Modelling Sentiment and Topics in Letters Written by 19th Century Immigrants in North America

Suzanne A. Moody

Department of Languages, Literatures and Cultures

McGill University, Montreal

April 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Arts

Table of Contents

List of Tables

List of Figures

Abstract

This thesis answers the call by the International Organization for Migration for researchers to listen to migrants. Based on the assumption that contemporary and historical migrants have similar experiences, this study takes as its subject 915 letters written by 218 immigrants in North America during the long 19th century. Within the framework of exploratory data analysis and using the tools of natural language processing and Bayesian linear regression, it measures sentiments, models topics and examines relationships between narrative features and temporal, structural and biographical variables. It also tests whether sentiment and topics predict cessation of correspondence, a potential indication that a migrant has successfully integrated into a new country. The hypothesis is that positivity in letters, signifying good experiences, correlates with an increased probability that correspondence will end. Sentiment was found to be mildly positive and topics mostly oriented around practical matters and relationship maintenance. These narrative features varied by time and writer traits, but they did not predict cessation of correspondence, which was mostly related to gender. Female migrants were more likely than men to continue writing. The findings are mostly in line with previous scholarship in the area of migrant correspondence, indicating that the methodology used here is valid. The quantitative techniques revealed subtle patterns, particularly around the influence of gender, and offered insight into how Bayesian multilevel modeling might address bias and representation in cultural datasets. Recommendations for future study include use of this technique for more nuanced modeling of migrant correspondence and for other work within the field of the digital humanities.

Resumé

Cette thèse répond à l'appel lancé par l'Organisation internationale pour les migrations aux chercheurs pour qu'ils soient à l'écoute des migrants. Partant de l'hypothèse que les migrants contemporains et historiques ont des expériences similaires, cette étude prend pour objet 915 lettres écrites par 218 immigrants en Amérique du Nord au cours du XIXe siècle. Dans le cadre d'une analyse exploratoire des données et en utilisant les outils du traitement automatique du langage naturel et de la régression linéaire bayésienne, ce mémoire mesure les sentiments, modélise les sujets et examine les relations entre les caractéristiques narratives et les variables temporelles, structurelles et biographiques. Elle vérifie également si les sentiments et les sujets prédisent la cessation de la correspondance, une indication potentielle qu'un migrant s'est intégré avec succès dans un nouveau pays. L'hypothèse est que la positivité dans les lettres, signifiant de bonnes expériences, est corrélée avec une probabilité accrue de la fin de la correspondance. Le sentiment s'est avéré légèrement positif et les sujets ont été principalement orientés vers des questions pratiques et l'entretien des relations. Ces caractéristiques narratives variaient selon l'époque et les traits de l’auteur, mais elles ne permettaient pas de prévoir la cessation de la correspondance. Surtout reliée au sexe, les femmes migrantes étant en effet plus susceptibles de continuer à écrire que les hommes. Les résultats sont pour la plupart conformes aux travaux antérieurs dans le domaine de la correspondance des migrants indiquant ici la validité de la méthodologie. Les techniques quantitatives ont révélé des modèles subtils, en particulier concernant l'influence du sexe et ont permis de comprendre comment la modélisation bayésienne multi-niveaux pourrait traiter les préjugés et la représentation dans les ensembles de données culturelles. Les recommandations pour des études futures comprennent l'utilisation de cette technique

pour une modélisation plus nuancée de la correspondance des migrants et pour d'autres travaux dans le domaine des humanités numériques.

Acknowledgements

I would like to thank my friends, colleagues and mentors in the Department of Languages, Literatures and Cultures, who have offered their kindness and support to me in many different ways and for many years now. In particular, I am grateful to Prof Andrew Piper, for his role in hiring me as a research assistant in 2013, for warmly welcoming me back after more than one Australian sojourn and for supporting my application to join the Digital Humanities program as a student. For the opportunity to be a student of Prof Stéfan Sinclair, I cannot adequately express my appreciation. This thesis would not have been possible without his excellent instruction and generous guidance. I am thankful for my supervisor, Prof Cecily Raynor, who has been a patient and understanding collaborator throughout my program, especially during the final year, when the COVID pandemic presented so many challenges. To Lynda Bastien, the graduate student affairs coordinator, and all the administrative staff at LLC, I wish to extend my gratitude for the important work that they do to make research and tertiary education happen. My final academic bow is to Dr Tully Barnett, who supervised my MITACS project in Australia and under whose direction I am excited to pursue a PhD. Endless thanks are due to my husband, Pippa, and my children, Leo and Audrey, for their love and encouragement over the last three years. I end with a nod to the loved ones who migrated to the hereafter while I worked on this degree: Chris Wethern, Leo Smetana, Dr David Holtzman, Gerry Printz and especially Kevin Halverson, in whose memory this thesis is offered.

Introduction

When W. J. B. Hamilton crossed the Atlantic Ocean in 1860, making his way from the north of Ireland to the southern United States, he wrote in a letter sent home about how his experience might compare to those of previous and future migrants:

> Although the journey appeared long to me it would not have been considered so 50 years since. Uncle Bones was 72 days on sea when he first came out. We were only ten days and a half. I could have reached Augusta from Ballymoney in fourteen days. If the next fifty years make as much difference, travelling will be performed with the speed of electricity. (Irish Emigration Database, PRONI D 1835/27/1; CMSIED 9310015)

We may not have achieved the electric-fast ocean crossings envisioned by the letter-writer, but migration does happen at a faster clip today than in Uncle Bones' time. Another element of the migrant experience—communication with family and friends—does travel at the speed of electricity. Paper and pen have been replaced by a parade of electrical devices that allow migrants to not only write to people back home but to talk to them and see them.

Whereas increasingly literate migrants of the past rode the wave of transportation technology and postal service expansion, contemporary migrants are techno-literate surfers of a global communications network. Historical migrants left a paper trail measurable by page, ounce and inch. By comparison, today's migrants are leaving traces countable by the word, byte and second. It is the paper trail that has been the subject of most scholarly inquiry into migration narratives, with historians and sociologists using qualitative and critical methods to explore them since the early 20th century. This thesis stands on those shoulders but joins pioneering scholars from linguistics in using computational and quantitative methods to examine these telling artifacts.

Why Study Migrant Letters?

