MORE THAN WORDS: A MULTI-DIMENSIONAL ANALYSIS OF ISLAMIC STATE LANGUAGE

Brian K. Ladd

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Linguistics.

Chapel Hill
2020

Approved by:

J. Michael Terry

Cori Dauber

David Mora-Marin

ABSTRACT

Brian K. Ladd: More than Words: A Multi-Dimensional Analysis of Islamic State Language
(Under the direction of Dr. J. Michael Terry)

In this thesis I will provide an exploratory study applying Multi-Dimensional analysis to the realm of Islamic State language. Systematic, observable variables that constitute IS communications can be identified through a built model. The idea is to make more of the type of information that a human analyst relies on available in the automated pre-processing and parsing phases, the types of information that they are looking for is something that a system like Bag-of-Words is missing.

The corpus studied was drawn from the Islamic State newsletter al-naba'. I collected five issues from January 2020 – February 2020. From those issues the editorial sections were focused on specifically. Two paragraphs were chosen from each editorial at random. Each paragraph was translated and coded for sixteen linguistic features, including the presence or absence of religious or political nouns and verbs. I ran a correlation matrix as well as a Factor Analysis and qualitatively analyzed each for the communicative functions.

The statistical analysis revealed that overall, the Islamic State uses significantly more political speech than religious. However, in longer sentences, political and religious speech is more likely to occur.

The qualitative analysis found that in longer sentences, religious and political speech are in service to each other. There is a political aim to the sentence punctuated with a religious admonition statement to reinforce the political goal.

I hypothesize that MDA can determine, via a seeded exploration method, a distinct IS register among the newsletters. This thesis does not fully realize this goal; however, it does provide support for the concept by identifying patterns that may very well help to determine an IS register.

ACKNOWLEDGEMENTS

First, I would like to thank my thesis advisor, Dr. J. Michael Terry of the Linguistics Department at the University of North Carolina at Chapel Hill. Dr. Terry entertained my ideas for several years before I even applied to the graduate program. Our long conversations have crystalized into this thesis. I consider Dr. Terry a mentor and friend. Thank you for all of the support.

Next, I would like to thank my committee members, Dr. David Mora-Marin of the Linguistics Department at the University of North Carolina at Chapel Hill, and Dr. Cori Dauber of the Department of Communication at the University of North Carolina at Chapel Hill. Both of whom I consider mentors. I have taken more courses with Dr. Mora-Marin than any other professor in the linguistics department. I have learned so much from him not just about Linguistics but also what it means to be a professor that cares deeply about his students and their success. Likewise, I consider Dr. Dauber as an inspiration. She is someone who cares deeply about those on her team. She allowed me to take a seat at the table on her research team and for that I will be forever grateful. I have learned so much about what it means to be a member of a research team and what that team can do when there is a steady hand at the wheel. Thank you both for being wonderful examples.

I would also like to thank my family for all of their love and support. My mother and father have especially been strong supports during this time. It has not always been easy, but they have always encouraged me to push forward. I would also like to thank my daughter, Cosette, for

being a ceaseless optimist and have the wide-eyed wonder of a budding scientist. Thank you for your support.

Last but by no means least, an enormous thanks to my partner Mara Howard-Williams. Mara has been a wellspring of love and support. From reading drafts, looking over data with me, to reminding me its ok to take a break every once in a while. Mara has been there for me. I simply could not have completed this thesis without her. Thank you for your love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

“If you **know the enemy** and know yourself, you need not fear the result of a hundred battles.”

-Sun Tzu

1.1. SETTING THE STAGE

On a warm summer day, two men climb to the top of a mid-size building in downtown Charlotte, North Carolina. One places three Avtomat Kalashnikova (AK) 47s with fully loaded banana clips around the edge of the building. He then takes a prone position on the ledge facing a major street with a Russian made Stronskiy SV-98 sniper rifle. His associate sets another AK-47 to his side as he kneels to assemble a Bazalt RPG-7 rocket propelled grenade launcher. Without uttering a word, the grenadier shoulders the RPG-7 and fires at the tallest building in Charlotte’s Uptown, the Bank of America tower. The side of the corporate bank tower explodes between the 47th and 55th floors. Flames lick the raw edges as glass and steel wind down its side. People flee in a panic from the blast area, where they are promptly targeted and shot by the sniper.

The above scenario demonstrates a hypothetical, but entirely plausible terrorist attack committed on United States soil. Immediately following a terrorist event, questions naturally abound. Some of these questions are merely rhetorical: How could this happen in America again? Some are political: Which policy X implemented by political party Y failed, and why? Some are critical to the immediate security of the nation. Questions like:

- Who are these specific terrorists?
- Are they lone actors or part of a larger group?
- If they are part of a group which one?
- Can we expect further imminent attacks*?*

Recognizing the importance of such questions, this thesis aims to answer a different, more limited, one. It asks can Multi-Dimensional Analysis, a computational tool often used in sociolinguist research, be used to extract statistical regularities from internal terrorist documents, and can those regularities be analyzed in terms of their communicative function. In demonstrating that the answer to both of these questions is yes, the thesis serves as first step in much larger project, the goal of which is generating a profile that will allow for automated identification of the group-level authorship and intent of suspected terrorist texts.

To demonstrate the process and utility of the method, let us extend the hypothetical. After the terrorist attack investigators find a car with plans for the attack and a computer in a backpack. The computer is taken to an agency that can break the encryption on the hard drive to obtain at the information within. On the computer the agency finds several emails all in Arabic, but without any identifying markers of what group the attackers are from, such as Islamic State or Al-Qaeda. Investigators then pre-process the texts to prepare for parsing by converting the text from Arabic to computer readable code such as ASCII, as well as conducting part-of-speech tagging. The pre-processed texts are then parsed using the current state-of-the-art method, Bag-of-Words (BoW) (described below). Bag-of-Words analysis aims to determine the likelihood of whether a text is terrorist related. A finding that the writing is likely terrorist in nature flags the writing for review by a trained human analyst. The human analyst then confirms or disconfirms the findings of the BoW analysis. Once the text is identified and deemed valuable this is considered intelligence and can be sent to the decision makers who will consider whether it is actionable or not.

The above described analysis is an exercise in Forensic Linguistics. In an event like the hypothetical situation, forensic linguists employ both human analysts and computational methods to answers two basic questions: 1) what, if any, is the group affiliation of a terrorist actor, and 2) what is the actors' intent? This thesis explores the use of Multi-Dimensional Analysis to extract language patterns from Islamic State documents that can potentially fuel an expert system designed to overcome the limitations of current computational techniques. .

In sociolinguistics, MDA has identified specific and distinct registers, notably in the language of academics. A register is "a [language] variety associated with a particular situation of use (including a particular communicative purpose)." (Biber & Conrad, 2019). This thesis adapts MDA's use to the forensic context.

The prospective utility of this tool could be seen most prominently in the aftermath to the very real and the deadly efficient 2019 Easter bombings in Sri Lanka. Investigators needed expedient answers to the key questions above as members of the *Jamā'at at-Tawḥīd al-Waṭanīyah*, a local Jihadist movement, coordinated eight attacks over several hours in several cities killing and wounding hundreds (*Terror in Sri Lanka*, n.d.) According to counterterrorism experts Bolz, Dudonis and Schulz (2011), "terrorism is [typically] a group-based activity in which several group members carry out specific tasks, which makes them vulnerable to detection." Coordination such as was seen in these attacks also means communication. Communication often leaves behind forensic texts. The rapid and accurate processing of forensic texts is crucial when the suspects are still at large when the suspect's affiliation and lethal capacity remains unknown. With analysis of collected texts, governments could more efficiently and effectively stop these actors and these plots.

This work is important for two reasons, first, there is a gap in the literature as to the application of computational forensic linguistic techniques to issues of terrorism. Second, there is currently a lack of capability. Natural Language Processing (NLP) has been used in these cases. NLP is an area of linguistic concerned with the interaction of computers and natural human language (Indurkhya & Damerau, 2010). Currently, the Natural Language Processing (NLP) tools that have been used to address terrorist language are ill-equipped to expediently identify terrorist actors at the group level, and at present do not identify intent.

Although the current state-of-the-art, BoW, provides information useful in identifying authorship, it does so at the word-level only. It counts how many and how often a set of "scary words" are used. The technique is ill-equipped to capture identifying patterns in the use of full sentences or combinations of sentences. MDA helps by analyzing linguistic traits drawn from a variety of linguistic levels: morphological, syntactic and semantic. From these features a pattern is developed that allows analysts to see the similarities and differences in texts under analysis. This thesis is a pilot study that seeks to identify whether MDA is a viable tool to find linguistic traits which may be unique to the Islamic State. This could facilitate building a specific language profile to distinguish Islamic State writings from other known terrorist groups.

Quite plainly, this work is an exercise in linguistic profiling. But profiling does not have to mean something discriminatory, or at least discriminatory in the sense of irrational bias. A goal of this work is to extract language patterns that are tied to communicative function and not just patterns that are tied to categories like race or religion, thereby avoiding language stereotyping that is unproductive, offensive, and dangerous.

1.1.1 Bless Your Heart

Linguistic profiling is something humans naturally do. To illustrate the point, consider the mundane phrase:

(1) Bless your heart.

What do we think we know about a person who says, "bless your heart?" First, we may assume the speaker is from the Southern United States. Some may assume this person is a woman. We may also guess something about their religious affiliation. In my mind, the picture painted here is of my mother. She sang in the choir on Sundays, cooked Southern fried chicken, and said bless your heart oh so sweetly. But I knew what that phrase meant. It wasn't sweet at all. What was she really saying or doing with this utterance? Let's consider a few of the following utterances:

(2) Bless your heart, you do try to fix more things than you break.

(3) Bless your heart, at least you're pretty.

We can see that there is something backhanded underlying the locution and revealing the intent of the speaker to be "catty."

This is rather profound. The analysis that we, as humans, implicitly conduct with each phrase we read or hear involves drawing on personal experience to assign characteristics to a speaker, whether we know the speaker or not. Humans are doing much more work when reading than just parsing words on a page. It's more than the simple parsing of collocations of a set of words {bless, your, heart}. We pragmatically code one short utterance, embedding numerous meanings that go beyond the conventional semantics of the phrase as written. Here, we saw encodings which allow us to glean something of group characteristics for the speaker, and we can infer something like intent from their use of a particular phrase. It is precisely this type of information that a human language analyst utilizes to ascertain certain characteristics from a forensic text, intuiting language data from a human perspective with a specialist's knowledge. The important factor is to bring the types of sociolinguistic information that a human analyst would draw on to "read" a text.

1.1.2 Linguistic Profiling

Ethical implications arise in linguistic profiling. As defined by John Baugh (2005) as the "audio equivalent of 'racial profiling.'" To Baugh's (2005) definition I will add the aspect of writing as well. The idea of profiling someone because they "write like a Muslim" can be just as damning as profiling someone because they "sound Black."

Unequivocally, MDA is linguistic profiling which requires careful consideration so as not to be used for discriminatory purposes. Baugh (2005) studied linguistic profiling as an extension of racial profiling in housing markets. He found that there is a discriminatory practice against those who use African American English (AAE) versus those who use the so-called standard variety. According to Baugh (2005), this varietal difference signals a racial difference to the hearer and it often triggers physical discrimination toward the AAE user.

I take as my charge Baugh's (2005) own challenge, "to have wisdom, patience, and sufficient tolerance of others whose linguistic backgrounds differ substantially from our own." The methods of this thesis could run the risk of enhancing discriminatory practices against communities practicing Jihadism. That risk must be weighed with the reality of Jihadism that affects all communities, *especially* those of Arab and Muslim populations who are disproportionally affected by attacks *and* who are targets of discrimination and often, with more deadly consequences, subject to erroneous military targeting. Being aware of these risks is what will guide the use of MDA in an effort to put a linguistically *and* culturally trained human analyst into the process so that discriminatory linguistic profiling is less likely to take place. This layer of professionalization of the tool mitigates the immediate prospects of discriminatory practices.

Professionalization allows us to move from linguistic profiling to creating a linguistic profile. The difference between the two is that a linguistic profile adheres to a set of standards and responsibilities that linguistic profiling does not. Linguistic profiling is essentially stereotyping. A linguistic profile does not reduce people to flattened stereotypes but attempts to build a full picture based on the aggregate of their linguistic "persona." For example, not everyone who speaks Arabic is a terrorist. That is linguistic profiling. However, if a terrorist speaks Arabic that can be a part of their profile. The buffer between profile and profiling is that professionalism requires a set of standards and responsibilities. These characteristics facilitate a principled extrapolation of common group traits to determine membership within a group, versus building a profile based on a singular stereotypical trait. The basis of these standards and responsibilities must take as their foundation Baugh's charge mentioned above. This combined with sociolinguistic theories such as Communities of Practice (described below) allow for a non-reductive linguistic profile.

1.2 SOCIOLINGUISTIC INFORMATION

Sociolinguistics is the study of the “relationship[s] between language and society” which adds a dimension to this endeavor by “[examining] the way people use language in different social contexts” (Holmes & Wilson, 2017). This thesis assumes that the language of Islamic State will pattern similarly due to shared social contexts, and likewise it should differ in distinct ways from the speech of other communities with different social contexts. This assumption undergirds the computational approach. The sociolinguistic inputs provide the probative value that allows us to contextualize the quantitative findings in a qualitative way. To do this, placing a linguistically trained analyst into the system at an earlier stage affords the ability to analyze results more expediently, but also to remain ethically conscious throughout the endeavor.

Important to this research is the sociolinguistic theory of communities of practice (CoP) developed by Eckert and McConnell-Ginet (1992). Communities of practice provide a strong framework from which to describe group affiliations. A community of practice is, “a group engaged in a shared enterprise, who interact regularly, and who have common attitudes, and values, and a shared repertoire” (Holmes & Wilson, 2017). The CoP models a more fluid set of relationships where members in the community join and leave the community voluntarily. Though, this thesis will show that voluntary is not such a distinct criterion and one must consider other forms of membership in communities. CoP accounts for a sociolinguistic description of linguistic variation, one that does not rely solely on the macro demographic categories of traditional speech communities, such as age, gender, and class. Speech communities are rigid: a white male between the ages of 30-40, does not change demographic categories until he ages out of the bracket. By contrast, in viewing Islamic State as a community of practice allows us to develop a picture of a fluid organization rather than searching for the monolithic linguistic markers of demographic groups. It also minimizes the ethical issues involving profiling.

Language attitudes inform communities of practice. Giles and Marlow (2011) state that, "the impressions one develops about another person's beliefs, capabilities, and social attributes are guided by many social cues, an important one of which is another's speech style." Attitudes about certain speech styles (or more broadly, certain languages) may affect what sort of shared repertoire the community uses. More formally, language attitudes are: "How people feel about different languages and dialects. This generally reflects their views of the people who use them," (Holmes & Wilson, 2017). For Islamic State (IS) as a CoP, the speech styles chosen are part and parcel of their particular media strategy. When communicating to an audience that exists outside of the borders of the "Caliphate" (when it was extant) IS used languages that targeted populations that they wished to persuade to join the fight: English, French, etc. When communicating amongst themselves through internal facing media inside the "Caliphate," or in personal communications they used Arabic.

The choice of Arabic, which continues to be their primary language, is not entirely a practical one. The number of foreign fighters from across the globe, including speakers of Arabic dialects that are not mutually intelligible, precludes Arabic as a lingua franca. However, the linguistic attitudes toward Arabic as the language of Islam and its association with the Holy Qur'an gives the language a preeminence that other languages lack in the so-called Islamic State. Grasping for an Islamic identity means grasping for Arabic as the language of the "Caliphate." This is why I will consider Arabic language texts in this thesis, because it says something specific about the values and beliefs of this particular CoP. The Arabic language texts analyzed in this thesis say something specific about the values and beliefs of IS as a CoP by the use of Arabic in and of itself.

Considering Arabic texts requires contending with the sociolinguistic phenomena of *diglossia*. Diglossia is traditionally conceptualized as a linguistic situation wherein a language is split between, “two varieties … [that] exist side by side in a community” (Ferguson, 1996). There is a High variety, “a variety used for formal functions and in formal contexts in a diglossic situation” (Holmes & Wilson 2017). In Arabic the High (**H**) register is classical or educated Arabic known by its Arabic name *fuṣḥa,* a catch-all term for a continuum of speech from Classical Arabic to the modern standard Arabic used by the media. There is also a Low variety, “the variety used as a colloquial means of everyday communication, and especially in the home” (Holmes & Wilson, 2017). The Low (**L**) register is commonly one of the many local dialects, called by its Arabic name -- *‘amiyah*.

1.3 SHIBBOLETHS

It would be a misconception, however, to claim that this work is merely the search for a *Shibboleth*. The story of the *Shibboleth* presented in the Hebrew Bible has fascinated many. It has been used most often in the search for a single linguistic marker that can categorize and stigmatize an entire people. In the story, the Gileadites would question passersby on how they pronounce the word *shibboleth.* The Gileadites would let one pass if they pronounced the word as the Gileadites did, but if one pronounced the word as *sibboleth* they were named an enemy Ephraimite and killed. Modern equivalents abound, Sri Lanka saw the use of a *Shibboleth* during its civil war, as did Lebanon. The *Shibboleth* test is, in essence, a “performance test” and is often looked to as a “fingerprint” or “DNA” that marks a speaker’s identity (McNamara, 2005).

Language is a marker of a speaker's identity, and, indeed, a *Shibboleth* can be shared by a speech community. However, this is not the length and breadth of what I am looking for here.

A *shibboleth*'s existence can point toward a member of a demographic speech community, such as male speakers of African American English over the age of 75. It is less adept at picking out loose communities of practice that patterns along ideological rather than simple demographic lines. It is harder to find a monolithic linguistic trait that one may characterize. Here a constellation of traits is needed and an understanding of how communities cohere, and the diffusion of language happens in a principled way.

1.4 COMPUTATIONAL METHODS

Recall the flow of information from collection, to preprocessing, parsing, and human analysis. There are computational linguistic tools that attempt to get at the same question of authorship. The current state-of-the-art is the Natural Language Processing (NLP) technique called Bag-of-words (BoW) which uses a statistical *n*-gram model that seeks out collocations of words within a text. Collocations are drawn from a validated list of keywords, such as {bless, your, heart}. In forensic cases of terrorism, keywords are often a bag-of-SCARE-words, for example: {bomb, attack, assassination}. BoW finds the collocations and correlates their cooccurrence to determine individual authorship, group membership, and even hints at intent in the case of scary words like "bomb." (Y. Zhang et al., 2010).

Bag-of-Words has been quite effective in the realm of authorship analysis; however, it has its limitations. The method only considers lexical items outside of their grammatical context, which can lead to much faster identification but searching for collocations of words regardless of the syntax of the words within sentences strips away much of the context that could be vital in establishing authorship at a group level. With the context missing, attempts at understanding the pragmatic content needed to derive intent is impossible. The Holy Grail in NLP is to create a computational method that "reads" like a human does. Many accomplished linguists and computer scientists are working diligently toward achieving this goal. But success in NLP does not solve the contextual problems that remain. As demonstrated above with the example in (1), a human is doing much more when they read than merely calling and parsing words on a page. A human brings in sociolinguistic and socio-cultural information that undergirds the locution and provides an extra layer of meaning that is conveyed pragmatically.

The alternative method is Multi-Dimensional Analysis, developed by Douglas Biber (1988). Initially used to study language data in academic contexts it has since been used in forensic contexts (Nini, 2017). Similar to Bag-of-Words, MDA is a computational corpus-based approach. In contrast to BoW, MDA employs multivariate statistical techniques, "to investigate the quantitative distribution of linguistic features across texts and text varieties," (Biber 1988). Biber states: "these patterns of linguistic variation are complex, with each linguistic feature being distributed in particular ways in accordance with its associated communicative functions." Features are clusters of co-occurring variables formed into structures called dimensions. Dimensions are the motivating factors for variation; they are observable and attributable to group level associations, such as terrorist organizations.

There are three bases as to why the MDA approach can determine terrorist group authorship. First, MDA has proven effective at identifying linguistic features on small corpora with limited text samples, much like our forensic contexts (Biber 1996). Second, MDA takes as its foundation two basic assumptions about groups and their communications. A group that operates regularly in a society as a functional element (e.g., in terms of physical location, marriage patterns, or economic, religious, or other interactional behavior) will tend to develop identifying markers of language structure and language use, different from the language of other social groups; a communication situation that recurs regularly in a society (in terms of participants, setting, communicative functions, and so forth) will tend over time to develop identifying markers of language structure and language use, different from the language of other communication situations. Third, since extremism is not only communicated in English, MDA has been effectively applied to languages with minimal technology support, such as Somali (Biber 1996). These types of languages are called low resource languages, and they include many languages relevant to the study of extremism: Arabic, Urdu, Pashto, etc. all languages that are problematic for current NLP tools.

