, O., ANDERSON, A., CORNEY, M. & MOHAY, G. (2002) E-mail authorship attribution for computer forensics. In *Applications of Data Mining in Computer Security.* Springer, Boston, pp. 229-250.

DOUGLAS, D. (1992) 'The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings', *Computers and the Humanities,* 26**,** pp. 331-345.

DU BOIS, J. W. & KÄRKKÄINEN, E. (2012) 'Taking a stance on emotion: Affect, sequence, and intersubjectivity in dialogic interaction', *Text & Talk,* 32**,** pp. 433-451.

ECKERT, P. (2012) 'Three waves of variation study: the emergence of meaning in the study of sociolinguistic variation', *Annual Review of Anthropology,* 41, pp. 87-100.

ECKERT, p. and RICKFORD, J. R. eds. (2001) *Style and sociolinguistic variation*. Cambridge University Press.

EDER, M. (2010) 'Does Size Matter? Authorship Attribution, Small Samples, Big Problem', *Digital Humanities,* pp. 132-134.

EDER, M. (2017) 'Short Samples in Authorship Attribution: A New Approach', In *Digital Humanities 2017: Conference abstract,* pp. 221-224.

EVANS, M. (2012) 'A sociolinguistics of early modern spelling: An account of Queen Elizabeth I's correspondence', *Outposts of historical corpus linguistics: From the Helsinki Corpus to a proliferation of resources,* 10.

EVANS, M. (2016) 'By the Queen: Collaborative Authorship in scribal correspondence of Queen Elizabeth I', *Women and Epistolary Agency in Early Modern Culture*. New York: Routledge, pp. 1450-1690.

EVANS, M. (2018) 'Styling Power: A Corpus-Linguistic Approach to the Correspondence of Queen Elizabeth I', In *Elizabeth I in Writing,* Palgrave Macmillan, Cham. pp. 59-82.

FAIRCLOUGH, N. (2013) Critical discourse analysis and critical policy studies. *Critical policy studies,* 7**,** pp. 177-197.

FATANY, S. (2013) *Modernizing Saudi Arabia,* Createspace Independent Publishing.

FATIMA, M., ANWAR, S., NAVEED, A., ARSHAD, W., NAWAB, R. M. A., IQBAL, M. & MASOOD, A. (2018) 'Multilingual SMS-based author profiling: Data and methods', *Natural Language Engineering,* 24, pp. 695-724.

FEIGUINA, O. & HIRST, G. (2007) 'Bigrams of syntactic labels for authorship discrimination of short texts', *Literary and Linguistic Computing,* 22 (4), pp. 405-417.

FERGUSON, C.A. (1959) Diglossia. *Word,* 15(2), pp.325-340.

FERRARA, K., BRUNNER, H. and WHITTEMORE, G. (1991) 'Interactive written discourse as an emergent register', *Written Communication,* 8(1), pp.8-34.

FISSETTE, M. (2010) Author Identifcation in Short Texts. Bachelor Thesis.

FLOWERDEW, L. (2004) The argument for using English specialized corpora to understand academic and professional language. In *Discourse in the professions: Perspectives from corpus linguistics,* 11, p.13-33.

FOBBE, E. (2020) 'Text-Linguistic Analysis in Forensic Authorship Attribution', *Journal of Language and the Law,* 9**,** pp. 93-114.

FOLEY, S. (2010) 'All I want is equality with girls: gender and social change in the twenty-first century gulf', *MERIA Journal,* 14(1), pp.21-37.

FRANTZESKOU, G., STAMATATOS, E., GRITZALIS, S. and KATSIKAS, S. (2006) 'Effective identification of source code authors using byte-level information', In *Proceedings of the 28th International Conference on Software Engineering,* pp. 893-896.

FREY, J.-C., GLAZNIEKS, A., and GLÜCK, A. (2022) 'Dialect or not? How to identify the use of dialect in written online communication', In *the 11th International Conference on Language Variation in Europe*, Vienna. [online]. Available from: https://iclave11.dioe.at/.