Letters have become understood by scholars as “a privileged source to learn about the experience of migration,” including motives and adaptations (Borges & Cancian, 2016, p. 282). According to Blegen (1931), they “betray the spirit, hopes, and aspirations of the humble folk who tilled the soil, felled the forest, and tended the loom” (p. vii). Barton (1975) celebrated them for documenting history from the viewpoint of people who had “humble origins and little schooling,” a privilege that until the 19th century was “the preserve of small upper-class elites” (p. 6).

In addition to recording the common person’s experience, migrant correspondents helped make history. When disseminated among friends and relatives in home countries, their letters served as “a great spur to mass emigration” (Gerber, 2006, p. 41). Houston and Smyth (1990) reported that the effect was so great that one government official claimed in 1827 to be able to use letters to predict emigration levels one year in advance: “they write home letters and if the season has been favourable, if there has been any great demand for labour…they send home flattering letters, and they send home money to assist in bringing out their friends” (p. 16; p. 94). As conduits for the exchange of information and ideas, letters influenced migration patterns and impacted the development of both sending and receiving countries. According to Fitzpatrick (1994a), they offer “unrivalled insight” into the phenomenon of chain migration.

Migrant letters also constitute an unconventional form of literature that has been described as “a literature of the unlettered" and “the beginnings of literature, the stuff out of which the *My Antonia*'s are eventually made” (Mulder, 1954, pp. 45-46). Personal correspondence intersects with testimonial literature, a genre heralded by Holocaust survivor, writer, Nobel Peace laureate and literary critic Elie Wiesel in 1977. According to Felman and Daub (1991): “It has been suggested that testimony is the literary—or discursive—mode par excellence of our times,” which itself “can precisely be defined as the age of testimony” (pp. 5-

6).

Testimonial literature (also known as survivor narrative) and the closely related Latin American testimonio, belong to the broad field of life writing, which comprises dozens of subcategories of self-referential narrative, including personal correspondence. According to Smith and Watson (2015), “letters seem to be private writings, but in the late eighteenth century they began to be understood as both private correspondence expressing the inner feelings of the writing subject and as public documents to be shared within a literary circle” (p. 273). They argue that including noncanonical works, that is extending the literary notion of autobiography to cover “the extensive historical range and the diverse genres and practices of life writing not only in the West but around the globe,” gives voice to “those whose identities, experiences, and histories remain marginal, invalidated, invisible, and partial” (Smith & Watson, 2015, p. 3). Thus, although Erickson (1972) noted of migrant letters “few...may be said to have literary merit," it can be argued that these life-writing forms are interesting literary objects with both historical and contemporary implications (p. 1).

Literary value aside, this type of correspondence is important because migration has become a high-profile topic in Canada and in many countries because of the increasing number of voluntary migrants and displaced persons in recent years. Since 1970, the number of international migrants in the world has more than tripled, rising from 2.3% of the global population to 3.5% (IOM, 2019). Projections show sustained migration levels over the next several decades, with the effects of global warming and climate change potentially raising the number of migrants motivated by environmental concerns alone to 1 billion in 2050 (IOM, 2015).

In its 2018 World Migration Report, the United Nations International Organization for Migration argued for research to support policy decisions, in particular research that involves

"listening to and learning from migrants" (IOM, 2017, p. 8), with the goal of achieving better outcomes for them as well as sending and receiving countries. A brief section in the report features techniques from the realm of computational text analysis to summarize migration research. Word frequencies and collocations were examined in the titles of 538 research articles published in seven migration journals between 2015 and 2016. The results showed "a dominant 'receiving country' perspective" and a collective preoccupation with employment and social issues, particularly interpersonal networking (IOM, 2017, p. 105). This approach to summarizing the body of social research on migration begs the question: If the same analysis was done on migrants' own narratives, would the results mirror those of the journal article analysis?

In an effort to widen the academic lens, I performed a similar word frequency and collocation assessment on migration related articles in arts and humanities journals.[1] Because of the small number of such articles in these types of journals, I expanded the timeframe from 2015 and 2016 to all years after 1999. This yielded a quantity of words similar to the dataset in IOM (2017) and suitable for computational text analysis, with the added benefit of bringing the results up to date. The search yielded 458 article titles, containing 6,122 word tokens and 2,322 types after stopwords were removed. The 75 most frequent types are presented with those of the IOM analysis in Figures 1 and 2.

My analysis revealed that the emphasis on receiving country perspectives was less pronounced but still present.[2] As in the IOM analysis, *social* was an important term (ranked #10) but it was not accompanied by *network*; instead, *science* was the most common collocate,

[1] Annotated code for this analysis is available at https://github.com/menyalas/ThesisPublic/blob/main/20200504_AM_HumanitiesReview.ipynb

[2] The IOM finding of a receiving country bias in social science articles rests on the higher frequency of word types containing the lemma *migra* with the prefix *im-* (inflow) rather than the prefix *e-* (outflow). In my analysis of the arts and humanities articles, all word types beginning with *immigra* were absent from the ten most frequent terms, but immigration did appear among the top 75 while types containing *emigra* did not, suggesting that receiving country perspectives were more common. Overall, terms beginning with *immigra* versus *emigra* occur at a frequency of almost three to one.

followed by structural terms, such as *cohesion*, *degeneration* and *change*. *Refugee* again makes it into the top 10, but *policy* and *labour* are absent along with semantically similar words. Instead, authors of articles in arts and humanities journals are interested in *religion*, as evidenced by that term plus *church*, *mission*, *theology*, *Christian* and *god* all appearing in the top 75 terms. After this, frequent terms indicated a focus on *identity* and things that might influence it – *home*, *memory*, *experience* – but also *belonging*, *community*, *diaspora*.

Objective, Research Questions and Dataset

In an effort to reconcile the perspectives of social scientists and humanists and to bring them into alignment with the people who are at the center of their research, this thesis looks at the words contained in letters written by immigrants in North America during the long 19th century, which is widely understood to be the period between the end of the 18th century and the start of World War I (Burke, 2000; Porter, 1999). The objective is to learn about migrants' experiences by identifying the concerns (i.e., topics) and emotions (i.e., sentiments) expressed in their letters and to examine how these factors correspond to migration outcomes, specifically social integration as indicated by the cessation of correspondence with the home country. My research questions are as follows:

1. What topics and sentiments are evident in the letters?
2. Do topics and sentiments vary by factors related to time, letter structure or author traits?
3. Do topics and sentiments predict cessation of correspondence?