1.5 PROCEEDING

My goal, therefore, is to provide an exploratory study applying Multi-Dimensional analysis to the realm of Islamic State language. Systematic, observable variables that constitute IS communications can be identified through a built model. The idea is to make more of the type of information that a human analyst relies on available in the automated pre-processing and parsing phases, the types of information that they are looking for is something that a system like Bag-of-Words is missing. Since Forensic trace evidence is not readily available in open source materials, at the time of writing, this thesis uses the IS newsletter al-Naba’ which is produced for and distributed (often by hand in broadsheet format) to an internal IS audience to more closely approximate forensic texts.

This model is only the first step, and future research should take several subsequent steps. First, the elements extracted by the MDA process should be compared against a testbed of other language corpora to examine if these elements could be a consistent pattern. One approach could be to compare a corpus of primary documents from IS, al-Qaeda, and non-violent Salafi groups. I hypothesize that MDA can determine, via a seeded exploration method, a distinct IS register among the newsletters. This thesis does not fully realize this goal; however, it does provide support for the concept by identifying patterns that may very well help to determine an IS register. In doing so these patterns become part of a useful profile. Subsequent work will have to test these, and other potential IS register elements' ability to distinguish between IS writing and writing by other groups. Second, coding of documents should be validated against other coders to assess inter-coder reliability. Both of these approaches are beyond the scope of this thesis, but this thesis proceeds with the understanding of its limitations in scope.

This thesis will proceed in the following manner. First, I will look at the background of Islamic State and the sociolinguistic implications of IS as a community of practice. Second, I will look at forensic computational tools, reviewing the state of the art and introducing Multi-Dimensional analysis. Next, I will introduce my methods and report my results. Finally, I will discuss my results and conclude the thesis.

CHAPTER 2. ISLAMIC STATE AS A COMMUNITY OF PRACTICE

"A language is everything you do."

— Margaret Atwood

2.1. SOCIOLINGUISTICS

Sociolinguistics is foundational to developing a profile of Islamic State language via a Multi-Dimensional method. One question is whether computational forensic linguistic methods can be used on linguistic trace evidence to identify a terrorist actor's group level affiliation? Sociolinguistics provides the tools to get at the group level affiliations.

As mentioned previously, sociolinguistics is the study of "relationship[s] between language and society" (Holmes & Wilson, 2017). In essence, it is as the Margaret Atwood quote above says: it is everything we do. The analysis of forensic linguistic evidence as an addition to sociolinguistics could increase effectiveness, especially when working with linguistic communities unfamiliar to the analyst. Awareness about social demographics is an important element that can help identify the speech of terrorist communities.

The theory of Communities of Practice (CoP) is the cornerstone of this work, building off of CoP developed in the so-called third wave of sociolinguistics. This wave of sociolinguistics moves beyond the monolithic macro-categories such as class, race, and sex to provide a more nuanced approach based on micro-categories that include fluid group associations as an expression of Identity with linguistic markers.

Penelope Eckert (1989) describes CoP in her seminal text “Jocks and Burnouts.” It is from this book that we can draw an example. Consider two children who share all of the same macro-characteristics even the same household. These children may join different cliques at school, and this will be reflected in their language. The same is true of those who voluntarily join the Islamic State. However, we need to be cautious because not everyone who lives under Islamic State rule has joined voluntarily to their cause. Many innocent people are under their purview against their will yet over time they share much of the same language.

2.2. COMMUNITIES OF PRACTICE

The concept of communities of practice was further elucidated by Eckert and McConnell-Ginet (1992) to account for a sociolinguistic description of linguistic variation which does not rely on macro speech communities such as age, gender, and class. A community of practice is defined as “an aggregate of people who come together around mutual engagement in an endeavor…[social] practices emerge in the course of this mutual endeavor” (Eckert and McConnell-Ginet 1992). Speech communities (Gumperz 1968) based on macro distinctions remain important and a relevant variable of study. CoP captures the fluid nature of human association. Through this lens sociolinguists can illuminate, “[a] speaker’s engagement in a matrix of interrelated social practices, the [community of practice] can provide a framework for understanding both the social and the linguistic facets of sociolinguistic variation,” (Meyerhoff, 2002). Essentially, it is important to understand how people naturally float in and out of many different communities. Individuals often “lean into” their social actions in an effort to consciously create and negotiate their social identities and perform these identities for others within different communities of practice. Ohers have their identities subsumed by a community that enforces it upon them.

The question remains, what formally comprises a community of practice? The criteria for what constitute a comminute of practice are adapted from Wegner (1998). These criteria are as follows:

Mutual engagement

- A member must consciously and voluntarily join a community of practice
- The members in the community must share goals with other members
- There is a reciprocal acceptance of membership between members

Jointly negotiated enterprise

- Members cohere for some purpose and are defined by this purpose
- They may not be able to articulate the specifics of this enterprise, but it is thought to be beneficial to the community as a whole

Shared repertoire

- Members are inducted via social learning process (either implicit or explicit)
- The social traits (linguistic or otherwise) develop a form of currency within the community
- The idiom of the community is continually negotiated and changed when the community deems fit.

An important take away from the categories is that communities of practice are permeable. Those individuals who participate in a community of practice must hold to and accept the shared goals of the group, even if they do so under duress. What this definition misses are that there are many shades of grey in group membership. In many cases there are members of a community who are loose connections in the network. They may be a "member" by close association with another group member but care little for the shared enterprise. Or, perhaps, a person may pick up the shared repertoire and work toward the shared enterprise online without ever formally joining a group. This can be seen in many of the recent white supremacist shooters that have arisen out of communication on online message forums such as 4chan.

A community of practice is not defined only by its membership, but also by the practices of those members. Deeper engagement with and usage of the shared repertoire can strengthen loose ties, upgrading someone from a peripheral to a core member (Meyerhoff, 2002). Under this view, languages change as a result of active use of language resources by speakers who are members of a community. According to Jucker and Kopaczyk (2013), "If we want to understand the processes of language change, therefore, we must consider them in the context of the networks of speakers who *use* specific linguistic resources."

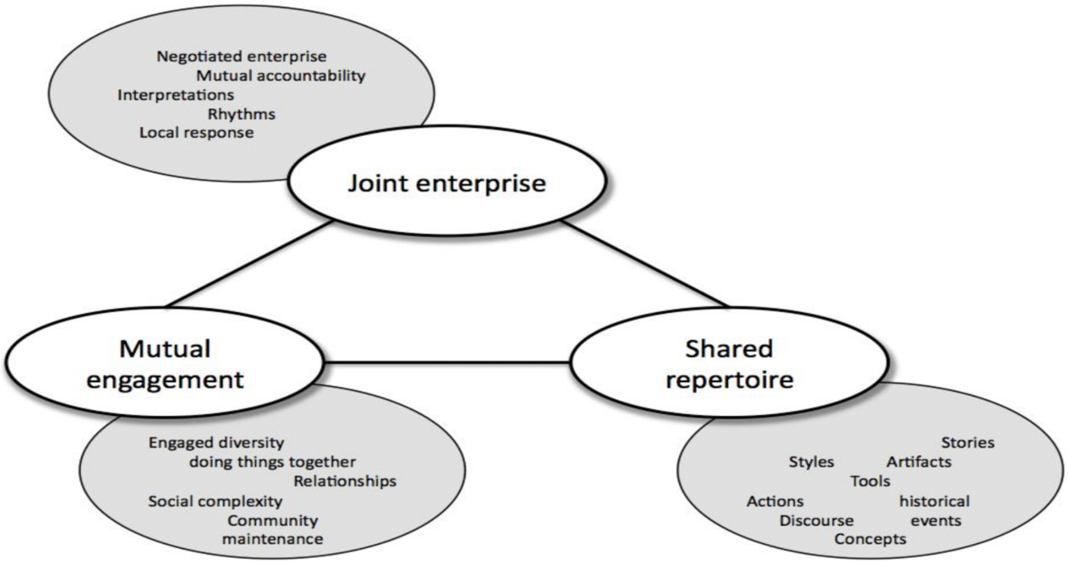


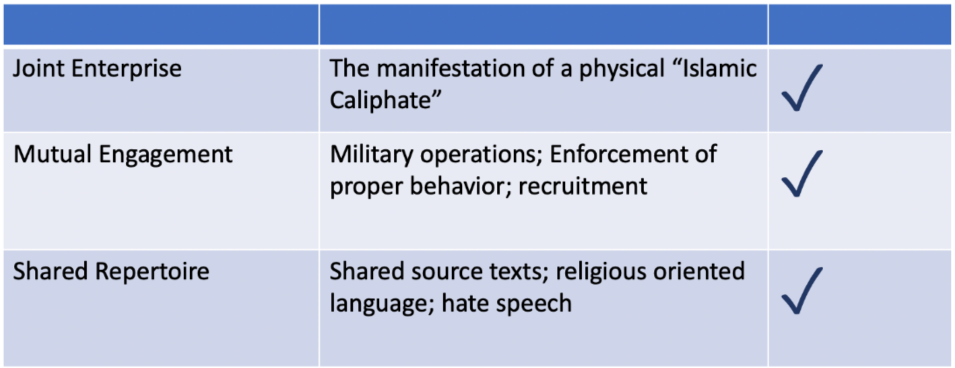
Figure 1: General dimensions of practice (Wenger 1998: 73)

Relevant for how linguistic resources are taken up by a community is how the Islamic State (IS) utilizes Arabic within the context of their "joint enterprise" of seeking to establish a physical "Caliphate." Jihadists are described by Hegghammer (2006) as constituting an "'epistemic community'. . . [that] share[s] common, or relatively similar, ideas and values that range between the fields of theology, politics and strategic matters." Jihadists construct identities that reach back to a halcyon past but operate in a modern context and utilize modern resources while seeking sanction via appeals to the very earliest days of Islamic jurisprudence. In fact, it is this appeal to religious authority from which IS derives their (presumed) authority.

Work by Berger (2018) on extremism quite clearly exposes the coherence of Jihadist organizations as a community of practice. He explains that extremists divide themselves into in-groups and out-groups where the out-group is seen as posing an existential threat that must be met with violence. Any individual who is a member of an eligible in-group -- for IS that is any Sunni Muslim – may consciously and voluntarily self-select to join the extremist group. In Berger's terminology, this form, "an identity collective . . . [is] defined by nation, religion, race, or some other shared trait, interest or concern," (2018). Another critical component of an extremist group is adherence to an ideology. This is the shared repertoire of a community of practice.

Again, we must pause here and urge caution. Though the literature emphasizes a voluntary mutual engagement, for IS this may not be the case. There is a large population in Iraq and Syria that did not voluntarily join IS, they were living in conquered lands. I suggest that these unwilling members of the Islamic State are also members of the community of practice. The shared enterprise is enforced upon them, and in the conduct of that they learn the shared repertoire of the group. This is an explicit social learning rather than the more implicit form undergone by voluntary members of a community of practice. Adhering to the community of practice is then a life and death situation for the conquered populace. This fact introduces many shades of grey into the idea of communities of practice that are novel. In this thesis I proceed with a definition of community of practice that does not include strict voluntary membership.

Berger notes that many scholars define ideology as ideas and concepts, but for him that is not concrete enough. Ideology is contained in texts. Thus, he defines extremist ideology as, "a collection of texts that describe who is part of the in-group, who is part of an out-group, and how the in-group should interact with the out-group. Ideological texts can include a wide range of media types, including books, images, lectures, videos, and even conversations," (2018). Finally, an extremist group seeks at all turns to gain legitimacy for their movement and ideology. In the sense of IS, this is the legitimacy derived from the Qur'an and Hadith about the establishment of a physical Caliphate under their control. According to Berger, this "quest for legitimacy is a key element in many extremist movements," (2018). In the terms of a community of practice, this quest for legitimacy and the violent actions undertaken to achieve it is tantamount to the shared enterprise that Wegner (1998) defines as a crucial component of a community of practice. We can see by this chart how the Islamic State functions under Wegner's (1998) model:



Joint Enterprise	The manifestation of a physical "Islamic Caliphate"	✓
Mutual Engagement	Military operations; Enforcement of proper behavior; recruitment	✓
Shared Repertoire	Shared source texts; religious oriented language; hate speech	✓

Table 1

2.3. BACKGROUND ON THE ISLAMIC STATE

Before we can study the language of Islamic State in any meaningful way, it is important to understand IS' background. There have been many studies released on the group, especially since the announcement of the so-called Caliphate in 2014. Most of these focus upon IS at one point in time, however, the history of the organization is longer and more storied. According to Ingram, Whiteside, and Winter, (2020), there are four clear stages of IS activity. The first stage encompasses the leadership of Abu Musab al-Zarqawi. He had been imprisoned in Jordan for seven years on charges of involvement with terror plots (Thurston, 2015). He spoke at his trial in a way that presaged Islamic State ideologies (Ingram, Whiteside, and Winter 2020). Al-Zarqawi said in court: "This is not about 'bombs, weapons and explosives,' rather, it is a call for unification and a call to religion. . . to help draw people out of disbelief and bring them toward monotheism, for a temporary life towards a permanent one, from injustice to fairness and to security, from hell to delightful gardens."

He was charismatic, leading his militant group to prominence in the years following the US invasion of Iraq. Under al-Zarqawi's leadership, however, the nascent IS (nominally associated with al-Qaeda) was more a cult of personality than a sustainable organization that could survive al-Zarqawi's eventual death (Ingram, Whiteside, and Winter 2020). A ruling council of leaders, the *Shura* council, knew there had to be a shift if the group was to persist. Al-Zarqawi was in fact targeted and killed by a US airstrike in 2006 (Thurston, 2015). The group shifted quickly, proving itself to be nothing if not protean in form.

The five years following al-Zarqawi's death was the second stage of IS. This was a stage of transformation that would give it the internal structures it would need to survive past a single leader and become a movement in its own right. This was a period where the strategic and logistical underpinnings were formed that would allow IS gain in ideological prominence in the minds of Jihadists, and in geographical dominance over large swaths of Iraq and Syria (Ingram, Whiteside, Winter 2020). Where the *Shura* council sought stability in new leadership that would not be too bold or driven by personal charisma as al-Zarqawi had been, they simultaneously sought new strategies to vigorously expand on al-Zarqawi's call for the establishment of a physical "Islamic caliphate" in "liberated" areas against the admonishment of al-Qaeda leadership (Roy 2017). The organization of IS was always predicated in a localized fight with the intent to capture and hold physical territory, to fight the caliphate into existence (Wilkinson, 2019).

This was contra al-Qaeda's focus on a global Jihad that could only come into being after waging war against the West *in* the West. The caliphate was "premature" in Osama bin Laden's opinion. After the death of Osama bin Laden in a US raid on his compound in Abottabad, Pakistan, his successor, Ayman al-Zawahiri, maintained his long-standing admonition that it was “premature” to name a physical caliphate (Roy 2017). For IS, Zawahiri was the wrong leader, and he certainly had all the wrong answers. This ideological rift grew until IS parted ways with al-Qaeda completely.

According to Ingram, Whiteside, and Winter (2020), the third stage of IS, “spans from 2011-2016 and is characterized by transnational expansion and the establishment of the Islamic State caliphate.” Here they elide much important history. Before becoming the proto state known as the Islamic State, it went through several stages of growth. In 2006 the nominally al-Qaeda affiliate in Iraq renamed itself the Islamic State in Iraq (Roy & Schoch, 2017). This was a period involving somewhat of an alliance between groups that sought to make the organization more completely Iraqi (Dauber, personal communication, June 15, 2020). The Islamic State in Iraq rebranded itself as the Islamic State in Iraq and Syria in 2011 following the rise of new leader Abu Bakr Al-Baghdadi. Baghdadi saw the importance of taking advantage of the civil unrest in Syria in 2011 (Laub, 2016). With territorial wins in Eastern and Northern Syria, and their holdings in Iraq, coupled with the withdrawal of US troop from Iraq, ISIS could now achieve its goal of proclaiming a Caliphate.

Its apogee came when Abu Bakr al-Baghdadi ascended the podium in the great al-Nuri mosque in Mosul, Iraq. There he delivered a powerful speech where he announced that the near contiguous territory captured by the Islamic State in Iraq and Syria (ISIS) over the past few years was to be their caliphate (Roy & Schoch, 2017). Within three years the al-Nuri mosque would be destroyed, by IS as they fled the city. A cultural victim of IS’ brutal campaign of terror and a fitting description, in microcosm, of IS’ meteoric rise and fall. As Ingram, Whiteside, and Winter (2020) state: “By mid-2016, the Islamic State’s advances stalled and began to reverse…. It was around this time that then spokesmen Abu Muhammad al-Adnani prepared the movement and its supporters for its imminent decline in (unbeknownst to him) his final address." This down-turn marked the fourth and current (as of this writing) stage of IS where the organization as a proto-state is in decline. IS has sustained campaigns on all sides, seen its physical territory significantly reduced, and its leadership decapitated with the raid against the self-proclaimed "caliph" of the Islamic State, Abu Bakr Al-Baghdadi's hideout in 2019.

Let us not count IS as defeated prematurely. It remains resilient. IS' media wing announced that the *Shura* council named a successor to Abu Bakr al-Baghdadi shortly after his death. The new leader is an IS fighter who now goes by the *nom de guerre* Abu Ibrahim al-Hashimi al-Qurashi (Seldin, 2020). Leadership succession is an important symbol for IS, as well as an operational necessity. Naming a new leader reaffirms the *Shura* council's policies, established in the rebuilding stage of the organization. Their policy is to choose leaders who, 1) stay out of the spotlight, 2) respect the organizational role over that of the self, 3) and prioritize strategic gains (Ingram, Whiteside, and winter 2020). In truth these leaders are expendable, which makes leadership decapitation so difficult to achieve in a meaningful way against IS.

In fact, IS may not be as degraded as recent losses might suggest. The UN has conducted an assessment ascertaining that IS' "administration, propaganda and recruitment" remain unchanged (Lederer, 2020). The resilience of IS and it's continued potential for lethality makes it an important object of study in this thesis.

2.3.1. The IS Brand

The rise of the Islamic State to prominence in the global consciousness was not by accident, "it [was] the outcome of a branding strategy" (Winter, 2015). Its media production network was vast. The varied modes through which it distributed media did not reflect a shotgun approach; nevertheless, the strategy was varied and intricately designed to target specific audience groups. The messaging is planned in great detail, and consists of a wide range of topics that extend beyond the brutality that most in the West have come to associate with IS. According to Winter (2015), “The full spectrum of its political messaging is vast – besides brutality, it is preoccupied with mercy, victimhood, belonging, militarism and, of course, apocalyptic utopianism.” Through its media production centers, IS can control the narrative. It presents the media to a Western audience with a message to emigrate to the Islamic State or if you cannot, then wage Jihad at home.

That message has been appealing to “tens of thousands of people from over 90 countries who have gone to join it thus far” (Winter, 2015). While the foreign fighter is not a new phenomenon, the sheer amount of *muhijreen* – emigrants -- and the alacrity with which they have joined far exceeds those in past wars (Winter, 2015). In fact, it exceeds all other modern jihads combined. The Caliphate also presents media for a localized audience, projecting to the population at home that they can govern efficiently and tend to the needs and concerns of their “citizens.” This insidious form of propaganda presents a sense of “’normal life’ in the Caliphate, the regular depiction of things like markets, service provision and agriculture,” (Winter, 2015). This type of local, often mundane, news is the focus of the newsletter al-Naba,’ which is disseminated by hand to people who live within the caliphate.

IS has come to be known, apart from its sheer brutality, for its sophisticated media strategy. The group crafts its propaganda carefully to reach specific audiences and to achieve specific goals (Sultan, 2016). The amount of propaganda it produced was considerable, and the regularity with which they produced material was surprising for a group so embattled. However, as Sultan (2016) elucidates, “Terrorist organizations now have sophisticated techniques of war and the media is the main tool of it, as it expands rather than reduces their rhetoric.”

Again, IS is masterful at segmenting their target audiences, “The group communicates different things to different people depending on whether they are friends or enemies, and whether they are inside or outside of its claimed territories,” (Mahlouly and Winter, 2018). Thus far the academic literature has largely focused on the media produced for an outside audience. This track is a fair one as we want to study the radicalizing effect of the propaganda. The media center most studied is al-Hayat, “well known for…sharing their messages, videos of violence, songs of their cause and motivational messages which appeal to others to join them in their war against the west (Becker, 2014 as cited in Sultan, 2016). The propaganda is sustained in online communities where it is shared and reposted when it is removed from platforms. As Mahlouly and Winter (2018) state, “the Islamic State’s online ecosystem is nebulous and ever-reconfiguring,” and, “crucially, only a tiny minority of its Internet cheerleaders are likely to engage in terrorism-related activities.” Al-Hayat also produces the English language magazine *Dabiq*, a centerpiece of study for some time now (see Dauber, forthcoming for a complete accounting of *Dabiq* studies). *Dabiq* is directed toward an external audience. Its character is quite distinct from that of internal communications. It is not a good proxy for understanding how IS communicates amongst those already under their control.

