Fall 12-10-2019

Essays on Technology in Presence of Globalization

Joshua Madden
jmadden4

Essays on Technology in Presence of Globalization

BY

Joshua Edward Madden

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree

Of

Doctor of Philosophy

In the Robinson College of Business

Of

Georgia State University

GEORGIA STATE UNIVERSITY
ROBINSON COLLEGE OF BUSINESS
2019

ACCEPTANCE

This dissertation was prepared under the direction of the Joshua Madden Dissertation Committee. It has been approved and accepted by all members of that committee, and it has been accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Business Administration in the J. Mack Robinson College of Business of Georgia State University.

Richard Phillips, Dean

DISSERTATION COMMITTEE

Ephraim McLean (chair)
Veda Storey (chair)
Richard Baskerville
Jonathan Trower

ABSTRACT

Essays on Technology in Presence of Globalization

BY

Joshua Edward Madden

June 14, 2019

Committee Chairs: *Ephraim McLean*
Veda Storey

Major Academic Unit: *Computer Information Systems*

Technology has long been known to enable globalization in ways previously not thought possible, with instantaneous communication allowing members of organizations all across the globe to communicate and share information with little to no delay. However, as the effects of globalization have become more prominent, they have in turn helped to shape the very technologies that enable these processes. These three essays analyze three examples of how these two processes – globalization and technological development – impact one another. The first looks at a national policy level, attempting to understand how increased possibilities for inside leakers can force governments to consider asylum requests. The second analyzes the issue at the level of corporations, attempting to understand how and why business leaders choose to hire individuals from other countries. The third and final essay analyzes the issue at the most micro level, studying a potential application that could help analyze linguistic factors that have taken a more prominent role in a more globalized society.

DEDICATION

To my uncle Mike Madden and my friend Lexie Pettit.

ACKNOWELDGEMENTS

I would like to thank my co-chairs, Dr. Ephraim McLean and Dr. Veda Storey, for their guidance and patience throughout this process. I would also like to thank those who served on my committee, Dr. Richard Baskerville, Dr. Dmitry Zhdanov and Dr. Jonathan Trower, for their support and feedback. It is no exaggeration to say that this would not have been completed without their support.

In addition, I would like to thank Hannah Peterson, Alicia Plemmons, Walker Rhine and Zirun Qi, as well as the rest of my colleagues and cohort at Georgia State University and Salisbury University, for their specific contributions and feedback.

I would also like to thank my father, Joe Madden, my mother, Tammy Madden, my sister, Kelsey Madden, my grandmothers, Lee Madden and Betha Regan, and my grandfather, Charlie Regan, as well as my entire extended family for their support and interest throughout the decade (plus some) that I have spent pursuing this degree. Finally, I would like to thank my crew of lifelong friends for the moments of humor provided throughout this process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The advancement of information technologies has enabled businesses to become more global than was previously possible. Using new technologies, individuals can be working on the same project simultaneously with other individuals who may live in different regions and who may hold different cultural values. This phenomenon has a great many benefits and yet also raises unique issues for individuals and corporations that now work in a more global environment (Scheve et al. 2001). This research proposes to study globalization and its interdependence with information technology. To do so, this research analyzes the phenomenon at three levels: the policy level, the corporate level, and the application level.

By conducting this research at the policy level, the corporate level, and the application level, this research is able to analyze the relationship between globalization and emerging technologies in a comprehensive way. These three studies combine to attempt to answer this overarching research question: In what ways are emerging technologies and the phenomenon of globalization interacting?

The first study, which researches the phenomenon at the policy level, studies a high-level phenomenon related to globalization and information technology. It analyzes two case studies concerning information leakers (Julian Assange and Edward Snowden) and the policy issues that they have created by seeking asylum in foreign nations in order to avoid prosecution. The objective of this research is to analyze these two case studies and provide a new theoretical analysis based on the evidence provided by their situations using Porra et al.'s historical research method technique (2014). A "two-good" theory is presented to explain why or why not nation states grant asylum requests. Ultimately, a game-theoretical model is proposed to help

understand the motivations of policy makers at the national level and to make predictions for future decisions.

The second study focuses on the corporate level and studies the impacts of globalization on hiring and funding decisions, specifically with regards to international presence by measuring the extent of offshoring and outsourcing. The study uses secondary data and analyzes it using regression analysis to control for the size and budgets of the companies overall and seeks to understand whether international companies make statistically different hiring and funding decisions as compared to American companies. The study also distinguishes between STEM (science, technology, engineering, and mathematics) based companies as compared to industries that do not self-identify as belonging to a STEM field.

The third study focuses on the application level, taking a design science approach to integrate linguistic factors more thoroughly into traditional forms of analysis. By creating a new algorithm based on existing research from linguistics and mathematics, this study generates a new method for studying linguistic factors within large bodies of text. The study then proposes and tests two applications for this algorithm. The first shows how the algorithm can be useful from a forensic linguistics perspective in order to determine common authorship of large bodies of text. Forensic linguistics is defined in a narrow sense as the use of language as evidence (Coulthard and Sousa-Silva 2016) and identifying the author of a text through the words they used would be useful for such an environment. The second shows how the algorithm could be used to improve error detection in automatic translation software.

1.1 Linguistic factors

Linguistic factors can have a profound impact on how technology is designed and used, particularly in this era of increasing globalization. However, as technology becomes more

integrated into daily life, technological factors are beginning to shape language as well. Researchers have theorized about the nature of this interrelationship since at least the 1940s (Gerr 1942). Cottrell and Sheldon (1963) noted the complex relationship between culture, language, and technology as early as the 1960s and this discussion has continued into this century (Van Pelt 2002). In addition to studying how language has shaped technology, research has increasingly noted the degree to which languages, such English, are being shaped by technology (Ho 2006). This paper will explore the impact of globalization and emerging technologies on each other and will show that the causality is not one way, but instead is a unique relationship where each factor continually shapes and changes the other.

A crucial reason that the relationship between technology and language is becoming more important is because of globalization. Technology has enabled an increase in direct foreign investment and involvement in other companies around the globe (Reid 1991), which has made language differences more relevant than they once were. This, in turn, has meant that designers must consider these linguistic differences when designing new technologies.

According to Reid (1991), internalization of technology in the 1950s and 1960s was primarily a result of American companies building manufacturing plants overseas. In the 1970s, this began to shift, with American companies beginning to increase their direct foreign investment and shifting other aspects of their companies (such as research and design) overseas as well. Then, during the 1980s, Japan and Western European nations began to capitalize on this trend and increased their foreign investment as well (Reid 1991). This meant that, in addition to Americans investing in other countries, other countries began to invest in American companies (Reid 1991). Globalization led to a steady increase in the proportion of traded goods in world output throughout the 20th century (Storper 1992). In more recent years, this interplay between

globalization, technology, and language led to innovations in supply-chain management and a change in how many companies operated, at a fundamental level, in terms of the technology they used in order to fulfill orders and create products (Johnson 2006).

In the broadest sense, language continues to shape technology by virtue of technology having a linguistic component. As many technologies – such as e-mail and other technology-based forms of communication - are at least somewhat dependent on written language – as opposed to spoken language or direct numbers – designers are frequently turning to language resources to shape the technologies they are designing. Mariani (2005) notes that language resources are now seen as "crucial for the development of written and spoken language processing systems." Bird and Simmons (2003) also note this pressure, emphasizing that "multiple communities depend on language resources, including linguists, engineers, teachers, and actual speakers."

Yet as technology becomes more essential to daily life, it is also shaping language. Kern (2006) explains that the "rapid evolution of communication technologies has changed language pedagogy and language." Cook (2004) finds that new technologies are correlated with changes in language. Squires (2010) explores this further, looking at the phenomenon of Internet-specific language patterns. Ambrose (2001) goes as far as to suggest that early technologies may have shaped human evolution with regards to how the brain processes language.

1.2 Emerging technologies

1.2.1 Emerging technologies in education

One area where the interplay between technology, globalization, and language is perhaps most evident is education. Globalization has enabled universities to work together as well as with educational institutions at other levels (such as K-12 in the United States) in ways that were

previously not possible, creating a unique situation where globalization is shaping the educative process (Abeles 1998). Globalization has impacted even the material that is actually being taught and the resources being used in education, as exemplified by impacts made on music education in Singapore (Lum 2008). However, the relationship actually factors in a third factor when it comes to technology and globalization, which is language. As academic communities become less homogenous linguistically, they are beginning to see linguistic factors play a much larger role in shaping the technology.

This relationship between language and technology within education is perhaps best described by Warschauer (2002), who asks, "Is technology a tool for language learning, or is language learning a tool with which people can access technology?" Research has emphasized the relationship between technology and language in education since at least the 1970s (Carlson 1976). Technology is an increasingly important tool for early-language learning, particularly for those beginning to acquire a second language (Cunningham and Redmond 2008) and for improving literacy (Wrigley 1993). Technology has also been used to further integrate literature into language education (Kraemer 2008).

This relationship between technology and language education has shown how technology can shape language. Lum (2008) notes that the impacts of technology and globalization on education are actually helping to shape the identities of the students using them, going as far as to look at the specific impacts on technology, media, and ethnicity. Language and education have both been noted to be impacted by a lack of technological development in nations and cultures less reliant on technology (Crawford 1990). Lack of technological development can lead to a widening gap between developed and developing nations in terms of language education (Batley 1991) – a factor that could potentially lead to slowed growth in terms of the number of

individuals learning certain languages and increased growth in terms of the number of individuals learning others, such as English.

Conversely, language is impacting the design of technologies for education. Waters and Gibbons (2004) note the relationship between design languages and instructional technologies. Lum (2008) discusses how Singaporean students in music see an interplay between their ethnicity and the technological environment in which they are learning. In related research, Saini (2009) explores the relationship between language and technology in artistic design for advertising, showing that the interplay between language, technology, and design is not necessarily limited to instructional technologies.

1.2.2 Emerging technologies and communication

Another area where technology is perhaps most shaping language is in the language preservation and revitalization movement – essentially efforts to preserve and expand the base of speakers for lesser-used languages. Technology is an increasingly important tool to help revitalize dying languages (Eisenlohr 2004).

Language and technology are also shaping each other by virtue of their joint relationship with globalization. While technology has facilitated instantaneous global communication, differences in language have become more problematic in the sense that individuals who once would have never met are now required to communicate with each other in certain contexts. On the other hand, this has led to the phenomenon of global languages, where certain languages – most notably English – are becoming the lingua franca for international communication. Researchers have previously noted the roles that language and technology are playing in this new need to reach a global public (Williams 2005).

In some perhaps extreme cases, technology is actually replacing the human voice as a means for communicating through language. For individuals with communication-related disorders, human language technology (HLT) applications are an increasingly viable option for replacing oral speech as a primary form of conveying language (Ruiter et al. 2012). Keeting and Mirus (2003) show how members of the deaf community are using computer-mediated video communication in order to interact with one another.

1.2.1.1. Relevance of education and communication technologies

Emerging technologies in education and communication can serve as a potential example for any research into emerging technologies, particularly research with any kind of linguistic component. The research presented in this paper will explore emerging technologies with a strong linguistic or communicative component. All three essays are concerned with emerging technologies within information systems – which means, by definition, they are systems designed to convey information. The first essay concerns a technology designed to provide education and information to mass audiences (wikis), the second concerns communication skills in a global environment, and the third looks at linguistic component theory within design science research.

1.3 Proposed Research

Globalization and technology have long been theorized to have an interdependent relationship, where advancements in technology have allowed for increased globalization and, in turn, increased globalization has led to the development and adoption of new technologies. This research proposes to explore this phenomenon at three levels: the policy level, the corporate level, and the application level.

At the policy level, this research proposes to study the phenomenon of information "leakers" such as Julian Assange and Edward Snowden and their need to request asylum in order to avoid potential prosecution by targeted nations. These asylum requests have created a unique policy problem for the nations receiving the requests – by granting the requests, they are rejecting an international norm upholding a general condemnation of information leaking. However, by granting these asylum requests, the nations can potentially make small utility gains relative to other nations – they have disincentivized a specific leaker from targeting them in the future, even if they have encouraged the behavior on a global level.

This research analyzes two case studies – Julian Assange and his attempt to gain asylum in Ecuador as well as Edward Snowden and his ongoing requests for asylum in Russia – and takes a game-theoretic modeling approach to show how these two case studies suggest that the policy issues faced by policymakers in government is similar to a well-known game-theoretic model known as the stag hunt.

At the corporate level, this research proposes to study whether or not the United States is, in fact, more open than other countries to hiring foreign workers and outsourcing labor and what impacts this openness might have on other budgeting concerns, such as disaster recovery spending, information technology [IT] spending, and spending on training with regards to communication skills. Using data collected on hundreds of companies representing a substantial percentage of global GDP, this research proposes to test several hypotheses suggested by the literature using a regression analysis to determine whether or not this increased openness exists and to determine the extent to which it is impacting budgeting and hiring habits.

Finally, at the application level, this research uses a design science approach to develop a new artifact – an algorithm that can be used to determine the degree of similarity between

disparate texts. The development of this algorithm is outlined and two unique applications are proposed.

The first application is for forensic linguistics, showing that the algorithm could be used to determine the authorship of different works. The second is an application for improving error detection in automatic translation software. By using the proposed algorithm to determine language-wide commonalities between two different languages (in the case of this research, English and German), we can determine the "true ratio" of similarity between two languages and use this to compare against the results from automatic translation software, allowing error detection to be improved in the future.

These three studies combine to form an analysis of globalization's impacts on technological development and to show in detail how globalization is changing technical processes at all levels of analysis – from nationwide policies to specific applications.

2 POLICY LEVEL

Information leakers, such as Julian Assange and Edward Snowden, two of the primary leakers behind some of Wikileak's major information releases, are polarizing figures, being praised by some for their efforts to bring transparency to governments while being criticized by others for revealing classified information and potentially placing operatives in danger. "Leaking" refers to situations in which individuals release information to the public that is not widely known or declassified for public consumption. For holders of sensitive information, such as governments, information leakers present a wide variety of new security risks to assess and solve. Assange and Snowden are part of a larger trend as the number of information leakers is likely to steadily increase in the future. This paper proposes that most of the behavior with regards to whether or not to grant Assange or Snowden asylum can be explained with the development of a "two-good" theory to explain the motivations of the concerned nation-states. After these two "goods" are presented and explained, an attempt is made to analyze these patterns using the stag hunt model, a game-theoretic model designed to analyze situations in which the maximum payoff requires cooperation from all participants.

This trend presents a variety of issues for government officials and policymakers, with information leakers creating unique security problems for nations attempting to store classified or sensitive information. Rather than presenting an external security threat, information leakers are generally an internal one – having been vetted and given access to the sensitive information, which can also make the security breaches more difficult to detect than traditional externally-based hacking. The fact that information leakers are requesting asylum from other nations adds

an additional layer to the issue, forcing nations to re-evaluate and adapt their current strategies with regards to existing digital policies.

Governments, citing national security concerns, are generally not supportive of information leaking, seeing it as a violation of existing laws designed to prevent the release of classified information. In some cases, such as with Snowden, this can result in the leaker being indicted with federal charges (NBC News, 2014). In other cases, such as the Assange case, there is a fear of being charged (BBC News, 2015; Rothwell & Ward, 2017; WikiLeaks, 2017). With regards to the cases of Assange and Snowden, both were either charged with crimes or sought asylum from other nations in order to avoid potential criminal prosecution. This causes a dilemma for the countries where asylum is being sought and a unique situation within the larger context of digital policy and eGovernance.

eGovernance is defined as the use of information and communication technologies by governments and other public organizations to enhance their ability to govern (Palvia and Sharma 2007, Bedi et al. 2001, Holmes 2001, Okat-Uma 2000). In order to use information technologies for these purposes, information must be stored in digital formats – and even less securely, occasionally the resources are stored directly online – thus making them susceptible to leaks from insiders with access to this information. Digitally stored information can be more easily copied and shared than was possible for previous generations, and governments must consider this fact when deciding what information to store in digital formats. Generally speaking, governments do not want their employees leaking information, thus this situation is an example of the larger context that needs to be considered in any serious look at eGovernance.

Granting asylum may encourage individuals to leak information, thus normalizing this behavior and encouraging leaks. This could encourage leakers within their own country, thus

causing problems for the country choosing to grant asylum. Other counties may retaliate by granting asylum to insiders who choose to leak information about their country, causing a spiral where countries are forced to continually grant asylum as retaliation for a previous asylum request granted. For example, Assange was granted asylum by Ecuador, and until recently resided in the Ecuadorian embassy in London. The United States, as the subject of many of WikiLeaks' revelations, and the United Kingdom, are forced into an awkward position by Ecuador's decision and may decide to retaliate in some way in the future.

Granting asylum, however, likely discourages that specific individual from leaking information about the country where they have been granted asylum. For example, since Snowden has been granted temporary asylum in Russia, one would assume that, even if Snowden were to somehow gain access to classified information about Russia, he is unlikely to leak information about Russia and risk jeopardizing his legal status within Russia. In the case of Assange, Ecuadorian officials have publicly stated that his current asylum is conditional on him "not making statements that could impact their relationship with their allies" (BBC News, 2015). These countries are explicitly and publicly taking steps to reduce the security risk that these information leakers pose to their own internal systems.

More broadly, however, the country granting asylum gains a large degree of control over the specific leaker. They might do this for a variety of reasons, including to embarrass the home nation of the leaker or in order to have a "bargaining chip" for future negotiations. In this sense, the asylum granting nation gains at the margin by now having a great degree of control over a potentially prominent foreign national. Subsequently, it appears that the two concerns nation-states have when deciding whether or not to grant asylum to an information leaker are whether or not they can gain control over the individual leaker and whether or not they can gain relative to

other nation-states. By presenting a theory that analyzes these two "goods," much of their behavior can be explained.

This general situation closely resembles a game-theoretic model known as the stag hunt. The stag hunt game is based on a situation in which all players must cooperate in order to get the maximum payoff (the "stag") and if any player defects, he will receive a smaller payoff (the "hare") while all other players who attempted to cooperate will receive no payoff at all. In the stag hunt, all players are best off when they all cooperate and gain maximum payout by doing so. However, players who choose to defect are relatively better off than some other players in games where at least one player attempts to cooperate (Skyrms, 2004).

This is similar to the situation where all countries are better off if they all choose not to grant asylum to any information leakers, upholding an international norm, which discourages the behavior overall. Should potential information leakers be aware that they are unlikely to be granted asylum, this has the effect of decreasing the internal security risk that authorized individuals present to policy makers. However, some countries can gain at the margin by granting individual leakers asylum. While these countries have prevented the international community as a whole from gaining by discouraging the behavior, they have at least ensured that an individual leaker will be unlikely to leak their own information and will have made a small gain relative to other countries who have chosen to discourage information leakers by rejecting all asylum requests. This situation likely applies more to individuals like Assange, who run a leaking organization, than it does to individual leakers such as Snowden or Chelsea Manning, the former U.S. Army member who first sent several videos and diplomatic cables to Assange and Wikileaks. For example, it is possible that Assange might be sent information regarding Ecuador, where he has been granted asylum. In a sense, countries that grant asylum to information leakers

may have increased security risks globally, but they have decreased the security risk that a specific individual poses to them.

In a broader sense, the nations that choose to grant asylum to individual leakers have gained control over a prominent individual that they could then use for their own gain. For example, if the leaker is able to continue leaking information about their home country without fear of prosecution, this information might hurt a geopolitical rival. For example, with Snowden being able to continue to leak information about the United States, this may benefit Russia's aim to hurt the United States, a geopolitical rival. In addition, the asylum granting nation now has control over an individual that could be used as a "bargaining chip" in future negotiations – for example, Ecuador could potentially agree to extradite Assange as part of a larger agreement in the future.

The objective of this research is to model the situation of information leakers who seek asylum. To do so, the case studies of Assange and Snowden are first analyzed, showing how these situations resemble a stag hunt game-theoretic model. The contribution of the research is to present the theoretic model and to show that, using the stag hunt model, the behavior of the countries granting asylum can be explained and future behavior predicted. In addition, another contribution of the paper is to present a new, two-good theory that can be used to explain this behavior. Once this theory is developed and outlined, an attempt is made to exemplify it using game theory. By adopting a model from existing game-theoretic work and applying it for the first time within information systems research, this research analyzes the potential impacts from choosing whether or not to grant asylum and presents recommendations for eGovernment policy makers and analyzes potential security concerns.

2.1 Related Research

This research is based on the stag hunt, a game-theoretic model designed to help explain social dynamics (Skyrms, 2004). This model does not appear to have been applied within the context of information systems, offering a new contribution to information systems research. An overview of the stag hunt model is presented below, followed by relevant research on game theory-based research in information systems and security research within information systems.

2.1.1 The Stag Hunt

The stag hunt is a game-theoretical model often used to model social dynamics (Skyrms, 2004). The game is based on a story told by Jean-Jacques Rousseau in his "A Discourse on Inequality," in which he proposes a situation where several men go on a hunt (Rousseau, 1755; Skyrms, 2004). The hunt begins with the goal of catching a deer, where each of the hunters must stay focused at their post in order to successfully catch the deer (Rousseau, 1755; Skyrms, 2004). However, "if a hare happened to pass within reach of one of them, we cannot doubt that he would have gone off in pursuit if of it without scruple" (Rousseau, 1755).

This anecdote has been subsequently taken into game theory and shaped into something now referred to as the stag hunt game, in which two (or more) players are given a choice to either hunt stag (cooperate) or to hunt hare (defect) (Skyrms, 2004). Cooperation amongst all players leads to the successful capture of the stag, which is the highest possible payoff for all participants. However, should any player defect and hunt hare, the stag will elude capture, resulting in no payoff for those who remained focused on the stag and a lesser payoff for those who chose to hunt the hare. This results in the model shown in figure 1.

	Stag (Cooperate)	Hare (Defect)
Stag (Cooperate)	2 , 2	0 , 1
Hare (Defect)	1 , 0	1 , 1

Figure 1 Example of potential Stag Hunt payoffs (adapted from Skyrms 2004))

The payoffs can vary from model to model, but several principles remain the same: the highest possible payoff can only result from total cooperation; defecting always results in a payoff that is less than the total cooperation but may be more than zero; and those who cooperate when at least one other player defects receive a payoff of zero. The game is usually shown as a two-player game, but it can be played with an infinite number of players provided these other principles remain the same (Skyrms, 2004). Cooperation must be universal in order for the stag to successfully be hunted – if even one "hunter" leaves their post in search of the hare, the stag eludes capture.

2.1.2 Game Theory in Information Systems

Game theory has been used to study a variety of contexts within information systems (Zhu, 1999; Papadimitriou, 2001; Kaser, 2002; Dellarocas, 2003; Zhu, 2004). Research in information systems that uses game theory often draws on the relevant research from economics, amongst other fields (Papadimitriou, 2001; Dellarocas, 2003). However, as a whole, the field has been relatively slow to adopt and exploit game-theoretic approaches within its research (Li & Whang, 2002).

Dellarocas (2003) uses a game-theoretical approach to explain online feedback mechanisms. Citing Papadimitriou, Dellarocas argues that the Internet creates social dynamics that require a more game-theoretic approach (Papadimitriou, 2001; Dellarocas, 2003). This echoes an argument made by Skryms (2004), who argues that game-theoretic approaches are necessary for the explanation of social dynamics and phenomena.

Dellarocas (2003) emphasizes the value of game-theoretic models within the field of information systems to study behaviors and "how these behaviors evolve over time if all players are simultaneously pursuing their own interests" (Dellarocas, 2003). Players do change their own strategies based on what they believe other players are likely to do (Wilson, 1985; Dellarocas, 2003). Game-theoretic models are uniquely suited to this context of ongoing decision making; and information systems, as a social science, is not immune from this need, allowing researchers in information systems to use a game-theoretic approach when there is a clear social component to the problem (Dellarocas, 2003).

Dellarocas (2003) also argues that "paradigms, such as decision theory and simulation, and empirical and experimental studies, are natural complements to game theory, both for qualifying these models and for adapting them to account for the complexities of the 'real world' and the bounded rationality of actual human behavior" (Roth, 2002; Dellarocas, 2003). This overlap between approaches commonly used in information systems and game theory further illustrates the value of using game-theoretic approaches to study information systems contexts (Dellarocas, 2003).

Evidence within the information systems literature suggests that during the first iteration of a repeated game, it is not uncommon for some players to accept lower or even negative profits while the other players learn whether they other players will cooperate or defect (Dellarocas, 2003). It is worth noting here that the stag hunt can be played as a repeated game and that many findings from game theory have come through the use of repeated games (Skyrms, 2004).