GALES, T. (2011) Identifying interpersonal stance in threatening discourse: An appraisal analysis. *Discourse Studies,* 13**,** pp.27-46.

GARCÍA-BARRERO, D., FERIA, M. & TURELL, M. T. (2013) 'Using function words and punctuation marks in Arabic forensic authorship attribution', In *Proceedings of the 3rd European Conference of the International Association of Forensic Linguists*, pp.42-56.

GAUNTLETT, D. (2008) *Media, gender and identity: An introduction*, Routledge.

GEORGE, A. (1990) 'Whose language is it anyway? Some notes on idiolects', *The Philosophical Quarterly (1950-),* 40**,** pp. 275-298.

GEORGAKOPOULOU, A. (2011) 'Computer-mediated communication'. In Östman, J. and Verschueren, J. eds. *Pragmatics in Practice,* 9, p.93-110.

GILES, H., COUPLAND, N. & COUPLAND, I. (1991) 'Accommodation theory: Communication, context, and consequence', In *Contexts of accommodation: Developments in applied sociolinguistics,* pp.1-68.

GÓMEZ-ADORNO, H., POSADAS-DURÁN, J.-P., SIDOROV, G. & PINTO, D. (2018) 'Document embeddings learned on various types of n-grams for cross-topic authorship attribution', *Computing,* 100**,** pp. 741-756.

GRAFF, D., BUCKWALTER, T., JIN, H. & MAAMOURI, M. (2006) *Lexicon Development for Varieties of Spoken Colloquial Arabic*.

GRANT, T. and BAKER, K. (2001) 'Identifying reliable, valid markers of authorship: a response to Chaski', *International Journal of Speech, Language and the Law,* 8(1), pp.66-79.

GRANT, T. (2007) 'Quantifying evidence in forensic authorship analysis', *International Journal of Speech, Language & the Law,* 14(1), pp.1-25.

GRANT, T. (2012) 'TXT 4N6: method, consistency, and distinctiveness in the analysis of SMS text messages', *Journal of Law and Policy,* 21(2), p.467-494.

GRANT, T. and MACLEOD, N. (2018) 'Resources and constraints in linguistic identity performance–a theory of authorship', *Language and Law/Linguagem e Direito,* 5(1), pp.80-96.

GRANT, T. & MACLEOD, N. (2020) *Language and Online Identities: The Undercover Policing of Internet Sexual Crime,* Cambridge University Press.

GREEN, R. M. & SHEPPARD, J. W. (2013) 'Comparing frequency-and style-based features for twitter author identification', In the *Twenty-Sixth International Florida Artifical Intelligence Research Society Conference,* pp. 64-69.

GRIES, S. T. (2012) 'Corpus Linguistics: Quantitative Methods', *The Encyclopedia of Applied Linguistics***,** pp.1-6.

GRIES, S. T. (2013) ‘Corpus linguistics: quantitative methods. *In:* CHAPELLE, C. A. (ed.) *The Encyclopedia of Applied Linguistics.* Blackwell Publishing Ltd.

GRIEVE, J. (2007) Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing,* 22**,** pp.251-270.

GRZYBEK, P. (2006) *Contributions to the Science of Text and Language: Word Length Studies and Related Issues,* Springer Netherlands.

GUMPERZ, J. J. (1982) Fact and inference in courtroom testimony. *Language and social identity***,** pp.163-195.

HALL, K., BORBA, R. & HIRAMOTO, M. 2020. Language and gender. *The International Encyclopedia of Linguistic Anthropology***,** pp.1-22.

HALL, K., BORBA, R. & HIRAMOTO, M. (2021) ‘Thirty-year retrospective on language, gender and sexuality research: Special focus: Practice’, *Gender and Language,* 15(3)**,** pp. 394-395.

HALVANI, O., WINTER, C. & PFLUG, A. (2016) Authorship verification for different languages, genres and topics. *Digital Investigation,* 16**,** pp. S33-S43.