For this study, my goal is to study internal communications that could approximate the textual trace evidence of IS private communication. There is a gap in the literature on the study of IS internal communications (Mahlouly and Winter, 2018). Only now are studies being produced on the internal newsletter al-Naba’ (see Winkler 2016; Mahlouly and Winter, 2018)

2.3.2. al-Naba'

Al-Naba' is the Islamic State's Arabic language weekly newsletter. It was designed for an audience that lives and operates inside the caliphate, and while it has an online presence, it is largely distributed offline (Mahlouly and Winter, 2018). The newsletter currently has over 230 issues online hosted on *Jihadology.net*, a research repository for IS and other jihadist primary sources. Al-Naba' has existed since 2010 in various forms (Mahlouly and Winter, 2018). It released its first issue in its current form in 2014 (Winkler 2016). Al-Naba' is 12-16 pages of short news articles, longer editorials, and infographics. It is written mostly in Media Arabic, a form of Modern Standard Arabic used by the mass media in the Arab world. It presents a "more pragmatic" view of the situation on the ground than other publications such as *Rumiyah* and *Dabiq,* online sources for a readership outside of the caliphate (Mahlouly and Winter, 2018). Al-Naba' has a focus on text over images, according to Mahlouly and Winter (2018), who argue "the relative de-emphasis on visual messaging in the internally targeted publication is consistent with the traditional oral culture of the MENA region." That being said, I make no claims that al-Naba' is more widely consumed than IS' video and image-based propaganda. Nor is anyone reading al-Naba' and becoming radicalized. It serves a different purpose as one piece in IS' larger marketing plan.

My reason for choosing al-Naba' is that it serves as a proxy for actual internal forensic texts. A newsletter designed to speak to an internal population brings us closer to understanding how IS will communicate among themselves inside the caliphate than any propaganda that is outward facing. Winter (2015) argued: "propaganda that focuses on 'everyday' life in the 'caliphate' rarely makes it into the mainstream press due to its subject matter; disengaged publics are not interested in Islamic State's administrative efforts." This makes Al-Naba both an understudied resource and a viable proxy for studying their internal communications. At least, that is, until a readily available corpus becomes available parallel to the CTC's Harmony corpus for Al-Qaeda documents.

2.4. BACKGROUND ON ARABIC

Arabic is a Semitic language spoken by roughly 300 million people (Owens, 2013). A "standardized" version of the language, Modern Standard Arabic (MSA), is used as an (arguably) mutually intelligible form across the Arab World. In truth there are no modern-day native speakers of the Classical or Standardized forms of Arabic. Every speaker is native in one of the dialects that are spread across the region and are taught MSA in school. This has developed what sociolinguistic scholars have called a diglossic situation.

2.4.1. Diglossia

Any discussion of the Arabic language must contend with Ferguson's (1996) seminal paper on diglossia. Since that time, Arabic has been held up as an exemplar of the prototypical diglossic situation. Diglossia is traditionally conceptualized as a linguistic situation wherein a language is split between, "two varieties of the language [that] exist side by side in a community" (Ferguson, 1996). The registers are delineated by the prestige associated with them. In Arabic the High (**H**) register is classical or educated Arabic known as *fuṣḥa.* The Low (**L**) register is commonly one of the many local dialects, called *'amiyah*. Because the term diglossia has been tied to the Arabic language here the terms *fuṣḥa* and *'amiyah* will be represented by **H** and **L** respectively.

The **H** and the **L** are said to appear only in designated speech domains where the appearance of the other form is precluded from being uttered in that domain (Ferguson, 1996). In

Arabic the **H** register is reserved for domains such as politics, religion, and education, while the **L** register is used in the in the home and among close companions outside of formal settings. In essence it is a split, and there is a distinction between the two registers. There are several other essential criteria that denote a language in a diglossic state:

- an important and venerable literary tradition
- access to literature which is controlled by a small elite class
- a sufficient span of time in which a language separates out into an H and an L register

Though in recent years this has been challenged (Albirini, 2015), diglossia is still the point from which sociolinguists of Arabic start. For the purposes of this paper, it will be assumed that the use of the **H** and **L** registers are subject to the shared repertoire of the community practice in which they are employed. This is similar to Caton's (1991, cited in Albirni, 2015) assertion that the differing registers are based in linguistic communities. As has been shown, a community of practice negotiates its shared repertoire and new forms can arise independently within a community that is not necessarily diffused more widely.

2.4.2. The Islamic State and the Arabic Language

There is an obvious gap in the literature. Little information has been presented on how IS uses language in the Arabic context. Most studies have focused on *Dabiq*, IS' English language magazine, as the basis for analyses. One might think that studying IS in Arabic is a moot point, with such a plethora of sources of IS propaganda developed in English having been studied so thoroughly (Ingram, 2018).

The organization's success in attracting foreign fighters from non-Arabic speaking countries has been exceptional. So exceptional, that even IS may agree Arabic is unimportant as a criterion for Jihadist fighters as one headline suggests: "[The Islamic State] demands many things of its top commanders, but good Arabic isn't always one of them," (Erard, 2016). Non-Arab fighters have risen to high position within IS, including the current leader Abu Ibrahim al-Hashimi al-Qurashi, who is of Turkmen descent (Seldin, 2020). Perhaps, in its pursuit of foreign fighters, IS has shrugged off one of the traditional features of Islamic identity, its association with *lisan al-arab* -- the Arabic tongue. It is possible, though not likely, especially given the ideological primacy of the language within Salafism.

IS clings to a decidedly Islamist identity which bears strong associations with the Arabic language. The Qur'an spends a great deal of time on the nature of the Arabic language (Owens, 2013). To many Muslims, Arabic is considered sacred, inimitable, and untranslatable. The Qur'an states that the Arabic language itself is revelation. This revelatory Arabic is free of pollution (Suleiman, 2013). Thus, describing an Arabic variety that is not elegant and sacred becomes, if not taboo, at least uncomfortable for many native speakers. Appealing to this association, one of IS' leading religious scholars, Abu Abdullah al-Masri, wrote on the role of Arabic in the life of the foreign fighters -- *muhajireen.* He states:

> . . .[T]he interest in the Arabic language and its use in daily life for the individual is an important matter in the Islamic State as is distancing from vulgar expressions that were

> put forward in society in a well-considered plan to guarantee the forgetting of the Islamic identity for society.

Erard (2016), commenting on the statement of al-Masri, notes that: "the *muhajireen* . . . should cultivate the 'Arabic character' and 'lay aside the foreign identity that bears in its hidden nature hostility to Islam, its culture and roots.'" Because IS proclaims to draw its authority from the Qur'an, members will "often speak in codes and allusions. . .[that] refer to specific traditions and texts of early Islam" (Wood 2015). Even the fighters on the ground will readily quote basic verses of the Qur'an. Thus, to obtain the authority it desires over the Islamic ummah -- the totality of believers -- IS' propaganda machine must employ Arabic when speaking to the in-group. IS spokesperson Abu Muhammad al-Adnani al-Shami set the emblematic language of IS, and "his speech was laced with theological and legal discussion" (Wood, 2015).

Of course, not every Arabic speaker identifies with religion so intimately, but Islam is the majority religion in most Arab speech communities and religion often "interacts with language in complex ways," (Albirni, 2016) one of which is the construction of identity. For centuries, Arabic has served as the vehicle for both religion and politics, especially as it carried the "Islamic mission" in the early 7th and 8th centuries (al-Wer, 1997). This created an environment that fostered the linguistic construction of religious and political identities. Still, in the Middle East, the two are often inextricably intertwined. Arabic of the classical variety has been used by both religious and political groups throughout the years as a point of resistance against colonial encroachment so much so, that, al-Wer (1997) observed:

> one is accused of treason for pointing out the need to re-standardise Arabic to incorporate the linguistic changes of the past twelve centuries, and of ignorance for acknowledging the fact that Arabic speakers from geographically distant regions are often compelled to resort to a European language in order to communicate.

Erard (2016) details the experience of an individual who was chastised for using "Egyptian street Arabic," with a group of conservative *salafi's.* He surmised that these *salafi's,* "have a very strong presumption that speaking high Arabic is a part of being a good Muslim." While that not every Salafi is a Jihadist, their religious ideology flows along similar veins. We can presume, then, that this is the type of attitude about Arabic that IS is drawing on when they produce media in the Arabic language. If this is true it should mean that all of IS media production is in the **H** variety, but that is a question beyond the scope of this thesis, but a valuable pursuit for further research.

CHAPTER 3. COMPUTATIONAL METHODS

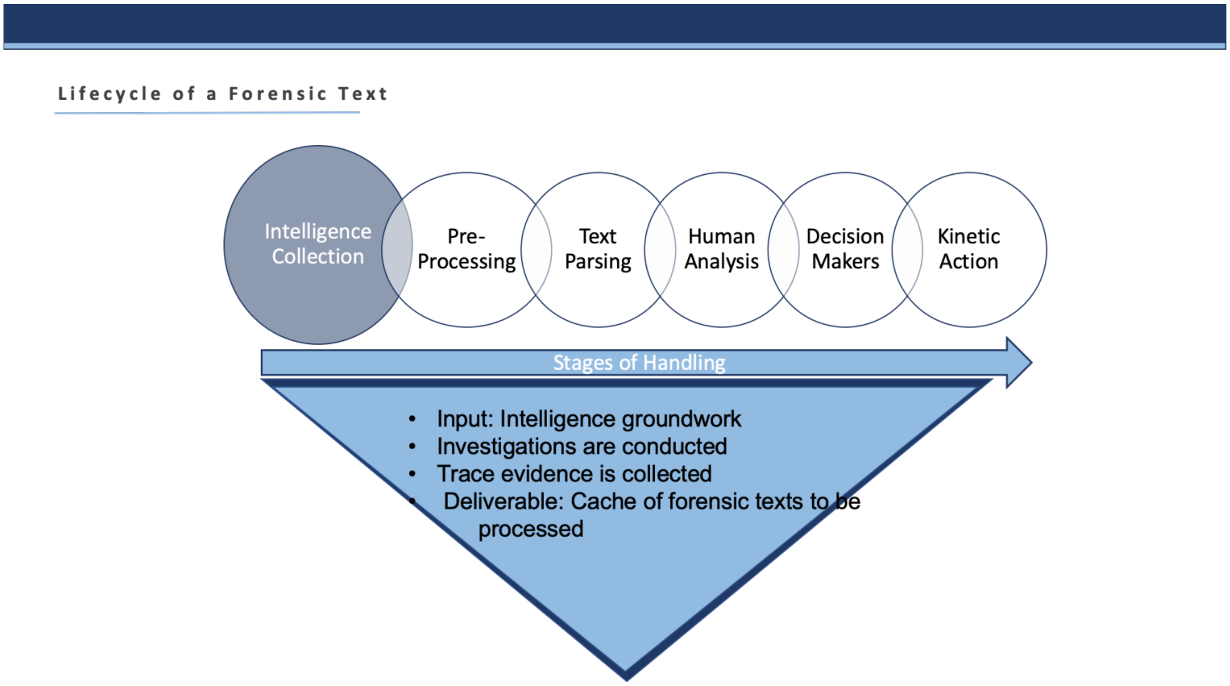
"Whenever I fire a linguist our system performance improves" -Fred Jelnik

3.1. FORENSIC TEXT

Texts, such as those found by the investigators in our hypothetical scenario that began this thesis are by nature what Nini (2017) terms, a malicious forensic text. He defines these as, "a text that is a piece of written evidence in a forensic case that involves threat, abuse, defamation or a combination of the above" (Nini, 2017). These texts present problems for conducting computational analyses because, "[these] kind of texts are neither clearly defined nor thematically unified" (Spranger & Labudde, 2013). This differentiates a forensic corpus from other corpora. The difficulties of working with forensic texts are threefold: First, a forensic text is always going to be found *in situ* in the context of an investigation but *ex situ* from the context from the communicative purpose for which it was produced. We will always find a forensic text outside of the situational characteristics in which it was produced, these will have to be reconstructed. Second, a forensic text will typically be degraded in some way. Audio will not be optimal for analysis, written texts may be only partial, or may only be one part of a larger string of discourse. We are lucky in some instances, for example Osama bin Laden's letters to Mullah Omar which we have in a complete form. Finally, any cache of forensic texts is going to be small comparatively to other corpora. How do we contend with these issues and still have confidence in computational forensic techniques? In the case of the current state-of-the-art, this is asking the wrong question. The tool simply isn’t suited to the task of processing forensic texts with regard to 1) identification at the group level, and 2) intent.

3.2. LIFECYCLE OF A FORENSIC TEXT

The lifecycle of a forensic text begins at the point of intelligence collection. Operators in the field collect a cache of texts as part of an investigation or other kinetic action. This cache may be on encrypted hard drives or they may be loose leaf papers. This cache is then sent to an intelligence organization with the forensic capability to decrypt the data such as the Defense Intelligence Agency (DIA), National Security Agency (NSA), or the Federal Bureau of Investigation (FBI). If the information is encrypted these agencies will use a technique called digital forensics to retrieve the data. The aim of Digital Forensics is defined as the attempt, “to acquire courtroom evidence from digital devices (e.g., servers, personal computers, laptops, mobile devices) that are used in some activity of interest, such as cyber-crime… and physical crime” (X. Zhang & Choo, 2020). Once the data is extracted, our investigators recover the cache of documents that require manual pre-processing, computational parsing under the current-state-of-the-art method, and then human analysis for potential intelligence value.



- Input: Intelligence groundwork
- Investigations are conducted
- Trace evidence is collected
- Deliverable: Cache of forensic texts to be processed

3.2.1. Preprocessing

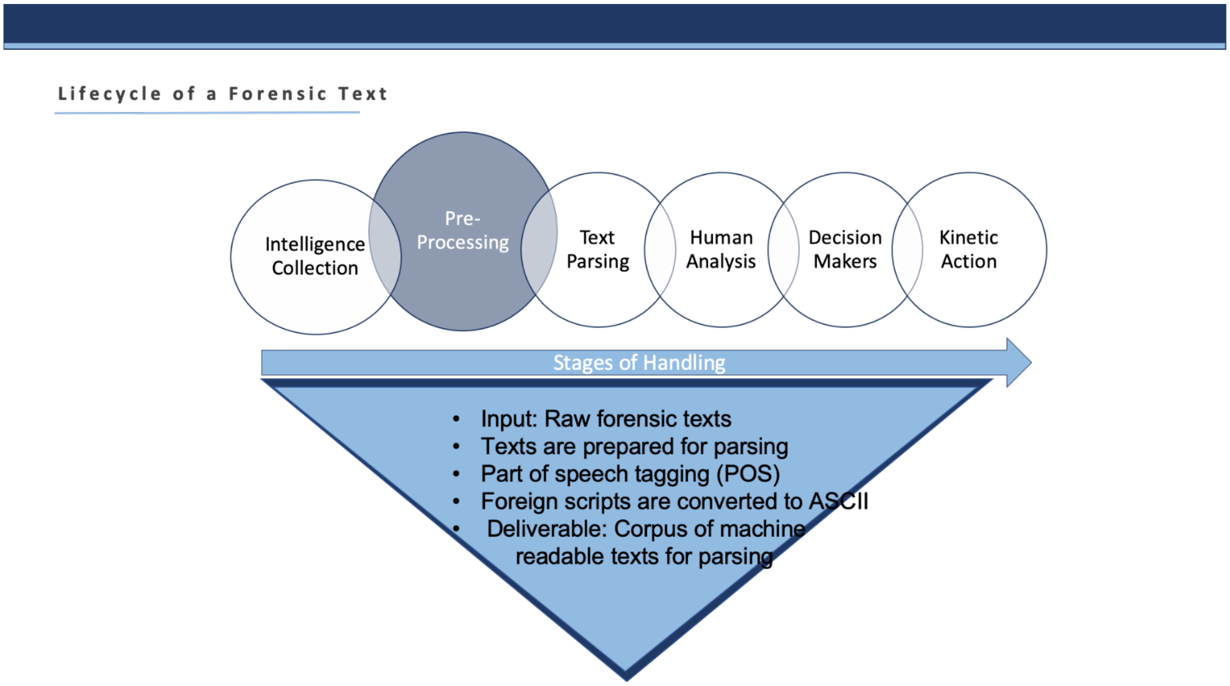
Text preprocessing is the necessary first step before a document can be machine-readable. According to Indurkhya and Damerau (2010), “text preprocessing, [is] the task of converting a raw text file, essentially a sequence of digital bits, into a well-defined sequence of linguistically meaningful units.” This is absolutely essential because the linguistic units identified during preprocessing are used as the basis for other processing steps. There are two steps to preprocessing: document triage, and text segmentation (Indurkhya & Damerau, 2010).

During Document Triage texts are converted from digital files into a set of “well defined text documents” (Indurkhya & Damerau, 2010). This used to be a manual process that was time consuming, however, machine learning techniques have automated the process. A major part of the triage process is to make the documents machine readable. The individual characters must be coded. This is a process, “in which one or more bytes in a file maps to a known character” (Indurkhya & Damerau, 2010). Next algorithms are applied to identify what natural language is used in the document. This can be helped by the character encoding but is not dependent on it.

Finally, the document is subject to text sectioning, which “identifies the actual content within a file while discarding undesirable elements such as images, tables, headers, links, and HTML formatting” (Indurkhya & Damerau, 2010). Stripping these elements leaves a clean document ready for further processing.

The next step in preprocessing is text segmentation. This is “the process of converting a well-defined text corpus into its component words and sentences” (Indurkhya & Damerau, 2010). The document is subjected to tokenization, where a token is each individual instance of a word. Tokenization defines words into normalized tokens. Different forms of words are normalized into a single, canonical form (e.g. Ms., Ms, Miss).

Sentence segmentation comes. According to Indurkhya & Damerau (2010), “sentence segmentation is the process of determining the longer processing units consisting of one or more words.” Sentences are bounded by punctuation which allow the system to know where the boundaries of each sentence lie. This is important to set the boundaries within which the tokens are found. Identifying collocations of tokens in a sentence necessitates knowing the extent of the string.


Lifecycle of a Forensic Text

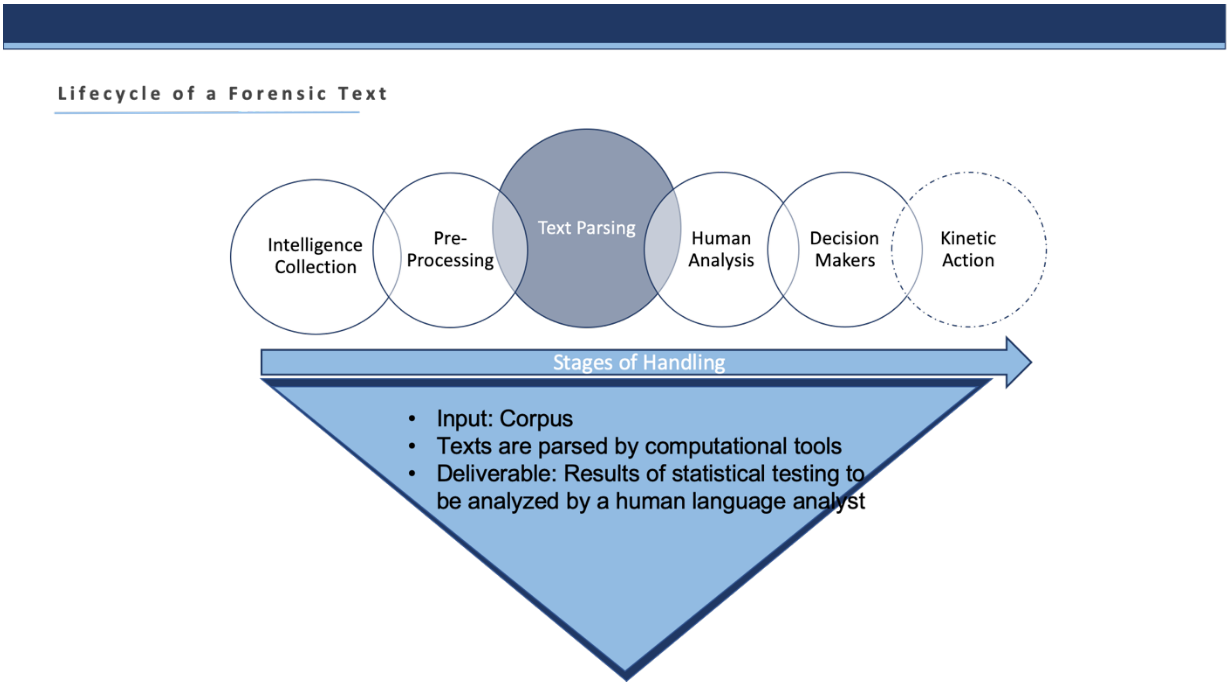
- Input: Raw forensic texts
- Texts are prepared for parsing
- Part of speech tagging (POS)
- Foreign scripts are converted to ASCII
- Deliverable: Corpus of machine readable texts for parsing

3.2.2. Text Parsing

Once preprocessing has been completed a document may then be parsed. One parsing technique is the previously described Bag-of-Words. Any document has the frequency of terms in the document mapped to a "positive value" (Manning et al., 2008). The order in which the terms appear is immaterial to BoW, whatever their frequency is. Thus, according to Manning et al. (2008), "the document, 'Mary is quicker than John". . .is identical to the document 'John is quicker than Mary.'" (Manning et al., 2008). Despite the fact that the context is ignored the model provides the probability of an ordering of the words.