All of this combines to suggest that game theory can be a valuable approach to understanding and explaining phenomena within information systems. Game theory has been shown to help explain current and future behavior within appropriate contexts. To the best of our

knowledge, however, the stag hunt model has not been applied in the context of information systems, offering a new contribution to information systems research.

2.2 Case Studies

This research presents two case studies of individuals who have revealed damaging information and have sought political asylum from other countries in order to avoid criminal prosecution. The first is the case of Australian citizen Julian Assange, who until recently resided in the Ecuadorian embassy in London, having been granted asylum by the Ecuadorian government. He has since been released by the Ecuadorian government and was arrested by British authorities and remains in their custody. The other is American citizen Edward Snowden, who is living at an undisclosed location in Moscow after having received temporary asylum status in Russia.

2.2.1 Julian Assange and Ecuador

Julian Assange was born in Townsville (The-Courier-Mail, 2010), a city in Queensland, Australia, and is an Australian national. Known originally as a hacker in Australia, Assange later rose to more prominence as the founder of WikiLeaks (The-Courier-Mail, 2010), an organization dedicated to anonymously revealing information provided by information leakers from around the world.

Founded in 2006 (Rothwell & Ward, 2017), WikiLeaks has published a variety of classified and leaked information concerning the United States government and its officials. In 2007, the organization revealed documents related to the United States' protocols for dealing with prisoners in Guantanamo Bay (The Telegraph, 2011). In 2010, WikiLeaks worked with

Bradley Manning[1], an enlisted member in the United States Army, to release leaks that included details of alleged war crimes in Iraq by United States forces (Sanchez, 2013; Rothwell & Ward, 2017). In 2013, WikiLeaks released documents obtained by Edward Snowden regarding the extent of the surveillance activities engaged in by the United States' National Security Agency and its international partners (Harley, 2015; Rothwell & Ward, 2017).

In August 2010, Assange was accused of assaulting women in Sweden (BBC News, 2015). Swedish authorities issued an arrest warrant for Assange on August 20, 2010, although the warrant was later withdrawn; and, instead, a request to detain for questioning was approved in November 2010. An international arrest warrant was later issued (BBC News, 2015). Assange claims the allegations are false (Rothwell & Ward, 2017) and that the accusations are part of a politically motivated move against him (BBC News, 2015).

Sweden did not formally charge Assange because the investigation was ongoing, and instead sought to extradite Assange for questioning (BBC News, 2015). Assange fought the extradition attempt in the United Kingdom, where he was residing at the time. He was unsuccessful, with the courts ruling in May 2012 that Assange could be extradited to Sweden to face questioning (BBC News, 2015).

Assange subsequently sought asylum from Ecuador in June 2012 by entering the Ecuadorian embassy in London. This request was granted in August 2012 (BBC News, 2015). Ecuador's Foreign Minister, Ricardo Patino, stated that the request was granted because of the potential for Assange's human rights to be violated, were he to be extradited (BBC News, 2015). Patino also stated that the asylum request was granted conditionally, with one of the conditions

[1] Bradley Manning is now known as Chelsea Manning.

being that Assange would not make “political statements that could affect our relations with friendly countries" (BBC News, 2015).

Assange’s primary concern was that if he were extradited to Sweden, he could eventually be sent to the United States, where he claims he could face the death penalty for his involvement with WikiLeaks (BBC News, 2015). This is despite the fact that the United States has not yet publicly charged Assange with any crimes. Furthermore, authorities in Sweden have said that they will not extradite individuals to a country where they could face the death penalty (BBC News, 2015). The United States Justice Department has also expressed skepticism as to whether they would be able to effectively prosecute Assange for espionage, which is the most serious charge he could potentially face (Rothwell & Ward, 2017). According to WikiLeaks’ official Twitter account, authorities in the United Kingdom have not revealed whether they have an extradition warrant for Assange from the United States (WikiLeaks, 2017).

The United Kingdom refused to grant safe passage for Assange to leave the country while he was living in the Ecuadorian embassy (BBC News, 2015). The authorities in the United Kingdom indicated that they would arrest him for violation of his bail conditions (BBC News, 2015; Rothwell & Ward, 2017) if he ever tried to leave the Ecuadorian embassy on his own[2]. WikiLeaks expressed concern that the United Kingdom had not revealed whether it had an extradition warrant for Assange from the United States (WikiLeaks, 2017), which left them concerned (at the time) that Assange could be extradited to the United States if he were arrested by authorities in the United Kingdom[3].

[2] This did not occur as, instead, on April 11, 2019, the Ecuadorian government formally withdrew Assange’s asylum and invited Scotland Yard into the embassy (BBC News 2019).
[3] These concerns were eventually shown to be valid as Assange is now facing a potential extradition to the United States (PA Media 2019).

On May 17, 2017, Swedish authorities dropped the charges against Assange, saying that they concluded the investigation because "at this point, all possibilities to conduct the investigation are exhausted" and that they could not expect to receive the assistance from Ecuador necessary to continue the investigation (Adley and Travis, 2017). WikiLeaks has claimed that the United Kingdom has not revealed whether it has an extradition warrant for Assange from the United States (WikiLeaks, 2017).

In March 2018, the Ecuadorian embassy cut off Assange's Internet access (Henley 2018). Ecuador's government said it did so "because Assange had breached 'a written commitment made to the government at the end of 2017 not to issue messages that might interfere with other states'" and that "Assange's recent behaviour on social media "put at risk the good relations [Ecuador] maintains with the United Kingdom, with the other states of the European Union, and with other nations" (Henley 2018). This was done after Assange had tweeted skepticism of a British accusation that Russia was responsible for a recent poisoning attack of a former agent (Henley 2018). A former Greek finance minister, Yanis Varoufakis, and music producer Brian Eno, blamed international pressure, stating that "Only extraordinary pressure from the U.S. and the Spanish governments can explain why Ecuador's authorities should have taken such appalling steps in isolating Julian" (Henley 2018).

In July 2018, Greenwald (2018) reported that Ecuador was withdrawing Assange's asylum and handing him over to British authorities. This led to speculation that the Ecuadorian government was doing so in response to pressure from the United States and in hopes of improving its relationship with the U.S. (Long 2018). In April 2019, relations between Assange and the Ecuadorian government further deteriorated after Ecuadorian President Lenín Moreno

blamed WikiLeaks for allegations of corruption and publishing personal photos of his (Associated Press 2019).

A few days later, on April 11, 2019, the Ecuadorian government formally withdrew Assange's asylum and invited Scotland Yard into the embassy (BBC News 2019). Assange was then arrested and charged in the United Kingdom with failing to surrender to the court (BBC News 2019). On May 1, 2019, Assange was found guilty and sentenced to 50 weeks imprisonment within the United Kingdom (BBC News 2019). As of September 14, 2019, Assange is still imprisoned within the United Kingdom and is facing a potential extradition to the United States (PA Media 2019).

2.2.2 Edward Snowden and Russia

Edward Snowden was born in North Carolina on June 21, 1983 (NBC News, 2014; NBC News, 2014b) and is an American citizen. Both of his parents worked for the United States government. Snowden himself attempted to enlist in the United States Army Special Forces, although he was unsuccessful due to an injury in a training incident where he broke both legs (Greenwald, MacAskill & Poitras, 2013; NBC News, 2014; NBC News, 2014b).

In 2006, Snowden was hired by the United States' Central Intelligence Agency as a computer systems administrator (NBC News, 2014). As part of the process of being hired by the CIA, Snowden received a top-secret clearance from the United States government (NBC News, 2014). From 2007 to 2009, Snowden was posted in Geneva, Switzerland in a diplomatic position for the CIA as an information technology and cyber-security expert (Greenwald, MacAskill & Poitras, 2013; NBC News, 2014b).

After serving in the diplomatic position, Snowden later worked for two different private intelligence contractors, Dell and Booz Allen Hamilton (the latter of which Snowden claims to

have worked for only to gain access to additional documents (Lam, 2013) who were engaged in work alongside the National Security Agency (Greenwald, MacAskill & Poitras, 2013). During this time, Snowden maintained his top-secret security clearance (NBC News, 2014).

While employed by these contractors, Snowden had access to, and downloaded, secret documents that revealed that the extent of some of the National Security Agency's surveillance programs, which included the monitoring of American citizens' Internet and telephone activity (NBC News, 2014). Snowden downloaded approximately 20,000 documents; he was undetected by exploiting security loopholes (NBC News, 2014).

In 2012, Snowden began to contact journalists, eventually leaking a series of documents to multiple journalists in 2013 (NBC News, 2014; NBC News, 2014b). Snowden took a leave of absence from work, telling his employer that he needed time off to receive treatment for epilepsy, and left for Hong Kong in May 2013 where he met with several journalists (Greenwald, MacAskill & Poitras, 2013; NBC News, 2014b). Snowden stated that he initially chose Hong Kong "because 'they have a spirited commitment to free speech and the right of political dissent'" (Greenwald, MacAskill & Poitras, 2013). The *Guardian* – a major British newspaper – elaborated on this, saying that Snowden "believed that it was one of the few places in the world that both could and would resist the dictates of the U.S. government" (Greenwald, MacAskill & Poitras, 2013). Snowden claimed that the documents revealed that the National Security Agency had hacked into computers in Hong Kong as well as mainland China (Lam, 2013).

In June 2013, The *Guardian* and the *Washington Post* revealed information from the documents provided to them by Snowden. Later that month, Snowden was revealed as the source for the documents and was subsequently fired by his employer (NBC News, 2014b). The *Guardian* revealed Snowden's identity at his request, with Snowden himself saying "I have no

intention of hiding who I am because I know I have done nothing wrong" (Greenwald, MacAskill & Poitras, 2013).

In late June 2013, Snowden was charged with “unauthorized communication of national defense information” and “willful communication of classified communications intelligence information to an unauthorized person” (NBC News, 2014; NBC News, 2014b). These are violations of the 1917 Espionage Act, which may also be used to charge Assange. Snowden faces up to 30 years in prison if convicted of all counts, although additional charges could still be added (NBC News, 2014).

Snowden left Hong Kong on June 23, 2013, intending to reach Ecuador after a stopover in Russia (NBC News, 2014b). The United States government criticized Hong Kong authorities for allowing Snowden to leave Hong Kong (Lam, 2013).

Because Snowden’s passport was rescinded by the United States government while he was in Russia, he was unable to leave for Ecuador and lived for a month in the airport while he applied for asylum in Russia (NBC News, 2014b). Since August 2013, Snowden has been living at an undisclosed location in Russia, where he was granted a one-year stay for temporary asylum (NBC News, 2014). Eventually, Snowden was granted asylum to stay in Russia until 2020 and he will soon be eligible to apply for Russian citizenship (Oliphant, 2017). A spokesperson for the White House criticized Snowden for these actions, saying that he had “fled into the arms of an adversary and has sought refuge in a country that most recently made a concerted effort to undermine confidence in our democracy” (Oliphant, 2017).

It remains unclear how many documents Snowden downloaded – estimates range from 200,000 to 1.7 million. NSA officials have stated that they remain unsure of exactly how many documents Snowden took and what they are (NBC News, 2014).

2.3 Historical research method development

Using the historical research methodology framework as outlined by Porra et al. (2014), this research attempts to find overarching patterns throughout historical events to guide the development of a theory. Previous literature in information systems has suggested that scholars might benefit from looking "in the field" at history that might help for us to understand IS phenomena (Bryant et al. 2013).

Porra et al. (2014) outline four key areas within their framework that historical researchers should consider and explain. The first is at the paradigmatic level, where researchers should analyze and state the meta-theoretic assumptions guiding their research (Porra et al. 2014). This research is most inspired by the functionalist paradigm, which Porra et al. (citing Burrell and Morgan 1979) describe as being most "concerned with providing explanations of the status quo, social order, social integration, consensus, need satisfaction, and rational choice. It seeks to explain how the individual elements of a social system interact together to form a working whole." This research seeks to understand why the leakers and nations have acted as they do, assuming that they are acting rationally and in response to an existing social order.

Porra et al.'s (2014) second identified level is that of the approach level, which is slightly more concerned with how the actual research is taking place than the paradigmatic level. This research takes what Porra et al. (2014) describe as a pragmatist approach, as our focus is on developing theories with practical value that could be used to predict future behavior, rather than dealing in abstract ideas with minimal real-world application.

The third level in Porra et al.'s (2014) framework is that of the method level, which, citing Iivari et al. 1998 is defined as "a codified set of goal- oriented 'procedures' which are

intended to guide the work and cooperation of the various parties involved in the process." As mentioned previously, this study uses two historical case studies.

The fourth level is concerned with the techniques used to gather, critique and write about the evidence (Porra et al. 2014). Because of the multiple tiers inherent with this process, this author has found that the technique level is more fluid and harder to define or classify than the other levels listed above. This research ultimately was most concerned with gathering evidence for the case studies, determining patterns, and then communicating what was found to others. Finally, it did not fit the 8-step outline suggested by Porra et al. (2014) as this research took a relatively uncritical view of the evidence that was collected, choosing instead to focus more on the events themselves than possible biases in the reporting.

Ultimately, taking inspiration from Palmer and Morgan's *A Theory of Foreign Policy* (2011), this paper proposes that the behavior taken by nation-states at the policy level with regards to deciding whether or not to grant asylum can be framed in terms of a "two-good" theory – meaning that the nation-states will make their decision based on the pursuit of two different goals. In both case studies outlined above, the nations deciding whether or not to grant asylum – Ecuador and Russia – did so based on two concerns. The first was whether or not they could gain control over the individual leaker. The second was whether they could gain an advantage relative to other nations. The rest of this paper will outline these two goods, then present a potential model using a game-theoretic approach.

2.3.1 The first "good" - control over the individual leaker

In both of the case studies outlined above, the nation-state granting asylum did so at least in part to gain control over the individual leaker. This motivation manifested itself in a variety of ways. Ecuador was quite explicit in stating this as one of their major motivations. Ecuador's

Foreign Minister Patino outright stated that the asylum request was granted conditionally, with one of the conditions being that Assange would not make “political statements that could affect our relations with friendly countries" (BBC News, 2015). It is clear that part of the reason Ecuador chose to grant Assange asylum was in order to gain further influence and control over Assange and, subsequently, the Wikileaks organization as a whole. This was further illustrated by the fact that Ecuador cut off Assange’s internet access once they believed it was negative to their international relationships (Henley 2019) and that they ultimately revoked his asylum after blaming his organization for publishing negative stories about the Ecuadorian president (Associated Press 2019).

Russia’s attempts to gain with regards to this good are less overtly stated, but it is reasonable to suspect that at least part of the reason for granting Snowden asylum is further access to the documents that he had stolen from the United States. As mentioned above, it remains unclear how many documents Snowden downloaded – estimates range from 200,000 to 1.7 million. NSA officials have stated that they remain unsure of exactly how many documents Snowden took and what they are (NBC News, 2014). There can be little doubt that one of Russia’s primary motivations was to gain control over Snowden in order to get access to these documents – Snowden’s ability to resist such demands is severely limited given his potential need to gain Russian citizenship (Oliphant, 2017).

2.3.2 The second “good” – gains relative to other nation-states

The other motivating factor appears to be a desire to make gains relative to other nation-states by gaining further bargaining power and harming potentially unfriendly countries. It is clear, for example, that the United States government sees Russia’s decision to grant Snowden asylum as something that comes at their expense. As mentioned previously, a spokesperson for

the White House criticized Snowden for his pursuit of asylum, saying that he had "fled into the arms of an adversary and has sought refuge in a country that most recently made a concerted effort to undermine confidence in our democracy" (Oliphant, 2017). Russia and the United States have had, at best, a tenuous relationship and often see each other as adversaries, as evidenced in the statement above. By gaining access to previously classified American documents and controlling the future movement of an individual that the United States wants to prosecute, Russia has made gains in overall utility relative to the United States.

Similarly, Ecuador's decision to grant Assange asylum was motivated at least in part to improve their own reputation and power in the international arena. Ecuador's Foreign Minister, Ricardo Patino, stated that the request was granted because of the potential for Assange's human rights to be violated, were he to be extradited (BBC News, 2015). This statement implies that by granting Assange asylum, Ecuador was attempting to increase their own credibility with regards to human rights and to weaken the credibility of the United States on the issue. It also appears that the decision to revoke Assange's asylum was motivated at least in part by a desire to improve bilateral relations between Ecuador and the United States (Long 2018).

2.3.3 The "two-good" theory

By analyzing these situations using Porra et al.'s historical research method approach (2014), one begins to see that these are the two most commonly recurring motivations in what are otherwise relatively unique cases with very different decision makers. Ecuador and Russia have little in common in terms of governmental structure or history, yet they behave in remarkably similar ways in this context. This suggests that inferences can be drawn from these two cases to present a theory that could be used to explain and predict future behavior by other nation-states.

Subsequently, this paper takes inspiration from Palmer and Morgan's *A Theory of Foreign Policy* (2011), and proposes that the behavior taken by nation-states at the policy level with regards to deciding whether or not to grant asylum can be framed in terms of a "two-good" theory – meaning that the nation-states will make their decision based on the pursuit of two different goals. The two-goods are gaining control over the individual leakers themselves and making gains relative to other nation-states in the international arena.

A two-good theory such as this one is not necessarily linear – instead of proposing that decisions are motivated by the pursuit of one-factor, a two-good theory proposes that nation-states are motivated by two potentially competing factors to pursue the highest overall utility gains (Palmer and Morgan, 2011). In this case, the decision on whether or not to grant asylum is concerned with how much utility can be gained from gaining control over an individual leaker and how much can be gained relative to other nation-states.

In both of these case studies, the nations granting asylum arguably made utility gains with regards to both "goods," but the primary motivation appears to be unique to each case. Ecuador's motivation appears to be more aligned with the first good, gaining control over Assange and Wikileaks in order to prevent him from leaking information that might harm Ecuador's relationship with other nations. Russia's motivation seems more aligned with the second good, attempting to use Snowden's asylum request to embarrass the United States, gain access to its classified information and to prevent them from prosecuting someone they believe to be a threat to their government.

In both cases, however, both "goods" were clearly considered and it is not inconceivable to think of potential situations where a gain with regards to one good might be outweighed by losses with regards to the other. In Russia's case, Russia and the United States have an

adversarial relationship, so there was little to be lost for Russia with regards to its relationship with the United States. If Russia had a much more friendly relationship with the United States, however, it is easy to imagine a very different outcome for the asylum request. In this hypothetical situation, Russia might still gain control over the individual leaker (Snowden) but the utility gained here is minimal – other than the documents Snowden already has, he is unlikely to gain access to anything else of value to Russia, nor is he likely to ever have access to sensitive internal Russian documents. The utility lost by harming a good relationship with a major world power would easily outweigh these marginal utility gains and, in this situation, Russia would likely deny Snowden's request.

It is possible that this is part of the reason that Snowden left Hong Kong, fearing that friendly relationships between the United States and Hong Kong might outweigh any utility gains the region hoped to make by controlling Snowden. Snowden initially chose Hong Kong, stating that he did so "because 'they have a spirited commitment to free speech and the right of political dissent'" (Greenwald, MacAskill & Poitras, 2013) but later left the region, causing the United States government to criticize Hong Kong authorities for allowing Snowden to leave Hong Kong (Lam, 2013). It is possible that part of the reason that Snowden chose to leave Hong Kong, a region with relatively friendly ties to the United States, and to move towards Russia, a nation with a far less friendly relationship with the United States, may have been because he believed Hong Kong would act in its own interest and extradite him to the United States in order to avoid losses on the "second good" – even if they might have made marginal gains on the "first good."

Ultimately, all of this combines to suggest that nation-states, when deciding whether or not to grant information leakers asylum, do so by analyzing their gains along two goods – control

over the individual leakers and gains relative to other nation-states – calculate the overall utility and then grant asylum only if they believe that the gains (or losses) on these two goods combine to a net positive utility gain.

2.4 Example of a potential model

Both Assange and Snowden, facing stated or potential charges from the United States government, have chosen to seek asylum from other nations in the hopes of avoiding criminal prosecution. This presents an interesting dilemma for the countries where asylum is being requested. Given the theory presented above, it seems that it closely models some of the motivations in a game-theoretic model known as "the stag hunt." Subsequently, this paper will attempt to provide an example of what a game-theoretical model of this two-good asylum request theory might look like.

On one side, granting the asylum request could help to ensure that the asylum seeker does not leak damaging information about the country where asylum is being sought. In the case of Assange, the Ecuadorian government has overtly stated that Assange's asylum is conditional on him not making "political statements that could affect our relations with friendly countries" (BBC News, 2015). Whether or not this statement would apply to publishing a compromising leak regarding Ecuador or its allies remains somewhat unclear. In the case of Snowden, he has only been granted temporary asylum and may attempt to gain Russian citizenship in the future (Oliphant, 2017), two factors that may discourage him from revealing any information that is potentially damaging to the Russian government. These countries are explicitly and publicly taking steps to reduce the security risk that these information leakers pose to their own internal systems.

However, by granting asylum requests for any information leakers, countries run the risk of creating an international norm where information leaking is acceptable. If information leakers understand that they are likely to be granted asylum somewhere in the world, it could help to encourage the behavior. All countries have some degree of classified information that they do not choose to reveal to the public. This is generally for national security reasons, but sometimes for more wide ranging reasons. Because of this, the best scenario for all countries is for no one to grant asylum, thus discouraging anyone from revealing any damaging information. The maximum payout is universally decreasing the security risk posed by information leakers. An example of the various potential payoffs is shown in figure 2 below.

	Stag (Cooperate)	Hare (Defect)
Stag (Cooperate)	2 , 2	0 , 1
Hare (Defect)	1 , 0	1 , 1

Figure 2 Stag Hunt example (adapted from Skyrms (2004))

This situation closely matches the game-theoretical approach known as the stag hunt, where the highest possible payout for any player is if all players cooperate. However, relative gains can be made by defecting in situations where at least one player cooperates. This leads to three possible situations: 1) there can be universal cooperation; 2) universal defection; or 3) a mixed result where some players cooperate and others defect. Each of these situations closely parallels a possible situation in the dilemma as to whether or not asylum requests for information leakers should be granted and will be explored in detail below.

2.4.1 Universal Cooperation

Figure 3 represents the universal cooperation payoff, which is what results when all players choose to cooperate and reject all asylum requests from information leakers.

	Cooperate (Reject Asylum Request)
Cooperate	**All information Leakers discouraged**

Figure 3 Universal cooperation payoff

In this first situation, all countries choose not to grant asylum requests for any information leakers. This leads to the highest possible payoff (analogous to the stag in Rousseau's hunt (Rousseau, 1755; Skyrms, 2004)) for all players. When no country grants any asylum requests for information leakers, this discourages all information leakers from engaging in this sort of behavior. Because of this, the risk of any country facing a leak of damaging or sensitive information is greatly reduced. In a sense, global security has been improved because all potential information leakers have been discouraged.

This is the best situation for all countries, because, instead of having to deal with each information leaker individually, all information leakers are discouraged, and global security risks are decreased. However, it is important to note that this payoff can only be achieved through total cooperation. If any one country chooses to defect by granting an asylum request, this international norm is no longer upheld and information leaking is encouraged on an individual basis. Therefore, countries will be forced to deal with each individual information leaker.

2.4.2 Universal Defection

Figure 4 represents the universal defection payoff, which is what results when all players choose to defect, granting asylum to individual leakers on case-by-case basis. In this context, no players reject any asylum requests.

	Defect (Grant Asylum Request)
Defect	**Individual leakers discouraged**

Figure 4 Universal defection payoff

In this second situation, all countries choose to grant asylum requests for at least one information leaker. The highest possible payoff of all information leakers being discouraged and global security norms being upheld is lost as information leakers now know there is a chance their asylum requests may be granted by at least one country.

However, there is a still a payoff – albeit a less significant one – for all countries because they have at least discouraged the individual leakers to whom they have granted asylum from leaking information specifically about them. For example, in the case of Assange, the Ecuadorian government has outright commented that Assange's asylum is conditional on him not making "political statements that could affect our relations with friendly countries" (BBC News, 2015). Ecuador is making it clear that they will not accept any increased security risk from Assange himself and that if he does present one, his asylum may be revoked.

This has the effect of discouraging Assange personally from leaking any damaging information concerning Ecuador or its allies – Assange is now in situation where his asylum (and

subsequently his ability to avoid criminal prosecution) would be revoked if he revealed sensitive information about Ecuador. Assange has not necessarily been discouraged from the behavior as a whole, however, and is apparently free to continue leaking information about other nations, so long as that information does not constitute a political statement that could impact Ecuador's relations with "friendly countries" (BBC News, 2015). In a sense, global threats to security presented by Assange have not been decreased, but the security threat that Assange specifically poses to Ecuador has been greatly reduced.