HARDAKER, C. (2010) ‘Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions’, *Journal of Politeness Research*, 6(2), pp.215-242.

HARERI, R. H. (2018) ‘Saudi Women and the Expression of Identity in the Domestic Living Space’, *Humanities and Social Sciences,* 6**,** 7.

HAZA'AL RDAAT, S. & GARDNER, S. (2017) An analysis of use of conditional sentences by Arab students of English. *Advances in Language and Literary Studies,* 8, pp. 1-13.

HEINI, A. and ALAMR, M. (2022) Global interest meets new perspectives: a report on the 15th Biennial Conference of the IAFL. *International Journal of Speech, Language, and the Law,* 28 (2), pp. 251-265.

HERRING, S. C. (2007) A faceted classification scheme for computer-mediated discourse. *Language@ internet,* 4.

HERRING, S.C. (2013) ‘Discourse in Web 2.0: Familiar, reconfigured, and emergent’, *Discourse*, 2(0), pp.1-26.

HEYDON, G. (2019) *Researching Forensic Linguistics: Approaches and Applications,* Routledge.

HOLES, C. (1996) Bruce Ingham: Najdi Arabic: Central Arabian.(London Oriental and African Language Library, Vol. I.) xvii, 215 pp. Amsterdam and Philadelphia: John Benjamins Publishing Company, 1994. $62. *Bulletin of the School of Oriental and African Studies,* 59**,** 561-563.

HOLMES, D. I. (1994) ‘Authorship attribution’, *Computers and the Humanities,* 28**,** pp. 87-106.

HOWEDI, F. & MOHD, M. (2014) ‘Text classification for authorship attribution using Naive Bayes classifier with limited training data’, *Computer Engineering and Intelligent Systems,* 5**,** pp.48-56.

HUNSTON, S. & THOMPSON, G. (2000) *Evaluation in text: Authorial stance and the construction of discourse: Authorial stance and the construction of discourse,* Oxford University Press, UK.

INGHAM, B. (1994) *Najdi Arabic: Central Arabian*, John Benjamins Publishing.

IQBAL, F., KHAN, L. A., FUNG, B. C. & DEBBABI, M. (2010) ‘E-mail authorship verification for forensic investigation’, In *Proceedings of the 2010 ACM Symposium on Applied computing,* pp. 1591-1598.

IRVINE, J. T., GAL, S. & KROSKRITY, P. V. (2009) ‘Language ideology and linguistic differentiation’, *Linguistic anthropology: A reader,* 1**,** pp.402-434.

ISMAIL, M. A. (2012) ‘Sociocultural identity and Arab women's and men's code-choice in the context of patriarchy’, *Anthropological Linguistics,* 54**,** pp.261-279.

ISMAIL, S. M., ALSHAYHAN, N. R., ALWAFAI, S. & ESSAM, B. A. (2019) ‘Frequency of Using Najdi Arabic Words Among Saudi College Male Students’, *International Journal of English Linguistics,* 9(2)*,* pp.24-29.

JAFFE, A. (2009) *Stance: sociolinguistic perspectives*, OUP: USA.

JOHNSON, A. & WRIGHT, D. (2014) ‘Identifying idiolect in forensic authorship attribution: an n-gram textbite approach’, *Language and Law/Linguagem e Direito,* 1**,** pp.37-69.

JOHNSTONE, T.M. (1967) *Eastern Arabian Dialect Studies* (17). Oxford University Press.

JOHNSTONE, B. (1996) *The linguistic individual: Self-expression in language and linguistics*, Oxford University Press.

JOHNSTONE, B. (2009) Stance, style, and the linguistic individual, In JAFFE, A. (ed.)*, Stance: sociolinguistic perspectives*. New York: Oxford.

JOHNSTONE, B. (2017). *Discourse analysis,* John Wiley & Sons.

JOHNSTONE, M.-J. (2017) ‘Honesty and Integrity in Authorship Attribution’, *Australian Nursing and Midwifery Journal,* 24(10)**,** p. 30.