The parsing of BoW attempts to differentiate between two documents by calculating the frequency of words within a vector space. Where documents are viewed as vectors, a collection of objects which can be mathematically operated on (Manning et al., 2008). Algorithms are designed to work on the frequency of terms measuring their importance by providing a weighting system, ignoring non important words (stop words), and accounting for document length. It is a given that, "the more times a term *t* occurs in document *d* the more likely it is that . . .[it] is relevant to the document" (Manning et al., 2008). Terms are weighted based on importance based on a logarithmic formula that balances weight and frequency.

Bag-of-Words is a "two-class classifier" system returning results that are either relevant or not relevant. The system "retrieves the subset of documents which it believes to be relevant" (Manning et al., 2008). However, the accuracy of the system is critical because the stakes are so high that false negatives are a great concern. According to Manning et al. (2008), "normally 99.9% of the documents are in the nonrelevant category." This can lead to misleading results. A BoW algorithm can appear to be running optimally by assigning all documents a label of nonrelevant. In acknowledgement of that, human analysts look at more than just those texts flagged by the system in order to test the system and avoid these false negatives. This is precisely why human analysis is needed.


Lifecycle of a Forensic Text

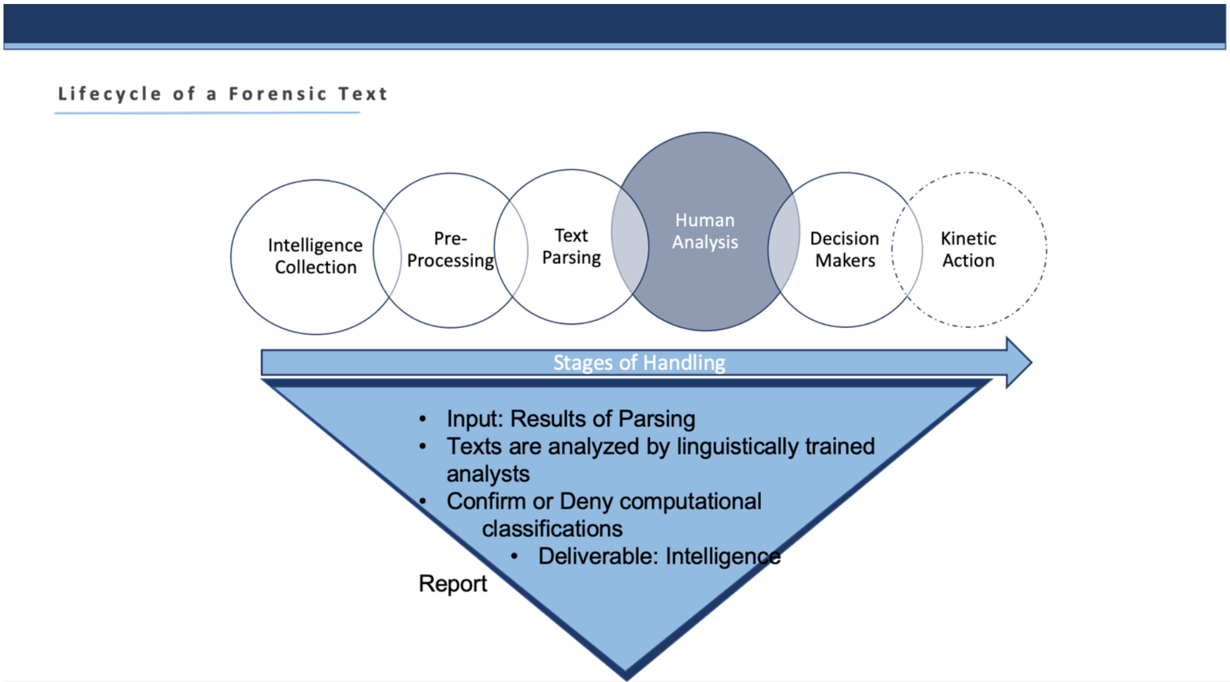
- Input: Corpus
- Texts are parsed by computational tools
- Deliverable: Results of statistical testing to be analyzed by a human language analyst

3.2.3. Human Analysis

The data provided by the BoW model needs to be validated by a human analyst. The analyst can be seen as the “professional consumer” of the data (Indurkhya & Damerau, 2010). The analyst will want to see results, so the “accuracy” of 99.9% of documents being categorized as nonrelevant is unsatisfactory. According to Manning et al (2008), “[analysts] are always going to want to see some documents and can be assumed to have a certain tolerance for seeing some false positives providing that they get some useful information.” To be clear, the human analyst is more tolerant of false positives than false negatives. As seen before false negatives can be quite detrimental.

For our purposes the BoW system would be categorizing documents as either Islamic State related or non-Islamic State related. So, the analyst would need to make calls like determinations of group affiliation. In such cases precision may not be the best measure, but recall may be more helpful. Precision and recall are the measures of the return of true positives, “asking what percentage of the relevant documents have been found and how many false positives have also been returned” (Manning et al., 2008).

Analysts are interested in high recall, or “a non-decreasing function of the number of documents retrieved,” choosing to tolerate a low degree of precision (Manning et al., 2008). The goal would be to develop an algorithm that will provide accuracy with a high recall that reduces the percentage of false positives. Until that is achieved, the analyst will deal with the false positives on the back end, either confirming or denying the system’s classification. Once the documents are classified by the analyst the intelligence received will be sent to the decision makers. They will make the determination whether the intelligence is actionable or not. If it is, there may be a determination for further intelligence gathering operations, or in some cases they may call for kinetic action.

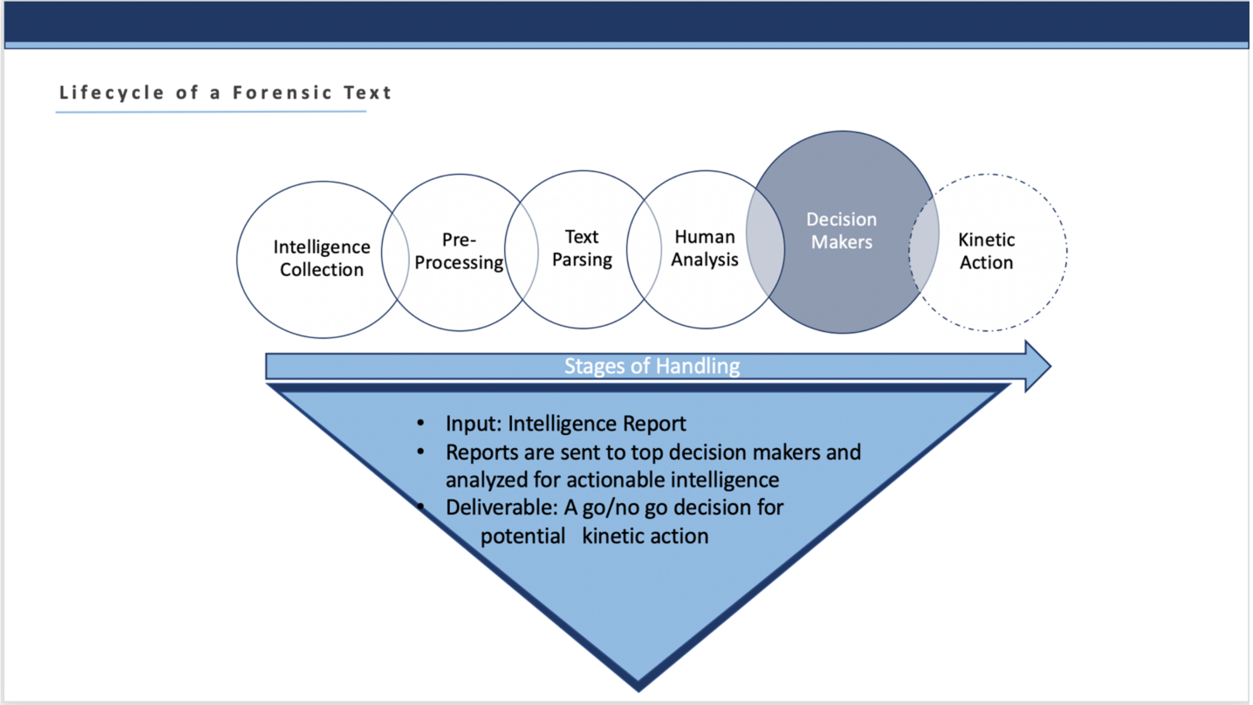


Lifecycle of a Forensic Text

Intelligence Collection
Pre-Processing
Text Parsing
Human Analysis
Decision Makers
Kinetic Action

Stages of Handling

- Input: Results of Parsing
- Texts are analyzed by linguistically trained analysts
- Confirm or Deny computational classifications
- Deliverable: Intelligence Report



Lifecycle of a Forensic Text

Intelligence Collection
Pre-Processing
Text Parsing
Human Analysis
Decision Makers
Kinetic Action

Stages of Handling

- Input: Intelligence Report
- Reports are sent to top decision makers and analyzed for actionable intelligence
- Deliverable: A go/no go decision for potential kinetic action

3.3. MORE ON BAG OF WORDS

For many years, the intelligence community has understood that manually parsing large amounts of textual data is time and resource intensive, in a word, impractical. For maximum efficiency the analysts utilize computational methods known as Natural Language Processing (NLP). The most common method used is statistical *n-gram*, where *n* is the number of collocated words, or Bag-of-Words (BoW) (Nadkarni et al., 2011). That is to say that words in a sentence can be searched for based on their proximity to one another in a string. While theoretically any length of a string of words can be parsed, often only bigrams, two words at a time, are considered. A bigram model is a model that, "predict[s] each word based on the immediately preceding word" (Wallach, n.d.). The BoW method by contrast "reads" the texts and counts the frequency of words co-located in a specific document (Boulis & Ostendorf, 2005). These words are often validated from a set of pre-selected keywords shown to bear significance to the domain under consideration. In the terrorism domain, "scare words" such as "detonate bomb," or "jihad against America" are viable choices of *n*-grams

One limitation of this methodology is that it ignores grammar such as syntactic order. As Le and Mikolov (2014) detail, "[in Bag-of-words] the word order is lost, and thus different sentences can have exactly the same representation, as long as the same words are used." Thus, the algorithm will flag "America against jihad" as confidently as it will "jihad against America." However, there is a complex layered process of determining the meaning of a sentence. There is the meaning of the words themselves (lexical meaning), the meaning of the words put together in the sentence (phrase and sentential) meaning, and the meaning of the sentences in context (pragmatic meaning). Pragmatic meaning is highly important in this case. In fact, semantics and discourse-level understanding is a key goal of NLP but it is very poorly understood (Nadkarni et al., 2011). The difficulty can further be seen in the following example. Consider that we have two forensic texts, found in possession of a suspected terrorist. Utilizing the current methods, they would be processed for frequency of words related to the domain of jihadism. I have selected an arbitrary word set: {Allah, bomb, mujahideen, kill, witness, victory, murder, loyalty, support, infidels}

(1) May **Allah** the Almighty bear **witness** that we will not let you down until we achieve **victory** [sic] or we taste what Hamza bin Abdul Mutallib (may **Allah** be pleased with him) tasted.

(2) **Allah** the Exalted, offer your **loyalty** and **support** to Him and He will grant you **victory**; obey Him and He will compensate you.

Under this technique, both of the statements would be identified as equally Jihadist. The problem with this classification is that (1) is an excerpt from an announcement by the Jihadist militant group *Jamaat Nusrat Al-Islam Wa al-Muslimeen* (*Jihadology,net*). However, (2) is an excerpt from the *Aqeedah*, or Islamic Creed, taken from a book edited by prominent anti-extremist Imam Abdul Malik Mujahid (2014).

The provenance of (1) would be analyzed in the human analysis portion of our lifecycle. For the purpose of capturing Jihadist ideology, the analyst would focus on the name Hamza bin Abdul Mutallib. This reference is to the martyrdom of Hamza, who wielded two swords and swore that he was a lion of Allah. It references both his steadfast resolve to fight, but also his death at the hands of traitors who slit his throat and eviscerated him while his back was turned (al-islam.org). There is no dearth of context to unpack in an appeal to a single name.

This sort of contextual appeal is not unique to this limited example. The current method would be hard pressed to parse the following excerpt from a captured letter from Al-Qaeda second in command (at the time, now leader) Ayman al-Zawahiri to Abu Musab al-Zaraqwi former leader of al-Qaeda in Iraq, the precursor to ISIS:

> (3) You know well -what I am mentioning to you- that many of the most learned ulema of Islam such as Izz Bin Abdul Salam, al-Nawawi, and Ibn Hajar - may **Allah** have mercy on them - were Ashari. And many of the most eminent **jihadists**, whom the Umma resolved unanimously to praise such as Nur al-Din Bin Zanki and Salahal-Din al-Ayyubi - were Ashari. The **mujahedeen** sultans who came after them - who didn't reach their level - whom the ulema and the historians lauded such as Sayf al-Din Qatz, Rukn al-Din Baybars, al-Nasir Muhammad Bin-Qallawun, and Muhammad al-Fatih, were Ashari or Matridi.
>
> (CTC Harmony Corpus)

The problem becomes clear. Our set of “scare” words only receives three hits with the words “Allah,” “jihadists” and “mujahedeen.” Yet, here is one of the top leaders of Al-Qaeda instructing his subordinate via an appeal to a litany of Islamic scholars, some with an extremist bent and some without. He states, “you know well what I am mentioning to you…” as if to say everything here is understood to be common ground in the conversation. Each name is didactic, meant to teach an entire extra-linguistic lesson with its pragmatic invocation. There is a wealth of insight packed into this one paragraph that would be absent from a word frequency analysis, but easily picked up by a human analyst.

Clearly the current state-of-the-art has a hard time parsing documents whose meaning is heavily dependent upon context. It runs a high risk of misattributing innocuous texts as extremist ones, and vice versa. The promise of the current state-of-the-art is to have autonomous parsing, where the system "reads like a human." We are nowhere near that. Even if we did develop an autonomous parsing capability the problem described above would remain: the machine can't read the text like an analyst or glean the context.

The MDA approach must be able to supplement the BoW model, especially when considering texts with national security implications. I hypothesize that as, if not more important, as the frequency of scare words is the particular way in which they appeal to authority. It is not the invocation of Allah that concerns the human analyst, but how Allah is invoked. The better expert systems can become at this discernment, the closer we are to parsing more effectively.

While the BoW method provides the correlations between *n-gram* collocations, the human language analyst must interpret these results. The success of the method relies on human expertise and intuition to confirm or disconfirm the tool's classification of texts as terror related or non-terror related. Use of the BoW method is a powerful force multiplier that extends the analyst's reach. However, no automated method alone will replace careful and close reading of texts because the automated tool only looks within, what the legal profession calls, the four walls of the page.

However, I advise caution because there is no reason that we can't improve upon current methods to add more into the automated part of the process. The analyst remains important, though, because she brings to the text a human intuition of how human language functions. An algorithm is not skilled in reading like a human; at least not yet. The analyst must validate the classification of the text by noticing implicatures and pragmatic appeals, and correctly interpreting nuanced idiomatic and metaphorical meaning. Furthermore, the analyst who is both linguistically trained and who has a deep domain expertise will be able to glean a deeper meaning and identify coded language and inter-textual relationships associated with the documents that underpin group ideology. One of the problems with this process is that the analyst's classification is not fed back into the system, so the number of false positives and false negatives generated by the BoW method remains unknown.

There are several layers of complexity with this text attribution process, and it's important to lay them out clearly. First, extracting meaning from a text is a multi-layered endeavor. You have both word meaning, and phrase and sentence meaning (which requires syntax). Beyond that there is the pragmatic meaning of the sentence, and sentences are placed in larger discourse structures. Next, as we have seen, BoW works almost exclusively at the level of word meaning. Though, there is some syntax required for concatenation of *n*-grams.

There is no reason to expect that these current limitations cannot be overcome. In fact, attempts to address these limitations have led to a rather verdant field of research with many new methods being studied as I write. NLP Projects such as TIDES, REFLEX LCTL, Babel and LORELEI hold great potential for breakthroughs to address these issues. However, even if these projects achieve the Holy Grail of efficient and accurate automated parsing of low resource languages, the problem remains of limited parsing capabilities when it comes to encoded meaning.

There is a lot of effort being put into getting computers to interpret these other higher orders of linguistic function. Most of this work is in sentence level processing, though work in complex morphology is taking place as well. There are real impediments to this, which is why I call it the Holy Grail. We may or may not get there.

Even now, we should be capable of improving our current systems by building the higher level into the systems early on. That is what I am attempting to do with MDA. Drawing from the resources I have already discussed, MDA will allow us greater latitude in understanding these higher order systems, finding complex patterns of use, and attributing qualitative meaning to those patterns.

3.4. MULTI-DIMENSIONAL ANALYSIS

We have seen that the current state-of-the-art in Natural Language Processing (NLP) needs a supplemental approach to address 1) understanding the group affiliation of terrorist authors, 2) gleaning something about their intent. I am not suggesting that this is a discourse analysis, or that it is expressly going to derive a formal pragmatic analysis, but what Multi-Dimensional Analysis allows us to do is cast a wider net than just the lexical approach of Bag-of-Words. MDA takes into account the complex syntactic and semantic patterns that we can view as independent levels that interact in order to serve a communicative function. There are many different approaches to getting at this higher order information in the system, but I have chosen MDA. This approach incorporates some of these other layers of meaning, albeit indirectly.

MDA was developed by Douglas Biber (1988). It is a methodology used to identify variation between two or more registers. The use of MDA is motivated by two theoretical assumptions:

- Generalizations concerning register variation in a language must be based on analysis of the full range of spoken and written contexts.

- No single linguistic parameter is adequate in itself to capture the range of similarities and differences among spoken and written registers.

(Biber, 1995)

Before the development of MDA, most inclinations about register variation were described in qualitative terms. To address what he saw as a gap in the literature, Biber (1988) employed a statistical technique called Factor Analysis. A factor analysis is, "a multivariate [statistical] technique that analyzes linguistic co-occurrence patterns to reduce a large set of variables (individual linguistic features) down to a much smaller set of underlying linguistic parameters," (Egbert & Biber, 2018). Because, "features do not randomly co-occur in texts," the strength of co-occurring patterns can be informative by revealing the critical features that drives their related functional use, otherwise known as dimensions. (Biber 1988). Biber (1995), defines dimensions as, "distinct groupings of linguistic features that co-occur frequently in texts… identified statistically by a factor analysis, and they are subsequently interpreted in terms of the communicative functions shared by the co-occurring features." It is with respect to analyzing dimensions that the hybrid nature of MDA appears. The communicative purposes of the dimensions must be analyzed qualitatively to understand the function of each dimension and how it influences the variation within and between registers.

Egbert and Biber (2018) lay out this process succinctly, stating that MDA consists of three components:

- Each dimension is defined statistically by a distinct set of co-occurring linguistic features
- There are different patterns of register variation associated with each dimension
- Each dimension is associated with particular functions.

Since MDA has been developed, many hundreds of MDA studies have been done on register variation. Most notably, Biber (2006) has applied MDA to understand the differing registers in academic language. In his study, Biber found that overall university language is that professors use in the classroom, in journals, and among each other is highly systematic. He also found that one of the major factors of linguistic difference in the academic setting was between the written and spoken registers. Three out of four dimensions that he investigated revealed a spoken vs. written difference that distinguished patterns of register variation in academic language.

Biber's (2006) finding was important and highly relevant to the work here, because "these patterns of linguistic variation are complex, with each linguistic feature being distributed in particular ways in accordance with its associated communicative functions." The point is that MDA is generalizable. There is reason to believe that MDA can be used to distinguish IS and other terrorist speech and writing from other registers. That is because they both are communities of practice. As seen previously a community of practice will have a shared repertoire and engage around a shared goal. In academic language the shared purpose is the dissemination of intellectual and institutional knowledge. In IS the goal is the dissemination of ideological and institutional knowledge. The methodology is premised on two foundational assumptions about groups and their communication.

- A group that operates regularly in a society as a functional element (e.g., in terms of physical location, marriage patterns, or economic, religious, or other interactional behavior) will tend to develop identifying markers of language structure and language use, different from the language of other social groups.

- A communication situation that recurs regularly in a society (in terms of participants, setting, communicative functions, and so forth) will tend over time to develop identifying markers of language structure and language use, different from the language of other communication situations.

(Biber 1995).

To further this, it has been observed that, "a single speaker will make systematic choices in pronunciation, morphology, word choice, and grammar associated with different registers, reflecting the situational characteristics of those registers" (Berber-Sardinha and Pinto, 2019). If that speaker is a member of a certain community of practice, such as the Islamic State, they may reach for the linguistic resources emblematic of that community, revealing their associations.