Snowden faces a similar situation, where, if somehow he were to come into possession of classified information concerning Russia, he is less likely to reveal that damaging information about Russia as his asylum has been granted on a temporary basis (Oliphant, 2017) and could be revoked should he damage the reputation of the Russian state. More broadly, the Russian government has gained on the margins because it now has a high degree of control over an American national – the information Snowden leaked continues to be potentially damaging to the United States, which is a geopolitical rival to Russia, and Russia can use Assange as a "bargaining chip" in future negotiations if it chooses to do so.

Again, the behavior overall has not been discouraged – Snowden continues to leak documents regarding the conduct of the United States' National Security Agency. Again, global threats to security presented by Snowden have not been decreased, but the security threat that Snowden specifically poses to Russia has been greatly reduced.

If the game is modeled with Russia and Ecuador as the only two players, the two countries have not received the highest possible payoff in that they still have to be concerned about the security threat from information leakers globally, because information leakers are aware that they may be able to receive asylum somewhere. However, at the individual level,

security threats have been marginally reduced as Ecuador has discouraged Assange from revealing any damaging information about the Ecuadorian government and its allies and Russia has discouraged Snowden from revealing any information about the Russian government. Because of this, the two players may have missed on the highest possible payoff, but they do still receive a lesser payoff.

2.4.3 Mixed Result

Figure 5 shows the payoff from mixed results, which is what occurs when some players choose to defect (by granting at least one asylum request) and other players attempt to cooperate (by rejecting all asylum requests by information leakers).

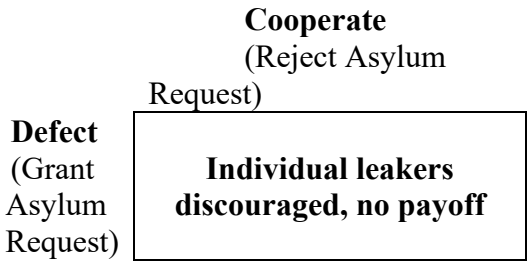

Figure 5 Mixed result payoff

The final possible payoff is a mixed result where some nations choose to cooperate and reject all asylum requests while other nations have chosen to defect and grant at least some asylum requests. Because the players have chosen different strategies, they will receive differing payoffs.

Similar to the universal defection scenario, no players receive the highest possible payoff. In this case, all information leakers are not discouraged because they know that an asylum request may be granted somewhere. The threat to security that information leakers present

globally is not decreased. Only one player needs to defect in order for the highest payoff to be prevented. All information leakers can only be discouraged if all countries choose to cooperate. In this mixed result scenario, because at least one player has chosen to defect, the highest possible payoff is still lost, despite the fact that other players had attempted to cooperate.

Unlike the universal defection scenario, however, not all countries receive the lesser payoff of discouraging individual information leakers. Those countries that shelter a leaker still receive this benefit. Although they may not have discouraged all information leakers, they have at least discouraged those leakers that were granted asylum in their country. While the global threat to security from information leakers remains, the players who defected have marginally reduced the security threats to themselves.

However, those who chose to cooperate by granting no asylum requests have not even received this lesser payoff. They have not discouraged any individual leakers from revealing damaging information about them because they have not actually granted asylum to any individual leakers. As a result, they receive no payoff at all. The international norm of discouraging information leaks has not been upheld, nor have any individual information leakers been discouraged. The global threat to security remains and as no individuals have been discouraged from leaking their information, these countries have not even marginally reduced the security threats they face.

This is a situation where one player can make a gain relative to the others. Even in the universal cooperation scenario, where all countries gain the maximum possible benefit, there are no relative gains as all players gain equally. However, in this case, a player who chooses to defect makes a relatively higher gain than those who chose to cooperate. This is because those who chose to cooperate received no payoff, whereas those who defected make a small gain.

2.4.4. *Stag as player*

The most commonly used game-theoretic model is that of the prisoner's dilemma, which differs slightly from the stag hunt with slightly different playoffs although the motivations can, at times, be similar. The stag hunt, however, applies better to this scenario because the stag can actually be modeled as a player in their own right – the asylum seekers are making decisions as well, thus making this more of a three-dimensional model than a two-dimensional model. Thus, while the first two dimensions were presented above, we now need to expand the model to be three dimensional to reflect the fact that the asylum seeker also has a decision to make and a role to play.

In our initial, two-dimensional model, we have three possible outcomes: universal cooperation amongst countries, universal defection, and a mixed result. Now we must model what would happen to the stag (the asylum seeker) in each of the three scenarios. The payoff for the stag is relatively straightforward – we must assume that they want to leak, but do not want to go to prison. Thus, they will first avoid choosing any situation in which they go to prison (which we will treat as a -2 payoff) and then will only choose to leak in a situation in which they are granted asylum. We will treat the first asylum being granted as a +1 payoff (as the first asylum is the most important for avoiding prison) and every subsequent asylum granted as a +0.5 payoff (as it gives them the benefit of choosing where to live and prevents a single country from having universal control over them).

Ultimately, the asylum seeker really has two choices: leak or not leak. We assume, for the sake of this model, that the leaker will always attempt to seek asylum and avoid imprisonment. Thus, if the asylum seeker chooses to leak, this is the first possible set of outcomes:

	Cooperate	Defect
Cooperate	Payoff: 2, 2, -1	0, 1, 2
Defect	Payoff: 1, 0, 2	0, 0, 2.5

Figure 6 Asylum seeker chooses to leak

In this case, the asylum seeker only goes to prison if all countries cooperate and choose not to grant asylum. The countries see the international norm upheld and the asylum seeker gains one point for leaking, but loses two for imprisonment. In the mixed outcomes, the asylum seeker gains one point for leaking and another for being granted asylum. One country gains marginally by having exclusive control over the leaker, while the other gains nothing as the international norm is upheld. In the final situation where all parties defect, the asylum seeker gains one point for leaking, one for the first asylum and 0.5 for the second asylum. All other nation players lose as international norms are not upheld, nor do they gain exclusive control over the individual.

The only other option the asylum seeker has is to not leak in the first place, thus creating the following options:

	Cooperate	Defect
Cooperate	2, 2, 0	2, 2, 0
Defect	2, 2, 0	2, 2, 0

Figure 7 Asylum seeker chooses not to leak

In this case, because no leaking ever occurred, it does not matter what the national policies are with regards to asylum requests as no asylum request will ever be made. The international norm of not granting asylum to information leakers is upheld in all scenarios (as no nations have an opportunity to violate the international norm even if they wished to) and the asylum seeker remains at zero as they do not leak, do not go to prison or do not seek asylum. In this case, the policies and decisions of the nations to do not matter as all possible outcomes offer the same solution.

2.4.4. *Nash equilibria*

It is worth noting that depending on how one structures this example game, the number of resulting Nash equilibria may vary from model-to-model. This is because there is not necessarily a "correct" application for this theory, but instead one can produce a game-theoretic model that they believe best fits the context in question. As some areas will emphasize the decision making of the leaker more than others, for example, some modelers may choose to include the stag as a player, while others may be dealing with a context that is less concerned with the decision making of the leaker, perhaps as a result of them having already made a decision.

2.5 Implications

Modeling these case studies against the stag hunt model is useful from an explanatory perspective to help understand why the various players (nations such as Russia and Ecuador) have acted in the ways they have. However, this research has implications beyond the obvious explanatory benefits.

There are three important aspects of this research worth exploring further. First, the relationship between information systems and game theory will be further explored. Secondly, the implications for future predictive research will be discussed, using this proposed model as an example as to how predictions could be made for future behavior. Finally, there will be a brief discussion of this study's limitations and its subsequent impacts for future research.

2.5.1 Information systems research and game theory

Previous research in information systems has shown that the social aspects of information systems create social dynamics that require a more game-theoretic approach (Papadimitriou, 2001; Dellarocas, 2003). This argument is echoed in philosophy, where scholars such as Skyrms

have argued for the value of game-theoretic approaches in terms of explaining social dynamics and phenomena related phenomena (Skyrms, 2004).

Information systems researchers attempting to study reputation phenomena have an advantage in that they can draw from over 20 years worth of research from economics and game theory in general (Dellarocas, 2003). Dellarocas (2003) states that "in most settings where reputation phenomena arise, equilibrium strategies evolve over time as information slowly leaks out about the types of the various players," which appears to be the case here. In addition, Dellarocas writes that "most game-theoretic models of reputation formation assume that stage game outcomes (or imperfect signals thereof) are publicly observed," which is also the case here, as notable information leakers and their asylum requests are often widely publicized events, giving all players the change to observe the actions of other players (Dellarocas, 2003).

Therefore, approaching this situation and others like it from the context of a game-theoretic approach allows researchers to learn more about reputation phenomena within eGovernance and information systems.

2.5.2 Predicting future behavior

Dellarocas highlights the value of game-theoretic models within the field of information systems to study behaviors and "how these behaviors evolve over time if all players are simultaneously pursuing their own interests" (Dellarocas, 2003). Players do change their own strategies based on what they believe other players are likely to do (Wilson, 1985; Dellarocas, 2003). This is important to note here because the decision about whether or not to grant asylum requests for information seekers is not a one-time decision – it evolves over time, with nations adjusting their strategies based on their own interests. Subsequently, this means that models such as the stag hunt can help to predict future behavior.

In this case, the world is essentially entering a second iteration of this game. In the first iteration, multiple players (such as Russia and Ecuador) chose to defect from upholding an international norm discouraging information leaks. Now that these early decisions have been made, the other players (the rest of the international community) now have an opportunity to adjust their strategies.

Skyrms (2004) argues that, in many situations, the most rational strategy for players who have cooperated in a mixed result game is to shift strategies and defect in the next round. That seems to be reasonable here – countries such as the United Kingdom and the United States, which have taken a strong stance against information leakers and subsequently received no payoff (reduced security risks), have little reason to continue attempting to cooperate. All players missed out on the highest possible payoff where all information leakers are discouraged because at least two players have defected. Information leakers now know that they may be able to get asylum in Ecuador or Russia if they are potentially facing criminal prosecution in their home countries.

The countries that choose to cooperate are also at a relative disadvantage – Russia and Ecuador have received minor payoffs, while countries that attempted to cooperate have received none. Because of this reasoning, it seems likely that in the future, an increasing number of countries will choose to defect by granting asylum requests to information leakers. There are also likely to be more requests – information leakers are no longer discouraged on the aggregate level as they know that they may be able to receive asylum somewhere. With more requests for asylum and less reason to attempt to cooperate, it seems likely that, barring some major change in the behavior of the defectors, nations will no longer collectively attempt to discourage

information leakers but will instead race to make marginal gains with individual leakers by granting them asylum.

This study presented just one example of how game-theoretical approaches can be used to model current behavior and predict future behavior within an eGovernance (and larger information systems) context. Other examples are likely to be explored in further research.

2.5.3 Implications for security

Skyrms' (2004) work suggests that those players who did not defect in the first round will do so in later rounds. This is due to the fact that they believe they are unlikely to "capture the stag" in future rounds as they see no reason that Russia and Ecuador will begin to cooperate and cease defecting. If this is the case, the logical step for other countries is to attempt to reduce their own security risks by granting asylum to specific leakers, particularly prominent ones who might be interested in leaking their information in the future.

However, this could change if other players believe that the defectors will change their behavior. It is possible, although unlikely, that the international community could create some sort of enforceable treaty with one another stating that they will not grant asylum to information leakers from another country. This hypothetical situation is unlikely to happen due to a variety of obvious concerns – whether or not the treaty will be effective, how it would be enforced, whether all countries would sign such a treaty, and so forth.

The global security threat posed by information leakers appears to be here to stay and may constitute a new global standard. Unless other players believe that those players who defected will not do so in the future, they are all likely to defect themselves sometime in the future. If this becomes an international norm, the best situation for each country is to grant asylum to any information leaker that they see as being a potential threat to themselves in the

future; they may not be able to decrease the security risk that information leakers create globally, but they can at least discourage security threats to themselves that these individuals pose.

2.5.4 Limitations and future research

One limitation of this study is the relatively small number of cases – this study focuses on two well-documented and highly-publicized cases of information leakers being granted asylum and their subsequent behaviors. Cases of such prominence are relatively rare – not all information leaks have as widespread impacts as those facilitated by Assange and Snowden. Subsequently, there are few cases with the amount of information necessary to thoroughly analyze in this context.

However, this presents an opportunity for further research. So far, there has been a relatively limited number of players – relatively few information leakers have sought asylum in another country. As new cases arise – which seems likely, given current trends – they can be further analyzed to see if they fit the necessary specifications to analyze through the lens of the stag hunt model.

Another limitation is that it is largely this is the first iteration of the game that has occurred and what may occur in later iterations is still a matter of speculation. Whether or not the predictions outlined in this study come true remains to be seen. Because this game is still in its early stages, it is possible that we have not yet seen the true impacts on security policy and that the attitude of security experts may change in the future as they receive more information about the threat that information leakers actually pose.

This presents another opportunity to prove the value of using game-theoretic approaches to analyze eGovernance issues and issues in the larger field of information systems as a whole. Continuing to study this phenomenon of information leakers and their attempts to gain asylum

could help to further refine how scholars within the field of information systems study how "behaviors evolve over time if all players are simultaneously pursuing their own interests" (Dellarocas, 2003). With eGovernance becoming more prominent (Palvia and Sharma 2007, Bedi et al. 2001, Holmes 2001, Okat-Uma 2000), there are likely to be more contexts worth exploring where the interests of various nations must be understood and analyzed. Game-theoretic modeling, as done here, presents a potentially worthwhile approach for conducting such an analysis.

2.6 Conclusion

The number of information leakers is likely to increase in the future, given current trends. This presents a variety of issues from an eGovernment perspective, with information leakers creating unique security problems for nations attempting to safeguard classified or sensitive information. The fact that information leakers are requesting asylum from other nations adds an additional layer to the issue, forcing nations to re-evaluate and adapt their strategies for eGovernance.

Not only are the information leakers themselves forcing nations to re-evaluate their strategies for information security but nations are now also being forced to react to the decisions of other nations. Game-theoretic approaches are necessary to study contexts such as this, where multiple players are making decisions based on their own motivations and the motivations of other players and nations are continually adapting and refining their strategies based on the actions of the other players.

This study shows how a real world situation where information leakers are seeking asylum closely parallels the game-theoretic model known as the stag hunt. The behavior of

certain nations, such as Ecuador and Russia, can be explained using this lens. In addition, the future behavior of other nations can be predicted and explained using this model.

Game-theoretic models have the potential to be of great value for information systems research, particularly within the area of eGovernance, which has an inherently social component. Future study of eGovernment phenomena using game-theoretic approaches in addition to what has been done here is warranted and may yield further findings that can explain current policies and help to predict and shape future ones.

3 CORPORTATE LEVEL

In an increasingly globalized world, jobs are more and more frequently being done by employees in a country other than the one in which their employer is based. This phenomenon of filling positions with employees in other countries is most commonly referred to as offshoring[4]. As education standards and literacy rates around the world rise, technology makes distance between an employer and its employees less of an issue than it would have been in previous years, especially when the technology levels of the countries in question are comparatively similar (Rodríguez-Clare 2010). Because of this, offshoring has become a much more common phenomenon and is a more viable option for employers looking to decrease their labor expenses (Feenstra and Hanson 1996, Olney 2012). The exponential growth of IT employment in India, for example, is seen to be, at least partially, as a result of American outsourcing of those same information systems positions (Hashmi 2006). Mass immigration and offshoring have had such an impact on cultures that some scholars have suggested that the very meaning of the term "going abroad" has changed (O'Brien 2007). Offshoring has become an integral part of many companies' overall strategies (Pla-Barber et al. 2018).

This study proposes to test the impacts of a company's home base country and industry and their likelihood of offshoring jobs within their IT departments, their concern with communication skills, and budgeting factors. The research objective is to understand whether these factors increase the likelihood of offshoring IT positions and, if so, to find out how much of an impact they are having. This will be tested by using data from the Society for Information

[4] In this paper, "offshoring" will refer to American companies filling positions with employees from other countries – including, if applicable, other countries in North America such as Canada and Mexico.

Management's IT Issues and Trends Study. The Society for Information Management (SIM) is an organization of over 5,000 IT professionals "including CIOs, senior IT executives, prominent academicians, consultants, and other IT leaders" (Society for Information Management, 2019).

The organization conducts an annual survey of its members. Respondents were asked about whether or not their companies are based in the United States[5] and whether or not they consider their primarily field to be within the STEM [science, technology, engineering, and mathematics] domain[6]. Results will be studied using regression analysis to find whether or not these two factors have had a significant impact on the percentage of IT department positions that have been outsourced to foreign countries.

3.1 Literature Review

Mass immigration and offshoring have had such an impact on cultures that some scholars have suggested that the very meaning of the term "going abroad" has changed (O'Brien 2007). Offshoring and immigration have even been shown to be impacting the overall power of the state itself (Harris 2010). Entire nations, such as Canada and Australia, have been largely built on successive waves of immigration meeting relatively small native populations, shaping the future of the countries themselves (Iacovetta et al. 1995).

[5] Question 16 ("Based on the context that you defined, where is your organization's corporate or main headquarters located?") was used to determine whether or not the organization self-identified as an American corporation

[6] Question 77 ("What is your organization's primary industry or economic sector?") was used to determine whether or not the organization was a STEM-based organization or an IT-based organization. As all respondents were IT professionals, this question was concerned more with the self-identified industry of the organization rather than the individual. Those who responded as being an organization in "IT Hardware / Software," "IT Services / Consulting," and "Telecommunications" were classified as IT-based organizations. Those who responded as being an organization in "Agriculture," "Aerospace / Defense," "Automotive," "Chemical Industry / Chemical Manufacturing," "Construction / Architecture," "Electronics / Semicondutor," "Energy," "Engineering," "Healthcare / Medical / Medical Technology / BioMedical," and "Mining / Minerals" were classified as a STEM-based organization.

The past few decades have seen education standards and literacy rates around the world rise, and technology has made the distance between an employer and its employees less of an issue than would have been true in previous years. This means that offshoring has become a much more common phenomenon and is a more viable option for employers looking to decrease their labor expenses (Feenstra and Hanson 1996, Olney 2012). This is partially due to the fact that foreign outsourcing – filling positions with employees based out of a country that is not the declared home of the company hiring them – is a more feasible option when the technology levels of the countries involved are similar. Rodríguez-Clare (2010) and Pattnaik (2013) directly link the rise of software development industries in Asia with the parallel advances in information technology that have made globalization possible. As global communication has become easier, foreign outsourcing within the IT industry has become more common.

Even with technology making distance less of an issue, however, proximity has historically been seen as a potential consideration with regards to outsourcing IT positions (Boudreau et al. 1998) and is still seen by some as a key factor when determining what provider firms one should hire when engaging in foreign outsourcing. Gonzalez et al. (2006), for example, recommend that European and American firms should look more towards Spain as a potential foreign outsourcing location due to quality, security and proximity. This proximity issue is an interesting one. Rai et al. (2009) find that cultural differences can play a significant role in the success or failure of foreign outsourcing in information systems. Increased proximity – which could lead to reduced cultural differences – may mitigate this issue to some degree. This study, however, is not so concerned with whether or not proximity has had an effect on the success or failure in outsourcing in information systems – as in whether or not employees or companies are satisfied with the arrangement – but instead is concerned with what impacts it has had on the

overall *rate* of outsourcing relative to the number of employees at the company. This paper is concerned with what percentage of positions are in a company's home nation as compared to positions in other nations abroad.

As offshoring has become more common, academic research has increasingly begun to look at what the impacts of outsourcing are for both the country exporting positions and for the countries that are receiving the new jobs, often within the context of information systems (Dibbern et al. 2004). Foreign outsourcing has been shown to have an impact on the labor expenses in the country from which jobs are being exported (Olney 2012). Within the context of the United States, foreign outsourcing is generally seen as having more benefits for the upper and middle classes in America than it does for the lower classes (Chakravorti 1996). It has been generally seen as particularly harmful to the number of manufacturing jobs available in the United States (Ottaviano et al. 2013).

It also can impact the country to which jobs are being exported. The huge rise in IT positions in India is seen by many to be the result of American outsourcing (Hashmi 2006). This increase in outsourcing positions in India has led to a situation where American and European companies are either acquiring or being acquired by Indian corporations in order to establish more credibility across cultures (Henley 2006). This has led to Indian immigrants being increasingly present within American IT departments and even starting their own companies in Silicon Valley (Varma and Rogers 2004). Hira (2004) further highlights that the Indian IT industry is now somewhat dependent on the immigration laws in the United States and has to guard its interests accordingly.

Globalization has also led to companies looking towards other countries to fill key positions, particularly within the context of science and engineering (Manning et al. 2008),

which is seen by some as being potentially threatening to American economic leadership in the world (Freeman 2006). All of this combines to suggest that outsourcing is having profound impacts on both the nation exporting jobs and the nations that are receiving them.

3.1.1 Immigration

Immigration patterns have been previously studied in a wide variety of contexts within the academic literature and have been shown to have impacts on a variety of different business situations, even impacting how the President of the United States views macroeconomic trends in official reports (Hanson 2005) and the overall power of the state itself (Harris 2010). For example, Burgoon et al. (2010) find that immigration generally decreases the ability of natives to unionize. Hickman and Olney (2011) find that immigration increases can also lead to higher post-secondary education rates in the targeted country.

More specifically, immigration habits have been previously shown to impact outsourcing behavior. Yomogida and Zhao (2010) describe a phenomenon known as "two-way outsourcing," which involves countries mutually exchanging different types of labor. This can be an exchange of two differing forms of highly skilled labor or one country outsourcing skilled labor while the other outsources relatively unskilled labor. Mithas and Lucas (2010) find that outsourcing behavior and immigration patterns are often linked, within an increase in foreign outsourcing often leading to a subsequent rise in immigration from the other nation. This argument shows some similarities to a concept from Vertovec (2006) known as "circular migration." Adam Smith wrote of the "comparative advantages" of nations in his *Wealth of Nations.*

Interestingly, the development of information systems itself has had an impact on immigration patterns across the world. Lin et al. (2011), for example, find that e-government has had positive effects in Gambia and could have positive implications in terms of getting additional

information to citizens in a relatively timely manner. Canada has made an effort on this issue as well, attempting to develop online portals to share information about immigration processes (Reitz 2005). It seems likely that these types of development could streamline immigration processes, facilitating an easier movement of people across national borders, although to this author's knowledge, this link has not definitively been shown in the literature.

This is a finding echoed by Venkatesh et al. (2011) who found that information systems processes can play a role in facilitating automatic immigration clearances and, again, potentially allowing for the easier movement of people across borders. In light of recent events, such as the United Kingdom leaving the European Union (an event commonly referred to as "Brexit") where the future status of immigration across borders for work purposes was a major concern both within the European Union (Bendel 2005) and outside of it (Lyons 2016), the focus on automating immigration processes is likely to only increase in the future, which should further increase immigration around the globe.

3.1.1.1 Immigration and company leadership

An increasing number of information technology departments within American companies are being led by someone who immigrated to the United States and who is not a natural-born American citizen. Globalization has led to companies looking towards other countries to fill key positions, particularly within the context of science and engineering (Manning et al. 2008). These individuals bring new perspectives to their positions, having lived and been educated in other countries, occasionally having served in the militaries of other countries and often speaking other languages besides English, many times even speaking multiple other languages. Increasingly, individuals who speak English but may not see it as their

native language are leading American IT departments, which could be having an impact on what individuals those in leadership positions then seek to hire.

This study proposes to understand how these two phenomena interact with one another. Previous studies have shown that offshoring is often a two-way street, with an increase in offshoring often being linked with a subsequent increase in immigration from that country (Mithas and Lucas 2010). Globalization has also led to an increase in companies looking towards other countries to find individuals who can assume key positions within the context of science and engineering (Manning et al. 2008). Because of this, there is reason to believe that key stakeholders in American organizations such as CIOs may be more frequently coming to the United States from other countries rather than having been natural-born American citizens, as was generally the case in previous decades. Indian immigrants are increasingly common in the American IT field and are starting their own companies and assuming leadership positions here in the United States (Varma and Rogers 2004). This leads to a question of what new perspectives these individuals bring to their jobs and how these new perspectives impact the decisions they make in their positions. One key area where this may be having an impact is on the likelihood of outsourcing positions to other countries.

Are immigrants more likely to offshore jobs for American IT departments than those who are natural-born American citizens? Are non-native English speakers more comfortable with hiring others who also are non-native English speakers? The literature has long suggested that experiences such as military service can impact the future of individuals' professional lives and migration habits (Kanter 1985, Baker 1998) and some notable papers have explored why companies are beginning to outsource IT positions more frequently (Carmel and Agarwal 2006). There is also research that suggests that cultural differences can positively impact the success of

foreign outsourcing (Rai et al. 2009). Despite all of this, the research within the IS field has not yet explored the impacts of these experiences on how likely individuals in key IT positions are to outsource positions within the IT departments that they are tasked with filling.