JUOLA, P. (2006) ‘Authorship attribution for electronic documents’, In *IFIP International Conference on Digital Forensics*, Springer, pp.119-130.

JUOLA, P. (2007) ‘Future trends in authorship attribution’, In *IFIP International Conference on Digital Forensics,* Springer, pp. 119-132.

JUOLA, P. (2008) *Authorship attribution*, Now Publishers Inc.

JUOLA, P. (2020) ‘Authorship Studies and the Dark Side of Social Media Analytics’, *Journal of Universal Computer Science,* 26(1)**,** pp.156-170.

JUOLA, P. (2021) ‘Verifying authorship for forensic purposes: A computational protocol and its validation’, *Forensic Science International,* 325**,** p.110824.

KHALIFA, S., HABASH, N., ABDULRAHIM, D. & HASSAN, S. (2016) A large scale corpus of Gulf Arabic. *arXiv preprint arXiv:1609.02960*.

KHAN, M.E. and KHAN, F. (2012) A comparative study of white box, black box and grey box testing techniques. *International Journal of Advanced Computer Science and Applications, 3*(6), pp.12-15.

KHOMYTSKA, I., TESLYUK, V., HOLOVATYY, A. & MORUSHKO, O. (2018) 'Development of methods, models, and means for the author attribution of a text’, *Восточно-Европейский журнал передовых технологий***,** pp.41-46.

KHONJI, M. & IRAQI, Y. (2018) ‘Attributing authors of emirati tweets’, In 2018 IEEE Global Communications Conference (GLOBECOM), IEEE, pp.206-212.

KHONJI, M. & IRAQI, Y. (2020) Evaluating author attribution on Emirati tweets. *IEEE Access,* 8**,** pp.149531-149543.

KIESLING, S. F. (2009) Style as stance. pp.171-194.

KINGSTON, J. & STALKER, K. (2006) ‘Forensic stylistics in an online world’, *International Review of Law Computers & Technology,* 20**,** pp.95-103.

KOESTER, A. (2012) ‘Corpora and workplace discourse’, *Corpus Applications in Applied Linguistics: Current Approaches and Future Directions,* pp.47-64.

KOORS IV, G. B. (2014) *Saudi Arab Stereotype and Culture in News Media and Literature,* Truman State University.

KOPPEL, M., SCHLER, J. & ARGAMON, S. (2009) ‘Computational methods in authorship attribution’, *Journal of the American Society for information Science and Technology,* 60**,** pp.9-26.

KOPPEL, M., SCHLER, J. & ARGAMON, S. (2011) ‘Authorship attribution in the wild’, *Language Resources and Evaluation,* 45**,** pp.83-94.

KOPPEL, M., SCHLER, J., ARGAMON, S. & WINTER, Y. (2012) 'The "fundamental problem" of authorship attribution', *English Studies,* 93**,** pp.284-291.

KREDENS, K. (2002) 'Towards a corpus-based methodology of forensic authorship attribution: a comparative study of two idiolects', In PALC, 1, pp. 405-437.

KREDENS, K. (2006) On the status of linguistic evidence in litigation. In Nowak, P., Nowakaowski (eds). *Language, Communication, Information,* 1, pp.23-30.

KREDENS, K., PERKINS, R. & GRANT, T. (2019) 'Developing a framework for the explanation of interlingual features for native and other language influence detection', *Language and Law/Linguagem e Direito,* 6**,** pp.10-23.

KREDENS, K. J. & COULTHARD, R. M. (2012) Corpus linguistics in authorship identification. *Oxford Handbook of Language and Law.* Oxford University Press.

KUHL, J. W. (2003) *The idiolect, chaos, and language custom far from equilibrium: Conversations in Morocco.* University of Georgia Athens, GA.

LABOV, W. (1972) Sociolinguistic patterns (No. 4). University of Pennsylvania press.

LABOV, W. (1990) 'The intersection of sex and social class in the course of linguistic change', *Language variation and change,* 2**,** pp.205-254.