3.4.1 Methods

Multi-Dimensional analysis requires that registers be compared, "with respect to the 'dimensions' of variation identified through a statistical factor analysis" (Biber, 1995). Factor analysis is defined by Egbert and Biber (2018) as a, "multivariate technique that analyzes linguistic co-occurrence patterns to reduce a large set of variables (individual linguistic features) down to a much smaller set of underlying linguistic parameters." These parameters are also called dimensions. Features are seen to occur together because they share a common discourse function (Biber, 1995).

There is an established procedure for conducting a MDA. Found in Egbert and Biber (2018) the steps are as follows:

1. Design an appropriate corpus. Describe the situational characteristics of each register
2. Develop programs to tag all relevant linguistic features in texts.
3. Tag the corpus automatically.
4. Develop additional programs to compute frequency counts of linguistic features in texts.
5. Identify the set of linguistic features to include using commonalities.
6. Perform factor analysis on the rates of occurrence to identify co-occurrence patterns.
7. Interpret the factors functionally as underlying dimensions of linguistic variation.
8. Compute [function] scores for each text. Use mean [function] scores to compare registers.

3.4.2. Corpus

In a typical Multi-Dimensional Analysis, the corpus will be robust, and (hopefully) representative. Egbert defines a representative corpus as, "a principled sample of texts that represents a well-defined target domain or linguistic population." Meaning that the texts are chosen based on their connection to the community of practice under observation. Many linguists define corpora based on size, the bigger the better. However, MDA has also proven effective at identifying linguistic features on small corpora with limited text samples and (Biber 1995;

Berber-Sardinha & Pinto, 2019). I agree with Nini (2017) when he says that Malicious Forensic Texts are difficult to access making large corpora difficult as well. This is not a problem for MDA because Egbert (2019) claims that small corpora are adequate, "where it is unfeasible to collect a large number of texts or words" what he calls, "specialized domains." This is precisely the environment that I am proposing the MDA will be operating in when attempting to identify forensic texts created by Islamic State.

3.4.3 Tagging

Most MDA studies utilize some form of tagging for grammatical information on the corpus. Part-of-Speech tagging is most commonly used in MDA (Gray, 2019). With a tagged corpus searching by grammatical category is allowable not just by "individual word forms" (Gray, 2019). However, with extremely small corpora, Biber and Conrad (2019) state that hand counting is allowable.

3.4.4. Linguistic Features

Establishing a set of linguistic features is crucial for MDA. In his original work Biber developed a list of 67 linguistic features that had been previously shown to be representative across registers (Nini, 2017). It is believed that features "do not randomly co-occur in texts" (Biber, 1988). Where they co-occur, there is typically some type of functional underpinning that drives their appearance

3.4.5. Factor Analysis

A factor analysis is a statistical method that reduces the number of linguistic variables into a set of factors – that is, dimensions (Berber-Sardinha & Pinto, 2019). This takes the linguistic features and reduces them to a small set of "super variables." Factor analysis will be discussed more in the methods section of this thesis.

3.4.6. Interpreting the Dimensions

There is necessarily a qualitative interpretation of the dimensions that reveals the communicative functions underlying them. The question to answer is what do the co-occuring features "do" together (Firginal & Hardy, 2019). One must look for patterns in the data and combine that with their statistical frequency to come to a hypothesis about what the features do. This is then compared against text samples to draw out the functional interpretation.

3.5. SITUATIONAL CHARACTERISITICS

Registers have both a situational and a functional quality. Since IS is a clandestine community it is difficult for us to know precisely the situational characteristics of a particular text. As previously mentioned, forensic texts are always found *ex situ* of their communicative purpose. That is, they are divorced in many ways from their original situational characteristics. This particular corpus side-steps some of this problem by virtue of it being an "official" newsletter produced by the media arm of IS. However, texts such as those found in the Countering Terrorism Center's Harmony project would take much more research to reconstitute the situational context of the document's production.

A register analysis will benefit from understanding these situational characteristics. Biber and Conrad (2019) provide a framework for analyzing the situational characteristics of texts. Their framework can be seen below.

Situational Characteristics of Registers and Genres

I. Participants

- a. Addressor(s) (i.e. speaker or author)
    - 1. Single/plural/institutional/unidentified
    - 2. Social characteristics e.g., age, education, profession
- b. Addressees
    - 1. Single/plural/unenumerated
    - 2. Self/other
- c. Are there onlookers?

II. Relations among participants

- a. Interactiveness
- b. Social roles: relative status or power
- c. Personal Relationship: e.g., friends, colleagues, strangers
- d. Shared Knowledge: personal, specialist

III. Channel

- a. Mode: speech/writing/signing
- b. Specific medium
    - 1. Permanent: e.g., taped, transcribed, printed, handwritten, email
    - 2. Transient: e.g., face-to-face, telephone, radio, TV

IV. Processing Circumstances

- a. Production: real time/planned/ scripted/ revised and edited
- b. Comprehension: real time/ skimming/ careful reading

V. Setting

- a. Are the time and the place of communication shared by the participants?
- b. Place of communication
    - 1. Private/ public
    - 2. Specific setting
- c. Time: contemporary/ historical time period

VI. Communicative Purpose

- a. General purpose: e.g., narrate/report, describe, inform/explain/interpret, persuade, how-to/procedural, entertain, edify, reveal self.
- b. Specific purpose: e.g., summarize information from numerous sources, describe methods, present new research findings, teach moral through personal story
- c. Purported factuality: factual, opinion, speculative, imaginative
- d. Expression of stance: epistemic, attitudinal, no overt stance

VII. Topic

a. General topic domain: e.g., domestic, daily activities, business/workplace. Science, education/academic, government/legal/politics, religion, sports, art/entertainment
b. Specific topic
c. Social status of person being referred to

3.5.1. Participants

Communication does not happen in a vacuum. It requires participants, both a "speaker" and a "hearer." I will follow Biber and Conrad (2019) and call these participants addressor and addressee. The addressor creates a text. The addressee consumes the text. Often times the addressor and addressee are very evident. For example, a personal conversation between a husband and wife has well defined participants. In written discourse, however, the addressor and addressee can be less apparent (Biber & Conrad, 2019). For our purposes, in al-Naba', the authors of the articles are not identified. This is not an unusual circumstance as noted by Biber & Conrad (2019), "some…texts have an 'institutional' addressor: they can be attributed to some institution, but there is no indication of who actually wrote the text." The addressor in al'Naba' *is* the Islamic State.

Similarly, the addressee in this particular type of text is not apparent. There is no way of knowing who is consuming the material in al-Naba'. In this case the addressee is said to be "unenumerated" (Biber & Conrad, 2019). According to Biber and Conrad (2019) this is like, "a novel [which] can exist physically for decades or even centuries, and there is no obvious way to identify who the set of readers will be over that time." We can make some inferences about who the general addressee is as two sub-groups, 1) the eligible in-group for IS membership, that would be the young, Sunni, males; 2) the individual in the Caliphate by dint of being conquered, not there voluntarily.

We must consider a third category by dint of who the institutional addressor is, the onlooker. There is no conceivable world in which IS media is producing these materials and not considering that the eyes of the global Intelligence community are on them. Onlookers are, “participants who observe but are not the direct addressees of the register,” (Biber & Conrad, 2019). There is a sense in certain registers where the onlooker is more important that the addressee. Biber and Conrad (2019) mention in the courtroom where a lawyer may be addressing a witness, but they are always aware of their duty to sway their onlooking jury to their side.

3.5.2. Relations among participants

We must next consider the relationships shared by the addressor and addressee with the text. The first question we must ask is how interactive the exchange is. To what degree do the addressor and addressee engage with each other in this register. This is called interactiveness, In our circumstance the register has a low interactiveness. We can liken this to Biber and Conrad’s (2019) description of a low interactive register of university catalogs since, “it is very difficult to even identify the exact authors of this text, it is virtually impossible to have an interactive dialogue with the authors.” The anonymous articles of al-Naba’ with their anonymous authors and with their unenumerated addressees have a very arm’s length relationship with low interactiveness.

Next, the framework asks us to consider the social roles and personal relationships between the addressor and addressee. The fact that the articles are written by an ‘institutional’ addressor means that they present their arguments from the stance of greater social standing. The addressee group is characterized as the eligible in-group. This social dynamic should appear in the language used as IS appeals to certain authoritative texts and presents information in a didactic manner.

The didactic character of the articles in al-Naba' is very important because it shows that, there is some shared background knowledge necessary to understand the message: knowledge of Arabic, knowledge of Islam. However, the addressee is not expected to have a specialist's knowledge of all of the topics presented unlike an Academic article presented to peers, and more like an introductory textbook (Biber & Conrad 2019).

3.5.3. Channel

Further developing our situational characteristic framework, we have seen already that the channel in which the text is conveyed is quite crucial. Speech versus writing is going to drive much of the variation in linguistic traits as well as the type of relations between addressor and addressee. Spoken and written registers, "differ in the typical production circumstances and even their typical communicative purposes," (Biber & Conrad 2019). Since we are dealing solely with written registers in this thesis, we will not be dealing with the more interactive interpersonal elements that might arise from conversation.

However, we will need to dig deeper than just "writing" to understand the specific sub-category of writing that we will be dealing with. The newsletter format, specifically the article format within the newsletter comes with its own influences on the linguistic features that are used.

3.5.4. Processing Circumstances

The processing circumstances of the text is where our knowledge of a forensic text starts to break down. While we know several things about the production of the articles in al-Naba', we cannot know the entire context in which it was produced. As writing, we know that these texts are planned, structured, and edited to present a precise and persuasive argument. What we cannot know is under what conditions the text was produced. Was there a bombing on an IS area while the author was writing, is the author a scholar who is writing under duress or threat by the IS regime, are all questions that we can never know about the production. Those open questions do not, however, change the addressor's mastery over the written text. They are in complete control of how and when they produce the material.

3.5.5. Setting

The written nature of al-Naba' means that the setting is a less important of a measure. Written texts mean that, "participants do not have to share the same time and place," (Biber & Conrad, 2019). As a part of a "regularly" produced newsletter, the articles have a sense that they are meant to be read in real time. But, as distinct from the news items in the newsletter, the articles have a sense of being able to be read out of the time in which they are produced and released.

3.5.6. Communicative Purpose

What has been described until now has been rather tangible aspects of situational context. It is also important to get at the question of *why.* Communicative purpose allows us to investigate the reasons a document is produced (Biber & Conrad, 2019). The articles from al-Naba' have the general communicative purpose of instruction via admonition. Each individual text has a specific communicative purpose of instructing eligible in-group members in a specific topic with the assumption that the author (IS) is an authority on the topic, and that the information presented is true.

3.5.7. Topic

Where the majority of these situational factors will be the same for each text in the corpus, topic will be different for each text. Topic is the main driver for lexical variation, “the words in a text are to a large extent determined by the topic of the text.” While the general topic will be didactic in admonishing the eligible in-group into doing the “right thing,” according to IS, the individual topics will take us a level deeper which will presumably change the language used. According to Biber and Conrad (2019), grammatical change is not influenced by topic in the same way, even though certain constructions (i.e. passive verbs) may seem to be correlated.

Situational Characteristics of the al-Naba’ Newsletter

I. **Participants**
   a. Addressors
      1. Institutional (ISIS as an entity)
   b. Addressees
      1. Unenumerated
      2. ISIS in-group located within the boundaries of the “Caliphate”
   c. Onlookers
      1. Counterterrorism experts
      2. Researchers

II. **Relations between participants**
   a. Interactiveness
      1. Low interactiveness
   b. Social rules (relative status or power)
      1. The publisher has a higher status and power, while the readership has low status and minimal power.
   c. Personal Relationship

1. Strangers, yet they share a relationship as fighters for a particular ideological cause.
2. Involuntary people within the borders of the caliphate.

d. Shared Knowledge

1. The readership is part of an eligible in-group of Sunni Muslim, as well as being indoctrinated members into ISIS' brand of Islam. They would have a specialist's knowledge of the ideological underpinnings of the ISIS message.

III. Channel

a. Mode

1. Writing

b. Specific medium

1. Permanent: Written newsletter

IV. Processing Circumstances

a. Production: Revised and edited

b. Comprehension: Ranges from skimming to careful reading

V. Setting

a. Time of the communication is asynchronous

b. Setting is distanced, from production and dissemination to private reading

c. Timing of the content is contemporary due to its purported news worthiness

VI. Communicative Purpose

a. General purpose

1. Report and inform

b. Specific purpose

1. To report on the news happening within the "Caliphate," and to provide opinion and editorials on issues of interest to the population residing within its borders

c. Purported factuality

1. News items are represented as being factual, but there is a biased spin in the favor of the ISIS brand
2. Opinion pieces and editorials claim to be as such, but often appeal to religious authority such as the Qur'an and religious scholars to support the "factuality" of the opinion presented

d. Expression of stance

1. Attitudinal – directed toward presenting a victorious image of ISIS in battle and present the Islamic State as a legitimate governing body to a population already living in its captured territory

VII. Topic

a. General topic domain

1. Domestic news items, religion, political commentary

b. Specific topic: varies

c. Social status of person being referred to: the addressee is referred to as a noble soldier in the cause, and this information is presented from the top down to make that soldier better informed and a more efficient combatant, and become a "better citizen"

CHAPTER 4. METHODS

4.1. SETTING EXPECTATIONS

This thesis takes a novel approach not found in the literature to utilize Multi-Dimensional Analysis for forensic linguistic analysis of register variation in Jihadist language. First, it is important to set expectations correctly. There are a few important considerations:

- This study is preliminary.
- The data set is small.
- This analysis only considers IS language in isolation without a comparison data set of different registers and different groups.

Thus, the findings herein will not be generalizable beyond this study; however, we can treat this work as a proof of concept that can inspire future work. It is intended to be an iterative study, so this is the first stage in a larger comparative work. With these principles in mind, we can discuss the study design itself.

4.2. STUDY DESIGN

This study was designed to use MDA to extract statistical regularities from a small number of inwardly focused IS texts. Those regularities were analyzed in terms of their potential communicative function, with the goal that in subsequent work they might define a register for terrorist writing and become part of a profile of IS writing in particular. A second benefit might be to use a validated pattern to seed further MDA studies. Since this study is preliminary it is designed to act as an exploratory technique only and while an interesting pattern exists, there is no external validation via comparative data and work would need to be done on a larger corpus.

For this thesis, the IS newsletter al-Naba’ was selected as the object of study. This newsletter was chosen because it is written in Arabic to an internal IS audience. The idea is that this internal facing document will help me closely approximate what real internal forensic texts might look like. As mentioned previously, al-Naba’ is a newsletter produced by the caliphate since 2010, however, only issues from January 2020 – February 2020 were considered. Each issue is broken down into current news, articles, editorials, and infographics. To date, only the infographics have been studied thoroughly in the literature (Winkler, 2016).

4.3. CORPUS DESIGN

The primary concern is that the corpus must be representative of a target population. A representative corpus is defined by Egbert (2019) as, “a principled sample of texts that represents a well-defined target domain or linguistic population.” Egbert (2019) clarifies, small corpora are adequate “where it is unfeasible to collect a large number of texts or words” what he calls, “specialized domains.” Furthermore, according to Biber and Conrad (2019), linguistic features are pervasive throughout a register. Thus, “complete texts are not required to analyze register characteristics.” This exemplifies the case of analysis of texts analyzed in a forensic context, which we attempt to replicate here with the al-Naba’ newsletters.

The data set is limited to 5 issues of al-Naba’ out of (currently) 233. These issues of al-Naba’ were collected from *Jihadology.net,* an open source aggregator for Jihadist content made available to scholars, from January 23, 2020 – February 27, 2020. This is the period of IS insurgency so the messaging of these issues will be different from those of earlier periods where a more institutional language would be used. Focus was placed on one section at a time. In this study, the editorial section was considered only. Two paragraphs were randomly selected from the editorial section of each issue. Eighty-six total sentences were translated from Arabic, and linguistic features were hand coded.

4.4. FEATURES

The features were selected from a list of 67 features outlaid in Biber and Conrad (2019). One primary feature is the number of words in a sentence. I conducted lexical counts of nouns, verbs, adjectives, and adverbs in each sentence. Finally, I coded whether the sentence was overtly religious, implied religious, or non-religious. I developed a Qualtrics survey to aid in the coding of my features and streamlining of my data. The survey questions are as follows:

- **Question 1: Number of words in the sentence**

  The number of words in the entire sentence was counted to provide a syntactic accounting of intra-sentential frequency.

- **Question 2: Number of nouns in the sentence**

  The number of nouns in the entire sentence were counted to provide a syntactic accounting of intra-sentential frequency.

- **Question 3: Number of verbs in the sentence**

The number of verbs in the entire sentence were counted to provide a syntactic accounting of intra-sentential frequency.

- **Question 4: Number of adjectives in the sentence**

  The number of adjectives in the entire sentence were counted to provide a syntactic accounting of intra-sentential frequency.

- **Question 5: Number of adverbs in the sentence**

  The number of adverbs in the entire sentence were counted to provide a syntactic accounting of intra-sentential frequency.

- **Question 6: Is the sentence political in nature?**

  I coded whether the sentence was overtly political, implied political, or non-political. For coding purposes, “overt” political is defined as a sentence that contains an overt political noun or verb, for instance “coup” or “militant.” “Implied” is defined as an inclination toward politics in context, though no overt language appears in the sentence. For instance, a sentence that alludes to a lock down in a city but does not call to political action would be coded as implied political. “Non-political” is defined as a sentence that has nothing to do with politics within the sentence or in context.

- **Question 7: Is the sentence religious in nature?**

  For coding purposes, “overt” religious is defined as a sentence that contains an overt religious noun or verb, for instance “prayer” or “worship.” “Implied” is defined an inclination toward religion in context, though no overt language appears in the sentence. For instance, a sentence that alludes to a Qur'anic passage, but does not quote it would be coded as implied religious. "Non-religious" is defined as a sentence that has nothing to do with religion within the sentence or in context.

- **Question 8: Is the sentence both religious and political in nature?**

  When coding whether sentences were both religious and political in nature, I coded as "yes" if the sentence had already been coded as an overt or implied religious sentence and an overt or implied political sentence.

- **Question 9: Are there overt religious nouns?**

  This question was coded "yes" if there was the identifiable presence of a religious noun. If there was not, it was coded "no."

- **Question 10: What are the nouns?**

  This is a space to record the exact religious nouns used.

- **Question 11: Are there overt political nouns?**

  This question was coded "yes" if there was the identifiable presence of a political noun. If there was not, it was coded "no."

- **Question 12: What are the nouns?**

  This is a space to record the exact political nouns used.

- **Question 13: Are there overt religious verbs?**

  This question was coded “yes” if there was the identifiable presence of a religious verb. If there was not, it was coded “no.”

- **Question 14: What are the verbs?**

  This is a space to record the exact religious verbs used.

- **Question 15: Are there overt political verbs?**

  This question was coded “yes” if there was the identifiable presence of a political verbs. If there was not, it was coded “no.”

- **Question 16: What are the verbs?**

  This is a space to record the exact political verbs used.

4.5. STATISTICAL TESTING

Basic descriptive statistics were conducted. For questions 1-5 on the survey, the mean, median, mode, and range with minimum and maximum values were calculated. For questions 6 and 7 the percentage of overt, implied, and “non” were calculated. Next, I ran a correlation matrix to reveal which of the variables are correlated with each other.

Then, I conducted a factor analysis. According to Yong and Pearce (2013), “Factor analysis assembles common variables into descriptive categories.” It is most commonly used when there are a large number of variables that need to be reduced to a manageable number. In instances such as this thesis, factor analysis does not make much sense due to the small number of variables. However, factor analysis is a core component of MDA. I conducted the factor analysis for the sake of building the machinery of the expert system.

CHAPTER 5. RESULTS

These results are based on a data set comprised of 5 issues of al-Naba’ an internal Arabic language Islamic State newsletter. These issues were issued from January 23, 2020 – February 27, 2020. Within these issues I focused solely on the editorial sections, choosing two paragraphs at random to analyze from each editorial in each issue under consideration. The result was a corpus of 86 sentences that I translated from the original Arabic into English. These sentences were then coded based on a set of 16 features and collected via a Qualtrics survey used solely by me to collect and organize the data. I then conducted statistical analyses on this data using descriptive statistics, a correlation matrix, and a Factor Analysis. The results of which are detailed below.