3.1.2 Military Service

There is a large volume of research discussing the potential link between military service and patriotism. Within the United States, the link between patriotism and military service has been discussed as far back as 1862, particularly whenever there was a prominent movement for universal military service amongst men aged 18 – 21, which was, in turn, often related to an ongoing war (Storey 2004). This debate has continued since the 1800s (Teigen 2006, Gore 2014). This trend of viewing patriotism and military service as being intrinsically linked has held fairly consistent even across vastly different groups and contexts within the American populace, including African-Americans (Nasmyth 1916), Italian-Americans (Gürsel 2008), Japanese-Americans (Sokolowski 2009) – including Japanese immigrants who were not born in the United States (Salyer 2004) - Latino-Americans (Leal 1999), and Native Americans (Denetdale 2008). All of the above groups viewed service in the American military as being linked to greater degrees of patriotism and to being an expression of it.

The impacts of prior military service on the lives, migration habits, and professional choices of those who served have been studied in other nations as well, including Canada (MacKinnon 1997) and France (Kanter 1985, Baker 1998, Propes 2011). This literature goes back to the nineteenth century in France (Kanter 1985, Baker 1998) and back to World War I in Canada (MacKinnon 1997). This service has been shown to have impacts beyond a reported increase in patriotism. For example, Qari et al. (2012) find that more patriotic individuals are more likely to live and work in the country where they were born than to relocate to others.

3.1.3 Language Knowledge

Language knowledge has been shown to have an impact on business-related decisions. The study of the influence of language and linguistic factors on information systems processes goes back several decades (Lyytinen 1985). Grinblatt and Keloharju (2000) find that individuals are more likely to buy, sell, and hold stocks of firms that are nearby, which could be due to similar language usage which in turn allows easier evaluation of the companies. Finnish individuals, for example, were more likely to purchase stock from geographically close companies due to having the same native language and a similar cultural background to the CEO of the company.

With regards to hiring decisions, Rubini and Menegatti (2008) find that linguistic biases can even impact the language usage within descriptions of open positions, which could be leading to subsequent impacts on who is eventually hired to fill that position. It seems likely that the description of an open position may be having an impact on who even applies for the position in the first place – individuals who are more familiar with the language being used in the description of the position may subsequently be more likely to apply for the position.

Neckerman and Kirschenman (1991) find that even differences in dialect within the same language can lead to discriminatory practices when a manager is hiring a new employee. The manager may find that a potential employee's dialect is "inappropriate" and subsequently not hire him or her. Purkiss et al. (2006) echo this finding, reporting that individuals with a different accent applying for a position are more likely to be viewed unfavorably by those making the hiring decision.

3.2 Hypotheses

The above factors combine to suggest that companies based in the United States will have different concerns from companies based outside of the United States, particularly with regards to budgeting and foreign staffing of certain positions. While it may seem somewhat unrelated to outsourcing, the United States' view of the military and the historical factors of what groups that have served within the military strongly impacts on how Americans view foreigners as well as their own domestic work force. This, in turn, impacts Americans' views about outsourcing and budgeting concerns. In addition, the strong history of immigration in the United States further emphasizes some of these resulting views about working with others of a different national identity. Finally, this resulting openness to working with foreign nationals – resulting from largely positive views of foreigners as well as increased budget flexibility due to national stability – suggests that Americans may be presented with unique challenges with regards to linguistic and cultural factors that more homogenous groups in other countries may not typically deal with as frequently. These structures of unified military service and relatively widespread immigration create an environment that is unique to the United States. This leads to several resulting hypotheses.

First, as mentioned previously, the development of information systems has been found to support and encourage immigration and globalization processes more generally (Venkatesh et al. 2011). This suggests that companies dealing with IT processes may be more open to globalization and, subsequently, more open to outsourcing as well. This suggests the following hypotheses:

Hypothesis 1a: Companies based in STEM fields will be significantly more likely to outsource than companies in other industries.

Hypothesis 1b: Companies based in STEM fields will be significantly more likely to dedicate a higher percentage of their budget to IT outsourcing than companies in other industries.

Hypothesis 1c: Companies based in STEM fields will spend a significantly lower percentage of their budget on IT outsourcing domestically than companies in other industries.

Hypothesis 1d: Companies based in STEM fields will be significantly more likely to dedicate a higher percentage of their outsourcing budget to offshore IT than companies in other industries.

Hypothesis 1e: Companies based in STEM fields will have a significantly higher percentage of their IT FTEs located outside their home country than companies in other industries.

Using survey data from a variety of corporations – including companies both based out of the United States and those based elsewhere – these hypotheses will be tested.

3.3 Method and Design

3.3.1 Data

This research uses data collected from the Society for Information Management's 38th Annual IT Trends Study. In April 2017, a survey was sent to the 4,213 members of the Society for Information Management and 1,178 surveys were completed, resulting in a response rate of 28.64%. Of these 1,178 completed surveys, 769 unique organizations were represented. These 769 organizations and these organizations, when combined, represent 19.3% of the $18.56 trillion GDP of the United States. Roughly 96% of these 769 organizations are based in the United States, leaving only about 4% based elsewhere.

3.3.2 Evaluation

Using these data, this research proposes to analyze the results using regression analysis to determine the validity of the hypotheses outlined above. Each of the hypotheses is mapped to a specific series of questions from the survey. Using this information, we are able to test the hypotheses by determining whether or not a statistically significant difference can be determined between companies based in the United States and companies based abroad.

3.3.2.1 Data Limitations

The major limitation faced by this study is the availability of data from countries based outside of the United States. Of the companies represented in the Society for Information Management's 38th Annual IT Trends Study, only approximately 4% reported being outside of the United States, meaning that data with regards to international trends is somewhat limited.

It is worth noting that the fact that only 4% of major corporations surveyed within the United States reported themselves as being based outside of the United States is itself a finding. As previously mentioned, Americans tend to believe their business environment is relatively global and welcoming to foreigners. This finding, that only 4% of corporations within the United States are based abroad, casts some doubt on that assumption and is worth exploring. However, it is difficult to extend these results to any larger trend as the results presented here are exclusively based on the results from the SIM sampling, which was done entirely based on U.S.-based respondents.

3.3.2.2. Questions used for evaluation

These hypotheses were tested by using data from the Society for Information Management's IT Issues and Trends Study. Several questions were used to determine the independent variables used in the regression analysis, a few of which will be explained in detail

here. Question 16 (“Based on the context that you defined, where is your organization's corporate or main headquarters located?”) was used to determine whether or not the organization self-identified as an American corporation. Question 77 (“What is your organization's primary industry or economic sector?”) was used to determine whether or not the organization was a STEM-based organization or an IT-based organization. As all the respondents were IT professionals, this question was concerned more with the self-identified industry of the organization rather than the individual. Those who responded as being an organization in “IT Hardware / Software,” “IT Services / Consulting,” and “Telecommunications” were classified as IT-based organizations. Those who responded as being an organization in “Agriculture,” “Aerospace / Defense,” “Automotive,” “Chemical Industry / Chemical Manufacturing,” “Construction / Architecture,” “Electronics / Semicondutor,” “Energy,” “Engineering,” “Healthcare / Medical / Medical Technology / BioMedical,” and “Mining / Minerals” were classified as a STEM-based organization.

3.4 Results

A regression analysis was completed using the aforementioned data. The results of this analysis are presented below. In most of the regression analyses completed, binary variables were included as to whether or not the organization is within STEM fields (as categorized above) and/or definitively within IT. Total revenue and IT budget were also included as regression variables. This was done as a control in case the size or type of the company had an impact on its propensity to outsource. Generally speaking, a larger company is going to have employees in more geographic areas, thus they are more likely to outsource or offshore regardless of their country of origin.

For **Hypothesis 1a**, "companies based in STEM fields will be significantly more likely to outsource than companies in other industries," a regression was done using the self-reported likelihood of outsourcing as the dependent variable and was regressed against whether or not the company was a STEM-based organization (IT or otherwise), whether or not the company was based in the US, the total revenue of the organization (in USD), and the amount of the budget dedicated to IT spending. The results are as follows:

Source	SS	df	MS	Number of obs	=	496
				F(4, 491)	=	2.02
Model	1.51979677	4	.379949193	Prob > F	=	0.0908
Residual	92.4721387	491	.188334295	R-squared	=	0.0162
				Adj R-squared	=	0.0082
Total	93.9919355	495	.189882698	Root MSE	=	.43397

Outsource	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
All_STEM	.0959821	.0422197	2.27	0.023	.0130285	.1789357
US	.1434782	.092837	1.55	0.123	-.0389285	.3258849
Total_Revenue	-1.22e-16	2.71e-16	-0.45	0.653	-6.54e-16	4.10e-16
IT_Budget	-4.61e-09	7.64e-09	-0.60	0.547	-1.96e-08	1.04e-08
_cons	1.088724	.0923331	11.79	0.000	.9073078	1.270141

Table 1 Results of Hypothesis 4a testing

Whether or not the company identifies as a STEM-based organization was found to be statistically significant with regards to how likely the company was to outsource and was trending in the theorized direction. Hypothesis 1a was supported.

For **Hypothesis 1b**, "companies based in STEM fields will be significantly more likely to dedicate a higher percentage of their budget to IT outsourcing than companies in other industries," a regression was done using the percentage of budget dedicated to IT outsourcing as the dependent variable and was regressed against whether or not the company was a STEM-based organization (IT or otherwise), whether or not the company was based in the US, the total

revenue of the organization (in USD), and the amount of the budget dedicated to IT spending. The results are as follows:

Source	SS	df	MS	Number of obs	=	495
				F(4, 490)	=	0.28
Model	153.009003	4	38.2522507	Prob > F	=	0.8932
Residual	67806.1213	490	138.379839	R-squared	=	0.0023
				Adj R-squared	=	-0.0059
Total	67959.1303	494	137.56909	Root MSE	=	11.763

IT_Outsourc~B	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
All_STEM	-.0080884	1.147025	-0.01	0.994	-2.261782	2.245605
US	-1.941274	2.516663	-0.77	0.441	-6.886057	3.003509
Total_Revenue	-9.94e-16	7.36e-15	-0.14	0.893	-1.54e-14	1.35e-14
IT_Budget	-1.27e-07	2.11e-07	-0.60	0.547	-5.42e-07	2.87e-07
_cons	8.917127	2.503044	3.56	0.000	3.999103	13.83515

Table 2 Results of Hypothesis 4b testing

No variables were found to be statistically significant. The coefficient for STEM-based organizations was not only not statistically significant, but also was not trending in the theorized direction. Hypothesis 1b was not supported.

For **Hypothesis 1c**, "companies based in STEM fields will spend a significantly lower percentage of their budget on IT outsourcing domestically than companies in other industries," a regression was done using the percentage of IT outsourcing budget dedicated to domestic spending as the dependent variable and was regressed against whether or not the company was a STEM-based organization (IT or otherwise), whether or not the company was based in the US, the total revenue of the organization (in USD), and the amount of the budget dedicated to IT spending. The results are as follows:

Source	SS	df	MS	Number of obs	=	253
				F(4, 248)	=	5.98
Model	39910.6597	4	9977.66492	Prob > F	=	0.0001
Residual	413765.218	248	1668.40814	R-squared	=	0.0880
				Adj R-squared	=	0.0733
Total	453675.877	252	1800.3011	Root MSE	=	40.846

IT_Outsourc~m	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
All_STEM	-18.14822	5.832213	-3.11	0.002	-29.6352	-6.661234
US	36.4164	11.24859	3.24	0.001	14.26146	58.57135
Total_Revenue	-4.00e-14	2.56e-14	-1.56	0.120	-9.04e-14	1.04e-14
IT_Budget	.0000115	.0000102	1.13	0.261	-8.62e-06	.0000316
_cons	27.62394	11.11352	2.49	0.014	5.735009	49.51286

Table 3 Results of Hypothesis 4c testing

Two variables were found to be statistically significant. Whether or not the company was based in the U.S. was found to be statistically significant. Organizations considering themselves as STEM-based companies were also found to be a statistically significant different in spending less on domestic outsourcing. Thus, hypothesis 1c is supported.

For **Hypothesis 1d**, "companies based in STEM fields will be significantly more likely to dedicate a higher percentage of their outsourcing budget to offshore IT than companies in other industries," a regression was done using the percentage of IT outsourcing budget dedicated to domestic spending as the dependent variable and was regressed against whether or not the company was a STEM-based organization (IT or otherwise), whether or not the company was based in the US, the total revenue of the organization (in USD), and the amount of the budget dedicated to IT spending. The results are as follows:

Source	SS	df	MS	Number of obs	=	134
				F(4, 129)	=	0.85
Model	6173.43916	4	1543.35979	Prob > F	=	0.4948
Residual	233664.053	129	1811.34925	R-squared	=	0.0257
				Adj R-squared	=	-0.0045
Total	239837.493	133	1803.28942	Root MSE	=	42.56

Off_Outsour~P	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
All_STEM	-.0379355	7.692841	-0.00	0.996	-15.25841	15.18254
US	.6602273	14.77519	0.04	0.964	-28.57284	29.8933
Total_Revenue	4.02e-14	2.68e-14	1.50	0.135	-1.27e-14	9.32e-14
IT_Budget	.0353641	.0318196	1.11	0.268	-.0275918	.0983201
_cons	34.05471	14.99949	2.27	0.025	4.377853	63.73156

Table 4 Results of Hypothesis 4d testing

No variables were found to be statistically significant, although this is unsurprising given the low number of respondents (134) who provided all the necessary data for the analysis. Also, the coefficient for whether or not the organization identifies as a STEM-based company is not trending in the theorized direction, nor is it statistically significant. Hypothesis 1d is not supported.

For **Hypothesis 1e**, "companies based in STEM fields will have a significantly higher percentage of their IT FTEs located outside their home country than companies in other industries," a regression was done using the percentage of IT FTEs based outside of their home country as the dependent variable and was regressed against whether or not the company was a STEM based organization (IT or otherwise), whether or not the company was based in the US, the total revenue of the organization (in USD), and the amount of the budget dedicated to IT spending. The results are as follows:

Source	SS	df	MS	Number of obs	=	411
				F(4, 406)	=	9.62
Model	17273.6133	4	4318.40333	Prob > F	=	0.0000
Residual	182244.93	406	448.879138	R-squared	=	0.0866
				Adj R-squared	=	0.0776
Total	199518.543	410	486.630593	Root MSE	=	21.187

IT_FTE_F_Pe~t	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
All_STEM	7.617859	2.275607	3.35	0.001	3.144416	12.0913
US	-28.41178	5.779711	-4.92	0.000	-39.77368	-17.04989
Total_Revenue	2.89e-15	1.33e-14	0.22	0.827	-2.32e-14	2.89e-14
IT_Budget	-1.79e-07	3.80e-07	-0.47	0.638	-9.26e-07	5.68e-07
_cons	34.61964	5.775586	5.99	0.000	23.26585	45.97343

Table 5 Results of Hypothesis 4e testing

Two variables were found to be statistically significant, with U.S. companies being significantly less likely to have a higher percentage of their IT FTEs located outside of their home country. STEM companies, in general, were found to be significantly more likely to have

IT FTEs outside of their home country and the effect was large. Hypothesis 1e is strongly supported.

3.5 Discussion

Hypothesis	Trending	Significant	Supported
1a	**Yes**	**Yes**	**Yes**
1b	No	No	No
1c	**Yes**	**Yes**	**Yes**
1d	No	No	No
1e	**Yes**	**Yes**	**Yes**

Table 6 Results of hypotheses testing

The results of the analysis were mixed, with only Hypotheses 1a, 1c and 1e – all concerning trends in STEM-based companies – being supported. Hypotheses 1b and 1d were not found to be supported, although they were not statistically significant in either direction.

3.6 Conclusion

Using the data described above, this research is designed to analyze data collected from a wide variety of largely U.S.-based companies representing a large percentage of global GDP on hiring and budgeting trends. Because SIM studies companies primarily based in the United States, the sampling was inherently limited with regards to any regressions concerning the home nation of the company. Only approximately 4% of the companies were identified as being based outside of the United States. This is such a small sample size that finding significance for any potential hypotheses concerning differences between U.S. and non-U.S. companies proved to be infeasible. This is an area that certainly warrants further research, but they are not questions to that can be answered with this dataset. This research has thus identified areas where findings have been suggested, but cannot be tested or verified. Further research, with larger datasets, are necessary.

This research represents an important early step in determining whether these effects exist and to what extent they may be present. In three cases regarding STEM-based companies, the effects were found to be present with statistical significance. The hypotheses with support all showed a general trend that companies that identify themselves as being STEM-based are more likely to outsource positions than companies of a similar size in other industries. The findings suggest that the decision on whether or not to outsource positions is dependent on more than budgetary factors. This also raises interesting questions about how companies identify themselves – in a world where most companies are heavily reliant on technology, why do companies that identify themselves as being part of certain industries show more likelihood to outsource positions than others? Are STEM-based companies different from companies in other industries in other significant ways? All of this combines to suggest that there are interesting findings here that are worth exploring in future research.

4 APPLICATION LEVEL

Linguistic factors, such as what words are used and the sentiments they imply, have long been important in terms of understanding business processes but have only recently been able to be digitally analyzed and quantified in a meaningful way. Researchers have used linguistic data to analyze business processes as varied as the effectiveness of ERP systems (Chang 2012), intellectual capital performance (Chen and Tai 2005), and business strategy design (Stašák et al. 2015). Big Data analysis techniques, including text analysis and text mining, have grown and enable faster and more precise understanding of large volumes of text. This is because Big Data analysis techniques allow individuals to quantify and measure large bodies of text quickly. Linguistic factors for analyzing text have also progressively become more important, although less adopted, in information systems research, than in other fields such as computational linguistics and human-computer interaction (Mastora et al. 2017). Computational linguistics is defined as "the scientific and engineering discipline concerned with understanding written and spoken language from a computational perspective, and building artifacts that usefully process and produce language, either in bulk or in a dialogue setting" (Schubert 2014). New methods for analyzing text have resulted in increasing quantification of large bodies of text (e.g., counts of numbers of terms) (Gao and Beling 2003; Ghosh et al 2012; Rockwell and Sinclair 2016). Yet despite all of these advances, information systems researchers have not yet sufficiently mixed linguistic analysis with more traditional models. By introducing an MIS corollary to existing findings within linguistics, this research hopes to facilitate further integration between information systems research and linguistic data.

Big Data analysis techniques now allow for a thorough integration of linguistic factors into other forms of research (Lewis et al. 2013) and this study outlines one possible algorithm that could be used to benefit information systems researchers when looking for useful data that could be drawn from large bodies of text. It does so by outlining a new set of assumptions known as linguistic component theory, which shows why information systems research could benefit from the integration of quantified, linguistic data. The research then outlines a new algorithm to analyze bodies of text and then provides two potential applications for the output the algorithm generates.

Within methodological literature, text analysis has been defined as "a method of data analysis that closely examines either the content and meaning of texts or their structure and discourse" (Given 2008). Traditional text analysis emphasizes the actual text, but often disregards the underlying linguistic factors or lack some key linguistic aspect needed for in-depth analysis (Rau et al. 1989; Binali et al. 2010). In addition to conducting sentiment and other forms of text analysis that consider the meanings of words, the actual patterns of language and word usage can provide useful data that is often ignored (Binali et al. 2010). By including these linguistic components in any analysis, we can expand the scope of analysis. Rather than being something to work around, linguistic components can now be a major and useful part of any analysis of business processes and behaviors.

The overall objective of this research is to develop and test a linguistic approach to generating useful information from text. To do so, we develop a new method to quantify the similarities between texts as a digital innovation in the sense of Fichman et al. (2014). Specifically, this research takes a design science approach to creating a method for identifying commonalities in separate texts (even from different languages). The method is based upon an

algorithm and implemented in a prototype for testing, thus serving as an instantiated artifact (Gregor and Hevner 2013). It also incorporates previous findings of extensibility, linguistic component theory and Zipf's law. The underlying logic of Zipf's law is used to create a new algorithm. The contribution of the research is to provide a new algorithm that facilitates the integration of linguistic data into more traditional models used in information systems research, such as econometric models. This research proposes a new set of assumptions known as linguistic component theory, which shows why information systems research could benefit from the integration of quantified, linguistic data into more traditional research approaches, such as econometric modeling. The research then outlines a new algorithm – which is inspired by, but is unique from Zipf's law – to analyze bodies of text and then provides two potential applications for the output the algorithm generates. This output provides an example of the type of quantified linguistic data that could be placed into other models within information systems research.

This paper proceeds as follows. Section 2 reviews related research on design science and linguistics within information system. Section 3 presents the new algorithm, outlines the method and its implementation. Section 4 applies the method to identify common authorship amongst texts. This research attempts to isolate linguistic components of texts with a similar structure to comparatively test one against another, even when the authors are different. It is intended to identify the presence of specific authors based on analysis of their writings, even when the linguistic components are held constant. Section 5 uses the method to improve error detection in automatic translation software. Section 6 discusses the results and suggests areas for future research. Section 7 concludes the paper.

4.1 Theoretical Background

Research dealing with text in unstructured forms is important in areas such as big data analytics, sentiment analysis, and social media analytics. Approaches to dealing with corpus of texts usually include natural language parsing techniques (Collins & Duffy 2002) as human beings do not naturally communicate in structures that are ideal for computers to understand. This research is primarily built upon the information systems literature – although it also builds upon work in computer science, literary criticism and computational linguistics (Mastora et al. 2017). Computational linguistics is defined as "the scientific and engineering discipline concerned with understanding written and spoken language from a computational perspective, and building artifacts that usefully process and produce language, either in bulk or in a dialogue setting" (Schubert 2014).

Much of our understanding of word patterns and our ability to quantify patterns in unstructured text is due to the influence of Zipf's law, an algorithm explaining a group of phenomenon as diverse as word frequency, the distribution of city size and the distribution of income (Zipf 1949, Hill and Wooduffe 1949, Hill 1970, Woodruffe and Hill 1979). The principles behind Zipf's law, particularly those related to the exponentially increasing rarity of less commonly used words, serves as the foundation of part of the algorithm in our method.

4.1.1 Digital innovation

Finchman et al. (2014) define digital innovation as an expansion of traditional information systems or technology innovation. Within information systems research, digital innovation has been given an increasing focus from 2009 – 2015 (Fielt and Gregor 2016). Yoo et al. (2012) analyze the translation of physical products into digitalized forms. Crossan and Apaydin (2010) define digital innovation as "both a process and an outcome" that occurs within

organizational contexts. Information systems research has focused on digital innovation within organizational contexts (Fielt and Gregor 2016). This suggests that taking a digital innovation-oriented approach to linguistic factors within information systems research may be useful to understand organizations and their behavior. One example of such an approach is using digitally generated linguistic data to improve localization for global businesses (Lommel 2006).

4.1.2 Extensibility

Mastora et al. (2017) argue that "Natural language is both fundamental and complicated as a communication system; therefore, it has been the subject of many disciplines" and that it has "rules, norms and patterns concerning its morphology and syntax" (pg. 496). They (2017) quote Portner (2005), who argues that "the theory of [meaning] holism claims that the meaning of a word or phrase or sentence depends on its relationships with other words, phrases, and sentences" (pg. 496). In other words, the full meaning of a word cannot be determined without considering the context within which it is used.

Human language is dynamic and constantly changing. Therefore, any method designed to analyze human language must feature *extensibility,* the ability to indefinitely expand without any barriers*,* in its design. The relationship between human language and culture is well-established in academic literature (Goodenough 1981, Schieffelin and Ochs 1986, Hinkel 1999, Stubbs 1996). Human language is anchored in culture, and cultures comprising a potentially infinite variety of combinations. Therefore, any artifact that is designed to analyze text in a meaningful way must accommodate a wide variety of *linguistic components.*

Natural language is indefinitely extensible (Cook 2007, 2009; Schleckner 2010, Luna 2013), so it can be continually extended, and changed, existing in a state of impermanence. No true form of permanent modeling for language studies can ever really exist (Luna 2013). A

similar concept, relative indefinite extensibility, can be explained through several examples (e.g., Luna (2013)), including, most notably, the fact that there is no complete, written set of all possible existing numbers (due to the infinite number of possible and valid combinations). Therefore, any information system artifact that attempts to model language must also be indefinitely extensible. No system can be pre-programmed to include an infinite number of possible (and valid) numeric combinations, but there are still contexts within which these terms can be used. Therefore, systems should be applied within many contexts to adapt to the changing circumstances surrounding the language being studied.

4.1.3 Linguistic component theory

In this research, a new set of assumptions known as linguistic component theory is presented, which proposes that authors will exhibit regularities in their language use, and that these regularities will be comparable both with language usage in general, and with the author's language usage, in particular. Therefore, our proposed method will operate within the context of these aforementioned linguistic regularities, of which *Zipf's Law* (Zipf 1935, 1948) is a well-known example due to the patterns it identifies across languages.