LABOV, W. (2002) 'Driving forces in linguistic change', In Proceedings of *the 2002 international conference on Korean linguistics*.

LAMBERS, M. & VEENMAN, C. J. (2009) 'Forensic authorship attribution using compression distances to prototypes', *International Workshop on Computational Forensics,*13-24. Springer.

LAYTON, R., WATTERS, P. & DAZELEY, R. (2012) 'Recentred local profiles for authorship attribution', *Natural Language Engineering,* 18**,** pp.293-312.

LEWIS, R. (2013) Complementizer agreement in Najdi Arabic (Doctoral dissertation, University of Kansas).

LI, J. S., CHEN, L. C., MONACO, J. V., SINGH, P. & TAPPERT, C. C. (2017) 'A comparison of classifiers and features for authorship authentication of social networking messages', *Concurrency and Computation: Practice and Experience,* 29**,** e3918.

LOUWERSE, M. M., MCCARTHY, P. M., MCNAMARA, D. S. & GRAESSER, A. C. (2004) Variation in language and cohesion across written and spoken registers. In Proceedings of *the Annual Meeting of the Cognitive Science Society,* 26(26), pp.843-848.

MACLEOD, N. and 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,* pp.210-224.

MADINI, A. A. & NOOY, J. D. (2013) 'Disclosure of gender identity in Internet forums: A case study of Saudi Arabian forum communication', *Gender, Technology and Development,* 17**,** pp.233-257.

MANSOUR, M.A. (2013) 'The absence of Arabic corpus linguistics: a call for creating an Arabic national corpus', *International Journal of Humanities and Social Science,* 3(12), pp.81-90.

MARKO, K. (2021) 'Exploring the Distinctiveness of Emoji Use for Digital Authorship Analysis', *Language and Law/Linguagem e Direito,* 7(1-2), pp.36-55.

MARUKATAT, R., SOMKIADCHAROEN, R., NALINTASNAI, R. & ARAMBOONPONG, T. (2014) 'Authorship attribution analysis of thai online messages', In *International Conference on Information Science & Applications* (ICISA), IEEE, pp.1-4.

MARWICK, A. E. (2013) *Online identity*.

MCENERY, T. (2018) *Arabic corpus linguistics*, Edinburgh University Press.

MCINTYRE, D. & WALKER, B. (2019) *Corpus stylistics: Theory and practice,* Edinburgh University Press.

MCMENAMIN, G.R. (2001) 'Style markers in authorship studies', *International Journal of Speech Language and the Law*, 8(2), pp.93-97.

MCMENAMIN, G. R. (2002) *Forensic linguistics: Advances in forensic stylistics*, CRC press.

MEHLER, A., HEMATI, W., USLU, T. & LÜCKING, A. (2018) A multidimensional model of syntactic dependency trees for authorship attribution. *Quantitative analysis of dependency structures*, 72, p.315.

MENAI, M. E. B. (2012) 'Detection of plagiarism in Arabic documents', *International Journal of Information Technology and Computer Science,* 10**,** pp.80-89.

MARTON, Y., WU, N. and HELLERSTEIN, L. (2005) 'On compression-based text classifcation', In *European Conference on Information Retrieval*, Springer, Berling, Heidelberg, pp. 300-314.

MOORE, E. (2004) 'Sociolinguistic style: A multidimensional resource for shared identity creation', *The Canadian Journal of Linguistics*, 49 (3/4), pp. 375-396.

MOSTELLER, F. & WALLACE, D. L. (1963) 'Inference in an Authorship Problem', *Journal of the American Statistical Association***,** pp.275-309.

MURRAY, D.E. (1990) 'CMC', *English Today,* 6(3), pp.42-46.

NEAL, T., SUNDARARAJAN, K., FATIMA, A., YAN, Y., XIANG, Y. & WOODARD, D. (2017) 'Surveying stylometry techniques and applications', *ACM Computing Surveys (CSUR),* 50**,** pp.1-36.

NEVO, J. (1998) 'Religion and national identity in Saudi Arabia', *Middle Eastern Studies,* 34**,** pp.34-53.