The 915 letters in this study were extracted from the Alexander Street Press database entitled *North American Immigrant Letters, Diaries and Oral Histories*. They represent 218 writers who emigrated from a variety of countries and settled mostly in the United States or Canada. The language of the letters is English; however, this does not mean that the writer's first language was English. Some letters were translated, dictated or written in an acquired language.

All letters were originally written on paper but have been transcribed into a machine-readable, digital format, making them suitable for computational text analysis.

Research Paradigm

This thesis will apply computational and statistical methods, typically the tools of quantitative researchers operating within a positivist paradigm in the social sciences, to material that is the domain of humanists grounded in interpretive, constructivist or critical paradigms. Assumptions about the nature of reality (ontology) and the nature of knowledge (epistomology) underpin the methodologies used within these paradigms. With its orientation around a single, objective, observable reality, the positivist paradigm is compatible with quantitative methods. My topic (i.e., migrant experiences) and the data (i.e., personal letters) are subjective and therefore better suited to an interpretive, constructivist or critical paradigm that accepts reality as mutable. The combination of my methods and my data make traditional paradigms an awkward fit for this thesis, which will operate instead under the less established but more appropriate paradigm of pragmatism.

Associated with mixed methods social research, pragmatism assumes that reality is constrained by nature but interpreted through human experience. Thus, "ontological arguments about either the nature of the outside world or the world of our conceptions are just discussions about two sides of the same coin" (Morgan, 2014, p. 1048). The orientation around human experience extends to researchers, who are understood to conduct inquiry within social contexts that lead them to think and behave accordingly. Their experiences, beliefs and actions interact continuously to produce different kinds of knowledge in different ways. Thus, epistomological concerns about "the nature of knowledge" and the relationship between "knower and known" are supplanted by the idea of knowledge production as a social and "active process of inquiry," characterized by "a continual back-and-forth movement between beliefs and actions" (Mertens,

2010, p. 470; Morgan, 2014, p. 1049).

Inherent components of pragmatism, which are integrated into other research paradigms as axiology, include ethics, morality, values, and by extension politics and social justice. These follow from conceptualizing research as the interaction of experiences, beliefs and actions, with an integral factor being judgements about "which goals are most meaningful and which methods are most appropriate" (Morgan, 2014, p. 1050). A central tenet of pragmatism is that the iterative process of active inquiry leads to social growth for all people. For this reason, Morgan (2014) argues that there is "a natural fit between pragmatism and many versions of transformative or emancipatory research through a shared emphasis on openness, fairness, and freedom from oppression" (p. 1050). As such, pragmatism fits well with the feminist and postcolonial aspects of this thesis. According to Mortari (2015), the active process of inquiry makes researcher reflexivity a key component of pragmatism.

Reflexivity Practice

The English language focus of this migrant letter study demands a moment of researcher reflexivity. This choice stems from my personal background, experiences, qualifications and perspective, which I will now reflect upon and attempt to make transparent. I am an anglophone descendent of white European settlers who, according to Native Lands (https://native-land.ca) and First Peoples' House (n.d.), lives and studies on unceded Indigenous territory in Montréal, Québec, Canada. This area was used and occupied prior to colonization by the St Lawrence Iroquoian people and the Mohawk, Huron-Wendat and Anishinabeg nations. I am also a contemporary migrant, having moved from the United States, where I was born and raised, to Australia, where I became a parent, to Canada, where I now live with my family as a naturalized citizen. These movements were voluntary and motivated by economic and personal interests; as such, I am an English-speaking migrant on the upper end of the agency axis described by

Manning (2013).

Aside from the challenges presented by my lack of fluency in French, the official language of Québec, I have had an easy experience compared to most of the world's migrants. I believe that migration is a natural human phenomenon that requires supportive policy ensuring safe passage, smooth integration and adequate agency for all migrants. At the same time, I am aware of the devastating impacts that migration, particularly European colonization, has had on Indigenous peoples and on the environment. As an activist for climate justice and human rights, I recognize the importance of Indigenous knowledge and belief systems, ways of living and land management practices. With this in mind, I stand in solidarity with Indigenous peoples seeking transitional justice and the restoration of land rights.

I am aware that my personal identity and beliefs interact with my thesis topic in problematic ways. To control for this, elements of reflexivity and decolonisation, as described by Seganti (2010) and Held (2019), are integrated into this thesis, particularly self-reflection to identify biases resulting from my personal experiences and my western, anglophone perspective as well as measures to ensure that non-western, non-anglophone viewpoints are considered. An important way to begin this process is to place the English language migrant letters at the center of this inquiry into a global and historical context.

Three Centuries of Human Migration: A Brief Overview

While humans have always migrated, the 18th century saw a significant increase because of the transatlantic slave trade, with migrant numbers ranging from tens of thousands per year in the early part of the century to nearly 100,000 per year by the end of the century (Manning & Trimmer, 2013). The movement of slaves, mainly from Africa to the Americas, largely accounts for migration during this century. During this time, Britain's naval capacity developed through warfare with the French, Spanish and Dutch, yielding by the 19th century steamship technology

fueled by British coal, offering “inexpensive, dependable travel and the British leadership of the sea lanes” (Manning & Trimmer, 2013, p. 160).

The industrialization of Europe during the 19th century contributed to colonial expansion into other parts of the world because it required the movement of labour and materials. This century saw a massive expansion in voluntary migration and the final throes of slavery and servitude in Europe and the Americas. During this century, European migration to the Americas peaked at about one million people per year (Fisher, 2013). Colonialism not only created migrant flows away from Europe toward occupied lands, but also within and away from those places, as local people were displaced, conscripted to colonizers’ armies or otherwise “reoriented themselves toward European culture and migrated accordingly” (Fisher, 2013, p. 81).

Various gold and silver rushes attracted hundreds of thousands of migrants and helped expand the footprint of settler communities in the Americas, Australasia, Africa and elsewhere. The mid-century failure of the potato crop in northern Europe motivated millions of people to leave their homelands. In Ireland, a flood of emigrants began with the potato famine but continued until the end of the century, by which point famine deaths and migration had cut the population in half, in part because “as many Irish [had] emigrated to the Americas as [had] remained in their homeland” (Fisher, 2013, p. 89).