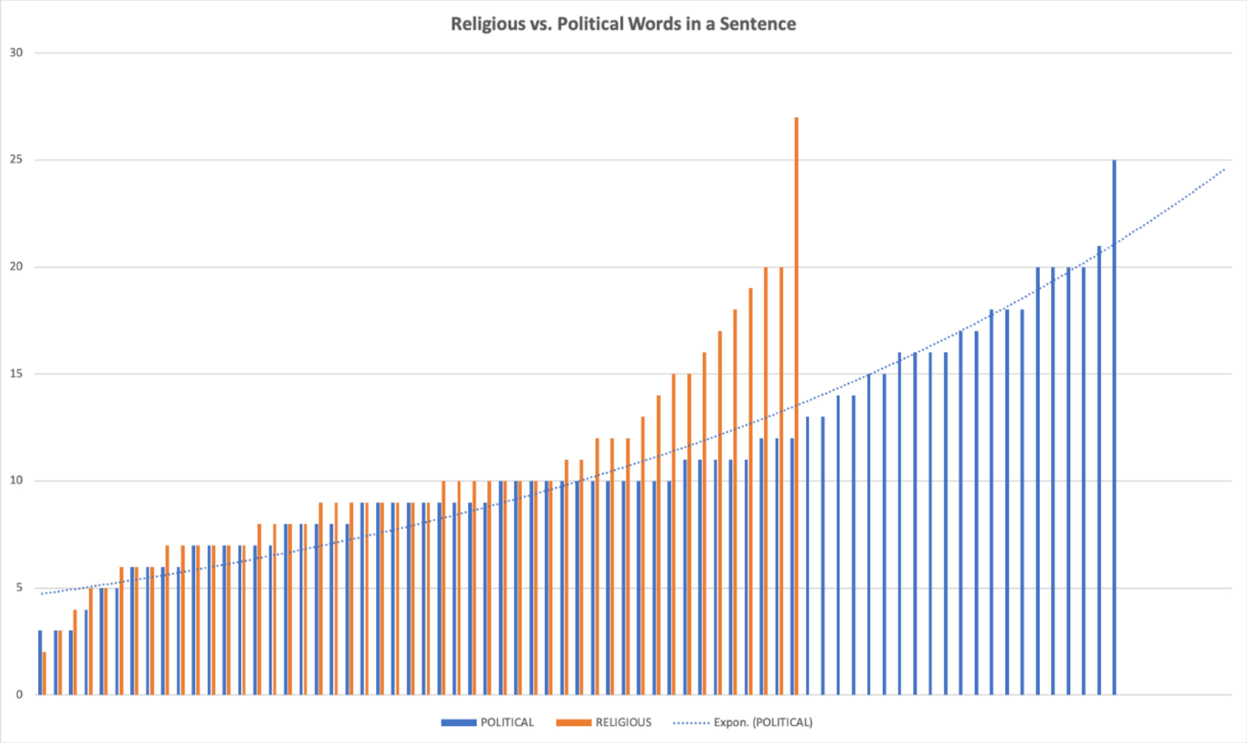
5.1. DESCRIPTIVE STATISTICS

Descriptive statistics were run on the data the five lexical categories, number of words in the sentence, number of nouns in the sentence, number of verbs in the sentence, number of adjectives in the sentence, and number of adverbs in the sentence. Most importantly the mean of number of words in the sentence that is 10.8. This serves as a breaking point for what are termed here long sentences: sentences whose length are above the mean. Out of the 86 sentences 39% have word lengths above the mean. Ten percent are long sentences of 20 words or more. All sentences of 20 words or more are political in nature while religious sentences make up 42% of the sentences of 20 words or more.

	# Words in Sentence	# Nouns in sentence	# Verbs in Sentence	# Adjectives in Sentence	# Adverbs in Sentence
Average	10.8	4.5	1.5	0.4	0.2
Median	10	4	1	0	0
Mode	10	3	1	0	0
Range	25	11	6	3	2
Minimum	2	0	0	0	0
Maximum	27	11	6	3	2

	# Overt	% Overt	# Implied	% Implied	# Overt &Implied	% Overt & implied	# No	% No
Religious Sentence	29	33%	21	24%	50	58%	36	42%
Political Sentence	41	48%	30	35%	71	83%	15	18%

This factors into the shape of the data. The number of sentences were charted with respect to the number of words in the sentence, showing a sharp increase above the mean between sentences coded as religious and sentences coded as political. While there are fewer religious sentences, their increase begins more sharply above the mean than does the political sentences. However, the number of political sentences above the mean far outstrips the religious sentences. In the following chart the Y axis represents the number of words in the sentence.



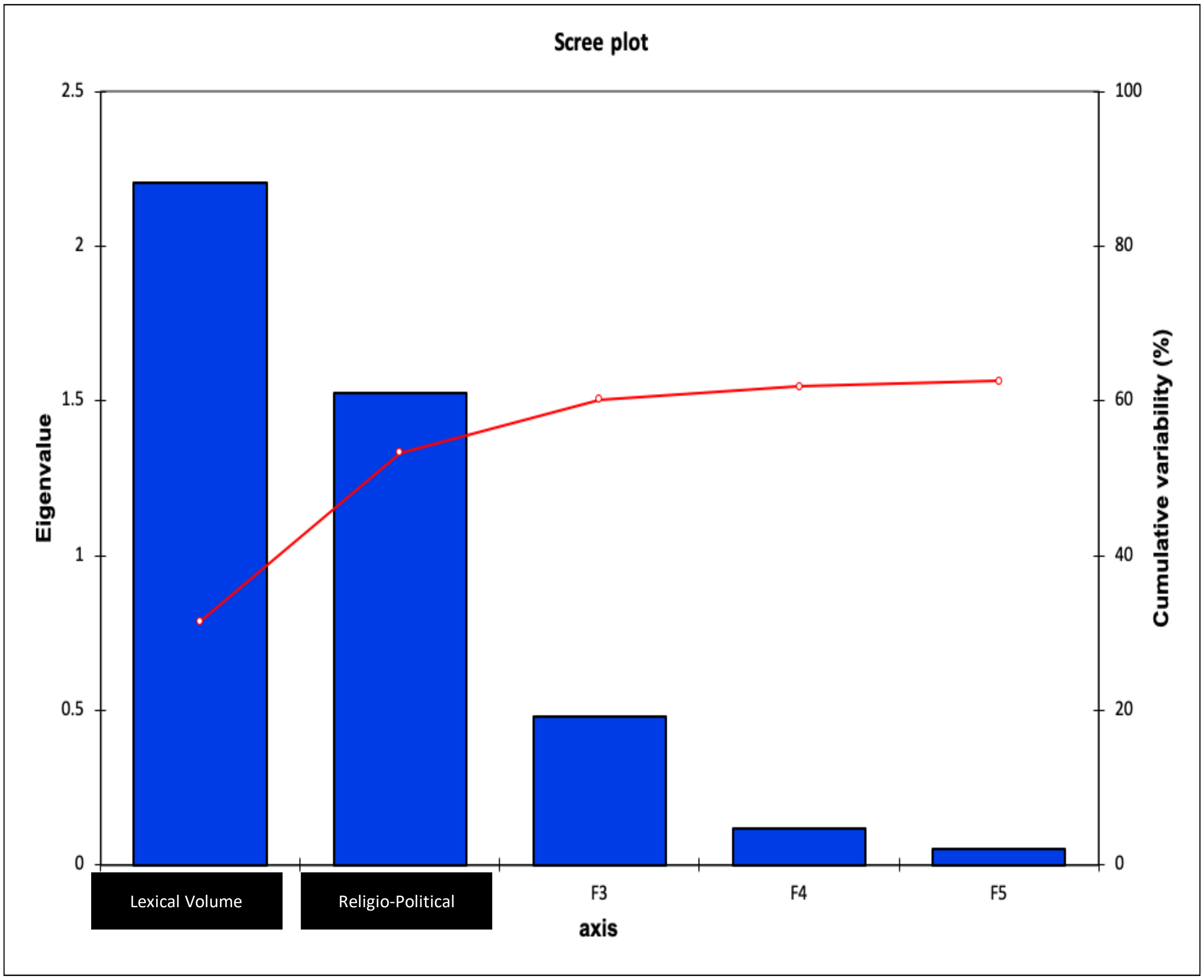
5.2. CORRELATION MATRIX

The features were subjected to a correlation test. The results of this were mixed. Most importantly the correlation matrix revealed a negative correlation between several of the data points. These negative correlations are consistent throughout the data. All values are reported in their Pearson r value, and the p value giving us the statistical significance. The related variables *are there overt religious verbs* r(86)=-0.51, p=.00001 and *if yes, how many verbs* r(86)=.51, p=.00001 exhibit a strong negative correlation. For the variable *number of verbs in the sentence* the correlated variables *are there overt political verbs* negatively correlates r(86)=-.40, p=.000136 while *if yes how many verbs* r(86)=.40, p=.000136. The variable *are there overt political verbs* r(86)=.63, p=.00001 and the variable *if yes, how many verbs* r(86)=-.63, p=.00001 are also strongly and negatively correlated with each other. This same negative relationship holds between the related variables *are there overt religious nouns* and *if yes how many nouns* and *are there overt political nouns* and *if yes, how many nouns.*

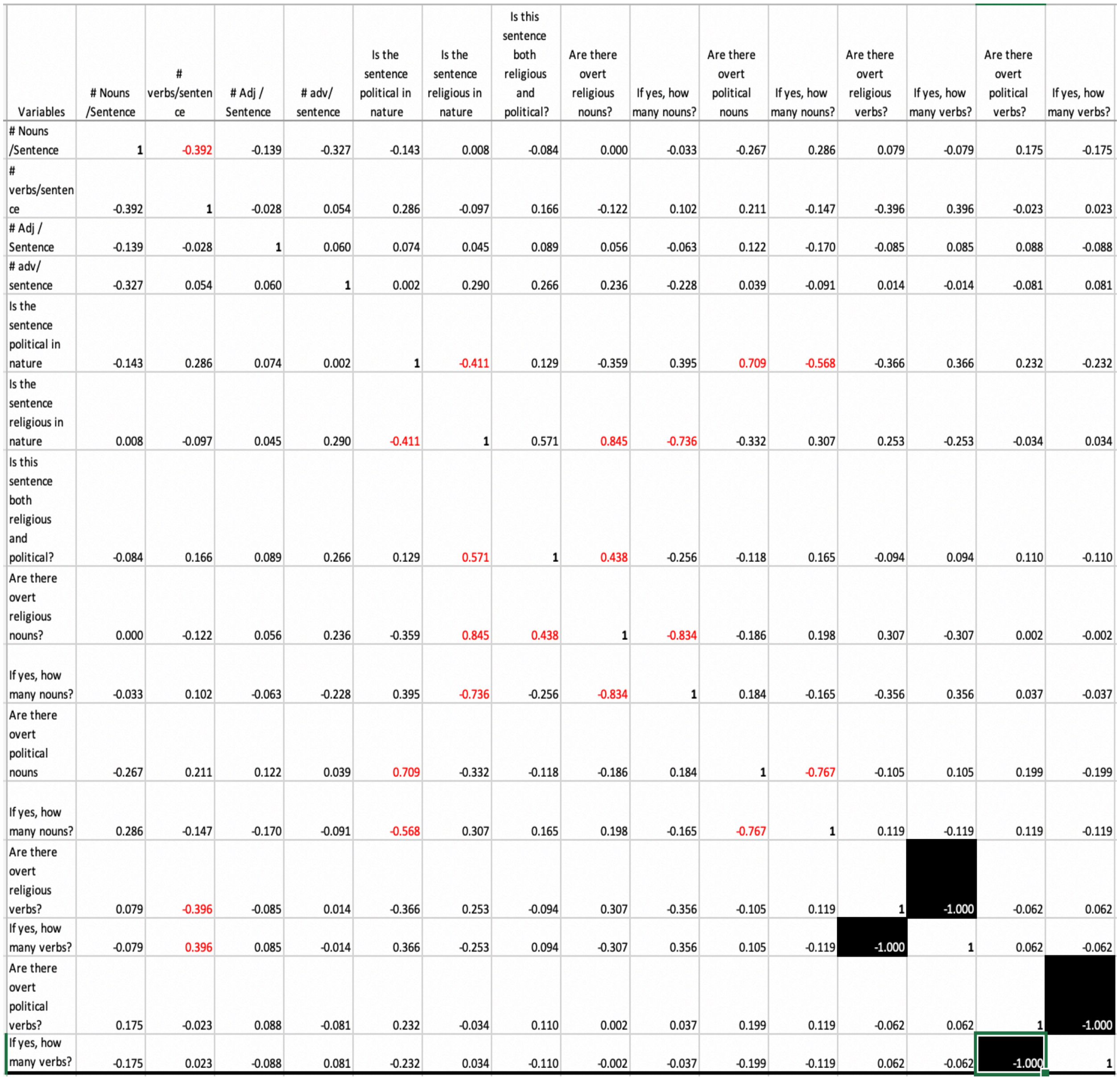
This relationship resulted in the r(86)=-1 relationships between *are there over religious verbs* and *if yes, how many verbs* as well as *are there over political verbs* and *if yes, how many verbs.* These values were excluded as "noise" in the data because the because the relationship cannot be one and be meaningful. This may mean that all of these relationships are noise as well, or they could be indicative of a larger trend. We cannot know until we run this data against a comparison set in further research.

5.3. FACTOR ANALYSIS

The features were reduced for the factor analysis. They were the lexical features plus *is the sentence religious* and *is the sentence political*. The result was a factor analysis with two predominant factors. Factor 1 is comprised of *number of words in the sentence* and *number of nouns in the sentence.* This factor does the majority of the "work" identifying which variables are most important in determining register. With an eigenvalue of 2.2 and a variability of 32%. Factor 2 is interesting because its most important variables are the factor loadings for religious speech and political speech. With an eigenvalue of 1.5 and a cumulative variability of 53%. The remainder of the factors were too minimal to be further analyzed. In the following chart, the factors are on the X axis, and the eigenvalues and the cumulative variability are on the Y axis.

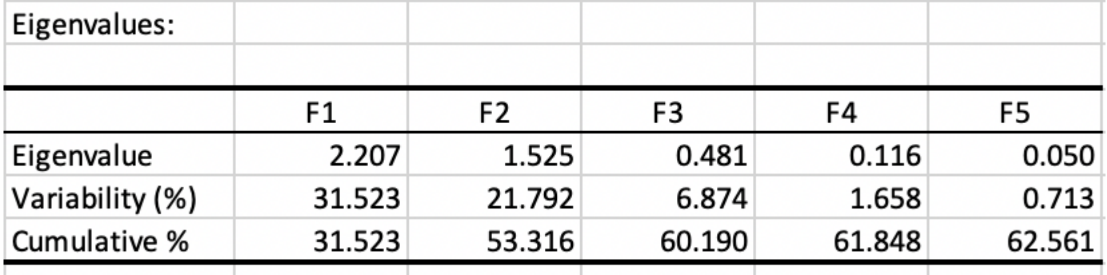

Scree plot
Eigenvalue
Cumulative variability (%)
2.5
2
1.5
1
0.5
0
100
80
60
40
20
0
Lexical Volume
Religio-Political
F3
F4
F5
axis

	Issue 223	Issue 222	Issue 221	Issue 220	Issue 218	Totals
Total Paragraphs	2	2	2	2	2	10
Total Sentences	21	13	13	27	12	86
Total Words	170	134	140	366	119	1015
Total Nouns	71	46	55	164	53	389
Total Verbs	28	24	23	37	19	131
Total Adjectives	4	3	5	21	1	34
Total Adverbs	1	1	4	9	2	17
Total Overt Political Sentences	2	2	7	21	9	41
Total Implied Political Sentences	10	5	6	6	3	30
Total Non-political Sentences	9	6	0	0	0	15
Total Political Nouns	1	0	4	42	25	72
Total Political Verbs	3	0	1	1	1	6
Total Overt Religious Sentences	9	9	4	5	2	29
Total Implied Religious Sentences	10	4	5	0	2	21
Total Non-religious Sentences	2	0	4	22	8	36
Total Religious Nouns	16	13	6	7	2	44
Total Religious Verbs	1	2	0	0	0	3
Total Sentences both Religious and Political	10	6	9	5	3	33

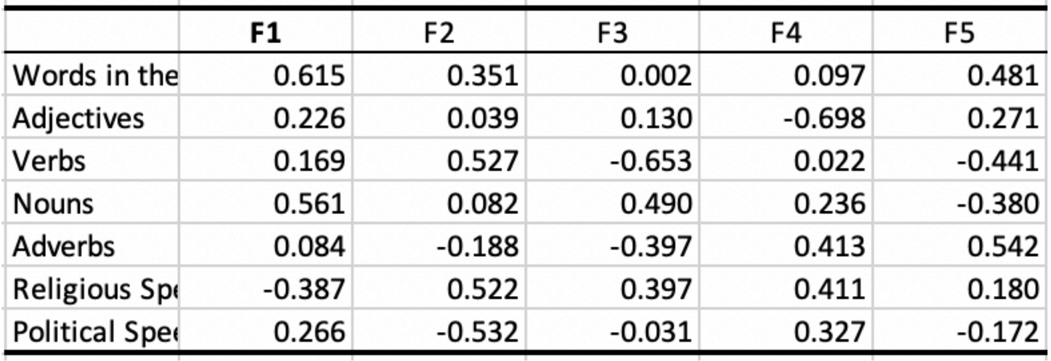


Variables	# Nouns /Sentence	# verbs/sentence	# Adj / Sentence	# adv/ sentence	Is the sentence political in nature	Is the sentence religious in nature	Is this sentence both religious and political?	Are there overt religious nouns?	If yes, how many nouns?	Are there overt political nouns	If yes, how many nouns?	Are there overt religious verbs?	If yes, how many verbs?	Are there overt political verbs?	If yes, how many verbs?
# Nouns /Sentence	1	-0.392	-0.139	-0.327	-0.143	0.008	-0.084	0.000	-0.033	-0.267	0.286	0.079	-0.079	0.175	-0.175
# verbs/sentence	-0.392	1	-0.028	0.054	0.286	-0.097	0.166	-0.122	0.102	0.211	-0.147	-0.396	0.396	-0.023	0.023
# Adj / Sentence	-0.139	-0.028	1	0.060	0.074	0.045	0.089	0.056	-0.063	0.122	-0.170	-0.085	0.085	0.088	-0.088
# adv/ sentence	-0.327	0.054	0.060	1	0.002	0.290	0.266	0.236	-0.228	0.039	-0.091	0.014	-0.014	-0.081	0.081
Is the sentence political in nature	-0.143	0.286	0.074	0.002	1	-0.411	0.129	-0.359	0.395	0.709	-0.568	-0.366	0.366	0.232	-0.232
Is the sentence religious in nature	0.008	-0.097	0.045	0.290	-0.411	1	0.571	0.845	-0.736	-0.332	0.307	0.253	-0.253	-0.034	0.034
Is this sentence both religious and political?	-0.084	0.166	0.089	0.266	0.129	0.571	1	0.438	-0.256	-0.118	0.165	-0.094	0.094	0.110	-0.110
Are there overt religious nouns?	0.000	-0.122	0.056	0.236	-0.359	0.845	0.438	1	-0.834	-0.186	0.198	0.307	-0.307	0.002	-0.002
If yes, how many nouns?	-0.033	0.102	-0.063	-0.228	0.395	-0.736	-0.256	-0.834	1	0.184	-0.165	-0.356	0.356	0.037	-0.037
Are there overt political nouns	-0.267	0.211	0.122	0.039	0.709	-0.332	-0.118	-0.186	0.184	1	-0.767	-0.105	0.105	0.199	-0.199
If yes, how many nouns?	0.286	-0.147	-0.170	-0.091	-0.568	0.307	0.165	0.198	-0.165	-0.767	1	0.119	-0.119	0.119	-0.119
Are there overt religious verbs?	0.079	-0.396	-0.085	0.014	-0.366	0.253	-0.094	0.307	-0.356	-0.105	0.119	1	-1.000	-0.062	0.062
If yes, how many verbs?	-0.079	0.396	0.085	-0.014	0.366	-0.253	0.094	-0.307	0.356	0.105	-0.119	-1.000	1	0.062	-0.062
Are there overt political verbs?	0.175	-0.023	0.088	-0.081	0.232	-0.034	0.110	0.002	0.037	0.199	0.119	-0.062	0.062	1	-1.000
If yes, how many verbs?	-0.175	0.023	-0.088	0.081	-0.232	0.034	-0.110	-0.002	-0.037	-0.199	-0.119	0.062	-0.062	-1.000	1

Factor Analysis



Eigenvalues:					
	F1	F2	F3	F4	F5
Eigenvalue	2.207	1.525	0.481	0.116	0.050
Variability (%)	31.523	21.792	6.874	1.658	0.713
Cumulative %	31.523	53.316	60.190	61.848	62.561



	F1	F2	F3	F4	F5
Words in the	0.615	0.351	0.002	0.097	0.481
Adjectives	0.226	0.039	0.130	-0.698	0.271
Verbs	0.169	0.527	-0.653	0.022	-0.441
Nouns	0.561	0.082	0.490	0.236	-0.380
Adverbs	0.084	-0.188	-0.397	0.413	0.542
Religious Sp	-0.387	0.522	0.397	0.411	0.180
Political Spe	0.266	-0.532	-0.031	0.327	-0.172

CHAPTER 6. DISCUSSION

6.1. THE PATTERN

The most compelling finding arising out of the results was a clear pattern of the data. As Biber (1995) says that, “the analytical goal of the MD approach is to provide comprehensive descriptions of the patterns of register variation…” As we can see in figure 7 the general shape of the data suggests that sentences with words above the mean (10.8) see a parallel increase of both religious and political language. Since there is an overall higher proportion of political sentences in the data set, the landscape above the mean words in the sentence shows that political and religious language patterns together. My study showed all sentences of 20 words or more were both political and religious in nature. Despite this fact, all sentences above the mean are more overtly religious than political at 9 sentences versus 6 sentences. This is seen in my factor analysis, where the second dimension which I call Religio-political Valence is shown to have a negative correlation with the number of verbs in a sentence. If we rotate this data point, we see that as nouns increase so do the religious and political speech.