Linguistic component theory is a new set of related assumptions proposing that models can be improved by factoring in linguistic components (such as the analysis of text-based data). In a global economy, understanding "new signals" from other cultures is important, particularly where data can be taken from countries all over the world and integrated into one project (LaVelle et al. 2011). A deeper integration of linguistics, which can only result from a deep understanding of the linguistic components inherent in the data, will facilitate the understanding of these signals. Senior executives now strive to run their companies on data-driven insights (LaVelle et al. 2011). However, this approach cannot be effective if the insights from this data do

not accurately reflect the linguistic context within which it exists. To understand the data that drives the insights, one must consider the larger linguistic context.

Previous deep structural work within information systems shows that information systems can be viewed and modeled as independent artifacts that reflect the real-world context it is intending to model (Wand and Weber 1995). These contexts include a linguistic component inherent in all informational transactions due to the universal usage of language by human beings. The inclusion of linguistic-based data (Rau et al. 1989; Binali et al. 2010) can help to represent this real-world context accurately. Although surface-level structure, such as the actual content of the text being analyzed, can change with social context, the underlying deep-structure is more consistent and can, potentially, provide more useful data, even across different genres of works or languages (Wand and Weber 1995).

4.1.4 Zipf's Law based algorithm

Zipf's Law (Zipf 1935, 1949) is a well-known linguistic algorithm which predicts that the frequency with which a word is used is inversely proportional to its ranking overall within the corpus. Zipf's law shows that the frequency in a word's usage decays at an exponential rate, based on its ranking against other words within the language as a whole (Ferrer i Cancho and Solé 2003). This means that the word used second most in a language is used half as much as the first, the third most used is used one-third as often as the first, etc. Zipf's Law, as well as modified forms of the algorithm, have been used within the field of computational linguistics for some time (Baayen 1992).

Zipf's law has been shown to apply in a wide variety of contexts (Powers 1998, Wyllys 1981), including with regards to city populations (Ioannides and Overman 2003, Gabaix 1999, Marsili and Zhang 1998, Hill 1974), income distribution (Okuyama et al. 1999), the Internet

(Adamic and Huberman 2002) and text (Li 1992). Zipf's law is commonly written in equation form as exemplified below, where *N* is the number of elements, *k* is the rank of the elements and *s* is the value of the exponent that characterizes the distribution (Adamic 2000, Zipf 1949, Zipf 1935):

$$f(k;s,N) = \frac{1/k^S}{\sum_{n=1}^{N}(\frac{1}{n^s})}$$

Figure 8 Zipf's Law

In more practical terms, this is generally applied to show that when elements are ranked in order, the frequency or number of an element is inversely related to their ranking in reference to the largest element in the series. Within city populations, one of the most common applications, it means city populations are generally inversely correlated with their ranking relative to other cities (Hill 1974, Giesen and Südekum 2010), meaning that the second most populous city is approximately half the population of the most populous city, the third most populous is approximately one third the population of the most populous city, etc. Another example is within natural language. Zipf's law has been found to explain word frequencies within natural language, with words appearing at roughly inverse frequency with regards to their ranking of usage within a language, meaning that the second most used word within a language is used approximately half as often as the most frequently used word, the third most used word within a language is used approximately one-third as often as the most frequently used word, etc. This pattern is not limited to a specific language – it has been found to hold true in a variety of languages, including English and Chinese (Dahui et al. 2005) and Greek (Hatzigeorgiu 2001) and even computer command languages (Ellis and Hitchcock 1986).

4.1.4.1. Zipf's law and city populations

One of the earliest and most well-documented findings with regards to Zipf's law is its usefulness in terms of explaining urban populations (Ioannides and Overman 2003, Gabaix 1999, Marsili and Zhang 1998, Hill 1974, Giesen and Südekum 2010). City populations can be explained neatly by Zipf's law – city populations are generally inversely correlated with their ranking relative to other cities (Hill 1974, Giesen and Südekum 2010). This has been repeatedly found to hold true for naturally forming cities, although defining what constitutes a "city" can be a challenge for scholars (Jiang ang Jia 2011). Research in Germany has found that this pattern holds true at both the national and regional level (Giesen and Südekum 2010) and the pattern has been found to hold true in Malaysia as well (Soo 2007). When populations do deviate from Zipf's law, they can generally be easily explained by other well documented phenomena (Ioannides and Overman 2003).

4.1.4.2. Zipf's law and income

Zipf's law has applications within income distribution in a variety of contexts. One study found that income distribution of companies in Japan (studied over three decades) had a clear power-law distribution similar to that of Zipf's law (Okuyama et al. 1999). This pattern with regards to income-distribution suggests that Zipf's law may be helpful in understanding human behavior in a more general sense than has been studied previously.

4.1.4.3. Zipf's law and the Internet

Zipf's law can be used to understand highly divergent phenomena in a variety of contexts, including on the Internet (Adamic and Huberman 2002). Within information systems literature, it has been shown to apply to websites, with a millions of web sites containing just one page, while only a few sites contain millions of pages (Adamic and Huberman 2002). Other patterns on the Internet appear to follow a similar pattern to Zipf's law – for example, many

websites contain only a few hyperlinks, while a much smaller number contain millions of hyperlinks (Adamic and Huberman 2002). Further research with regards to Zipf's law and hyperlinks suggests that Zipf's law also applies to the number of links a web surfer will follow - large numbers of users will follow a small number of links while a much smaller number of users will follow a larger number (Levene 2001). In addition, traffic patterns online follow a similar trend – only a few websites gain millions of views, while millions of websites gain a relatively small number of views (Adamic and Huberman 2002). Zipf's law has also been shown to be one of the best formulas for predicting web caching behavior (Breslau 1998, Serpanos 2000). It has also been found to be a useful explanatory tool for identifying frequencies within linux open source development (Maillart 2008).

4.1.4.4. Zipf's law and text

Zipf's law has been found to explain word frequencies within natural language, with words appearing at roughly inverse frequency with regards to their ranking of usage within a language. This pattern has been found to hold true in a variety of languages, including English and Chinese (Dahui et al. 2005) and Greek (Hatzigeorgiu 2001) and even computer command languages (Ellis and Hitchcock 1986). Even beyond just specific words, word phrases have been found to follow this pattern as well (Ha et al. 2002). In addition to the well-documented correlation between Zipf's law and word frequencies occurring in natural language, Zipf's law has also been found to apply to random text generation (Li 1992). In a wider-sense, it "has been observed that the rank statistics of string frequencies of many symbolic systems (e.g., word frequencies of natural languages) follows Zipf's law in good approximation" (Troll and Graben 1998). While scholars generally agree that Zipf's law accurately predicts the frequencies with

which specific words are used in natural language, there is some disagreement as to an explanation for the phenomenon (Dahui et al. 2005).

4.1.5. Plagiarism detection

Plagiarism is a concern even for information systems academics and the IS field has analyzed various possibilities for detecting plagiarism (Kock and Davison 2003). There is significant research work done on plagiarism detection methods that analyzes word frequencies and patterns in order to detect plagiarism (Lancaster and Culwin 2001). Plagiarism detection systems are often built on methods that attempt to detect unusually similar phrasing or word choice (Lancaster and Culwin 2001). One aspect of plagiarism detection is the need to detect similarities between works submitted within the same context. For example, there is a need to detect whether submissions from different students for the same assignment contain unusual similarities that may indicate one student plagiarized the work of another student or the two students engaged in unauthorized collaboration (Jones 2001). In this sense, any attempt to detect plagiarism on an assignment must not look only at existing literature and work, but also new submissions within the same context (Jones 2001). This type of work can be done by attempting to detect identical or near-identical documents within a specialized environment (Seshasai 2009).

Some students are even using "back-translation" methods to avoid detection. In order to do this, students take the quote they intend to use, translate the quote into another language and then translate the "translated" results back into the submission language (Jones and Sheridan 2015). This has the effect of disguising the plagiarized material by changing the wording with minimal effort on the part of the student. Because of this, work has been done to detect

plagiarism within the contexts of specific languages, such as Arabic (Alzahrani 2009) and even within computer programming languages (Cebrian 2009).

4.2 Methodology

This research takes the design science approach of Peffers et al. (2007) as summarized in table 19 and detailed below.

Table 19. Design Science Research

Component	Task
Problem identification and motivation	Show how the lack of linguistic sensitive analysis within text analysis prevents some analyses from being sufficient
Objectives of a solution	Create a method for addressing linguistic components.
Design and development	Create a method to analyze linguistic factors within differing bodies of text by adapting and extending an algorithm.
Demonstration	Implement the method in a prototype.
Evaluation	Evaluate whether the prototype answers potential research questions and/or tests appropriate hypotheses.
Communication	Document the development of the method and the resulting calculations in proof-of-concept applications.

Table 7 Design Science Research

Problem identification and motivation. Because Big Data methods have allowed for an increased ability to quantify text (Gao and Beling 2003, Ghosh et al 2012, Rockwell and Sinclair 2016), this research attempts to identify issues and challenges that could benefit from emphasizing the linguistic components of text. By identifying potential areas where this could be helpful, such as authorship identification and error detection in automated translation software, a solution can be developed.

Objectives of a solution. We focus on authorship identification and error detection in automated translation software and present hypotheses and research questions that can be tested and/or answered by analyzing the linguistic structure of bodies of texts. For the authorship identification application, we present several hypotheses centered around a central notion. This is, given the choice between three pairs of works (e.g., book), where one pair represents a pair of works by the same author, and the other two pairs represent works by two different authors, a linguistic sensitive analysis should be able to identify which pair of works were written by the same author more successfully than by random chance. For the application for error detection in automatic translation software, rather than hypothesis testing, we attempt to find the "true ratio" of words held in common between languages (e.g., English and German). This can help to identify when words have been translated incorrectly and improve the accuracy of automatic translation software.

Design and development. We develop a method based on an algorithm that can analyzes the underlying linguistic structure of differing bodies of text. Based upon Zipf's Law (Zipf 1935, 1948), we develop a new algorithm focused on word frequencies within texts and show what this can reveal about authorship or languages as a whole, such as patterns between English and German.

Demonstration. This algorithm is incorporated into a program that can take as input bodies of text (placed in .txt files, and ranging from short poems to entire novels) and can run the algorithm using the words provided within these .txt files. The program calculates relative measures of commonalities across the bodies of text, showing the similarities between different works.

Evaluation. The relative comparison values are used to test hypotheses and/or to answer research questions. Two applications are used in the evaluation: authorship identification and automated translation.

Communication. The results of this process are presented in this paper.

4.2.1 Method

4.2.1.1. The MIS Corollary to Zipf's Law

In order to further integrate Zifp's law into information systems research, we have developed a new algorithm based on the underlying logic of Zipf's law. This is algorithm, while inspired by the underlying logic of Zipf's law, is entirely separate and unique from it. The math presented is new and will be tested here in this research using two sample applications. We are calling this the MIS Corollary to Zipf's law and the method that is developed to compare text is comprised of a set of steps that generate the data needed to make the comparisons. The steps of the method are as follows.

Step 1: Generate a set of corpus values for the entire data set.

Calculate the total number of words in the corpus. For the purposes of this paper, we use three works in each dataset. This can be a set of any three works (for example, three separate novels) that are tested together. Then, calculates the total number of words, which need not be unique:

$$\sum_{i} N(w_i \in Corpus) = \sum_{i,j} N(w_i \in T_j) \quad (1)$$

For each unique word, a value based upon the number of times a unique word occurs is calculated. Less frequently used words are valued more highly than more frequently used ones, a principle borrowed from the underlying logic of Zipf's law.

For $w_i \in \bigcup_{j \in J} T_j$, we have,

$$N(w_i \in Corpus) = \sum_j N(w_i \in T_j) \quad (2)$$

Step 2: Perform individual word analysis and values.

Once this value is created for every word in the corpus, it will be converted to a proportional value that shows the frequency of the usage within the context of the data set, and which can be adapted based on the structure of the text being analyzed. The analysis is primarily based on word counts and frequencies, rather than the structure of the actual work, as in prior.

$$F(w_i \in Corpus) = N(w_i \in Corpus) / \sum_i N(w_i \in Corpus) \quad (3)$$

$F(w_i \in Corpus)$ represents the relative frequency for each unique word in the corpus (datasets containing a wide variety of texts).

$C(w_i \in Corpus)$ is the complementary value of the relative frequency for each unique word in corpus.

$$C(w_i \in Corpus) = 1 - F(w_i \in Corpus) \quad (4)$$

This value is generated for every word in the corpus. The number of times each word is used within two texts being compared (versus the corpus overall) is expressed as follows.

For each word $w_i \in T_1 \cap T_2$:

$$N(w_i \in T_1 \cap T_2) = N(w_i \in T_1) + N(w_i \in T_2) \quad (5)$$

For the word $w_i \notin T_1 \cap T_2$, $N(w_i) = 0$.

The commonalities between the texts are expressed using a unique word value.

$$I(w_i \in T_1 \cap T_2) = N(w_i \in T_1 \cap T_2) \times C(w_i \in Corpus) \quad (6)$$

This is generated for each word present in the two texts. It is the total word count from each of the two texts from each genre selected for comparison and is totaled to obtain what is referred to as the "comparison value."

Step 3: Generate comparable values.

Comparison of words is performed by:

$$Comp(T_1, T_2) = \sum_i I(w_i \in T_1 \cap T_2) \quad (7)$$

where $Comp(T_1, T_2)$ is comparable value of Text 1 and Text 2.

The total number of words in the comparison is:

$$\sum_i N(w_i \in T_1 \cap T_2) = \sum_i N(w_i \in T_1) + \sum_i N(w_i \in T_2) \quad (8)$$

However, this "comparison value" does not yet take into account the total number of words, so a "relative value" must be generated using the following formula:

$$R(T_1, T_2) = Comp(T_1, T_2) / \sum_i N(w_i \in T_1 \cap T_2) \quad (9)$$

The process is repeated to obtain $R(T_2, T_3)$ and $R(T_1, T_3)$ as relative comparison values. This process of using the combined inputs of the bodies of text themselves as well as previous results from within the algorithm is reflected in figure 9 below.

Step 4: Create relative comparison values.

All of the steps are repeated for all possible combinations, to obtain the following (final) values:

RelativeComparisonValue(1) = The relative comparison value between text 1 and text 2.

RelativeComparisonValue(2) = The relative comparison value between text 2 and text 3.

RelativeComparisonValue(3) = The relative comparison value between text 1 and text 3.

4.2.2 Differences from previous research

One unique element of this corollary to Zipf's law is the lack of a "kill list," which is a technique traditionally used in text mining to remove commonly used words such as "the," "and" or "or." This was done for several reasons. First, Zipf's law is entirely based around the idea of relative frequencies – thus, the frequency of usage with regards to the most commonly used words is essential to determining the frequency of all other words in the language. Subsequently, this means that any corollary to Zipf's law should include an emphasis on these frequencies. The frequency of usage of certain words is essential to determining key factors about the language and tendencies of the speaker/author. Thus, a "kill list" being included in this corollary has the potential to actually remove data that could be useful in highlighting surface-level trends.

Another unique element of this research is that while it is similar to previous work on plagiarism detection, it differs in a few fundamental ways. Firstly, as the algorithm presented here is new, it is fundamentally different from plagiarism detection methods that have previously been studied within the literature on information systems (Kock and Davison 2003). Plagiarism detection methods focus primarily on looking for "exact matches" between two bodies of text – these methods look for areas where two bodies of text use the exact same wording. This method looks for similarities between bodies of a text in a more comprehensive sense. It analyzes bodies of text for the frequency with which certain words are used in the overall text, rather than if specific phrases are used verbatim.

In this sense, this research has the potential to reveal larger, overarching patterns with regards to word frequencies rather than specific instances of identical text. Traditional research on plagiarism detection has focused primarily on word patterns – that is, looking for cases where the exact same words were used in the exact same order, thus indicating that the phrase or section was copied from another. This research instead looks at the frequencies of how often identical words are used and is not so concerned with the order in which they are used – the idea is that these frequencies, rather than revealing instances of plagiarism, can reveal patterns within an individual's use of language or the language as a whole. This emphasis on frequencies represents a key contribution of this method.

4.2.3 Implementation

An overview of the implementation is shown in figure 9. This Zipf's Law-based algorithm was designed to analyze large bodies of text and a program was then built using PHP to run these computations for the new software. This software analyzes three bodies of text and generates a value measuring the degree of similarity between all possible pairings, meaning that we are given a value for the degree of similarity between texts 1 and 2, texts 2 and 3, and texts 1 and 3. A higher value indicates a higher degree of similarity.

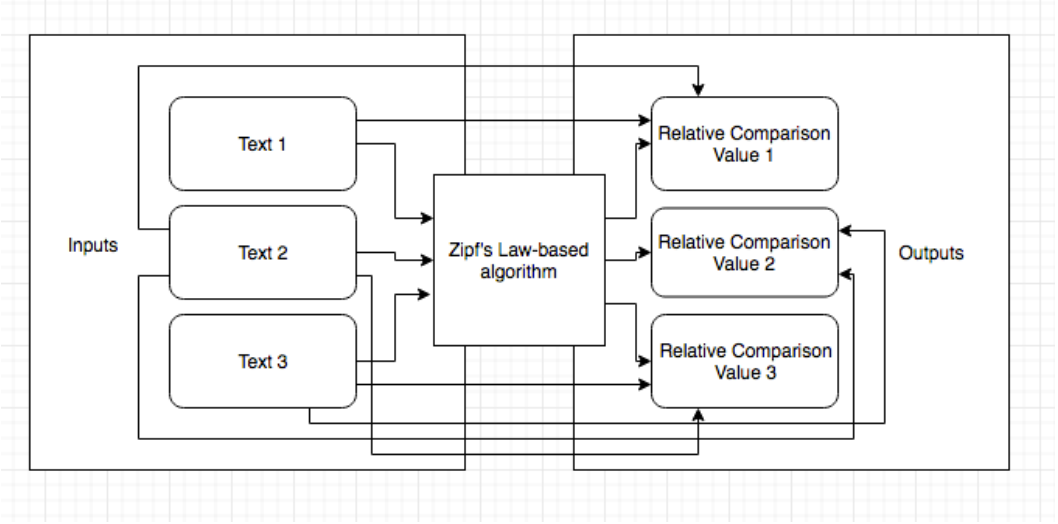

Figure 9 Zipf's law-based algorithm

To test the method, two cases studies were used (Venable et al. 2016).

4.3 Application 1: Identifying authorship

Identifying the true authorship of a text based upon linguistic components has a long history in a variety of fields, including for authors as famous as William Shakespeare (Barber 2009, Thomas 1932, Wells 2014). This tradition of identifying accurate authorship has applications beyond mere curiosity, with impacts within the national security and criminal justice system, where identifying authorship can be a key aspect of identifying suspects, with a famous example being the Unabomber (an American criminal who used the U.S. Postal Service to send explosives to victims), who was identified based upon the linguistic patterns in his manifesto (Houtchens 2001).

Although research suggests that authors can be identified based upon the linguistic patterns they employ (Sandal 1963), it is unclear how much similarity is dependent upon repeated structural patterns. Relying on only structural patterns could lead to misidentification of authorship, when one considers inherent common cultural structures amongst authors from similar geographical areas or ideologies. For example, cooperation between terrorist groups can

impact their longevity (Phillips 2014). Since terrorist organizations originate from similar cultural, religious and ideological backgrounds, this presents a potentially large problem for forensic linguistics because considering only structural patterns may not be sufficient. At the same time, identifying terrorists and others based upon their online presence is needed (Carr 2008). Incorrectly, unnecessarily, or too quickly identifying a group responsible for an attack can cause problems (Frey 1987, Conrad 2011). Research is needed that looks at overarching patterns in text in a more comprehensive way – that shows more than just whether or not two bodies of text copy one another, but instead reveals patterns found from looking at word frequencies throughout multiple works.

The applications for looking beyond structural components within forensic linguistics are even clearer. Carr (2008) emphasizes that the internet is serves as "an all-purpose communications network, surveillance medium, propaganda channel and recruiting tool." Researchers have retrieved audio messages, images of attack targets, covert terrorist websites and videos, highlighting the need for linguistic analysis from a forensic perspective (Carr 2008). The government has also funded research to identify authors of online text messages based upon the users' diction and syntax (Carr 2008). Besides law enforcement, there are many applications of big data analytics that could make use of an improved ability to identify a common author of multiple texts.

This research attempts to isolate linguistic components of texts with a similar structure to comparatively test one against another, even when the authors are different. It does so by looking at works within the same genre in order to control as best as possible for the structure and then seeing if patterns can still be found by focusing on the unique word frequencies. It is intended to

identify the presence of specific authors based on analysis of their writings, even when the linguistic components are held constant.

4.3.1 Application of Method

Zipf's Law suggests that while commonly used words (such as "the," "and," "or," etc.) will show up frequently in bodies of text regardless of authorship, other words will appear significantly less often (such as proper nouns or other less commonly used words). Because of this, less frequently used words have more value in identifying patterns, because, by definition, these words appear less often than common ones. For example, if the works of two authors are being analyzed, seeing the word "the" in their work tells very little that is specific to one of the authors, because we would expect that both authors to use the word frequently. However, if one of the authors tends to use a much less common word (for example, xylophone) more frequently than the other, the appearance of that word could suggest a great deal about the authorship.

The evaluation is comparative across different genres with different degrees of linguistic components. Our method is used on a variety of genres, including haikus. Since haiku poems have linguistic components that are narrowly defined with a smaller number of words, we expect author identities to be more difficult to detect via the linguistic components in haiku poems. For comparison, we analyzed songs, which have a higher degree of structure, but less than haiku poems. Third, we considered online reviews, which have a much lower degree of structure. Finally, we analyzed poems, which have a lower degree of structure as well.

4.3.2 Selection of Texts

To highlight extensibility and to isolate the structures present within text-based writings, texts were extracted from songs, haikus and online reviews. A variety of genres were chosen in order to emphasize the extensibility of the algorithm – this was done in order to generate results

from a variety of genres that, in turn, vary significantly with regards to structure and word count. By looking at a variety of types of texts, it helps to support the idea that the method presented here is extensible. The individuals who extracted the text were not involved in the actual analysis and instructed to select works randomly. Although some degree of non-randomness occurs, due to limitations on the data (such as the need for writings by the same authors and their availability) the intent is that the data set represents an accurate reflection of the real-world context within which this analysis takes place. Three genres of text were selected for analysis. 150 examples of each genre type -- meaning 150 songs, 150 haikus and 150 reviews - were selected and sorted into 3 unique datasets. In total, 450 works were selected for analysis and sorted into 150 different datasets.

4.3.3 Hypotheses

One goal of this research is to prove the value of this algorithm by using it to mirror existing findings within linguistics research, such as the linguistic principle the writings of the same author are more similar to one another than to different authors. Each dataset of 3 separate works of texts generates 3 unique comparison values (one for each pair of works), so there is a 1/3, or 33%, chance that random chance would accurately identify which two works were created by the same author. Findings that show this method's ability to correctly identify joint authorship across multiple genres, not just in books, would further highlight the extensibility of the method itself.

Because of this, we present the following hypotheses in order to appropriately test the method:

Hypothesis 1A: The songs written by the same author/artist should be correctly identified more than 33% of the time.

Hypothesis 1B: The haikus written by the same author should be correctly identified more than 33% of the time.

Hypothesis 1C: The reviews written by the same author should be correctly identified more than 33% of the time.

Hypothesis 1D: The books written by the same author should be correctly identified more than 33% of the time.

Hypothesis 1E: The poems written by the same author should be correctly identified more than 33% of the time.

Hypothesis 1F: The text written by the same author should be correctly identified more than 33% of the time across all genres.

The next question is whether the isolation of a specific structure has an impact on the accuracy of the comparisons. The issue is whether a specific structure will have a substantial impact on whether an author can be correctly identified. We test this by isolating the structure using haikus as a very rigorously structured and defined form of text.

The resulting hypotheses are:

Hypothesis 2A: The percentage of correctly identified haikus will be lower than or the same as the percentage of correctly identified songs.

Hypothesis 2B: The percentage of correctly identified haikus will be lower than or the same as the percentage of correctly identified online reviews.

Hypothesis 2C: The percentage of correctly identified haikus will be lower than or the same as the percentage of correctly identified online books.[7]

If the percentage of correctly identified haikus differs significantly from any of the other groups tested, it will support the general principle underlying the hypotheses that structure does have an impact on the accuracy of the analysis. If they do not substantially differ from one another, then, perhaps the focus on word usage and frequencies presented in this algorithm immunizes the analysis from this impact to some extent. Since the algorithm does not accommodate structure but, instead, focuses only on word usage and frequencies, then haikus may in fact be incorrectly identified more often than any other genre.