While European powers were colonizing the Americas, Australasia and Africa, the Russian Empire was becoming “the most expansive state in continental Eurasia,” incorporating much of Eastern Europe and creating conditions that prompted the permanent exile of more than 4.5 million Jews to North America and elsewhere by the early 20th century (Fisher, 2013, pp. 95-96). Imperialism and migration peaked around this time, by which point European powers governed more than half the world’s population. According to Manning & Trimmer (2013), these movements “set the scene for the divisions and conflicts” to come (p. 159).

Twentieth century migration was largely driven by war. According to Fisher (2013), "much of the peaceful civilian intercontinental migration halted" (p. 104). That is not to say that civilians were not on the move. Enormous and undocumented post-World War I migrations prompted the League of Nations to order the development of a system to track and manage such movements, giving us the passports and other travel documents we know today. Advances in military technology and the application thereof to the leveling of entire cities created floods of international refugees and internally displaced people, as did the redrawing of national boundaries and the reassignment of colonies at the conclusion of the wars. Because of the enormous effect of the wars on migration patterns, policies and experiences, this study concludes with the end of the long 19th century, generally understood to coincide with the start of World War I in 1914.

Decolonization, Equity and Representation

The extent of 19th century imperialism was so great that "with only a few exceptions, the entire world was divided between colonizing countries and their colonies" (Fisher, 2013, p. 102). Unfortunately, the migrant letters that have survived were written almost exclusively by people originating from the colonizing countries. Unheard voices include those of millions of slaves and indigenous people forcibly removed from their homelands as a result of European trade, colonization and expansion. The kind of personal correspondence at the center of this study is virtually nonexistent for these forced migrants.

One approach to approximating these lost voices is to "imagine that all the elements of migration can be formed into a multidimensional grid…extending in many different directions" (Manning & Trimmer, 2013, p. 195). One such axis—migrant agency—would range from free to unfree, replacing categorical notions, such as immigrant, refugee, indentured servant and slave. Lucassen et al. (2010) describe such typologies as "unproductive when they develop into

exclusive dichotomies…that makes [sic] comparisons irrelevant, thus obstructing the tracing of possible similarities, as well as uncovering what is indeed different" (p. 11). They point out that all migration has voluntary and coercive facets that are inextricably entwined.

With this in mind, the letters of bonded or exiled migrants, such as indentured servants or prisoners, and those experiencing significant social persecution can be understood to approach, albeit at varying distances, the experiences of non-letter writing slaves and indigenous people. For example, a Chinese worker in the 19th century Coolie Trade reported that:

> [We] were all beguiled on board the barbarian [American] ship as contract laborers by emigration agents and confined in the hold…After the ship sailed, the said barbarian gave each man in the hold a contract of servitude. If he did not accept he was flogged…More than ten who were sick in bed and could not walk were immediately killed and thrown into the ocean. (Irick, 1982, p. 33-35)

Manning and Trimmer (2013) observed that "the English [people], in speaking of 'the Irish race,' used terms as negative and deprecatory as any applied to other groups" (p. 147). Thus, it may be that the experiences of early Irish emigrants to North America, two-thirds of which were male and came as indentured servants, might begin to approach those of the African men brought to the same land as slaves. Similarly, the experiences of the estimated 160,000 men and 25,000 women sent to penal colonies in Australia between 1787 and 1868 might capture some of what Aboriginal people thought or felt when forced from their homelands. And finally, the voices of 19th century female immigrants in Canada and the United States, some of whom were prolific letter writers, may be able to offer some insight into the experiences of lower agency migrants.

On Methodology

Quantitative and computational methods can serve to resurrect and propagate social injustices because racism and sexism are woven into the cultural record (D'Ignazio & Klein,

2018; Risam, 2018). Counteracting this requires informed, critical and ethical reflection and decision-making throughout the research process. Rockwell and Sinclair (2016) presented a methodology for computer assisted humanistic inquiry that supports this kind of interplay between technology and interpretation and thereby offers a bridge between paradigms. While they did not associate *Agile Hermeneutics* to pragmatism, they did call it pragmatic as well as collaborative, interdisciplinary, open, iterative and reflective, all features of the pragmatism paradigm associated with mixed methods social research. Of particular importance is the mutual prioritization of problem solving above theory and the shared understanding of interpretation as a core component of research. Rockwell and Sinclair (2016) described quantitative and computational tools as *hermeneutica* that contribute to—rather than replace—critical or interpretive approaches to literary material. As such, they are methods that mix well with qualitative and critical methods and therefore fit nicely with the pragmatic paradigm guiding this thesis.

The methods I will use are quantitative and computational, but they have been calibrated according to information and ideas generated by researchers operating mostly within interpretive and critical frameworks. In this sense, my collaborators are historians and life writing scholars as much as they are computer scientists, linguists, statisticians and sociologists. Specifically, I will use some of the *hermeneutica* developed by Rockwell and Sinclair (2016) and supplied through the Voyant application. I will also develop tools in the Python programming language using the Natural Language Toolkit (NLTK) and other resources to acquire, explore and prepare the narrative data for analysis and to measure sentiment and identify topics.

The R programming language will be used with the brms and ggplot packages to develop and visualize Bayesian linear regression models, within the tradition of exploratory data analysis, to examine the relationships between key variables, specifically the sentiment and the topics in

the narrative data and the author- or letter-level information provided in the metadata. This modeling will be undertaken in the Bayesian framework for two reasons. First, the results of Bayesian analysis come in the form of conditional probabilities, which are straightforward to interpret and communicate. Nonintuitive notions, such as null hypothesis testing and p values, are not part of the Bayesian approach. My goal is to produce findings that humanists, who are generally not trained in statistics, can readily engage with and potentially integrate into their critical and qualitative studies. Second, the Bayesian framework is well suited to multilevel modeling, which can help address some of the core challenges in the digital humanities, particularly around sample bias and missing data. In this study, multilevel modeling will be used to deal with repeat observations (i.e., multiple letters by the same individual) and sample imbalance (i.e., more letters by certain genders or social classes). In 2012, Ted Underwood wrote that "if humanists were more familiar with Bayesian statistics, I think it would blow a lot of minds" (para. 6). I concur and hope that this thesis will serve as a persuasive demonstration of its potential for the digital humanities.