NINI, A. & GRANT, T. (2013) 'Bridging the gap between stylistic and cognitive approaches to authorship analysis using Systemic Functional Linguistics and multidimensional analysis', *International Journal of Speech, Language & the Law,* 20(2), pp.173-202.

NIRKHI, S., DHARASKAR, R. & THAKARE, V. (2016) 'Authorship verification of online messages for forensic investigation', *Procedia Computer Science,* 78**,** pp.640-645.

NIRKHI, S. & DHARASKAR, R. V. (2013) Comparative study of authorship identification techniques for cyber forensics analysis. *arXiv preprint arXiv:1401.6118*.

O'KEEFFE, A., MCCARTHY, M. and CARTER, R. (2007) *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

OMAR, A. and ALDAWSARI, B.D. (2019) 'Towards a linguistic stylometric model for the authorship detection in cybercrime investigations', *International Journal of English Linguistics*, 9(5), pp.182-192.

OUAMOUR, S. & SAYOUD, H. (2012) 'Authorship attribution of ancient texts written by ten arabic travelers using a smo-svm classifier', In International Conference on Communications and Information Technology (ICCIT), 2012. IEEE, pp.44-47.

PAGE, R. (2017) 'Ethics revisited: Rights, responsibilities and relationships in online research', *Applied Linguistics Review,* 8**,** pp.315-320.

PENG, J., CHOO, K.-K. R. & ASHMAN, H. (2016) 'Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles', *Journal of Network and Computer Applications,* 70**,** pp.171-182.

POSADAS-DURÁN, J.-P., GÓMEZ-ADORNO, H., SIDOROV, G., BATYRSHIN, I., PINTO, D. & CHANONA-HERNÁNDEZ, L. (2017) 'Application of the distributed document representation in the authorship attribution task for small corpora', *Soft Computing,* 21**,** pp.627-639.

POTTER, J. & WETHERELL, M. (1987) *Discourse and social psychology: Beyond attitudes and behaviour*, Sage.

POWERS, D. M. (1998) 'Applications and explanations of Zipf's law', In Powers, D.M. (ed.) *New methods in language processing and computational natural language learning,* pp.151-160.

PROCHAZKA, T. (1988) *Saudi Arabian Dialects* (Vol. 8). Routledge.

RAZA, A. A., ATHAR, A. & NADEEM, S. (2009) 'N-gram based authorship attribution in Urdu poetry' In *Proceedings of the Conference on Language and Technology,* pp.88-93.

REPPEN, R. (2010) Building a corpus: what are the basics. *The Routledge handbook of Corpus Linguistics*. London: Routledge, pp.31-38.

RICO-SULAYES, A. (2011) 'Statistical authorship attribution of Mexican drug traficking online forum posts', *International Journal of Speech, Language & the Law,* 18(1), pp.53-74.

ROCHA, A., SCHEIRER, W. J., FORSTALL, C. W., CAVALCANTE, T., THEOPHILO, A., SHEN, B., CARVALHO, A. R. & STAMATATOS, E. (2016) 'Authorship attribution for social media forensics', *IEEE transactions on information forensics and security,* 12 (1)**,** pp.5-33.

ROMA, K., JEFFEREY LA, R. & MIGLĖ, L. (2018) 'Stance Taking in Social Media: the Analysis of the Comments About Us Presidential Candidates on Facebook and Twitter', *Verbum,* 9, pp.21-30.

RYBICKI, J., EDER, M., & HOOVER, D. L., (2016) 'Computational stylistics and text analysis', In *Doing Digital Humanities*. Routledge, pp.159-180.

SADAT, F., KAZEMI, F. & FARZINDAR, A. (2014) Automatic identification of arabic language varieties and dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP),* pp.22-27.

SAHA, N., DAS, P. & SAHA, H. N. (2018) 'Authorship attribution of short texts using multi-layer perceptron', *International Journal of Applied Pattern Recognition,* 5**,** pp.251-259.