This presents a curious finding. The pattern of data seems to be driven by the Noun Phrase (NP) rather than the Verb Phrase (VP). In the factor analysis, the first dimension, which I call Lexical Volume, shows a strong correlation between the number of nouns in the sentence with the number of words in the sentence. There is no similar correlation in the number of verbs and words in the sentence. Descriptive statistics showed that the number of nouns per sentence (mean = 4.5) exceeds the number of verbs in the sentence (mean = 1.5). The correlation matrix indicates this as well. The variables *is the sentence religious in nature* and *does the sentence have religious nouns* are correlated strongly r= (0.85). Similarly, *is the sentence political in nature* and *does the sentence have political nouns* are strongly correlated r= (0.71). Generally, it can be said that, in the dimension of Lexical Volume, as sentence length increases so does the noun richness of the sentence. Verbs do increase as sentence length increases, but only negligibly. It is hard to say at this point why this is, but it remains a question that requires further research.

However, what I can tangibly see is that the religious and political sentences in my data have nouns in line with the above break down: 13 overtly religious nouns, 11 overtly political nouns. Six of the sentences have both overt religious *and* overt political nouns in them. Why might this be important? First of all, because for computational purposes noun phrases are good indicators of text content (K. Gharaibeh & K. Gharaibeh, 2012) but also, because the number of nouns in a sentence provides a good indication of the complexity of the noun phrases within. This nominal complexity is related to the communicative purpose of the al-Naba' editorials. This gives us insight into a potentially important feature of IS language, where complex nominal sentences with a religio-political content may be indicative to IS writing.

This type of granular data has been used to distinguish registers in other contexts. Biber (2006) was able to find complex nominalizations as an import factor in differentiating between academic language and nonacademic language. In discussing the differences, he notes that the "intuitive" conceptualization of written academic language is that it will be complex because of a Russian doll-like series of expanding embedded clauses is incorrect. Instead, Biber (2006) found that "a detailed study of noun use in spoken and written registers is informative because much of the referential information of academic language is packaged in noun phrases." The spoken register showed that while nouns in spoken registers are “relatively rare” their use in writing is complex. Biber elucidates further on the complex stringing of nouns in written academic language:

> [Nouns and nominal modifiers] often occur together to build very complex noun phrase structures. [One paragraph] begins with a very long sentence, which has only one main verb... Most of the sentence comprises a single noun phrase, functioning as the direct object of [that verb].

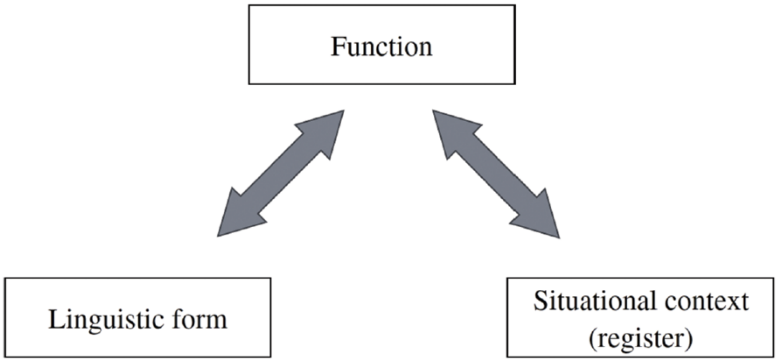
Different registers, such as textbooks for different disciplines, will use different types of nouns, such as the prevalence of animate nouns and group nouns in business textbooks and abstract/process nouns in social science textbooks. In essence, group affiliation can be differentiated via patterns derived by linguistic features.

6.2. QUALITATIVE INTERPRETATIONS

It is not enough to conduct a quantitative assessment to analyze register variation, a qualitative element is assumed by the Multi-Dimensional approach. Biber (1988) explains:

> The interpretation of the factors is based on the theoretical assumption that these co-occurrence patterns indicate an underlying communicative function shared by the features; that is, it is assumed that linguistics features co-occur frequently in texts because they are used for a shared set of communicative functions in those texts.

This is by nature interpretive. The goal is to get at the intersection of the linguistic features and the situational context of the register and explain how they function together (Biber and Conrad, 2019). There is no direct mapping from the linguistic feature to the communicative function, there are often multiple relations. For example, in my data I have identified two dimensions: Dimension 1: Lexical volume, and Dimension 2: Religio-Political valance. The functional bases of these two dimensions were derived by multiple linguistic features that map together in an asterism. It can often be hard to see the constellation for the stars. Similarly, it takes abstracting from the individual features in comparison to the situational context to derive the communicative function. For a concrete example, Biber and Conrad (2019) mention that in classroom settings, pronouns can serve many functions, however, they often refer to, "things [that] are present in the communication situation: oneself, the listener, other people, objects." The relationship between linguistic feature and register is its function. This can be seen in the following figure adapted from Biber (1995).



In university language, Biber's (2006) study was broad and he found many different aspects of language being used in differing registers by different groups of speakers. One of the areas he looked at was a grouping of multi-word sequences which he called lexical bundles. When looking at the functional underpinnings of lexical bundles served as "discourse framing devices" and they "differ dramatically across registers, academic disciplines, depending on the typical purposes of each" (Biber 2006). Despite the differences among internal registers, Biber (2006) found that lexical bundles are universally important for the structure of the university register as a whole.

Textbook	Lexical Bundle Type	Example
Natural Science	Place/location	Water is poured onto the tephra **in the vicinity of** the steam wells.
Engineering	Quantity/Mathematical	The resonant frequency depends on **the ratio of** the mass to the total spring rate of the tires.
Social Science	Stance	Federalism issues remain **an important part of** our political agenda.

6.3. THE COMMUNICATIVE PURPOSE

Now I return to my pattern. Remember that I said my pattern shows that longer sentences see a joint increase in religious and political language. However, overall, there is a higher proportion of political language than religious language throughout the corpus. While I cannot claim that this is a register of Islamic State language, it is the beginning of a pattern that is the beginning of a profile. The provocative question is: what might the purpose of this pattern be?

I suggest that there is a dual-purpose to be found in this pattern, suggestive of a communicative purpose where the political messages of IS are driven home with a religious admonition punctuating the political message with the authority of religion. I have shown that the al-Naba’ newsletter is an internal facing publication with an audience of already affiliated IS members, whether by choice or by force. Previously, I said that the internal IS audience is more concerned with immediate political concerns such as battleplans, troop movements, and updates from ongoing battles. Al-Naba’ appears to fit squarely within the pattern where there are predominately political sentences, punctuated with religious terminology. This terminology is specific, however, the religious punctuations are meant as admonitions against the in-group member, to remind them not to stray from the group or God will smite them.

Consider sentence (1) (Issue 218, Paragraph 1, Sentence 3):

(1)

فقدت تلك الفصائل المرتدة أهم مقومات وجودها ونموها وهو القرب من فلسطين

> These apostate factions have lost the most important determinants of their existence and growth, which is proximity to Palestine.

This sentence focuses on an out-group that IS calls “apostates.” This is a highly religiously charged word, however it has a political valence as well. One may call this word value-laden, in the sense that IS has placed specific pragmatic associations on the word which draw on in-group knowledge. In the Islamic State repertoire, apostates are Muslim factions that have either repudiated IS or are other Jihadist groups turned away from the IS ideology.

One of the tactics IS uses to deal with these Muslims is *Takfir* or excommunication (Wood, 2015). While this is controversial in mainstream Islamic discourse IS has used the principle with abandon, though it also has a long history in extreme Muslim views such as the

Wahhabis. IS has focused their ire on these so-called “apostates” and deride their efforts even if they share an end goal with IS. Concurrently, there is inherent politicization of the word Palestine that calls to mind images of the struggle between the State of Israel and the occupied Palestinian territories. This is a highly charged issue broadly in the Middle East but highly important as a motivating factor to Jihadist movements.

If the religious nature of IS language is truly punctuated with purely religious language, we should be able to see this in our corpus. This can be seen in what I call “admonition” sentences. One can be seen in sentence (2) (Issue 222, Paragraph 2, Sentence 1) below:

(2)

والغالب في أتباع الشيطان من كل الملل أنهم يُقدّمون ما يُمنّيهم به من الأكاذيبعلى ما يبْلُغُهم من الحقّ المبين الذي
أنزله الله تعالى على عباده المرسلين

> The majority of the followers of Satan from all their weariness, they present what lies in them and is lying to them, what he tells them is that to demean what God Almighty has revealed to his servants is clearly right.

Sentences like this one admonish members of IS if they do not follow Allah’s commands. Remember, what is said to be a commandment of Allah is different in Jihadist ideology than it is in mainstream Islam. This sentence, then, is clearly an entreaty for IS members to turn away from the ideologies of “apostates” who would be seen as servants of Satan. The moral lesson is to follow IS or run the risk of *Takfir.* In essence, being left out in the cold, with a death sentence over your head. However, as already mentioned “apostate” and *Takfir* are political tools of IS. There is a political resonance that is pragmatically embedded in this admonition sentence. Other admonition sentences like this one from the same issue has the same pragmatic valence.

Now consider sentence (3) (Issue 222, Paragraph 7, Sentence 9):

(3)

والله لا يهدي كيد الخائنين

God does not guide the hand of the treacherous.

The treachery in context here refers to those who deviate from God's rightly ordained path, as interpreted by IS. This admonition statement carries the same pragmatic function as the much longer sentence above. It reads: *Don't do it or you'll become an outcast from the Caliphate* which that is tantamount to a death sentence. These admonition sentences beg to be studied in more detail with a larger corpus, to see if more of the religious language in internal IS propaganda carries an implicit political nature.

With my goal in mind, to create a profile of Islamic State language, it is imperative to take our pattern and understand how it is used. We have just seen how political speech and religious speech are conflated, a political goal with value laden words meant to admonish. How does one come to know the meaning behind these words? To understand these value-laden words, it takes a certain "indoctrination" into a shared background to understand what is being conveyed. This is seen in our section on communities of practice where we saw that CoP require members to conform to a shared background and build a shared repertoire. Islamic State, as a CoP, appears to work in the same way. As one is further indoctrinated into the group, how they perceive their language and how they use that language will trend toward the broader group usage. Value-laden language can be a sign of indexing identity in relation to social norms. This is called Stance (Holmes &Wilson, 2017). Stance includes personal attitudes toward lexical items (Biber and Conrad, 2019). In *University Language*, Biber (2006) discusses stance in the sense of evaluative lexical items. To adapt his example, if one were to say, *"I hate America,"* they are only communicating any other information than their hate (Biber, 2006). What is more germane to my thesis is his description of what is required to parse an expression of stance. Biber (2006):

> At one level, almost any choice among related words can be seen as evaluative. Such lexical expressions of stance depend on the context and shared background for their interpretation. There is nothing in the grammatical structure of these expressions to show that they mark stance. Rather, stance is embedded in these structures, depending on the addressee's ability to recognize the use of value-laden words.

In fact, it can be very hard to "identify a closed set of words used to convey specific attitudes and evaluations" (Biber 2006). Leaving the outside reader just that, on the outside.

6.4. IS THIS PATTERN A PROFILE?

The pattern of longer sentences in the al-Naba' corpus containing both religious and political language, with a larger proportion of political sentences overall may not be exclusive to IS. However, I cannot discount at this time that it is indicative of something more inherent to their community of practice. There is still work that must be done to show if this pattern is part of an emerging profile where I could look at other coding schemes to find further identifying characteristics. Like Biber (2006), this work is “deliberately exploratory.” To see if a pattern would emerge that could be used to seed a more complete Multi-Dimensional Analysis. I cannot make the assertions that Biber can, because my analysis is missing a key component, a comparison group. MDA is comparative by nature (Biber 1995). Due to the exploratory nature and limitations on time and scope this study did not take up the comparative aspect, saving it for further research at the dissertation level.

It is not a stretch of the imagination, however, to speculate about the patterns that might arise from comparative data from other communities of practice that are non-Islamic State especially if such a CoP is a non-Jihadist Salafist organization. In this group’s attempt to speak in defense of the religion against violent Jihadism it is conceivable that the language that they use will be more religious than political. If this is the case then it would look like our pattern here, only in reverse where the largest proportion of sentences will be religious in nature rather than political.

Furthermore, it is not inconceivable to think of the communicative purpose of this language pattern to be different as well. Where my language pattern suggests that IS is using politically charged language with a punctuation of religious admonition, the non-Jihadist Salafists may use their religiously proportioned pattern to communicate an appeal to the authority of God as benevolent and kind in rejection of the Jihadists ideology.

Similarly, another community of practice such as al-Qaeda may have a different pattern to their language. It is possible that their language may hinge more on religious appeals that are not purely admonitions but further appeals to the authority of God to seek to legitimize their religious authority.

CHAPTER 7. CONCLUSION

There are big questions that beg to be answered, getting at group level affiliation and intent. While we have been discussing how to get closer to those in an incremental manner. There are many smaller steps that must be taken before we can hope to venture an answer to those questions. That is why this thesis is constrained tightly. What I *am* asking is: Can MDA be used to extract statistical patterns from IS texts, and can those patterns be plausibly analyzed in terms of their communicative functions. This is done in service to the larger questions, but we are not there yet.

MDA is a powerful tool and can help us uncover linguistic variables that can then be analyzed with regards to their communicative function. What are the texts trying to *do* with language? The broad goal is that we can get at the big questions, for instance the example of *bless your heart*. What was found when we analyzed this phrase, it said something about both group affiliation and intent. *Bless your heart* shows that there is a pragmatic element that is otherwise missed by BoW because BoW does not interpret the data as to the communicate purpose. MDA is suited to teasing out these pragmatic elements because in tandem with its quantitative approach there is a necessary qualitative approach. A human analyst takes the data and derives an understanding of the texts under consideration. A human analyst is doing much more than merely parsing words on a page. MDA takes that ability and moves it up in the process allowing texts to be pre-processed with an eye toward the linguistic and cultural purposes of the language under observation.

When I analyzed the data from the corpus using MDA, I found that a pattern emerged. The pattern suggests that there is a negative correlation between political and religious sentences, meaning where one is present the other is absent. However, in sentences above the mean of 10.8 words in a sentence we see an increase in co-occurrence of religious and political speech. I suggested that this arises because there is a dual purpose of these sentences that requires a political assertion backed by a religious admonition. A qualitative analysis of this pattern appears to confirm the claim that there is a political assertion backed by religious admonition. The question remains, can we use this pattern to sketch a profile of Islamic State language. The answer is a resounding *probably*. This method has been effective in drawing out nuance in university language in the work of Biber (2006), so by inference we can open a *window* on what we can expect when a corpus of IS language is subjected to a full-scale MDA approach.

There are interesting connections between university language and the language of IS, though not a 1:1 mapping, so we can speculate about the similarities. First, there is a level of indoctrination that comes with university language. Biber found that it is difficult for the uninitiated (i.e. students) who are not prepared for the language used by their professors to understand it. Linguistic variables were shown to be indicative of different groups within academia. In breaking down Biber's linguistic features one by one he is very clear to point out which group uses which feature more. Such as the following: "the class of attitudinal adverbs… are usually used by instructors to mark personal attitudes," or, "students use [communication verb] constructions to identify the source of information." This distinction establishes a connection between linguistic traits and group identity that may seem ancillary to Biber's point but is crucial for my work here. It stands to reason that a similar form of indoctrination will be needed for IS language. Specifically, with regard to internal communication such as al-Naba' that is not intended for the general public or for recruitment purposes. There is a uniformitarian bend to this logic, where we assume that the similar processes are happening in both domains, however, we expect a bit of catastrophism as the language of IS will need to adapt quickly due to the stress and duress of its unique social situation. Looking at this language in comparison to other groups is the only way to truly test this assertion.

While this study did not include a comparative element, it is easy to see how the above assertion might play out in comparison with other groups. While registers differ from one another, it is reasonable to believe that groups will differ from one another in the register they use. It is important to be able to differentiate between similar groups. Can we do any better than saying this or that text is Jihadist? Can we say this text is IS versus al-Qaeda, etc.? How do we do this? We would want to compare linguistic variables between a comparison set of IS documents and al-Qaeda documents to see if there is variation between them that mark them as a distinct group. This is not the search for a shibboleth, but the search for understanding if certain groups use certain types of language in systematic ways. I have postulated that IS language is religio-political in sentences above the mean of words in the sentence. Groups that are anti-Jihadist might appeal from a point of religion in seeking to mark IS as acting against what Islam stands for. The anti-IS groups might rely more heavily on religious language while relying on political language only sparingly. One might want to sample the works of prominent anti-Jihadist Imam, Abdul Malik Mujahid, to see if this hypothesis holds.

The other issue that this thesis tackles is that the corpus is completely in Arabic. There is a gap in the literature on studies conducted on IS in Arabic. The only exception is perhaps Ingram, Whiteside, and Winter (2020). The use of Arabic marks this newsletter as internal facing, since much of the media produced for the Western audience is in local languages for maximum efficiency in recruitment. Inside the Caliphate, there is an expectation (if not a reality) that Arabic is the lingua franca. That is why it is extremely important to study the documents of any extremist group in their native language. So much can be missed when only focusing on what is easy to obtain and understand (English propaganda) than doing the hard work of translating and annotating the language of operation (Arabic propaganda). This thesis is one step toward rectifying this oversight in the literature. It is my hope that future research will follow suit and study Arab Jihadist groups like IS in Arabic.

7.1. COMMUNITIES OF PRACTICE

The work on Communities of Practice becomes highly important when we think of linguistic profiles. Using CoP to build speaker profiles provides an entirely new rubric that centers on group affiliations rather than static and broad demographics such as race, gender, and age. Communities of Practice allow us greater insight into how identity shapes speech. Recall the example of two children raised in the same household turn out so differently; One becomes a doctor and community leader, and she speaks like a doctor and as a community leader, the other becomes a gang member with long stints in prison and speaks like her fellow gang members. That is because of the communities that they choose to associate with. This is particularly easy to see in youth cultures, but all linguistic communities have their own ways of speaking. Linguistics professors speak differently from communication professors, spies speak differently from management consultants.

Sociolinguistic research reveals that youth cultures consciously create subcultures that seek to differentiate themselves based on intragroup linguistic behavior, either in pronunciation of certain words, creating their own slang or isolating members of an out-group by consistent use of in jokes (Eckert, 1989). Fought (1999), studied the linguistic behaviors of Hispanic students in Los Angeles and found that these behaviors, "group together through their participation in other practices." In fact, a speaker who utilizes a certain linguistic behavior of their subgroup is often "demonstrating their social position in wider social networks" (Mendoza- Denton 1999). Thus, if a new member of an online community that supports racist ideology wants to signal identification and social belonging in that community, they may begin increasing their use of racial slurs in their online communication.

Eckert (1989) came to the same conclusion in the populations of Detroit area high schools in which she identified the linguistic practices and attendant social activities of the self- termed Jocks (those invested in the high school community) and Burnouts (those who are dissociated from the high school community). An individual's level of participation within the social fabric of a community often belies their social status within that community. Seeking a higher status within the community can lead one from mere speech to actual performances of increasingly dangerous activities, such as drug use or willingness to participate in crime (Mendoza-Denton, 1999; Fought, 1999; Eckert, 1989). However, there are often shades of grey where loose connections float between the groups without ever fully being a part of either. They are often a vector for language variation and change by disseminating innovations from the shared repertoire of one group to another.

Islamic State functions in many ways like the youth cultures described by Eckert (1989). Largely there is a voluntary affiliation, and because of that there is a movement toward a shared goal and a shared repertoire. Though the same members with loose affiliations to the group pick up the shared repertoire without being full members of the group. In IS this is more sinister than in youth cultures. The "members" of the group who are not full adherents are based in a population that was conquered by IS. Their affiliation was not voluntary but mandated. This conquered population is forced to follow in the shared enterprise of IS and in doing so they become exposed to and begin to use the shared repertoire. Recall, earlier this thesis stated that it was proceeding with a definition of communities of practice that did not require voluntary membership. The premise laid out is an expansion of the work by Eckert and McConnell-Ginet (1992) and Wegner (1998) and stands as an addition to that conversation

One of the most striking similarities between youth cultures and IS is the fact that there is a shared repertoire among members of the group that requires a certain level of indoctrination to understand completely. I want to be clear Islamic State is not unique in this way. All communities of practice gain a shared repertoire. For example, Biber (2006) showed that one learns to speak and write like an academic. Eckert (1989) states that the sharpest distinction between Jocks and Burnouts, “lies in their use of language.” She says that the Burnouts are known for “… obvious, conscious differences [from Jocks]” which includes, “specialized vocabulary” (Eckert 1989). As mentioned previously, members of a community of practice learn their shared repertoire (i.e. specialized vocabulary) by being inducted via a social learning process. For IS this social learning process is largely enforced explicitly and less implicit. The voluntarily joined members as well as the conquered populace are policed by a cultural authority that enforces the “proper” way of life. Everyone is expected to conform, and many do for fear of death.