4.3.4 Songs

Fifty datasets of three songs each were extracted by an individual instructed to select songs randomly from online sources. Each dataset consisted of three songs, with each in their own text file – the text files contained the song's lyrics in text form. An online database of publicly available works was used to generate random selections. Two works were randomly selected in order to ensure that each dataset had two authors represented. Once two works were selected, one other work by one of the authors was selected, resulting in a dataset that included three total works. Perfect randomness was not possible due to the availability of data and the requirement that at least two of the songs be written by the same author/artist. However, the data is intended to be representative of the real world context in which this type of analysis might take

[7] No comparison was made between haikus and poems as haikus themselves are a specific form of poetry.

place. Within each dataset, texts A and B are works by the same author whereas text C is always by a different author.

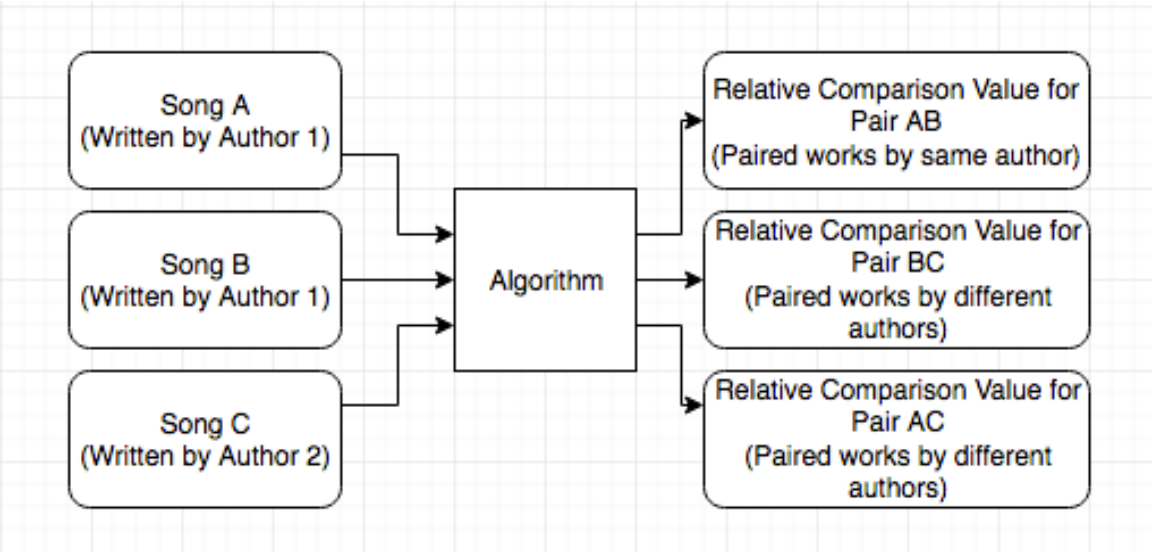

Figure 10 Examples of inputs and outputs

The highest value for the comparison indicates which two works the algorithm identifies as being the most similar. Thus, the comparison value for AB should the highest if the joint-authorship is properly identified, as a high value for AB and would indicate that texts A and B are the most similar. A result of BC or AC being the highest would be incorrect since C was written by a different author than A and B. The first five results of the analysis are given in table 20.

Dataset	AB (Same Author)	BC (Different Author)	AC (Different Author)
1	0.361044137	**0.437684826**	0.378155221
2	0.346523022	0.355811223	**0.369746338**
3	0.286866632	**0.407574696**	0.227307246
4	**0.421945449**	0.313599338	0.280545375
5	0.338712968	0.348575537	**0.477111064**

Table 8 Comparison of analyzed songs

In total, 22 out of 50 pairs were correctly identified as being the work of the same author/artist, resulting in a probability of 0.44 or 44%, which is indeed higher than the

probability that a correct result would have occurred through random chance. The results of the matched-pair analysis are displayed in table 21.

	AB (Same Author)	**BC** (Different Author)	**AC** (Different Author)
Amount	**22**	15	13
Probability	**0.44**	0.3	0.26

Table 9 Results of matched pairs for songs

Since 0.44 is greater than the 0.33 probability that a correct result would have occurred through random chance, Hypothesis 1A is supported.

4.3.5 Reviews

Similar to the selection process for songs, fifty datasets of three reviews each were retrieved. These were selected from an online retailer with publicly accessible comments. An individual other than the primary researchers was told to select two reviews using random selection. Once two works were selected, one other review by one of the authors was selected, resulting in a dataset that included three total reviews. The first five results of the analysis of each data set are shown in table 22.

Dataset	**AB** (Same Author)	**BC** (Different Author)	**AC** (Different Author)
1	0.541827597	0.570669104	0.47316592
2	0.476357447	0.384955598	0.38377702
3	0.369754309	0.351143506	0.383070977
4	0.297540945	0.389610949	0.335790336
5	0.241538866	0.269408117	0.356709767

Table 10 Comparison of analyzed reviews

In total, 16 out of 50 pairs were correctly identified as being the work of the same author/artist, resulting in a probability of 0.32 or 32%, surprisingly lower than the probability

that a correct result would have occurred through random chance. The results of the matched-pair analysis are summarized in table 23.

	AB (Same Author)	**BC** (Different Author)	**AC** (Different Author)
Amount	16	17	17
Probability	0.32	0.34	0.34

Table 11 Comparison of matched pairs for reviews

The results are nearly identical to what one would find by selecting the datasets randomly. The pair of reviews written by the same author was correctly identified only approximately one-third of the time and incorrectly identified approximately two-thirds of the time, suggesting that this method provided no support beyond that of random chance. The reasons for these results are unclear. Perhaps a larger sample would yield more conclusive trends, or this can be explained by the relatively small number of words commonly used in reviews. Hypothesis 1B is not supported.

4.3.6 Haikus

The Haiku Society of America (HSA) defines the structure of the Japanese haiku as either "an unrhymed Japanese poem recording the essence of a moment keenly perceived, in which Nature is linked to human nature. It consists of seventeen onji (Japanese sound-symbols)" or "a foreign adaptation of [the above]. It is usually written in three lines of five, seven, and five syllables" (Haiku Society of America 2004). Others have defined "haiku" similarly, highlighting the consistency of the structure (Hirschfield 2011). Since Hiakus have a brief and highly structured form, they are useful bodies of text with a consistent structure that can be used for comparisons.

Matsuo Bashō is a well-known haiku writer (Hirschfield 2011) whose haiku titled “Old Pond” is presented in figure 11 in its original Japanese form, the romaji transliteration, and an English translation.

Figure 11 "Old Pond" (Basho, n.d.)

These unique structural (and, to some extent, content-centric) characteristics provide an opportunity to eliminate the variance resulting from structure within forensic, and other, linguistic-type analysis. Since the structure of a haiku is rigidly defined, any author writing a haiku must produce a structure similar to that produced by all other authors who have ever written a haiku. Thus, this presents an opportunity for a technical analysis of the linguistic structural components within haikus while isolating other components. Haikus have been discussed for their unique structure and potential interplay with technology in speculative fictional works (Howey 2011). Haikus are one of the most rigidly defined forms of text. Since multiple haikus (at least those within the standard format) have the same structure and very similar word counts (due to the limitations on the number of syllables), identifying authorship of haikus is a unique challenge because one cannot rely only on the structural patterns that might be present, which further highlights the extensibility of the method.

Fifty datasets of three haikus each were selected, using publicly available works online. A limitation is that the number of haikus available in English is much more limited than the

number of available songs or reviews. Because of the limitation of the number of haikus available in English (particularly paired works by the same author) it was difficult to find a large number of haikus to analyze. However, 150 total haikus were analyzed, which is sufficient for a proof-of-concept type test. In future research, ideally more haikus will be available. The first five results from analyzing the haikus data set are shown in table 24.

Dataset	AB (Same Author)	BC (Different Author)	AC (Different Author)
1	**0.076576577**	0.149189189	0.149189189
2	0.1	**0.108**	0.064285714
3	0.107638889	**0.178888889**	0.149758454
4	0.141025641	0.141025641	**0.271634615**
5	0	**0.089093702**	0

Table 12 Comparison of analyzed haikus

Twenty out of 50 pairs were correctly identified as being the work of the same author/artist, resulting in a probability of 0.40 or 40%, which is higher than the probability that a correct result would have occurred through random chance. Interestingly, this group of datasets yielded a tie, likely due to the fact that, since the structure of haikus is so rigid and word usage is relatively limited, it is much more likely for three haikus to have no words in common than it is for three books, songs or reviews. This may partially be because the haiku structure allows the author to drop common grammatical patterns, but it is unclear why this impact is so strong. The results are given in table 25.

	AB (Same Author)	BC (Different Author)	AC (Different Author)	TIE
Amount	**20**	11	13	6
Probability	**0.4**	0.22	0.26	0.12

Probability (without ties)	**0.454545455**	0.25	0.295454545	N/A

Table 13 Comparison of matched pairs for haikus

Since 0.40 is greater than the 0.33 probability that a correct result would have occurred through random chance, Hypothesis 1C is supported. If ties are considered to be an "unable to identify"-type result rather than an "incorrectly identified"-result, they are excluded from the total and the probabilities recalculated. When this is done the probability for all other categories rises, resulting in an even higher probability of 0.4545, lending more support to Hypothesis 1C.

4.3.7 Books

Results of the analysis from the first five data sets, each comprised of three books, out of the fifty total datasets are shown in table 26, with the largest value (meaning the two bodies of text are found to be most similar) listed in bold.

Data Set	**AB** (Same Author)	**BC** (Different Author)	**AC** (Different Author)
1	**0.880278**	0.857968	0.841821
2	0.778082	**0.890425**	0.751863
3	**0.883169**	0.728292	0.71815
4	**0.825125**	0.789261	0.753744
5	**0.890047**	0.780765	0.77339

Table 14 Comparison of analyzed books

The highest (bolded) value marks the comparison found to be most similar. To test the results, the number of times AB had the highest value was calculated with the results summarized in table 27.

	AB (Same Author)	**BC** (Different Author)	**AC** (Different Author)
Amount	**43**	5	2
Probability	**0.86**	0.1	0.04

Table 15 Comparison of matched-pairs for books

Texts A and B are written by the same author, so our hypothesis would suggest that the value for AB should be the largest in the majority of cases. The value for AB was the highest 86% of the time, supporting Hypothesis 1. When combined, BC and AC were the highest only 14% of the time when random chance would have suggested around 66%. Hypothesis 1D is supported.

4.3.8 Poems

Another, more rigorous test of this method uses poems, which vary much more than other forms of text in terms of structural patterns and have smaller word counts than full prose-length works. Works with smaller word counts provide less data to analyze, thus creating a more difficult test for the algorithm. Testing our method on a less convenient form of text and using a larger sample size than used in previous research, helps demonstrates the robustness and value of the method.

Even within the context of two poems by the same author, the author may purposefully vary the structure of the poems for artistic emphasis or impact. A dataset comprised of sixty sets of three poems each (approximately 180 poems in total) was analyzed to test whether our method could correctly identify common authorship more frequently than random chance would have.

Poems, as a genre, were selected due to their differences from other previously studied genres. Poems are generally shorter than other genres and represent a more general type of literature than haikus (which are themselves a specific form of poetry). Poetry is particularly appropriate for robustness tests because it is both more specific and more general in key components. There are three possible pairings for each set (AB, which are the two poems written by the same author, and BC and AC, which would have been written by different authors) so the odds of correctly identifying authorship at random would be one-third.

Table 28 presents the results of the analysis for each of the sixty sets. The highest number indicates the greatest degree of similarity and is labeled with bold text in the table below. A value of 1 indicates that the poems are identical and a value of 0 that there is no content in common between the pair of poems being analyzed.

Poem Set	AB (Same Author)	BC (Different Author)	AC (Different Author)
1	**0.371132376**	0.294288849	0.318181818
2	0.198351262	0.235530835	**0.344696296**
3	0.267941659	0.269613485	**0.365639446**
4	0.138207161	0.110370781	**0.312677247**
5	**0.446448398**	0.145693753	0.105306575
6	**0.492280854**	0.332201211	0.361732075
7	0.130522648	0.213379791	**0.220847291**
8	0.226834225	**0.378165485**	0.239483678
9	**0.371325632**	0.106222065	0.128452877
10	0.239678066	0.256961674	**0.279742474**
11	0.306247814	0.354132564	**0.432573449**
12	0.203759746	**0.353837198**	0.148443627
13	**0.193415987**	0.171651053	0.187019145
14	0.293945042	**0.377279631**	0.30857458
15	0.343571246	**0.348590996**	0.283851902
16	0.258444726	0.23077447	**0.296682139**
17	**0.313271605**	0.247501329	0.292994333
18	0.266623938	0.280577638	**0.367869478**
19	0.389949887	**0.422516718**	0.317596064
20	**0.291851608**	0.274676443	0.259512814

21	**0.437360105**	0.364495798	0.297928732
22	0.36455043	0.426033719	**0.435595039**
23	**0.418026534**	0.302757969	0.380227242
24	0.18828892	0.189937013	**0.226742098**
25	0.319103822	0.311018329	**0.360485076**
26	0.363160484	**0.373380151**	0.360073222
27	0.201384033	0.22054705	**0.287910937**
28	**0.415415749**	0.317217074	0.343256154
29	0.266499015	**0.422455343**	0.183197564
30	**0.424436269**	0.381101856	0.359030536
31	0.32122033	0.293369951	**0.339204254**
32	0.287563235	0.290135001	**0.366967741**
33	**0.412723221**	0.241068218	0.234036107
34	0.223809524	0.185615949	**0.321601222**
35	**0.28477817**	0.208547774	0.241837136
36	**0.368357862**	0.322839745	0.225838395
37	**0.451255142**	0.394004843	0.423340431
38	0.33979431	**0.369412442**	0.276089225
39	0.362626268	**0.438686305**	0.328725964
40	**0.384563994**	0.294499949	0.307465619
41	**0.210810811**	0.023104545	0.162600426
42	0.249971938	**0.358502947**	0.301214718
43	0.237924935	0.232879785	**0.243884374**
44	0.049940547	0.154965683	**0.180368425**
45	**0.281721746**	0.193229831	0.183220231
46	0.357334612	**0.470184854**	0.356143804
47	0.126439704	**0.328326606**	0.235044141
48	**0.478794818**	0.443846006	0.388896179
49	**0.339852742**	0.319954249	0.263866785
50	0.237927162	**0.251489706**	0.231494468
51	0.279237066	0.206549884	**0.294472338**
52	**0.408740942**	0.238884739	0.218824618
53	**0.422144771**	0.347034198	0.307248494
54	**0.288970191**	0.255195225	0.274252594
55	0.316048489	0.241139192	**0.349069922**
56	0.437166332	0.418790954	**0.443604841**
57	0.193755809	0.176615488	**0.33050265**
58	0.251074181	**0.285535147**	0.262147032
59	0.078924305	0.150421017	**0.208757187**
60	0.238864724	**0.322581927**	0.177782401

Table 16 Results of poem pairings

To be useful, the probability that the pair AB is correctly identified as the most similar pair of poems would need to be greater than 0.33. If AB is the most commonly identified pair – not just above 0.33, but also greater than all other possible pairs – this would further support the validity of the method. Table 29 shows a summary of these results:

	AB (Same Author)	**BC** (Different Author)	**AC** (Different Author)
Amount	**23**	15	22
Probability	**0.383333333**	0.25	0.366666667

Table 17 Summary of results for poem pairings

The pair AB was correctly identified as the most similar approximately 38.3% of the time, which is greater than we would expect to have found through random chance. It is also the pairing found to be the most similar pair more often than any other combination of poems, which further yields support for the validity of this method. Combined, these two pieces of information help to give users new information that could be used to help further strengthen existing models, particularly those reliant on linguistic data. Overall, this test indicates further support for the method.

4.4 Application 2: Improving automatic translation software

Automatic translation software, although obviously helpful in many circumstances, still does not perform as well as manual translation in terms of accuracy and readability (Vilar et al. 2006). When a user enters text into automatic translation software, they discover that, although the translated result is helpful, it often contains errors or the result is less readable than the results from a manual translation, sometimes with wording that a native speaker would not use (Salkoff

1990). Automatic translation is an application of natural language processing, which is an area of widely recognized importance within information systems research (Chowdhury 2003; Liddy 2001; Allen 2003; Collobert and Weston 2008; Lewis and Jones 1996). This study proposes one way of identifying errors during automatic translation. Specifically, we want to identify errors in automatic translation software that are instances of inaccurate transliterations.

Inaccurate transliterations are often the result of systematic patterns in both languages overall as well as how translation software works. Within any two languages using the same alphabet, such as the Latin alphabet, certain words are the same. Sometimes these are loan words (words borrowed from one language and adopted into another), but, most commonly, these words are proper nouns (Kashani 2007, Babych and Hartley 2003). An English speaker named "John," for example, would still be named "John" in German. Because proper nouns are often the same across languages, automatic translation software generally treats an unrecognizable word as a proper noun and transliterates it directly into the translated result. While this can sometimes be accurate, it is not always so.

To identify instances of text where automatic translation software has introduced unnecessary error, a "true ratio" of similarity is needed, which is defined in this research as the actual percentage of identical words across two languages. A certain percentage of words will be identical across translations even in a "perfect" translation, as many proper nouns and other words will be the same in any work, regardless of the language presented. This is known in theory, but in practice is difficult to calculate. The true ratio is needed to compare the ratio of identical words present in the result of an automatic translation. Because automatic translation software generally treats an unrecognizable word as a proper noun and transliterates it directly

into the translated result, this paper proposes that the "true ratio" of identical words between two languages will generally be lower than the ratio present in automatically translated results.

First, can a method be developed to accurately analyze the translated version of the same work in two languages and generate a relative comparison value? Second, if this method is successful in analyzing the English version and the German version of the same work and generate a relative comparison value, what is the average relative comparison value of the English language and the German language?

German was chosen for two reasons. Firstly, English and German both use the Latin alphabet, which allows for direct comparison instead of relying on phonetic spellings. The second is that, because English and German are from the same linguistic family, we assume that there may be a higher percentage of loan words shared between the two languages and thus more opportunities for identical words to be found. By using the same work translated manually into two languages, this research attempts to find the "true ratio" between two languages. Manual translations are necessary in order to compare against the results automatic translation.

4.4.1 Application of method

For the evaluation, we selected texts and formulated research questions about the effectiveness of the method proposed above and its potential results. To perform the evaluation, pairs of identical texts manually translated into two languages were identified and placed into our prototype. By using the same work translated manually into two languages, this research is an attempt to find the "true ratio" between two languages.

Firstly, a dataset must be generated by selecting texts to analyze. Each dataset consists of two texts paired together. For the purposes of finding a "true ratio," the two texts should be the same work translated into two different languages. To provide an example of what a dataset

might look like, we might select the novel *A Tale of Two Cities* by Charles Dickens. The first text would be the English version, starting with "It was the best of times, it was the worst of times…" and the second would be a version of the work presented in German, starting with "Es war die beste und die schlimmste Zeit…" (Dickens 1859, 2011). Subsequently, this hypothetical dataset would be as shown in table 30.

File Name	Work	Language
1A	A Tale of Two Cities	English
1B	A Tale of Two Cities (Eine Geschichte aus zwei Städten.)	German

Table 18 Hypothetical dataset for English-German pairing

The method developed to compare text is comprised of a set of steps that generate the data necessary to make the comparisons between the two works in each dataset.

4.4.2 Selection of Texts

Texts were selected using Project Gutenberg, an open source database of long-form works in the public domain (Project Gutenberg 2017). Ultimately, 15 pairs of English and German long-form works were selected. Long-form works were chosen due to an assumption that longer works have the potential to yield more accurate results. Rather than using random selection, a purposeful selection method was employed to find works that were available in both English and German. These works were required to have been in one of the following categories: an original English work and its German translation, an original German work and its English translation or an English translation and a German translation of the same work originally written in another language. For the third category, only the English and German translations were analyzed and the original work was not. In the end, 15 works were found to be viable.

***Question #1*:** Can a method be developed to accurately analyze the translated version of the same work in two languages and generate a relative comparison value?

***Question #2*:** If this method is successful in analyzing the English version and the German version of the same work and generate a relative comparison value, what is the average relative comparison value of the English language and the German language?

4.4.3 Results

The 15 pairs of English and German works were jointly placed into 15 datasets and then entered into the prototype and analyzed, yielding a relative comparison value for each dataset. After the 15 relative comparison values were generated, an average of the 15 values was generated to serve as an estimate for the "true ratio" of identical words between English and German.

This average is labeled as an "estimate" due to a variety of factors, most notably that a much larger body of manually translated works would be necessary to find a more accurate result. However, this study is testing only the viability of determining the "true ratio" and does not seek to actually find the "true ratio" or to compare it with the ratio present in automatically translated software. The estimate generated using this method should be sufficient to prove our method as a viable way of attempting to find the "true ratio."

The results from the analysis showing the relative comparison values for the 15 datasets and the average estimate are given in table 31.

Dataset	**Relative Comparison Value**
1	0.418158736

2	0.282692847
3	0.394856512
4	0.366572456
5	0.394478023
6	0.385623587
7	0.387949669
8	0.374445009
9	0.377504222
10	0.457032907
11	0.502493545
12	0.502493545
13	0.405978139
14	0.364905481
15	0.411887444
Average	0.401804808

Table 19 Relative comparison value for English-German pairs

The results, which vary from being approximately 28.2% identical to 50.2% identical, are relatively high. This is due to three factors. The first is that English and German are both Germanic languages with a high degree of similarity, so a reasonably high ratio is to be expected. Second, these results indicate a potentially high frequency of proper nouns in the works selected, which is plausible for long-form works that may have used a proper noun with a relatively high frequency (such as the name of the protagonist in a work of fiction or the name of the subject in a biography). Third, because the data was aggregated using an open source database of works in the public domain, legal disclaimers, notes and other additional information may have been identical across the translations, artificially raising the ratio. Future research may seek to identify to what extent these three factors impacted the ratio of identical words across the two works and to isolate potential variance or errors they may have placed into the results.

4.5 Discussion

4.5.1 Identifying Authorship application

The method is intended to provide a new form of analysis that could be designed and implemented to add useful surface-level data, contributing to modeling and comparing unstructured text. This is intended to generate useful information from the structure and wording of a body of text that could be then be used in more traditional models. The research was motivated by work in computational linguistics and text analysis that recognizes the potential of massive amounts of text data for customer relations and other applications. The values generated represent structural data that is difficult to measure, thus, providing a comparison value that provides useful information beyond existing methods. With books, for example, the algorithm was tested against simple random chance and provided an accurate determination of authorship 53% more often than random chance. Application of the method identifies similarities between texts without necessarily having to read the content directly. This might be useful for linguistic forensics or translation software, if a big data-style sample of works, translated between two languages, were compiled and analyzed to assess the extent of the similarity.

Number	Hypothesis	Supported?
1A	The songs written by the same author/artist should be correctly identified more than 33% of the time.	**Yes**
1B	The haikus written by the same author should be correctly identified more than 33% of the time.	**No**
1C	The reviews written by the same author should be correctly identified more than 33% of the time.	**Yes**
1D	The books written by the same author should be correctly identified more than 33% of the time.	**Yes**
1E	The poems written by the same author should be correctly identified more than 33% of the time.	**No**
1F	The text written by the same author should be correctly identified more than 33% of the time across all genres.	**Yes**
2A	The percentage of correctly identified haikus will be lower than or the same as the percentage of correctly identified songs.	**Yes**
2B	The percentage of correctly identified haikus will be lower than or the same as the percentage of correctly identified online reviews.	**No**
2C	The percentage of correctly identified haikus will be lower than or the same as the percentage of correctly identified online books.	**Yes**

Table 20 Results of hypotheses

Hypotheses 1A, 1C, 1D and 1F being supported supports the claim that this method is extensible in its application. In addition to being able to correctly identify joint authorship of books more often than random chance, it appears this method is also more accurate in terms of correctly identifying joint authorship of songs or haikus. The authorship issues resulting from reviews are unclear, requiring more research. The support for extensibility goes beyond the fact

that the program can confirm well-known linguistic patterns. One of the challenges is whether it is possible to avoid the limitations on authorship identification based upon structural patterns. Whereas traditional authorship identification techniques rely on syntax and other such patterns, this analysis focuses only on word usage and frequencies.