Literature Review

Migrant letters captured the attention of Western scholars toward the end of World War I. It was a team of sociologists that brought them into the spotlight, but while the social science community debated the empirical value of such personal documents, historians embraced them as giving narrative life to statistical facts and figures. Concurrent with and part of the New Social History movement, these scholars began producing edited collections of migrant letters beginning in the mid 1950s. Typically, these volumes included the scholar's interpretation of the historical context for the correspondence, followed by the letters themselves, curated and edited to varying degrees.

As social scientists witnessed the narrative turn in their field and literary scholars the rise of the testimonial genre, historians studying migrant letters became more empirical in their approach. By the late 20th century, some were using methodologies inspired by narrative inquiry and content analysis to categorize texts, describe the structure of letters and identify thematic patterns and keywords. They were beginning to envision migrant letters as the object of statistical procedures rather than the dressing on them.

The computational turn in the humanities has further entwined these interdisciplinary approaches to studying the migrant letter, with the most technologically oriented work now being conducted in the fields of corpus linguistics and the digital humanities. Concurrently, qualitative and critical work on personal letters continued to be generated within the field of auto/biographical studies, a rising interdisciplinary field with major nodes in literary studies and sociology. Thus, in the 100 years that Western scholars have been collectively telling the story of migrant letters, the narrative has passed through academic domains and been influenced by their diverse research paradigms, theories and methods.

Once Upon a Time: The Social Psychologists

William I. Thomas and Florian Znaniecki's *The Polish Peasant in Europe and America*, originally published in five volumes between 1918 and 1920, laid the groundwork for the use of life documents, such as personal letters, as primary source material in social science research. By analyzing the life writings of peasant-class Polish immigrants in the United States, Thomas and Znaniecki (1918-1920/1927) found systemic community disorganization, which they attributed to disintegrating ties with family and traditions in the old country paired with an undeveloped sense of belonging and social order in their rapidly modernizing new country. Their study was concerned with group life and how members think and act relative to it, society-at-large and the forces of change in the world. In particular, they challenged prevailing ideas about assimilation, or Americanization, arguing that Polish immigrant society is "neither Polish nor American but constitutes a specific new product whose raw materials have been partly drawn from Polish traditions, partly from the new conditions in which the immigrants live and from American social values as the immigrant sees and interprets them" (Thomas & Znaniecki, 1918-1920/1927, p. 1469).

As described by Blumer and Bain (1939) and Thomas (1978), the ideas presented in *The Polish Peasant* had an enormous impact on theory and method within the social sciences, receiving accolades but also generating controversy, particularly with respect to the role, significance and measurement of internal, subjective factors, which Thomas and Znaniecki called attitudes, and external, objective factors, which they defined as circumstances and rules. A key assumption in their study was that behavior is the result of an interaction between subjective and objective factors, and they argued that life documents, such as personal letters, offered a window into the necessary subjective component. In general, social scientists of the era were moving toward the experimental approaches used in the physical sciences and therefore wrangled with whether and how to consider the subjective aspect of the human experience.

Historians had no such ambitions and for decades kept migrant letters under the light of scholarly inquiry, often referring back to Thomas and Znaniecki (1918-1920/1927) as they pondered how best to analyze and interpret this unwieldly source of information. As I recount the story of migrant letter scholarship, I too will return to the findings and problems raised by this foundational study to show how it influenced and remains in conversation with the interdisciplinary research it inspired.

Picking up the Thread: The Social Historians

At first, historians collected and contextualized essentially unaltered letters, which Blegen (1931) said "swing wide the door to the realization that immigrants are people, not lines in a graph or figures in a table" (p. vii). In 1955, he presented correspondence that amounted to "a diary on a grand scale, kept by people…undergoing the transition from one mode of life to another" (p. 5). Sent from all corners of the United States to Norway during the mid 19th century, these letters conveyed "hopes and heartaches, courage and fear, failure and success" but with a forward-looking perspective and a firm resolve in the decision to emigrate (Blegen, 1955, p. 14). This curated collection of letters, grouped into thematic chapters with scholarly introductions and a subject index, was pioneering. In particular, this technique of categorizing letters can be construed as a human form of topic modeling, one that would be replicated by other historians.

Conway (1961) followed with a study of 19th century Welsh-American immigrant correspondence, which had been printed in newspapers and other publications in the homeland, often with that end use in mind at the time of writing. As such, this collection demonstrates a central challenge in the study of migrant letters: representativeness. In addition to bias introduced by the scholar's own curatorial process, such as his omission of material that was "flowery" or "trivial, repetitive, or dealing with issues only remotely connected with emigration," this

collection was shaped by "editors long dead," who "acted as filters, keeping back much of the dross" (Conway, 1961, p. v).[1]

Because of the public nature of the letters in Conway's collection, their influence on driving emigration was central in his analysis, with tone being of particular interest. He reported that the attitude of letter writers toward prospective new emigrants changed during the course of the century, from "full of encouragement" to "lukewarm if not openly hostile" in later years (Conway, 1961, p. 9). This shift would be later noted by Hoerder (1992), who reported that the pro-labor press of the 1880s advised immigrant workers "not to lure their friends to the U.S. by private letters" because this served "capitalists…who needed cheap workers and strikebreakers" and by Benton and Liu (2018), who said that a "new tone of disillusion entered the letters" in the late 19th century (p. 14; p. 162). As in Blegen (1955), the bulk of Conway (1961) is relatively verbatim migrant correspondence, introduced by historical contextualization and grouped into thematic chapters, in this case oriented around somewhat more objective factors, such as time period, geography, occupation type and current events.

The Plot Thickens: Historians with Scientific Leanings

As a London School of Economics scholar who "taught history as a social science," Charlotte Erickson was well equipped to bridge perspectives (Harte, 2008, para. 1). In her analysis of private (i.e., unpublished) letters written by 19th century English and Scottish immigrants in the United States, Erickson (1972) proposed that life documents can help "complete the picture" of the migration story from this period, which is sparse due to a "paucity of good statistical records of the movement" (pp. 1-2). Among the information that can be gleaned from personal correspondence, Erickson (1972) counted socioeconomic status, motives

[1] According to Plummer (1983), the "dross rate" means that a document, in particular a life narrative, is not "focused enough to be of analytical value" to the social scientist (p. 24).

for migration, adaptation strategies and social resources used to find work and otherwise become settled in the new country.