Aspirationally, this thesis hopes that by looking at language as a product of CoP that one is able to guard against the worst discriminatory practices of linguistic profiling. The intent is to not flatten the subjects into stereotypes. There is potential to avoid this flattening in two ways: 1) by professionalizing the process, implementing rigorous standards and responsibilities for the human analysts, and 2) by building on CoP. This thesis has shown that BoW *unintentionally* builds into it a type of discriminatory bias by nature of the set of "scare" words that are used to parse the document. When dealing with Jihadist terrorism, many of the scary words are Arabizations. If all Arabizations are scary it is a small leap to saying all Arabs are scary. Luckily there are alternatives to this approach. MDA used in conjunction with a BoW method is important to help mitigate inherent biases in the system because it does not rely merely on scary words, but looks more granularly at the morphological, syntactic as well as lexical to identify communities of practice.

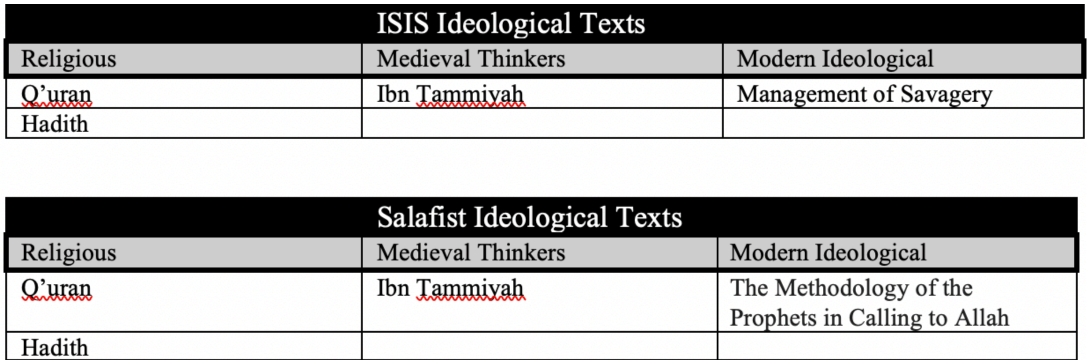
7.2. FUTURE WORK

While the pilot study has yielded some interesting results, the data is not robust enough to make any real generalizations. It needs to be confirmed by a full-scale MDA to confirm or falsify the findings. I have planned such a research study for my dissertation to carry on from this thesis. The model I have proposed in my discussion will be the guiding principles for that study. In it I will address my two big picture questions:

- Can we say something about group level affiliation?
- Can we say something about terrorist actor's intent?

The current study considers group level affiliation as a guiding principle, though nothing definitive was found. Intent has not been considered in this study; thus, it will be a novel addition in the dissertation.

MDA is a powerful technique, but it has a problem. As with Bag-of-Words, problems arise when the method is not constrained. That is why I plan to constrain my MDA study in two ways that will allow me to study my two research questions. First, group affiliation – when thinking about the structure of Jihadist groups, we notice that there is always an appeal toward authority, be it religious or political. Jihadists draw from a canon of religious texts, medieval scholars, and modern revolutionaries. To differentiate between group identities, I will constrain the search of forensic texts by using a corpus of authoritative texts as a comparative baseline. I will supplement these with several publicly available corpora such as the Arabic News Corpus (ANT) (Chouigui et al., 2017). I will supplement that with corpora built from the documents from several comparison groups. For the Islamic State I will take into account the Video and Audio propaganda as well as the written register of al-Naba' and (if available) forensic texts. I will replicate this corpus for al-Qaeda as well as a group of non-violent Salafists.



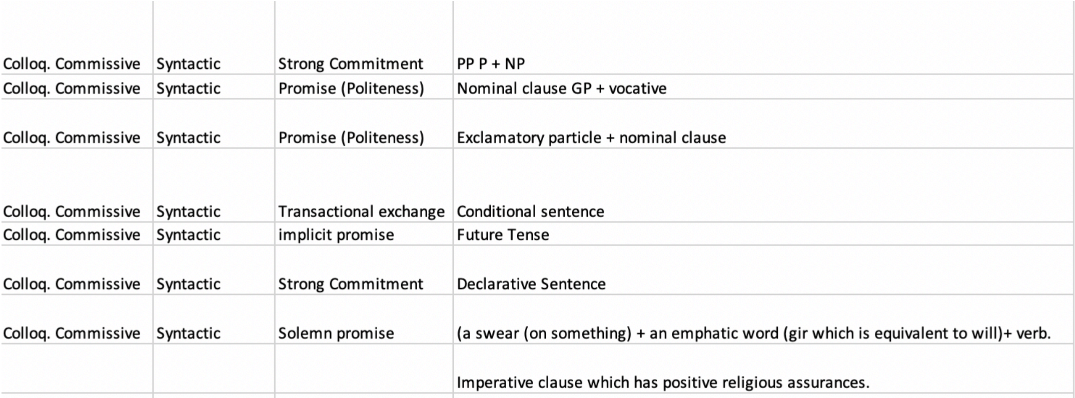
ISIS Ideological Texts

Religious	Medieval Thinkers	Modern Ideological
Q'uran	Ibn Tammiyah	Management of Savagery
Hadith		

Salafist Ideological Texts

Religious	Medieval Thinkers	Modern Ideological
Q'uran	Ibn Tammiyah	The Methodology of the Prophets in Calling to Allah
Hadith		

The second constraint on my methodology will allow me to attempt to identify intent in the language of Islamic State. I will utilize the Speech Act theory developed by Austin (1962) to search the corpus for the linguistic structure of Speech Acts. The theory has been used most often to speak about illocutionary force or perlocutionary effect, with locution being largely ignored. I will use the linguistic structure of the locution to uncover co-locations of use in the corpus. This should reveal the types of Speech Acts most prevalent in the corpus which should reveal the intent behind their use.



Colloq. Commissive	Syntactic	Strong Commitment	PP P + NP
Colloq. Commissive	Syntactic	Promise (Politeness)	Nominal clause GP + vocative
Colloq. Commissive	Syntactic	Promise (Politeness)	Exclamatory particle + nominal clause
Colloq. Commissive	Syntactic	Transactional exchange	Conditional sentence
Colloq. Commissive	Syntactic	implicit promise	Future Tense
Colloq. Commissive	Syntactic	Strong Commitment	Declarative Sentence
Colloq. Commissive	Syntactic	Solemn promise	(a swear (on something) + an emphatic word (gir which is equivalent to will)+ verb.
			Imperative clause which has positive religious assurances.

Biber (1988) states that MDA is by nature computational. Especially when considering a vaster set of linguistic features on a larger corpus, there is no way that one can effectively or efficiently hand count

The Multi-Dimensional analysis will follow very closely to the work of Berber-Sardinha and Pinto (2019) which updates the methods of Biber. Features will be selected based more on semantic domains such as tense and aspect in verbs, modal verb classes, semantic category of nouns (i.e. animate, concrete, etc.). Increasing the features will feed the functional analysis with enough inputs to be valuable in feature reduction. This will give a better sense of the dimensions of register variation. Finally, these data will be compared across registers within IS as well as other groups to more fully test the patterns and, in theory, create a profile of IS language.

7.3. FIN

In conclusion, this thesis has investigated if MDA can be used to extract statistical patterns from IS texts, and can those patterns be plausibly analyzed in terms of their communicative functions. It has taken steps toward using MDA to develop a profile of IS internal facing communication, but this was inconclusive. What was found was a pattern that is interesting and deserves further study. This research was limited due to the nature of the thesis, and the timelines involved. Due to this the corpus was much smaller than expected and the robustness of the data is of concern. With these limitations in mind, I did find that a certain pattern emerged where the majority of the sentences in the corpus were political in nature. However, sentences with words in the sentence above the mean of 10.8 show a prevalence of both religious and political speech. It was described qualitatively that this is because of political sentences are the primary goal, but they are punctuated by religious admonitions meant to encourage the political edict to be followed. I cannot generalize these findings to all internal IS language, the profile is worth studying further. I do think that this shows that internal facing IS communication has a different pattern than even external facing communication. This finding could be incredibly useful in the identification of malicious forensic texts post terrorist event.

REFERENCES

A Conversation with Barham Salih—YouTube. (n.d.). Retrieved November 16, 2019, from https://www.youtube.com/watch?v=WAqs2jTwb9Y

Abdelmeneim, S. (2008). *The Changing Role of Arabic in Religious Discourse: A Sociolinguistic Study of Egyptian Arabic*. Indiana University of Pennsylvania.

Adams, O., Makarucha, A., Neubig, G., Bird, S., & Cohn, T. (2017). Cross-Lingual Word Embeddings for Low-Resource Language Modeling. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 937–947.

al-Sayyid, K. (n.d.). *Hamza bin Abdul Muttalib*. Al-Islam. https://www.al-islam.org/hamza-bin-abdul-muttalib-kamal-al-sayyid/hamza-bin-abdul-muttalib

Al-Ayyoub, M., Alwajeeh, A., & Hmeidi, I. (2017). An extensive study of authorship authentication of Arabic articles. *International Journal of Web Information Systems*, *13*(1), 85–104.

Albirini, A. (2016). *Modern Arabic Sociolinguistics: Diglossia, variation, codeswitching, attitudes and identity* (0 ed.). Routledge.

Al-islam.org. (n.d.). https://www.al-islam.org/hamza-bin-abdul-muttalib-kamal-al-sayyid/hamza-bin-abdul-muttalib

Al-Taani, A., Msallam, M., & Wedian, S. (2012). A Top-Down Chart Parser for Analyzing Arabic Sentences. *The International Arabi Journal of Information Technology*, *9*(2).

Al-Wer, E. (1997). Arabic between reality and ideology. *International Journal of Applied Linguistics*, *7*(2), 251–265.

Asencion-Delaney, Y., & Collentine, J. (2011). A Multidimensional Analysis of a Written L2 Spanish Corpus. *Applied Linguistics*, *32*(3), 299–322.

Austin, J.L. (1975). *How to do things with words* (Vol. 88) Oxford University Press. Chicago.

Baugh, J. (2005). Linguistic Profiling. In *Black Linguistics: Language, Society and Politics in Africa and the Americas.*

Berber-Sardinha, T., & Pinto, M. (Eds.). (2019). *Multi-Dimensional Analysis: Research Methods and Current Issues*. Bloomsbury Academic.

Berger, J. M. (2018). *Extremism*. The MIT Press.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

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

Biber, D., & Conrad, S. (2019). *Register, Genre, and Style* (2nd ed.). Cambridge University Press.

Bolz, F., Schulz, D. P., & Dudonis, K. J. (2012). *The Counterterrorism Handbook* (4th ed.). CRC Press.

Botha, R. P. (2016). *Language evolution: The Windows approach*. Cambridge University Press.

Boulis, C., & Ostendorf, M. (2005). Text Classification by Augmenting the Bag-of-Words Representation with Redundancy- Compensated Bigrams. In *Proceedings of the Workshop on Feature Selection for Data Mining: Interfacing Machine Learning and Statistics*. SIAM Workshop.

Chouigui, A., Khiroun, O. B., & Elayeb, B. (2017). ANT Corpus: An Arabic News Text Collection for Textual Classification. *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 135–142. https://doi.org/10.1109/AICCSA.2017.22

Cohen, M. J. (2014). *Britain's moment in Palestine: Retrospect and perspectives, 1917-48*. Routledge/Taylor & Francis Group.

Combating Terrorism Center at Westpoint. (2020). *CTC Harmony Corpus.* https://ctc.usma.edu/

Craig, D. H., & Kinney, A. F. (2009). *Shakespeare, computers, and the mystery of authorship*. Cambridge University Press.

Dauber, C. E. (personal communication June 15, 2020).

Dauber, C. E, (nd.) Dabiq. [Forthcoming].

Deniz, A., & Kiziloz, H. E. (2017). Effects of Various Preproceesing Techniques to Turkish Text Categorization Using N-Gram Features. In *2nd International Conference on Computer Science and Engineering*.

Douglas, D. (1992). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, *26*(5–6), 331–345.

Eckert, P. (1989). *Jocks and burnouts: Social categories and identity in the high school*. Teachers College Press.

Eckert, P. & McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. *Annual review of anthropology, 21*(1), 461-488.

Egbert, J. (2019). Corpus Design and Representativeness. In *Multi-dimensional analysis: Research methods and current issues*. Bloomsbury Academic.

Egbert, J., & Biber, D. (2018). Do all roads lead to Rome?: Modeling register variation with factor analysis and discriminant analysis. *Corpus Linguistics and Linguistic Theory*, *14*(2), 233–273.

Egbert, J., & Staples, S. (2019). Doing Multi-Dimensional Analysis in SPSS, SAS, and R. In *Multi-dimensional analysis: Research methods and current issues*. Bloomsbury Academic.

Erard, M. (2016, August 5). ISIL demands many things of its top commanders, but good Arabic isn't always one of them. *Quartz*. https://qz.com/746732/isil-demands-many-things-of-its-top-commanders-but-good-arabic-isnt-one-of-them/

Everitt, B. (2006). *The Cambridge dictionary of statistics* (3rd ed). Cambridge University Press.

Ferguson, C. A. (1996). Diglossia revisited. In *Understanding Arabic* (pp. 49–67). American University in Cairo Press.

Firingal, E. & Hardy, J. A. (2019). From factors to dimensions: Interpreting linguistic co-occurrence patterns. In *Multi-dimensional analysis: Research methods and current issues*. Bloomsbury Academic.

Fought, C. (1999). A Majority Sound Change in a Minority Community: /u/- fronting in Chicano English. Journal of Sociolinguistics 3(1): 5-23.

Giles, H., & Marlow, M. L. (2011). Theorizing Language Attitudes Existing Frameworks, an Integrative Model, and New Directions [1]. *Annals of the International Communication Association*, *35*(1), 161–197.

Gumperz, J. (1968). The speech community. In *International Encyclopedia of the Social Sciences* (pp. 381–386). MacMillan.

Gray, B. (2019). Corpus Design and Representativeness. In *Multi-dimensional analysis: Research methods and current issues*. Bloomsbury Academic.

Haroro J Ingram, Whiteside, C., & Winter, C. (2020). Lessons from the Islamic State's 'Milestone' Texts and Speeches. *CTC Sentinel*, *13*(1), 11–21.

Hegghammer, T. (2006). Global Jihadism after the Iraq War. *The Middle East Journal*, 60(1), 11-32.

Holmes, J., & Wilson, N. (2017). *An introduction to sociolinguistics* (Fifth Edition). Routledge.

Indurkhya, N., & Damerau, F. J. (2010). *Handbook of natural language processing*. Chapman & Hall/CRC.

Ingram, H. J. (2018). Islamic State's English-language Magazines, 2014-2017: Trends & Implications for CT-CVE Strategic Communications. *Terrorism and Counter-Terrorism Studies*.

Jucker, A. H., & Kopaczyk, J. (2013). Communities of Practice as a Locus of Language Change. In *Communities of Practice in the History of English* (pp. 1–16). John Benjamins.

K. Gharaibeh, I., & K. Gharaibeh, N. (2012). Towards Arabic Noun Phrase Extractor (ANPE) Using Information Retrieval Techniques. *International Journal of Software Engineering*, *2*(2), 36–42.

Laub, Z. (n.d.). *CFR Backgrounders The Islamic State*. 6.

Le, Q. & Mikolov, T. (2014). Distrubuted representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning.* Beijing, China.

Lederer, E. (2020, February 3). *UN experts: Islamic State steps up attacks in Syria and Iraq*. https://abcnews.go.com/US/wireStory/experts-islamic-state-steps-attacks-syria-iraq-68734535

Liu, E. (2016). *How many words make a sentence* https://techcomm.nz/Story?Action=View&Story_id=106

Liu, H., Stine, R., & Auslender, L. (Eds.). (2005). *Proceedings of the Workshop on Feature Selection for Data Mining: Interfacing Machine Learning and Statistics*. SIAM Workshop.

Mahlouly, D., & Winter, C. (2018). *A TALE OF TWO CALIPHATES*. 49.

Makoni, S. (Ed.). (2003). *Black linguistics: Language, society, and politics in Africa and the Americas*. Routledge.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Marcellino, W. M., & Magnuson, M. (2019). ISIS Versus the United States Rhetorical Battle in the Middle East. In *Rhet Ops Rhetoric and Information Warfare*. University of Pittsburgh Press.

McEnery, T. (2019). *Arabic corpus linguistics*.

Mendoza-Denton, N. (1999). Sociolinguistic and linguistic anthropological studies of US Latinos. Annual Review of Anthropology 28: 375–95.

Meyerhoff, M. (2002). Communities of Practice. In *Handbook of Language Variation and Change,* 526–248. Wiley-Blackwell.

Migdadi, F., Badarneh, M. A., & Momani, K. (2010). Divine Will and its Extensions: Communicative Functions of *maašaallah* in Colloquial Jordanian Arabic. *Communication Monographs*, *77*(4), 480–499.

Mujahid, A. M. (Ed.). (2014). *Selected Friday Sermons*. Darussalam.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551.

Nini, A. (2017). Register variation in malicious forensic texts. *International Journal of Speech Language and the Law*, *24*(1), 99–126.

Owens, J. (2013). *The Oxford handbook of Arabic linguistics*. Oxford University Press.

Penelope Eckert, & McConnel-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology*, *21*(1), 461–488.

Qasim, S., & Shakir, A. (2016). *Linguistic Variation of Pakistani Fiction and Non-Fiction Book Blurbs: A Multidimensional Analysis*. 22.

Racial Profiling. (2020). https://www.aclu.org/issues/racial-justice/race-and-criminal-justice/racial-profiling.

Rashid, A., & Mahmood, M. A. (2019). Linguistic Variations across Disciplines: A Multidimensional Analysis of Pakistani Research Articles. *Global Social Sciences Review*, *IV*(I), 34–48.

Robinson, M. D., & Dauber, C. E. (2019). Grading the Quality of ISIS Videos: A Metric for Assessing the Technical Sophistication of Digital Video Propaganda. *Studies in Conflict & Terrorism*, *42*(1–2), 70–87.

Roy, O., & Schoch, C. (2017). *Jihad and death: The global appeal of Islamic State*. Oxford University Press.

Ryding, K. C. (2005). *A reference grammar of modern standard Arabic*. Cambridge University Press.

Sardinha, T. B., & Pinto, M. V. (Eds.). (2019). *Multi-dimensional analysis: Research methods and current issues*. Bloomsbury Academic.

Seldin, J. (2020). US Officials Uncover True Identity of New Islamic State Leader. *Voice of America*.

Sousa-Silva, R. (2018). Computational Forensic Linguistics: An Overview of Computational Applications in Forensic Contexts. *Language and Law*, *5*, 27.

Spranger, M., & Labudde, D. (2013). *Semantic Tools for Forensics: Approaches in Forensic Text Analysis*.

Suleiman, Y. (2013). *Arabic in the Fray*. Edinburgh University Press.

Sultan, K. (2016). The "ISIS" Online Media War: A Construction of Ideology through Terrorism. *Pakistan Journal Peace & Conflict Studies*, *1*(2), 1–14.

Terror in Sri Lanka. (2019). CNN.Com. https://www.cnn.com/interactive/2019/04/world/sri-lanka-attacks/.

Thurston, A. (2015). *The Islamic State's intellectual genealogy (and what you need to read to understand it)*. 7.

Wegner, E. (1998). *Communities of Practice*. Cambridge University Press.

Winkler, C. (2016). Visual Images: Distinguishing Daesh's Internal and External Communication Strategies. In *Countering Daesh Propaganda: Action-Oriented Research for Practical Policy Outcomes*. The Carter Center.

Winter, C. (2015). *The Virtual 'Caliphate': Understanding Islamic State's Propaganda Strategy* (p. 52). Quilliam.

Wood, G. (2015). What ISIS Really Wants. *The Atlantic*, 26.

Yong, A. G., & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94.

Zelin, A. Y. (2015). *Picture or It Didn't Happen: A Snapshot of the Islamic State's Official Media Output*. *9*(4), 13.

Zelin, A. Y. (2020). *Jihadology.net.* https://jihadology.net/.

Zhang, X., & Choo, K.-K. R. (2020). *Digital Forensic Education: An Experiential Learning Approach*.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1–4), 43–52.

Zitouni, I. (2014). *Natural language processing of Semitic languages*. Springer.