Thus, whether the structure had any impact on the accuracy was tested by isolating the structure using a set of fifty datasets that contained only haikus, a strict and rigid form of text. If differences in structure do have a positive impact on the accuracy of identifications, one would expect that haikus would be correctly identified in a lesser amount of the time. However, Haikus were more accurately identified correctly than reviews (40–45% of the time as compared to 32%), which supports Hypothesis 2C. Hypothesis 2B is also largely supported. If one includes the tie results, then the haikus are correctly identified a slightly lower amount of the time than songs (40% compared to 44%). When the ties are properly identified, the haikus were identified accurately a higher percentage of the time (45% to 44%), not supporting Hypothesis 2B. Hypothesis 2D is also supported, as the haikus are correctly identified significantly less often than books (44% of the time for haikus as compared to 86% of the time for books).

Since both of these numbers are higher than the 33% rate of correct identification suggested by random chance, Hypothesis 1F is supported. Finally, as haikus are correctly identified more often than the average across all genres – both with ties excluded (45.4% to 40.3%) and with ties included (40% to 38.6%). Therefore, while Hypothesis 2D is not supported, it is possible that, because the reviews were the least structured form of text, they may have had fewer words at times than the haikus, which would explain why hypotheses that included reviews in the analysis were not supported.

This is one such example beyond that of identifying common authorship and forensic linguistics in which this method may be useful. Having this additional data about the word patterns within bodies of text may be useful to integrate into a variety of models. Thus far, the data generated has primarily been presented as sufficient in its own right, but there is sufficient reason to believe that it could work well as supplementary data that serves not as a replacement to existing methods, but as a compliment to it. Other scholars may have the opportunity to adapt the method to other contexts beyond that which is described here or, as we have done, to new genres.

4.5.2 Language Commonalities Application

For the second application, this method analyzes the English version and the German version of the same work and generate a relative comparison value. However, to what extent it is accurate warrants further research and can only be determined by further testing. It is possible that one of the three factors presented above, particularly the third dealing with legal disclaimers, notes and other additional information that may have been identical across the translations, may have artificially raised the ratios of identical words between the texts. Further research can help to identify what impact, if any, these factors had on the accuracy of the ratios.

This method provided an estimate of the "true ratio" of identical words between English and German as approximately 40.2% (based on the average presented in Table 31 above), a relatively high ratio and almost certainly higher than the "true ratio" really is. However, because English and German are both Germanic languages, it is likely that these two languages may have one of the highest "true ratios" across language pairs and so the estimate proposed in this paper is likely reflecting this close linguistic relationship – English is a Germanic language and evolved out of the same linguistic family as German – even if it is potentially impacted by some

degree of error. However, factors such as the database used for work selection and the fact that long-form works were selected may have raised this ratio higher than is accurate. The results here show significant promise, but further testing on a wider variety of works from a larger number of sources is needed.

The results indicate potential for this method to determine the "true ratio" present between two languages. If such a "true ratio" can be calculated, it could have positive impacts for improving the accuracy of automatic translation software. By taking the "true ratio" and comparing it to the ratio present in automatically translated results, one can see exactly how often automatic translation software is transliterating results when it should not be doing so. This could provide a useful outlet for coders to identify where automatic translation software is introducing errors into its translation, which could help in debugging such software.

4.6 Conclusion

This research has proposed a corollary to Zipf's law and provides a new algorithm that can be used to analyze bodies of text. The primary purpose of this was to facilitate the integration of linguistic data into more traditional models used by information systems researchers, such as econometric models. In order to place this new algorithm in the larger context of information systems research, this research proposed a new set of assumptions known as linguistic component theory, which shows why information systems research could benefit from the integration of quantified, linguistic data. The research then outlined a new algorithm – which is inspired by, but is unique from Zipf's law – to analyze bodies of text and then provided two potential applications for the output the algorithm generates. These two potential applications produced quantified data that could be used in other models.

This algorithm was used for two applications, the first of which was comparing texts to identify those created by the same author. The method was implemented and tested. It is intended to be extensible and created for underrepresented applications such as haikus, arias, and foreign languages. The contribution is to successfully identify authorship without relying on traditional structural analysis. In contexts where the structure is uniform (e.g., homogenous groups) or not well-understood by outsiders (e.g., less frequently spoken languages), this could present opportunities for new forms of analysis and accuracy not previously possible if relying on structural patterns. The method has the potential to be effective in applications such as forensic linguistics. Additional genres (such as arias and blogs) will be analyzed in future research in order to further test the extensibility of the algorithm. In addition, sonnets might prove to be a useful area for future research as it is a rigidly defined genre with similarities to haikus. Further research is warranted in this area.

The second application was an attempt to find how frequently a word is correctly transliterated within automatic translations. By attempting to find a "true ratio" of words in common between languages, error detection within automatic translation software can potentially be improved. This research suggests that a "true ratio" of identical words may be found between English and German, but such analysis will require a substantially larger body of works to draw analysis from.

This research suggests two possible areas for future research and expansion. Firstly, the general trend within this research is that the predictive capabilities of the algorithm generally seemed to be positively correlated with works that were longer in nature, such as books being more accurately predicted than reviews. However, no direct comparison of subsets within a category (such as longer books as compared to shorter books) or a direct analysis of word count

was done, which could be a potentially fruitful area for future research but is beyond the scope of the MIS corollary to Zipf's law in its current form. In order to test this more fully, a new section calculating overall word count and would needed to be added to the algorithm that would consider word count as a factor for analysis in its own right. This new section added directly into the algorithm could be used to look at the overall word count of each body of text and use this to generate probabilities for accuracy.

Another area for expansion is to look into phonetic spelling of words rather than limiting the analysis to languages that use the Latin alphabet. In the current form, the algorithm could not be used to compare languages with differing alphabets, such as Chinese or Korean. By converting all words to a phonetic spelling, it may be possible to do this, but to do so would require a dramatic expansion of the MIS corollary that is currently beyond the scope of this research. To do this would require using far more than the Latin alphabet, which is currently the only form of input that can be handled by the application developed. However, it may be an area for future research that could yield useful additional information concerning these trends across languages – going beyond a comparison between English and German could yield valuable information for languages not reliant on the Latin alphabet, such as how similar these languages are or what words are most frequently dispersed across languages.

The algorithm presented in this chapter serves as a new method that can be used by information systems researchers to generate quantitative data about word frequencies that can integrated into more traditional models. More traditional econometric and quantitative models may benefit from being able to use linguistic data presented in a quantitative form, such as the exact number of times certain words are used within multiple bodies of text or the "true ratio" of

similarities between languages. This research represents a first step in the development of one such algorithm – research that will be hopefully expanded upon in the future.

5 CONCLUSION

This research has analyzed the relationship between emerging technologies and globalization by studying the relationship at three different levels; the policy level, the corporate level and the application level. These three studies combine to attempt to answer an overarching research question: in what ways are emerging technologies and the phenomenon of globalization interacting?

By analyzing the relationship at three different levels, this research is able to present specific ways in which emerging technology and the process of globalization are shaping each other. the following outlines what the research has found so far and what remains in this domain.

5.1 Findings

5.1.1 Policy Level

The first study analyzed information leakers and their requests for asylum. This was analysis was conducted from a high-level policy standpoint, seeking to understand the motivations behind nations' decisions on whether to grant asylum and then using those motivations to attempt to predict future behavior. In order to do this, two case studies were analyzed – Julian Assange and his request for asylum in Ecuador and Edward Snowden and his request for asylum in Russia.

By analyzing these two cases, this research was then able to present a proposed game-theoretical outlook for analyzing this issue in the future. The policy issue of whether or not countries should grant asylum is modeled as a "stag hunt" game in which each country is a player and must decide whether to cooperate by rejecting all asylum requests or to defect and grant an asylum request. If even one nation defects, an international norm of rejecting all asylum

requests is broken, but countries who did choose to defect gain at the margins by gaining control over a specific leaker.

Using this model, the paper makes predictions for future behavior on the part of policy makers. In addition, it shows how game-theoretic approaches might be useful for future studies concerning eGovernance. Finally, the paper takes a novel approach to the issue by adding the stag as a player to the game-theoretic model – usually the stag is presented as a passive participant, but this extension of the model presented here could be useful to researchers in information systems and elsewhere.

5.1.2 Corporate Level

The second study outlines a proposal for using corporate level data to determine the differences between groups of companies with regards to their hiring and budgeting habits, with a particular emphasis placed on their behaviors concerning offshoring and outsourcing labor. Companies are divided into different groups based on the proposed hypotheses, with divisions based on industry and country of origin.

The regression analysis has now been conducted. Early results from the data show a relatively small percentage of companies based outside of the United States, which did present a limitation for finding statistically significant results for the hypotheses concerning country of origin. The results were subsequently limited to trends regarding companies' declared industries, as presented in the following table:

Hypothesis	Trending	Significant	Supported
1A	**Yes**	**Yes**	**Yes**
1B	No	No	No
1C	**Yes**	**Yes**	**Yes**
1D	No	No	No
1E	**Yes**	**Yes**	**Yes**

Table 21 Results of corporate level hypotheses testing

This study has significant findings in some cases and suggests several others where significant findings may be found with a larger dataset. It represents a worthwhile first step to testing trends in the literature that have thus far been largely untested.

5.1.3 Application Level

The third study focuses on the most specific level – the application level – and takes a design science approach to show how emerging technologies can help to include linguistic structural data into more mainstream findings. In order to do this, an algorithm is developed to find structural similarities between large bodies of text. This algorithm is developed and implemented in a program that is then tested in two different contexts.

The first context is a forensic linguistic application where the algorithm is used to determine the likelihood of two large bodies of text having the same author. This was then tested on a variety of genres of text, with the hypotheses concerning the algorithm's effectiveness being largely verified and affirmed.

The second context is to use the algorithm as part of a larger method for improving error detection in automatic translation software. In order to do this, the algorithm is used to determine a ratio of similarity between the English and German languages. Once this ratio is determined, the study concludes by outlining how this could be used to improve error detection in automatic translation software in the future.

Number	Hypothesis	Supported?
1A	The songs written by the same author/artist should be correctly identified more than 33% of the time.	**Yes**
1B	The haikus written by the same author should be correctly identified more than 33% of the time.	**No**
1C	The reviews written by the same author should be correctly identified more than 33% of the time.	**Yes**
1D	The books written by the same author should be correctly identified more than 33% of the time.	**Yes**
1E	The poems written by the same author should be correctly identified more than 33% of the time.	**No**
1F	The text written by the same author should be correctly identified more than 33% of the time across all genres.	**Yes**
2A	The percentage of correctly identified haikus will be lower than or the same as the percentage of correctly identified songs.	**Yes**
2B	The percentage of correctly identified haikus will be lower than or the same as the percentage of correctly identified online reviews.	**No**
2C	The percentage of correctly identified haikus will be lower than or the same as the percentage of correctly identified online books.	**Yes**

Table 22 Results of application level hypotheses testing

The majority of the hypotheses for the proof-of-concept test using different genres were supported. The research questions regarding the potential for a "true ratio" using this method to analyze across languages remains somewhat unclear and warrants future study.

5.2 Future Research

5.2.1 Policy Level

The situation with Julian Assange continues to develop and, since historical research methods and techniques were used to develop the model in this paper, these future developments may yet shape the models being developed. As of September 14, 2019, Assange is still imprisoned within the United Kingdom and is facing a potential extradition to the United States (PA Media 2019). Although Ecuador has revoked Assange's asylum, it is conceivable that Assange might seek asylum elsewhere in order to attempt to avoid extradition to the United States. Should he do so, this would provide a new case study that could be used to test and refine the two-good theory developed in this paper.

5.2.2 Corporate Level

There are, however, limitations on what this dataset contained with regards to information on companies based abroad. This led to a difficulty finding statistical significance in many cases. Statistically significant conclusions were reached on 5 of the 17 hypotheses outlined, but the remaining 12 warrant being studied further with a larger dataset. The majority of the remaining hypotheses were trending in the hypothesized direction, but a larger dataset may be necessary in order to more clearly define the trends and to find statistical significance.

5.2.3 Application Level

This research has outlined many of the various applications of Zipf's law, developed an MIS Corollary consisting of a new algorithm based on Zipf's underlying logic and offered a proof-of-concept test of the MIS Corollary using a test for potential improvement of error detection in automatic translation software. Applying the underlying logic of Zipf's law to a

newly developed algorithm with an emphasis on practical application could be useful for future research in information systems. Many fields have effectively used Zipf's law to extend their own research and it is time that information systems began to apply these widespread findings as well.

More rigorous testing can be done in the future on both proofs of concept. With regards to the authorship testing, there are trends that suggest that including word count in the analysis may prove fruitful for future research. With regards to the second application on translation, a larger corpus with a wider variety of texts is necessary to truly define a "true ratio" of similarity between the two languages.

5.3 Conclusion

This research has proposed and conducted studies that attempt to answer an overarching research question: in what ways are emerging technologies and the phenomenon of globalization interacting? While progress has been made to show some of the ways that this is happening, further expansion is needed on each of the three studies. Each study represents a novel attempt to look at how and why technologies are shaping an increasingly globalized world, but new technologies continue to be developed and, in many ways, the world grows ever smaller. It is my sincere hope that this dissertation represents a small but notable contribution to understand why these things are occurring and how we might better prepare for them in the future.

REFERENCES

CHAPTER 1 – INTRODUCTION:

Abeles, P. "TECHNOLOGY, GLOBALIZATION AND THE UNIVERSITY." *International Journal World Peace* 15, no. 3 (September 01, 1998): 29-44.

Ambrose, S. H. "Paleolithic Technology and Human Evolution." *Science* 291, no. 5509 (2001): 1748-753.

Batley, Edward. "Language Learning and the Technology of International Communications." *International Review of Education / Internationale Zeitschrift FÃ¼r Erziehungswissenschaft / Revue Internationale De L'Education* 37, no. 1, Language Policy and Education (January 01, 1991): 149-62.

Bird, Steven, and Gary Simmons. "Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources." *Computers and the Humanities* 37, no. 4 (2003): 375-88.

Bryant, A., Black, A., Land, F., & Porra, J. (2013). What is history? What is IS history? What IS history? ... and why even bother with history? *Journal of Information Technology*, *28*(1) 1-17.

Carlson, Paul E. "PERSPECTIVES: Educational Adaptation to Language and Technology: An Anthropological Perspective." *Language Arts* 53, no. 7 (October 01, 1976): 815-21.

Cook, Susan E. "New Technologies and Language Change: Toward an Anthropology of Linguistic Frontiers." *Annual Review of Anthropology* 33, no. 1 (2004): 103-15.

Cottrell, L. S., and E. B. Sheldon. "Problems of Collaboration between Social Scientists and the Practicing Professions." *The ANNALS of the American Academy of Political and Social Science* 346, no. 1 (1963): 126-37.

Crawford, Kathryn. "Language and Technology in Classroom Settings for Students from Non-Technological Cultures." *For the Learning of Mathematics* 10, no. 1 (February 01, 1990): 2-6.

Cunningham, Ann, and Mary Lynn Redmond. "Instructional Design and Early Language Learning: Cognition, Creativity, and Technology." *Hispania* 91, no. 2 (May 01, 2008): 435-45.

Eisenlohr, Patrick. "Language Revitalization and New Technologies: Cultures of Electronic Mediation and the Refiguring of Communities." *Annual Review of Anthropology* 33, no. 1 (2004): 21-45.

Gerr, Stanley. "Language and Science the Rational, Functional Language of Science and Technology." *Philosophy of Science* 9, no. 2 (1942).

Ho, Caroline M. L. "Review: English Language Learning and Technology." *Language* 82, no. 1 (March 01, 2006): 191.

Johnson, M. Eric. "Supply Chain Management: Technology, Globalization, and Policy at a Crossroads." *Interfaces* 36, no. 3 (May 01, 2006): 191-93.

Keating, Elizabeth, and Gene Mirus. "American Sign Language in Virtual Space: Interactions between Deaf Users of Computer-mediated Video Communication and the Impact of Technology on Language Practices." *Language in Society* 32, no. 05 (2003).

Kern, Richard. "Perspectives on Technology in Learning and Teaching Languages." *TESOL Quarterly* 40, no. 1 (2006): 183.

Kraemer, Angelika. "Happily Ever After: Integrating Language and Literature through Technology?" *Die Unterrichtspraxis/Teaching German* 41, no. 1 (2008).

Long, Guillaume. “Ecuador's Case for Assange's Asylum Is Stronger than Ever.” *OpenDemocracy*, 27 July 2018, www.opendemocracy.net/en/democraciaabierta/ecuador-s-case-for-assange-s-asylum-is-stronger-than-ever/.

Lum, Chee-Hoo. "Home Musical Environment of Children in Singapore: On Globalization, Technology, and Media." *Journal of Research in Music Education* 56, no. 2 (2008): 101-17.

Mariani, Joseph. "Developing Language Technologies with the Support of Language Resources and Evaluation Programs." *Language Resources and Evaluation* 39, no. 1 (2005).

Reid, Proctor P. "The Globalization of Technology." *Issues in Science and Technology* 7, no. 4 (Summer 1991): 92-93.

Ruiter, Marina B., Lilian J. Beijer, Catia Cucchiarini, Emeil J. Krahmer, Tony CM Rietveld, Helmer Strik, and Hugo Van Hamme. "Human Language Technology and Communicative Disabilities: Requirements and Possibilities for the Future." *Language Resources and Evaluation* 46, no. 1, Linguistic Annotation (January 01, 2012)

Saini, A. "A Good Sign." *Science* 325, no. 5939 (2009): 391.

Scheve, K.F., Slaughter, M.J. and Slaughter, M., 2001. *Globalization and the perceptions of American workers*. Peterson Institute.

Squires, Lauren. "Enregistering Internet Language." *Language in Society* 39, no. 04 (2010): 457-92.

Storper, Michael. "The Limits to Globalization: Technology Districts and International Trade." *Economic Geography* 68, no. 1 (January 1992): 60-93.

Van Pelt, Tamise. "The Question concerning Theory: Humanism, Subjectivity, and Computing." *Computers and the Humanities* 36, no. 3 (2002): 307-18.

Warschauer, Mark. "A Developmental Perspective on Technology in Language Education." *TESOL Quarterly* 36, no. 3 (2002): 453.

Waters, Sandie H., and Andrew S. Gibbons. "Design Languages, Notation Systems, and Instructional Technology: A Case Study." *Educational Technology Research and Development* 52, no. 2 (2004): 57-68.

Williams, Jessica. "Reaching a Global Public: Language, Technology and the Practice of Interpretation." *Future Anterior: Journal of Historic Preservation* History, no. Theory, and Criticism, Vol. 2, No. 1 (July 01, 2005).

Wrigley, Heide Spruck. "Technology and the Language Classroom. Ways of Using Technology in Language and Literacy Teaching." *TESOL Quarterly* 27, no. 2 (1993): 318.

CHAPTER 2 – POLICY LEVEL:

Addley, E., and Travis, A. 19 May 2017. “Swedish Prosecutors Drop Julian Assange Rape Investigation.” *The Guardian*, Guardian News and Media,

www.theguardian.com/media/2017/may/19/swedish-prosecutors-drop-julian-assange-investigation.

Associated Press. “Ecuador President Blames WikiLeaks for Leak of Private Data.” *Fox News*, FOX News Network, 2 Apr. 2019, www.foxnews.com/world/ecuador-president-blames-wikileaks-for-leak-of-private-data.

BBC News. (13 March 2015) "Q&A: Julian Assange and the Law." Retrieved 14 June 2017.

Bedi, K., Singh , P.J. & Srivastava, S. 2001. Government net: New governance opportunities for India. New Delhi: Sage.

Burrell, G., & Morgan, G. (1979). *Sociological paradigms and organizational analysis*. London: Heinemann Books.

Charmaz, Kathy. *Constructing Grounded Theory*. SAGE Publications Ltd, 2014.

The Courier-Mail. (29 July 2010) “Wikileaks founder Julian Assange a born and bred Queenslander." Retrieved 14 June 2016.

Dellarocas, C. (2003) “The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms.” *Management Science* 49(10):1407-1424, 2003.

Greenwald, Glenn. “Ecuador Will Imminently Withdraw Asylum for Julian Assange and Hand Him Over to the U.K. What Comes Next?” *The Intercept*, 21 July 2018, theintercept.com/2018/07/21/ecuador-will-imminently-withdraw-asylum-for-julian-assange-and-hand-him-over-to-the-uk-what-comes-next/.

Greenwald, G, Macaskill, E, and Poitras, L. (11 June 2013). “Edward Snowden: the whistleblower behind the NSA surveillance revelations.” *The Guardian,* Web. 15 June 2017.

Harley, N. (14 June 2015) "British Spies Removed from Operations after Russia and China Crack Codes to Leaked Snowden Files." *The Telegraph*. Telegraph Media Group. Web. 14 June 2017.

Henley, Jon. “Ecuador Cuts off Julian Assange's Internet Access at London Embassy.” *The Guardian*, Guardian News and Media, 28 Mar. 2018, www.theguardian.com/media/2018/mar/28/julian-assange-internet-connection-ecuador-embassy-cut-off-wikileaks.

Holmes, D. 2001. eGov: eBusiness Strategies for Government. London, U.K.: Nicholas Brealey.

Iivari, J., Hirschheim, R., & Lyytinen, K. (1998). “Paradigmatic Analysis Contrasting Information Systems Development Approaches and Methodologies.” *Information Systems Research*, *9*(2) 164-193.

“Julian Assange: Wikileaks Co-Founder Arrested in London.” *BBC News*, BBC, 12 Apr. 2019, www.bbc.com/news/uk-47891737.

“Julian Assange: Wikileaks Co-Founder Jailed over Bail Breach.” *BBC News*, BBC, 1 May 2019, www.bbc.com/news/uk-48118908.

Kaser, C. (2002) Trust and Reputation Building in e-Commerce. *Scientific publication* 2002s-75, Center for Interuniversity Research and Analysis on Organizations (CIRANO), Montreal, Quebec, Canada, http://www.cirano.qc.ca/pdf/publication/2002s-75 .pdf.

Lam, L. (25 June 2013) "Snowden Sought Booz Allen Job to Gather Evidence on NSA Surveillance." *South China Morning Post*. Web. 15 June 2017.

Li, L., and Whang, S. (2002). “Game Theory Models in Operations Management and Information Systems.” *Game theory and business applications*, 95-131.

NBC News (26 May 2014) "Who Is Edward Snowden, the Man Who Spilled the NSA's Secrets?" *NBCNews.com*. NBCUniversal News Group. Web. 15 June 2017.

NBC News (26 May 2014b) "Edward Snowden: A Timeline." *NBCNews.com*. NBCUniversal News Group. Web. 15 June 2017.

Oliphant, R. (January 18, 2017) "Russia Extends Edward Snowden Asylum until 2020." *The Telegraph*.

Okot-Uma, R.W. 2000. Electronic Governance: Re-inventing Good Governance. London, U.K.: *Commonwealth Secretariat*.

PA Media. "Julian Assange to Remain in Jail Pending Extradition to US." *The Guardian*, Guardian News and Media, 14 Sept. 2019, www.theguardian.com/media/2019/sep/14/julian-assange-to-remain-in-jail-pending-extradition-to-us.

Palmer, Glenn, and T. Clifton. Morgan. *A Theory of Foreign Policy*. Princeton University Press, 2011.

Palvia, S.C.J. and Sharma, S.S., 2007, December. "E-government and e-Governance: Definitions/Domain Framework and Status around the World." *International Conference on E-governance*. 1-12.

Papadimitriou, C. (2001) "Algorithms, games, and the Internet." *Proc. 33rd Annual ACM Sympos. Theory Comput. Association for Computing Machinery*, Hersonissos, Greece, 749–753.

Porra, J., Hirschheim, R. and Parks, M.S., 2014. “The Historical Research Method and Information Systems Research.” *Journal of the Association for Information Systems*, *15*(9), p.536.

Roth, A. 2002. The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics. *Econometrica* 70(4) 1341–1378.

Rothwell, J. & Ward, V. 19 May 2017. "Julian Assange Emerges on Embassy Balcony to Say He Will Not 'Forgive or Forget' as Swedish Rape Investigation is Dropped." *The Daily Telegraph*. Retrieved 14 June 2017.

Roussaeu, J.J. (1755) *A Discourse on Inequality*.

Sanchez, R. (30 July 2013) "WikiLeaks Q & A: Who Is Bradley Manning and What Did He Do?" *The Telegraph*. Telegraph Media Group. Web. 14 June 2017.

Skryms, B. 2004. *The Stag Hunt and the Evolution of Social Structure*, Cambridge University Press, Cambridge.

The Telegraph. 28 Apr. 2011. "WikiLeaks: The Guantánamo Files Database." *The Telegraph.* Telegraph Media Group, Web. 14 June 2017.

Wikileaks. 19 May 2017. "UK Refuses to Confirm or Deny Whether It Has Already Received a US Extradition Warrant for Julian Assange. Focus Now Moves to UK." *Twitter*. Twitter. Web. 14 June 2017.