As a quantitative historian, Erickson (1972) expounded on the notion of representativeness. In addition to the kind of scholar-side selection bias evident in Conway (1961), Erickson (1972) noted two other "hazards" (p. 6). First, migrants who were illiterate, from the poorest classes or did not have family ties in the home country produced fewer letters. Second, recipient characteristics, such as personal interests and geographic mobility, could generate survivorship bias. As with other historical studies of the time, Erickson (1972) is largely comprised of curated letters with scholarly contextualization; however, this volume includes more analysis as well as a more complex structure, with multilevel clustering of letters, first by job type (i.e., agricultural, industrial or commercial-clerical-professional) and then by migrant writing group (i.e., centered on an individual or family). Focusing on motivations as well as social and economic adaptation, Erickson (1972) reported patterns that varied by post-migration occupation.

Among letter writers in the agricultural sector, Erickson (1972) discovered attitudes about land, work and independence that fit the traditional peasant model proposed by Thomas and Znaniecki (1918-1920/1927) for Polish migrants. That is, the decision to migrate was driven by an aversion to modernizing forces, such as industrialization and commercialization, and a yearning for a self-sustaining, agrarian livelihood. The letters written by industrial workers dated closer to mid-century, when Conway (1961) and others noted a negative sentiment shift. Similarly, Erickson (1972) recorded complaints, regrets and discouraging counsel. The motivations she detected in these letters were less ideological and lifestyle oriented and more economic in nature. The primary motivations for migrants in the commercial-clerical-professional class were different yet again. Economic pressures were almost nonexistent,

replaced by “personal, family and status reasons,” which could have the aspect of seeking opportunity or escaping trouble (Erickson, 1972, p. 395).

In terms of social adaptation, Erickson (1972) found that migrants in the commercial-clerical-professional class were more “isolated partly because of their ambition” and “expected to derive more satisfaction, status and companionship from their children than did most of the farmers and industrial workers,” who tended to be part of a broader community (Erickson, 1972, p. 407). With respect to economic adaptation, Erickson (1972) observed variation among the occupational groups, with immigrants in the agricultural sector—except those who were well educated but lacking “manual skills or aptitude”—most likely to settle permanently. For many, this meant letting go of pastoral dreams and joining the industrial revolution, but according to Erickson (1972), “farming in America whittled away their resistance to change, and they were able to accept changes they had hoped to avoid” (p. 61). Through her class-based analysis, her consideration of social and economic forces, and her focus on concerns from the subfield of migration studies (i.e., motivations, adaptations, permanence), Erickson (1972) blazed a new trail in the study of migrant letters.

Around the same time and in a similar vein, an interdisciplinary team specializing in history, sociology and economics examined 367 letters written by people from the northeastern part of Poland who migrated to the United States and Brazil in 1890 and 1891. The concentrated origins of the emigrants in Kula et al. (1973/1986) is a valuable characteristic of this collection because, as Alroey (2011) and Richards (2006) would later show, migrants of this era had local mindsets and thus local factors were key to their motivations and behaviors. Kula et al. (1973/1986) focused on the social, cultural and economic aspects of the letter writers’ adaptation, generally finding similarities among their expectations but variation in their experiences. For example, Polish migrants in Brazil, who tended to settle in rural areas, indicated

that they encountered less pressure to assimilate while the American contingent, which mostly settled in urban areas, felt compelled to adopt local language, clothing and food. Religion played a more important role in the adaptation experiences of the Polish-American writers, with Protestants reporting better experiences than Catholics and Jews. Kula et al. (1973/1986) found that letter writers had an expectation of ongoing post-migration family unity but almost none of the "nagging and negative family relationships" that Thomas and Znaniecki (1918-1920/1927) observed and interpreted as a sign of social disorganization (p. 5).

Scholarly attention to migrant letters would continue to oscillate between and at times blend a social science inspired positivism and the social historians' populist perspective. For example, in his edited collection of migrant letters from "every major Dutch-American settlement in the United States," Brinks (1986) offered little analysis, instead letting his subjects "speak for themselves," but then reducing these voices to "A Typical Immigrant" (pp. 7, 11). His profile of the average Dutch immigrant suggested that crucial networks were ethnic and religious. Motivations to migrate were fundamentally economic, as described by cartographer Ravenstein (1889), who laid the groundwork for contemporary models of migration and whose ideas have "withstood the test of time" (Greenwood, 2019, p. 269). But Brinks (1986) also found migrants to be motivated by the same avoidance of modernization (e.g., industrialization) described by Thomas and Znaniecki (1918-1920/1927) and Erickson (1972).

In another edited collection, this time focusing on Norwegian emigrants in the Americas, Africa and Australia, Hale (1986) grouped letters into chapters about ocean voyages, rural and urban life, ethnicity, politics and war, and religion. This grouping of letters was intended to show relationships between factors like destinations, themes and topics. This study was unique in its comparative approach, which highlighted overlooked voices, specifically the 10% of Norwegian emigrants who settled somewhere other than the United States or Canada. Hale (1986) observed

that the letters served to connect migrants with their ethnicity but also as "a barometer of its erosion," with letter series displaying, for example, the writer's frame of reference shifting from home country to new country (p. xxii). One manifestation of this dynamic was the low frequency of the term *Norway* in late letters, where it was used primarily for formulaic greetings to friends and relatives, versus in early letters, where it occurred often, particularly in comparisons with the homeland. Hale (1986) also remarked on a sentiment shift over time, with writers who had become established in the new country commenting "disparagingly on the conditions they had left behind," but these were partly matched by negative remarks about the new country including topics like unemployment, rural life, weather and politics (p. xxiii).

Historical geographers Houston and Smyth (1990) examined the letters of Irish immigrants in Canada, noting the important role of local family, friends and acquaintances, even early in the wave of mass migration. They focused on a small number of letter series, one covering three decades and others showing activity mostly during periods of crisis or diminishing altogether a few years after migration. In addition to reflecting on factors related to the frequency and cessation of correspondence, Houston and Smyth (1990) presented contrasting cases related to an important outcome: permanent settlement versus return migration. The first case featured a group of brothers who settled near each other and had close ties not only to each other but also with a wider community of Irish migrants. Their letters generally reflected a positive sentiment about their adopted country, in which they spent the rest of their lives. The other letter series featured a young man who migrated alone to a remote part of western Canada and whose letters showed that his "youthful optimism gave way to lonely bachelorhood and mounting private desperation" (Houston and Smyth, 1990, p. 307). This solo migrant eventually returned to Ireland, perhaps indicating a relationship between topics (e.g., social life), sentiments and outcomes.