SAPIR, E. (1927) 'Speech as a personality trait', *American Journal of Sociology,* 32(6), pp.892-905.

SARWAR, R., YU, C., TUNGARE, N., CHITAVISUTTHIVONG, K., SRIRATANAWILAI, S., XU, Y., CHOW, D., RAKTHANMANON, T. & NUTANONG, S. (2018) 'An effective and scalable framework for authorship attribution query processing', *IEEE Access,* 6**,** pp.50030-50048.

SAVOY, J. (2016) 'Estimating the probability of an authorship attribution', *Journal of the Association for Information Science and Technology,* 67**,** pp.1462-1472.

SAYOUD, H., KHENNOUF, S., BENZERROUG, H., HAMADACHE, Z., HADJADJ, H. & OUAMOUR, S. (2017) Automatic Authorship Attribution of Noisy Documents. In *the Thirtieth International Flairs Conference,* pp.202-205.

SCOTT, M. (2020) WordSmith Tools version 8, Stroud: Lexical Analysis Software.

SEARGEANT, P. & TAGG, C. (eds) (2014) *The language of social media: Identity and community on the internet*. Springer.

SEROUSSI, Y., ZUKERMAN, I. & BOHNERT, F. (2011) Authorship attribution with latent Dirichlet allocation. In *Proceedings of the fifteenth conference on computational natural language learning,* pp.181-189.

SEROUSSI, Y., ZUKERMAN, I. & BOHNERT, F. (2014) 'Authorship attribution with topic models', *Computational Linguistics,* 40**,** pp.269-310.

SHOLIKHAH, I. M. (2019) 'Linguistic study of stance-taking in online media', *KnE Social Sciences***,** pp.55–61.

SHRESTHA, P., SIERRA, S., GONZÁLEZ, F. A., MONTES, M., ROSSO, P. & SOLORIO, T. (2017) Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers,* pp.669-674.

SIDOROV, G. (2018) 'Automatic authorship attribution using syllables as classification features', *Rhema. Рема,* (1), pp.62-81.

SIDOROV, G., VELASQUEZ, F., STAMATATOS, E., GELBUKH, A. & CHANONA-HERNÁNDEZ, L. (2014) 'Syntactic n-grams as machine learning features for natural language processing', *Expert Systems with Applications,* 41**,** pp.853-860.

SILVA, R. S., LABOREIRO, G., SARMENTO, L., GRANT, T., OLIVEIRA, E. & MAIA, B. (2011) 'twazn me!!!;('automatic authorship analysis of micro-blogging messages. In *the International Conference on Application of Natural Language to Information Systems*, Springer, pp.161-168.

SMITH, S. S. & SHUY, R. W. (2002) Forensic psycholinguistics: using language analysis for identifying and assessing offenders. *FBI Law Enforcement Bull.,* 71**,** p.16.

SOLAN, L. M. (2012) 'Intuition versus algorithm: The case of forensic authorship attribution', *Journal of Law & Policy,* 21**,** p.551.

SOUSA SILVA, R., LABOREIRO, G., SARMENTO, L., GRANT, T., OLIVEIRA, E. and MAIA, B. (2011) "I didn't mean to steal someone else's words!": a forensic linguistic approach to detecting intentional plagiarism. In *International Conference von Application of Natural Language to Information Systems,* Springer, Berlin, Heidelberg, pp.161-168.

SOUSA-SILVA, R. (2014) Investigating academic plagiarism: A forensic linguistics approach to plagiarism detection. *International Journal for Educational Integrity,* 10 (1), pp.31-41.

SPASSOVA, M.S. and GRANT, T. (2008) Categorizing Spanish written texts by author gender and origin by means of Morpho-Syntactic Trigrams: some observations on method's feasibility of application for linguistic profiling. Curriculum, Language and the Law Inter-University Centre, Dubrovnik: University of Zagreb.

SRINIVASAN, L. & NALINI, C. (2019) 'An improved framework for authorship identification in online messages', *Cluster Computing,* 22**,** pp.12101-12110.