Wilson, R. 1985. “Reputations in Games and Markets.” A. Roth, ed. *Game-Theoretic Models of Bargaining*. Cambridge University Press, Cambridge, U.K., 27–62.

Zhu, K. (1999) Strategic Investment in Information Technologies: A Real- Options and Game-Theoretic Approach. Doctoral dissertation, Stanford University, Stanford, CA.

Zhu, K. (2004) “Information Transparency of Business-to-Business Electronic Markets: A Game-Theoretic Analysis.” *Management Science* 50(5):670-685.

CHAPTER 3 – CORPORATE LEVEL:

“About SIM.” Accessed 2019. *About SIM - Society for Information Management*, Society for Information Management, www.simnet.org/page/About_SIM.

Baker, A.R., 1998. Military service and migration in nineteenth-century France: Some Evidence from Loir-Et-Cher. *Transactions of the Institute of British Geographers*, *23*(2), pp.193-206.

Bazerman, M.H. and Moore, D.A., 2012. Judgment in managerial decision making. 8th edition. Wiley.

Bendel, P., 2005. Immigration policy in the European Union: Still bringing up the walls for fortress Europe?. *Migration Letters*, *2*(1), p.20.

Boudreau, M.C., Loch, K.D., Robey, D. and Straub, D., 1998. Going global: Using information technology to advance the competitiveness of the virtual transnational organization. *The Academy of Management Executive*, *12*(4), pp.120-128.

Burgoon, B., Fine, J., Jacoby, W. and Tichenor, D., 2010. Immigration and the transformation of American unionism. *International Migration Review*, *44*(4), pp.933-973.

Carmel, E. and Agarwal, R., 2006. The maturation of offshore sourcing of information technology work. In *Information Systems Outsourcing* (pp. 631-650). Springer Berlin Heidelberg.

Chakravorti, R., 1996. Labouring class stratification: US Style. *Economic and Political Weekly*, pp.1127-1128.

Denetdale, J., 2008. Carving Navajo national boundaries: Patriotism, tradition, and the Diné Marriage Act of 2005. *American Quarterly*, *60*(2), pp.289-294.

Dibbern, J., Goles, T., Hirschheim, R. and Jayatilaka, B., 2004. Information systems outsourcing: a survey and analysis of the literature. *ACM Sigmis Database*, *35*(4), pp.6-102.

Feenstra, R.C. and Hanson, G.H., 1996. *Globalization, Outsourcing, and Wage Inequality* (No. w5424). National Bureau of Economic Research.

Freeman, R.B., 2006. Does globalization of the scientific/engineering workforce threaten US economic leadership? In *Innovation Policy and the Economy, Volume 6* (pp. 123-158). The MIT Press.

Gonzalez, R., Gasco, J. and Llopis, J., 2006. Information systems offshore outsourcing: A descriptive analysis. *Industrial Management & Data Systems*, *106*(9), pp.1233-1248.

Gore, D.F. 2014. "Review: A band of noble women: Racial politics in the women's peace movement." *The Journal of African American History* 99.1-2, Special Issue: “Rediscovering the Life and Times of Frederick Douglass” pp. 138-40.

Grinblatt, M. and Keloharju, M., 2000. Distance, language, and culture bias: The role of investor sophistication.

Gürsel, B., 2008. Citizenship and military service in Italian-American relations, 1901-1918. *The Journal of the Gilded Age and Progressive Era*, *7*(03), pp.353-376.

Harris, N., 2010. Immigration and state power. *Economic and Political Weekly*, pp.8-11.

Hashmi, M., 2006. Outsourcing the American Dream? Representing the stakes of IT Globalisation. *Economic and Political Weekly*, pp.242-249.

Henley, J., 2006. Outsourcing the provision of software and IT-enabled services to India: emerging strategies. *International Studies of Management & Organization*, *36*(4), pp.111-131.

Hickman, D.C. and Olney, W.W., 2011. Globalization and investment in human capital. *Industrial & Labor Relations Review*, *64*(4), pp.654-672.

Hira, R., 2004. US immigration regulations and India's information technology industry. *Technological Forecasting and Social Change*, *71*(8), pp.837-854.

Iacovetta, F., Quinlan, M. and Radforth, I., 1996. Immigration and labour: Australia and Canada compared. *Labour History*, pp.90-115.

Kanter, S., 1985. Sacrificing national defense to class interest: The French Military Service Law of 1872. *The Journal of Military History*.

Leal, D.L., 1999. It's not just a job: military service and Latino political participation. *Political Behavior*, *21*(2), pp.153-174.

Lin, F., Fofanah, S.S. and Liang, D., 2011. Assessing citizen adoption of e-Government initiatives in Gambia: A validation of the technology acceptance model in information systems success. *Government Information Quarterly*, *28*(2), pp.271-279.

Lyons, G. 2016. *The UK Referendum: An Easy Guide to Leaving the EU.* Amazon Digital Services LLC.

Lyytinen, K.J., 1985. Implications of theories of language for information systems. *MIS Quarterly*, pp.61-74.

MacKinnon, M., 1997. Canadian railway workers and World War I military service. *Labour/Le Travail*, *40*, pp.213-234.

Stephan, M., Silvia, M., and Lewin, A.Y. 2008. A dynamic perspective on next-generation offshoring: The global sourcing of science and engineering talent. *The Academy of Management Perspectives*, *22*(3), pp.35-54.

Mithas, S. and Lucas Jr, H.C., 2010. Are foreign IT workers cheaper? US visa policies and compensation of information technology professionals. *Management Science*, *56*(5), pp.745-765.

Nasmyth, G., 1916. Universal military service and democracy. *The Journal of Race Development*, *7*(2), pp.208-219.

Neckerman, K.M. and Kirschenman, J., 1991. Hiring strategies, racial bias, and inner-city workers. *Social Problems*, *38*(4), pp.433-447.

Pattnaik, B.K., 2013. Globalization, ICT revolution in India and socio-cultural changes: Sociological explorations. *Polish Sociological Review*, pp.39-62.

Petersen, T., Saporta, I. and Seidel, M.D.L., 2000. Offering a Job: Meritocracy and social networks. *American Journal of Sociology*, *106*(3), pp.763-816.

Pla-Barber, J., Linares, E. and Ghauri, P.N. 2018. The choice of offshoring operation mode: A behavioural perspective. *Journal of Business Research*.

Propes, E., 2011. Re-thinking antimilitarism: France 1898–1914. *Historical Reflections*, *37*(1), pp.45-59.

Purkiss, S.L.S., Perrewé, P.L., Gillespie, T.L., Mayes, B.T. and Ferris, G.R., 2006. Implicit sources of bias in employment interview judgments and decisions. *Organizational Behavior and Human Decision Processes*, *101*(2), pp.152-167.

Qari, S., Konrad, K.A. and Geys, B., 2012. Patriotism, taxation and international mobility. *Public choice*, *151*(3-4), pp.695-717.

O'Brien, P., 2007. New-media art: An Irish context. *Circa*, (120), pp.34-39.

Olney, W.W., 2012. Offshoring, immigration, and the native wage distribution. *Canadian Journal of Economics/Revue canadienne d'économique*, *45*(3), pp.830-856.

Ottaviano, G.I., Peri, G. and Wright, G.C., 2013. Immigration, offshoring, and American jobs. *The American Economic Review*, *103*(5), pp.1925-1959.

Rai, A., Maruping, L.M. and Venkatesh, V., 2009. Offshore information systems project success: the role of social embeddedness and cultural characteristics. *MIS quarterly*, pp.617-641.

Ranganathan, C. and Balaji, S., 2007. Critical capabilities for offshore outsourcing of information systems. *MIS Quarterly Executive*, *6*(3), pp.147-164.

Rao, H.R., Nam, K. and Chaudhury, A., 1996. Information systems outsourcing. *Communications of the ACM*, *39*(7), pp.27-29.

Reitz, J.G., 2005. Tapping immigrants' skills: New directions for Canadian immigration policy in the knowledge economy. *Law & Bus. Rev. Am.*, *11*, p.409.

Rodríguez-Clare, A., 2010. Offshoring in a ricardian world. *American Economic Journal: Macroeconomics*, *2*(2), pp.227-258.

Rubini, M. and Menegatti, M., 2008. Linguistic bias in personnel selection. *Journal of Language and Social Psychology*, *27*(2), pp.168-181.

Salyer, L.E., 2004. Baptism by Fire: Race, Military Service, and US Citizenship Policy, 1918–1935. *The Journal of American History*, *91*(3), pp.847-876.

Sokolowski, J., 2009. Internment and post-war Japanese American literature: toward a theory of divine citizenship. *MELUS: Multi-Ethnic Literature of the US*, *34*(1), pp.69-93.

Storey, M.M. 2004. "Review: The war was you and me: Civilians in the American Civil War." *The Journal of Southern History* 70.1 pp. 155-56.

Teigen, J.M., 2006. Enduring effects of the uniform: Previous military experience and voting turnout. *Political Research Quarterly*, *59*(4), pp.601-607.

Tversky, A. and Kahneman, D., 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), pp.207-232.

Tversky, A. and Kahneman, D., 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), p.293.

Varma, R. and Rogers, E.M., 2004. Indian cyber workers in US. *Economic and Political Weekly*, pp.5645-5652.

Venkatesh, V., Thong, J.Y., Chan, F.K., Hu, P.J.H. and Brown, S.A., 2011. Extending the two-stage information systems continuance model: Incorporating UTAUT predictors and the role of context. *Information Systems Journal*, *21*(6), pp.527-555.

Vertovec, S., 2006. Is circular migration the way forward in global policy? *Around the globe*, *3*(2), p.38.

Yomogida, M. and Zhao, L., 2010. Two-Way Outsourcing, International Migration, and Wage Inequality. *Southern Economic Journal*, *77*(1), pp.161-180.

CHAPTER 4 – APPLICATION LEVEL:

Adamic, Lada A. "Zipf, power-laws, and pareto-a ranking tutorial." *Xerox Palo Alto Research Center, Palo Alto, CA, http://ginger. hpl. hp. com/shl/papers/ranking/ranking. html* (2000).

Adamic, Lada A., and Bernardo A. Huberman. "Zipf's law and the Internet." *Glottometrics* 3, no. 1 (2002): 143-150.

Allen, J.F. 2003. Natural language processing.

Alzahrani, S.M., Salim, N. and Alsofyani, M.M., 2009, April. Work in progress: Developing Arabic plagiarism detection tool for e-learning systems. In *Computer Science and Information Technology-Spring Conference, 2009. IACSITSC'09. International Association of* (pp. 105-109). IEEE.

Baayen, H. 1992. Statistical Models for Word Frequency Distributions: A Linguistic Evaluation. Computers and the Humanities 26, 347-363.

Babych, Bogdan, and Anthony Hartley. "Improving machine translation quality with automatic named entity recognition." In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*, pp. 1-8. Association for Computational Linguistics, 2003.

Barber, R. 2009. “Shakespeare Authorship Doubt in 1593, *Critical Survey,* (*21*:2, Questioning Shakespeare), pp. 83-110.

Bashō, Matsuo. N.d. Old Pond.

Binali, H., Wu, C., & Potdar, V. 2010. Computational approaches for emotion detection in text. In *Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on* (pp. 172-177). IEEE.

Breslau, Lee, Pei Cao, Li Fan, Graham Phillips, and Scott Shenker. *On the implications of Zipf's law for web caching*. Technical Report CS-TR-1998-1371, University of Wisconsin, Madison, 1998.

Carr, N. G. 2008. *The big switch: Rewiring the world, from Edison to Google*. New York: W.W. Norton &. Co.

Cebrian, M., Alfonseca, M. and Ortega, A., 2009. Towards the validation of plagiarism detection tools by means of grammar evolution. *IEEE Transactions on Evolutionary Computation*, *13*(3), pp.477-485.

Chang, T.H., Hsu, S.C., Wang, T.C. and Wu, C.Y., 2012. Measuring the success possibility of implementing ERP by utilizing the Incomplete Linguistic Preference Relations. *Applied Soft Computing*, *12*(5), pp.1582-1591.

Chen, C.T. and Tai, W.S., 2005. Measuring the intellectual capital performance based on 2-tuple fuzzy linguistic information. In *The 10th Annual Meeting of APDSI, Asia Pacific Region of Decision Sciences Institute* (Vol. 20).

Chowdhury, G.G. 2003. "Natural language processing." *Annual review of information science and technology* 37, no. 1: 51-89.

Collins, M., & Duffy, N. 2002. Convolution kernels for natural language. In *Advances in neural information processing systems* (pp. 625-632).

Collobert, R. and Weston, J. 2008. July. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167). ACM.

Conrad, J. 2011. "Interstate Rivalry and Terrorism: An Unprobed Link." *The Journal of Conflict Resolution,* (55:4), pp. 529-55.

Cook, R.T. 2007. "Embracing the Revenge: On the Indefinite Extensibility of Language," in J.C. Bealt, ed., Revenge of the Liar, Oxford: Oxford University Press, pp. 31-52.

Cook, R.T. 2009. "What is a Truth Value and How Many Are There?" *Studia Lógica*, (92), 183-201.

Coulthard, Malcolm, and Rui Sousa-Silva. "Forensic Linguistics." *What Are Forensic Sciences?: Concepts, Scope and Future Perspectives*, edited by Ricardo J Dinis-Oliveira and Teresa Magalhães, Pactor, 2016.

Crossan, M. M., & Apaydin, M. (2010). A Multi-Dimensional Framework of Organizational Innovation: A Systematic Review of the Literature. Journal of management studies, 47(6), 1154-1191.

Dahui, Wang, Li Menghui, and Di Zengru. "True reason for Zipf's law in language." *Physica A: Statistical Mechanics and its Applications* 358, no. 2-4 (2005): 545-550.

Dickens, C. 1859. *A Tale of Two Cities*. Chapman & Hall.

Dickens, C. 2011. *Eine Geschichte Aus Zwei Städten*. Insel Verlag.

Ellis, Stephen R., and Robert J. Hitchcock. "The emergence of Zipf's law: Spontaneous encoding optimization by users of a command language." *IEEE transactions on systems, man, and cybernetics* 16, no. 3 (1986): 423-427.

Ferrer i Cancho, R. and Solé, R.V. 2003. Least Effort and the Origins of Scaling in Human Language. Proceedings of the National Academy of Science of the United State of America 100, 788-791.

Fielt E, Gregor S (2016) What's new about digital innovation?. Information Systems Foundation Workshop, Canberra.

Fichman, R. G., Dos Santos, B. L., & Zheng, Z. E. (2014). Digital Innovation as a Fundamental and Powerful Concept in the Information Systems Curriculum. MIS Quarterly, 38(2), 329.

"Free Ebooks by Project Gutenberg." Project Gutenberg. Accessed May 06, 2017. https://www.gutenberg.org/.

Frey, B.S. 1987. "Fighting Political Terrorism by Refusing Recognition." *Journal of Public Policy* (7:2), pp. 179-88.

Gabaix, Xavier. "Zipf's Law and the Growth of Cities." *American Economic Review* 89, no. 2 (1999): 129-132.

Gao, L., Beling, P. A. 2003. "Machine quantification of text-based economic reports for use in predictive modeling," *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, pp. 3536-3541 vol.4.

Ghosh, S., Samanta, D., and Sarma, M. 2012. "Cost of error correction quantification with Bengali text transcription," *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, Kharagpur, pp. 1-6.

Giesen, Kristian, and Jens Südekum. "Zipf's law for cities in the regions and the country." *Journal of Economic Geography* 11, no. 4 (2010): 667-686.

Goodenough, W.H., 1981. Culture, language, and society.

Gregor, S., & Henver, A. R. 2013. “Positioning and Presenting Design Science Research for Maximum Impact,” *MIS Quarterly*, (27:2), pp. 337-355.

Ha, Le Quan, Elvira I. Sicilia-Garcia, Ji Ming, and F. Jack Smith. "Extension of Zipf's law to words and phrases." In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pp. 1-6. Association for Computational Linguistics, 2002.

Haiku Society of America, The Definitions Committee. 2004. *Official Definitions of Haiku and Related Terms* [Press release].

Hatzigeorgiu, Nick, George Mikros, and George Carayannis. "Word length, word frequencies and Zipf's law in the Greek language." *Journal of Quantitative Linguistics* 8, no. 3 (2001): 175-185.

Hill, B. and Woodroofe, M. 1949. Stronger forms of Zip's law. Journal of American Statistical Association 70, 212-219.

Hill, B. 1970. Zipf's law and prior distributions for the composition of a population. Journal of American Statistical Association 65, 1220-1232.

Hill, Bruce M. "The rank-frequency form of Zipf's law." *Journal of the American Statistical Association* 69, no. 348 (1974): 1017-1026.

Hinkel, E. ed., 1999. *Culture in second language teaching and learning*. Cambridge University Press.

Hirshfield, J. 2011. *The Heart of Haiku.* Amazon Digital Services.

Houtchens, B. C. 2001. "English in the News," *The English Journal*, (91:1), pp. 98-102.

Howey, H. 2011. *The Plagiarist*. Amazon Digital Services.

Ioannides, Yannis M., and Henry G. Overman. "Zipf's law for cities: an empirical examination." *Regional science and urban economics* 33, no. 2 (2003): 127-137.

Jiang, Bin, and Tao Jia. "Zipf's law for all the natural cities in the United States: a geospatial perspective." *International Journal of Geographical Information Science* 25, no. 8 (2011): 1269-1281.

Jones, E.L., 2001, April. Metrics based plagarism monitoring. In *Journal of Computing Sciences in Colleges* (Vol. 16, No. 4, pp. 253-261). Consortium for Computing Sciences in Colleges.

Jones, M. and Sheridan, L., 2015. Back translation: an emerging sophisticated cyber strategy to subvert advances in 'digital age'plagiarism detection and prevention. *Assessment & Evaluation in Higher Education*, *40*(5), pp.712-724.

Kashani, Mehdi M., Fred Popowich, and Anoop Sarkar. "Automatic transliteration of proper nouns from Arabic to English." In *Proceedings of the Second Workshop on Computational Approaches to Arabic Script-based Languages*, pp. 275-282. 2007.

Kock, N. and Davison, R., 2003. Dealing with plagiarism in the information systems research community: A look at factors that drive plagiarism and ways to address them. *MIS Quarterly*, pp.511-532.

Lancaster, T. and Culwin, F., 2001, June. Towards an error free plagarism detection process. In *ACM SIGCSE Bulletin* (Vol. 33, No. 3, pp. 57-60). ACM.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M., & Kruschwitz, N. 2011. "Big data, analytics and the path from insights to value," *MIT Sloan Management Review*, (52:2), pp. 21-3.

Levene, Mark, José Borges, and George Loizou. "Zipf's law for Web surfers." *Knowledge and Information Systems* 3, no. 1 (2001): 120-129.

Lewis, D.D. and Jones, K.S. 1996. Natural language processing for information retrieval. *Communications of the ACM*, *39*(1), 92-101.

Lewis, S.C., Zamith, R. and Hermida, A., 2013. Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of broadcasting & electronic media*, *57*(1), pp.34-52.

Li, Wentian. "Random texts exhibit Zipf's-law-like word frequency distribution." *IEEE Transactions on information theory* 38, no. 6 (1992): 1842-1845.

Liddy, E. D. 2001. Natural language processing.

Lommel, A., 2006. Localization standards, knowledge-and information-centric business models, and the commoditization of linguistic information. *2006), Perspectives on Localization*, pp.223-239.

Luna, L. 2013. "Indefinite Extensibility in Natural Language." *The Monist,* (96:2 Formal and Intentional Semantics), pp. 295-308.

Maillart, Thomas, Didier Sornette, Sebastian Spaeth, and Georg von Krogh. "Empirical tests of Zipf's law mechanism in open source Linux distribution." *Physical Review Letters* 101, no. 21 (2008): 218701.

Marsili, Matteo, and Yi-Cheng Zhang. "Interacting individuals leading to Zipf's law." *Physical Review Letters* 80, no. 12 (1998): 2741.

Mastora, A., Peponakis, M., & Kapidakis, Sarantos. 2017. *Journal of Information Science.* 43(4), pp. 492-508.

Okuyama, Kazumi, Misako Takayasu, and Hideki Takayasu. "Zipf's law in income distribution of companies." *Physica A: Statistical Mechanics and its Applications* 269, no. 1 (1999): 125-131.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. 2007. A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.

Phillips, B.J. 2014. "Terrorist Group Cooperation and Longevity," *International Studies Quarterly,* (58:2), pp. 336-47.

Portner, P.H. 2005. *What is Meaning? Fundamentals of Formal Semantics*. Malden, MA: Blackwell.

Powers, David MW. "Applications and explanations of Zipf's law." In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*, pp. 151-160. Association for Computational Linguistics, 1998.

Rau, L. F., Jacobs, P. S., & Zernik, U. 1989. Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management*, *25*(4), 419-428.

Rockwell, G., Sinclair, S. 2016. "The Measured Words: How Computers Analyze Text," in *Hermeneutica:Computer-Assisted Interpretation in the Humanities* , 1, MIT Press, pp.256-.

Salkoff, Morris. "Automatic translation of support verb constructions." In *COLNG 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*. 1990.

Sarndal, C. 1967. “On Deciding Cases of Disputed Authorship,” *Journal of the Royal Statistical Society*, (16:3), pp. 251-268.

Schieffelin, B.B. and Ochs, E. eds., 1986. *Language socialization across cultures* (No. 3). Cambridge University Press.

Schlenker, P. 2010. "Super Liars," The Review of Symbolic Logic, (3), pp. 374-414.

Schubert, Lenhart. “Computational Linguistics.” *Stanford Encyclopedia of Philosophy*, Center for the Study of Language and Information (CSLI), Stanford University, 26 Feb. 2014, plato.stanford.edu/entries/computational-linguistics/.

Serpanos, Dimitrios N., George Karakostas, and Wayne Hendrix Wolf. "Effective Caching of Web Objects using Zipf's Law." In *IEEE International Conference on Multimedia and Expo (II)*, pp. 727-730. 2000.

Seshasai, S., 2009. *Efficient near duplicate document detection for specialized corpora* (Doctoral dissertation, Massachusetts Institute of Technology).

Stašák, J., Vaníčková, R. and Grell, M., 2015. Business Process Modeling Linguistic Approach–Problems of Business Strategy Design. *Universal Journal of Management*, *3*(7), pp.271-282.

Stubbs, M., 1996. *Text and corpus analysis: Computer-assisted studies of language and culture* (p. 158). Oxford: Blackwell.

Soo, Kwok Tong. "Zipf's Law and urban growth in Malaysia." *Urban Studies* 44, no. 1 (2007): 1-14.

“Textual Analysis.” *The Sage Encyclopedia of Qualitative Research Methods*, by Lisa M. Given, Sage Publications, 2008.

Thomas, H. 1932. “The Shakespeare Authorship Controversy,” *The British Museum Quarterly,* (*7*:2), pp. 40-41.

Troll, Günter, and Peter beim Graben. "Zipf's law is not a consequence of the central limit theorem." *Physical Review E* 57, no. 2 (1998): 1347.

Vilar, David, Jia Xu, D'Haro Luis Fernando, and Hermann Ney. "Error Analysis of Statistical Machine Translation Output." In *LREC*, pp. 697-702. 2006.

Wand, Y., & Weber, R. 1990. Toward a theory of the deep structure of information systems. In J. DeGross, M. Alavi & H. Oppelland (Eds.), Proceedings of the Eleventh International Conference on Information Systems (pp. 61-72). Baltimore: ACM Press.

Wells, S. 2014. *Why Shakespeare WAS Shakespeare*. Amazon Digital Services.

Woodruffe, M. and Hill, B. 1975. On Zipf's Law. Journal of Applied Probablity 12, 425-434.

Wyllys, Ronald E. "Empirical and theoretical bases of Zipf's law." (1981).

Yoo, Y., Boland, R. J., Lyytinen, K., & Majchrzak, A. (2012). Organizing for Innovation in the Digitized World. Organization Science, 23(5), 1398-1408. doi:10.1287/orsc.1120.0771

Zipf, G.K. 1935. *The Psychobiology of Language*. Houghton-Mifflin.

Zipf, George. K. "Human Behaviour and the Principle of Least Effort Addison." Wesley Pub. (1949).

Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

CHAPTER 5 – CONCLUSION:

Porra, J., Hirschheim, R. and Parks, M.S., 2014. The historical research method and information systems research. *Journal of the Association for Information Systems*, *15*(9), p.536.

Rebaza, C., Doherty, L. and Hanna, J. 29 Mar. 2018. "Ecuador Cuts off Julian Assange's Internet Access at Embassy." CNN, Cable News Network, www.cnn.com/2018/03/28/europe/ecuador-julian-assange/index.html.