Interested in the topic of labor, social historian and migration specialist Walaszek (1992) returned to the Polish-American letters presented in Kula et al. (1973/1986) to take a closer look using a linguistic approach. By focusing on the term *robota*, meaning *work* or *job*, and its contexts, Walaszek (1992) found that Polish-American immigrants described working hard with very little leisure time. Although language was a major barrier to advancement, immigrants often wrote about comradery among friends and family in the workplace as a pleasure. Work was described as hazardous and unstable, with references often associated with seasonal terms, such as *summer* or *Easter.* The negative tone noted by Conway (1961) and others was evident in these late 19th century letters, with writers reporting "too many people in America" and suggesting that emigrants go elsewhere to look for work (Walaszek, 1992, p. 90). The minimal use of adjectives and the abundance of metaphors in passages about work served as indications to Walaszek (1992) that Polish emigrants in America "did not understand the fundamental character of the forces behind the factory system and the labor market" (p. 90).

Career Storytellers: Four Historians

Toward the end of the 20th century, some scholars, including Irish-American immigration scholar Kerby Miller, dedicated careers to studying letters and other forms of life history documents. Using 6,000 letters written by 18th and 19th century Irish-American emigrants, Miller and Boling (1990) explored the nature, origins and consequences of a prevalent myth, which was that the New World offered "an earthly paradise—a land of incredible and easily attained wealth" (p. 18). Although early letters from migrants who were willing and able to farm offered positive reports about rural life in America, most letters from later migrants, whose experiences were by choice or by necessity more industrial and urban, expressed disappointment and resentment over being deceived by their forerunners. According to Miller and Boling (1990), the latter sentiment was paradoxical because "letters which painted an extravagant, effusive portrait

of America were very few and far between" (p. 26). In fact, this myth persisted despite evidence to the contrary, which came in the form of letters, beginning around mid-century, containing "realistic, cautionary and even negative testimony" (Miller & Boling, 1990, p. 24). The historians argued that the myth was a "necessary and appealing" response to the "negative and apocalyptic" feelings that Irish peasants had about their prospects at home, where famine and social, economic and cultural factors were colliding to generate a sense of "wretchedness and hopelessness" (Miller & Boling, 1990, p. 28). Under such circumstances, Miller and Boling (1990) proposed that emigration "could not be halted by either logical or emotional arguments" (p. 20).

In a study of letters and memoirs by earlier Irish-American immigrants (circa 1675 to 1815), Miller et al. (2003) examined migration processes, such as identity construction and motivations, which ranged from religious and political to various forms of economic in "inseparable" ways (p. 5). Like Erickson (1972), they oriented their analysis around the occupations of the writers—farmers and planters; craftsmen, laborers and servants; merchants, shopkeepers and peddlers; clergymen and schoolmasters—with a special section dedicated to those involved in politics and war. According to Miller et al. (2003), these Irish migrants represented far greater social, ethnic and religious diversity than their 19th century compatriots. Their letters led to the "positive, even paradisaical visions of America" reported above as well as to "correspondingly negative opinions of Irish conditions" (Miller et al., 2003, p. 9).

Within the context of 19th century Irish-Australian immigration, David Fitzpatrick (1994b) interpreted letters in terms of their social function, for example as "a tool for sustaining solidarity among separated kinsfolk," or in terms of the transnational ideas they conveyed, such as "the differences between Ireland and Australia" and the "nature of 'Irishness'" (p. 36). As

with other edited collections, the chapters in Fitzpatrick (1994b) contain letters grouped thematically, in this case by letter function or the "problems raised by migration" (p. 28).

While Fitzpatrick (1994b) considered "the relative frequency of various themes," he mostly avoided "arithmetical 'content analysis'" because of its "spurious precision in a study not based on a representative sample," but rather on just 111 letters sent or received by 14 migrants between 1843 and 1906 (p. 28). Although material from the "letter-writing classes" was available, the sample was drawn from "the semi-educated, subliterate majority," with the expectation that "the preoccupations of the emigrants who chose Australia probably differed little from those of the Irish in America or Canada" (Fitzpatrick, 2006, p. 98). A high degree of variation in format, rhetorical structure and vocabulary in the migrants' letters, associated with factors like gender and class, "undercut all attempts to reduce their correspondence to uniformity," save for the presence of "a distinctive blend of ceremonial and conversational elements" (Fitzpatrick, 2006, p. 98).

In particular, Fitzpatrick (1994a) described ritualized salutations, often containing references to health and God. Also noted by Schrier (1958), the "standard opening sentence which seemed to characterize a great majority of the letters" could be modeled as follows:

> Dear Father and Mother, I take this favorable opportunity to write these few lines hoping the arrival of this letter finds you in good health as it leaves me at present, thanks be to God for his kind mercies to us all. (pp. 24, 175)

Four out of five salutatory elements identified by Fitzpatrick (2006) appear in this formulaic opening. These include introductory phrases, references to correspondence, discussion of health, and affirmation of faith. Thomas and Znaniecki (1918-1920/1927) described Polish-American emigrant letters similarly as having an "exactly determined composition" that included references to religion, health and "bows for the whole family" (p. 303). After the courteous

opening, a kind of intimate conversation would follow, “often achieving the effect of informal chat” (Fitzpatrick, 1994a, p. 18). The flow of a typical letter reminded Fitzpatrick (1994a) of the medieval rules of correspondence, called the *Ars Dictaminis*, in which writers offer the formal *salutatio* and *exordium*—that is, the “salutation and securing of good will”—followed by the *narratio*, or narrative, where they introduce and discuss subjects or try to amuse the reader (Perelman, 1991, p. 110). An example of the kind of commentary found in this section is described by Fitzpatrick (1994a) as comparison, “explicit or implicit,” between countries of origin and resettlement. Unlike the Norwegian-Americans in Hale (1986) and the British-Americans in Fender (1991), the Irish-Australians in Fitzpatrick (1994a) did not comment disparagingly about the old country while “extolling the virtues” of the new country; rather, they expressed generally positive sentiment, including “admiration for both ways of life” (p. 19).