STAMATATOS, E. (2009) 'A survey of modern authorship attribution methods', *Journal of the American Society for information Science and Technology,* 60**,** pp.538-556.

STAMATATOS, E. (2017) Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers,* pp.1138-1149.

STAMATATOS, E., FAKOTAKIS, N. & KOKKINAKIS, G. (1999) Automatic authorship attribution. In *the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pp.158-164.

SWAIN, S., MISHRA, G. & SINDHU, C. (2017) Recent approaches on authorship attribution techniques—An overview. In *International Conference of Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, pp.557-566.

SWEEL, A.I.A. (1992) 'Some aspects of Najdi Arabic phonology: Part II', *Zeitschrift für Arabische Linguistik*, (24), pp.82-90.

THURNER, S., HANEL, R., LIU, B. & COROMINAS-MURTRA, B. (2015) 'Understanding Zipf's law of word frequencies through sample-space collapse in sentence formation', *Journal of the Royal Society Interface,* 12(108)**,** p.20150330.

TOGNINI-BONELLI, E. 2001. *Corpus linguistics at work*, John Benjamins Publishing.

TURELL, M. T. 2010. 'The use of textual, grammatical and sociolinguistic evidence in forensic text comparison', *International Journal of Speech, Language & the Law,* 17(2), pp.211-250.

TURELL, M. T. & ROSSO, P. Computational approaches to plagiarism detection and authorship attribution in real forensic cases. IAFL Porto, 2012 Porto, Portugal. pp.19-30.

TURELL, M. T. & GAVALDA, N. (2013) Towards an Index of Idiolectal Similitude (Or Distance) In Forensic Authorship Analysis. *Journal of Law and Policy,* 21(2)**,** p.10.

VERSTEEGH, K. (2014) *Arabic language*. Edinburgh University Press.

WANG, B. & FENG, D. (2018) 'A corpus-based study of stance-taking as seen from critical points in interpreted political discourse', *Perspectives,* 26**,** pp.246-260.

WATSON, J. C. (2002) *The phonology and morphology of Arabic*. Oxford University Press on Demand.

WEHR, H., COWAN, J. M., & WEHR, H. (1966). *A dictionary of modern written Arabic*. Ithaca, Cornell University Press.

WEINREICH, U., LABOV, W. and HERZOG, M. (1968) *Empirical foundations for a theory of language change*. University of Texas Press.

WRIGHT, D. (2014) *Stylistics versus Statistics: A corpus linguistic approach to combining techniques in forensic authorship analysis using Enron emails.* University of Leeds.

WRIGHT, D. (2017) 'Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem', *International Journal of Corpus Linguistics,* 22**,** pp.212-241.

ZAPPAVIGNA, M. (2011) 'Ambient affiliation: A linguistic perspective on Twitter', *New media & society,* 13**,** pp.788-806.

ZAPPAVIGNA, M. (2014) 'Enacting identity in microblogging through ambient affiliation', *Discourse & Communication,* 8**,** pp.209-228.

ZHANG, C., WU, X., NIU, Z. & DING, W. (2014) 'Authorship identification from unstructured texts', *Knowledge-Based Systems,* 66**,** pp.99-111.

ZHANG, R., HU, Z., GUO, H. & MAO, Y. (2018) Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp.2742-2753.

ZHENG, R., LI, J., CHEN, H. & HUANG, Z. (2006) 'A framework for authorship identification of online messages: Writing-style features and classification techniques’, *Journal of the American society for information science and technology,* 57**,** pp.378-393.

List of Abbreviations

AA Authorship Attribution

CDA Critical Discourse Analysis

CMD Computer Mediated Discourse

CL Corpus Linguistics

FL Forensic Linguistics

FS Forensic Stylistics

MSA Modern Standard Arabic

NA Najdi Arabic

NASCoT Najdi Arabic Specialised Corpus of Tweets

ReCNAT Reference Corpus of Najdi Arabic Tweets

NB Naïve Bayes

NLP Natural Language Processing

SVM Support Vector Machine