Fitzpatrick (1994a) was especially interested in use of the term *home*, observing that “this word invites a statistical analysis of its connotations, which range from house to neighborhood and nation while roaming from Ireland to Australia and even ‘our heavenly home’” (p. 19). He noticed that writers associated home with Australia only when referring to “a place or dwelling” but with Ireland in “many an expression of nostalgia, regret or affection” (Fitzpatrick, 1994a, p. 19). He acknowledged that letters represent the “transition phase between displacement and disconnection” and that, therefore, the term *home* may not be associated with Australia because migrants stopped writing around the time they began to feel at home in the new country (Fitzpatrick, 1994a, p.19).

David A. Gerber (1997) placed edited migrant letter collections like those produced by fellow historians Fitzpatrick (1994b), Blegen (1955), Conway (1961) and Erickson (1972) in critical dialogue with the groundbreaking work by sociologists Thomas and Znaniecki (1918-1920, 1927). He described historians’ approach of presenting letters mostly verbatim with

scholarly contextualization as a populist treatment lacking systematic analysis. He contrasted this with the positivist approach of Thomas and Znaniecki (1918-1920, 1927), whose aim was a law explaining social change. By arguing that "all of the concerns addressed in the letters—property, work, education, raising children, relationships with relatives, loneliness, etc., may be neatly packaged as declining solidarity and increasing disorganization," the sociologists failed to "realize the potential" of personal letters by imposing "one key to interpretation upon texts that demand multiple frames" (Gerber, 1997, pp. 10, 12).

In his own analysis of letters written and received by 19th century Scottish, Irish and English immigrants in the United States and Canada, Gerber (2006) concentrated on two themes—identity formation and long-distance relationship maintenance—in a book that takes a different form than those of previous historians. Half analysis, half case study, *Authors of their Lives* nods to the positivist origins of migrant letter scholarship, while taking a critical sociohistorical approach to the subject. On the one hand, Gerber (2006) acknowledged the limitations presented by his non-representative sample; in particular, his focus on English-speaking Protestant immigrants, who had a "close cultural similarity to the core Anglo-American people they settled among," meant that the "study would have no implications…for thinking about other immigrant groups" (pp. 13-14, p. 14). However, he also noted, like world historians and migration experts Manning and Trimmer (2013), a "harshly negative evaluation" of Irish Catholics, which was pervasive in letters, to the extent that: "No people encountered by British immigrants, including American and Canadian native peoples and African Americans, who usually inspire sympathetic commentary, evoke [sic] this type of bigoted writing" (Gerber, 2006, pp. 22, 23). Also evident were negative sentiments about the resident white population in the United States. Such assessments "caused a degree of wariness in social interactions" and served as "powerful incentives to limit one's most significant interactions to the circle of fellow

ethnics," scenarios Gerber (2006) described as counter to "commonsense notions" about the easy integration of culturally similar migrants (p. 27). Thus, from a positivist perspective, Gerber (2006) could not extend his interpretation to groups outside his sample, but from a critical sociohistorical perspective, dichotomies between migrant types could be relaxed, comparisons drawn, similarities and differences explored, as proposed by Lucassen et al. (2010).

In terms of motivations, some letters fit the economic model described by Ravenstein (1889), but many indicated "highly individualized difficulties," such as family feuds, unhappy marriages or public humiliation, which were as important as the "structural push and pull forces we usually associate with the emigration decisions of large occupational cohorts" (Gerber, 2006, p. 15). From a structural standpoint, Gerber (2006) helped to describe epistolary narrative, categorizing text as regulative, expressive or descriptive. Regulative writing focused on relationships, with common themes being reciprocity (or lack thereof) in correspondence, letter conveyance, privacy and social networking. Expressive writing communicated "lived experience" and "emotional states," while descriptive writing portrayed "daily concerns, events, and routines" (Gerber, 2006, p. 101). In reference to the idea presented by Fitzpatrick (1994a) that immigrants may stop writing when they begin to feel at home in a new country, Gerber (2006) noted "no particular pattern" to the cessation of correspondence and "few, if any, suggestions in the previous letters that an end is drawing near" (p. 203). While most letter series ended suddenly without apparent reason, a few indicated some possible factors: alienation, boredom or dislike of writing, family conflict or reunification, and death or illness. The "natural close of life" is very rarely the rationale, leading Gerber (2006) to conclude that letter collections underrepresent "aging immigrants" (p. 208).

Another historian who devoted much of his life's work to migrant letters was Eric Richards. In 2006, he reported that 19th century migrant correspondents expressed "their

individuality, their localism and their attachment to family" but rarely their views about the "larger forces" (e.g., sociological, political or economic) that were so clearly impacting their lives (pp. 59, 60). With a "low self-awareness of their place in the wider world," migrants wrote about "their domestic world…the prospect of returning home…ways of coping, borrowing and repatriating money," resulting in "limits to their [letters'] explanatory value" for the researcher (Richards, 2006, pp. 60, 61). Nevertheless, Richards (2006) asserted that the "emigrant letter is most effective where it can be linked to a hypothesis, where it can be made to yield evidence toward explanation" (p. 70). Appropriate lines of inquiry, according to Richards (2006), include the psychology of emigration and adaptation, historical literacy rates in countries of origin and settlement, mortality, the flow of money, attitudes about race, interactions among groups, family unity and social networks. In near dialogue with IOM (2017), which highlighted "the importance of understanding migration from the migrants' perspectives" to understand decision making in "life and death scenarios," Richards (2006) observed that letters show "the way in which people calculated the risks" of migration (p. 172; p. 69). He believed that "dealing with emotional dislocation" was the "central purpose" of the migrant letter and that experiences followed a "U-shaped curve" that could be traced through time (pp. 67, 68). Echoing the concerns of Lucassen et al. (2010), Richards (2006) noted that letters were "too often considered in isolation, in segregated ethnic clusters, which tended to mask the sense of similarity or difference among migrants" (pp. 69-70).

Interludes: Interdisciplinary Dabbling

As part of an ongoing project in social history and geneology, Cameron et al. (2000) examined letters written by poor English labourers who resettled in Canada in the 1830s under an assisted passage scheme sponsored by a wealthy landowner and coordinated by a parish rector, the Reverend Thomas Socket, who went to extraordinary measures to encourage, collect

and circulate letters. Eighteen hundred men, women and children from the same part of rural southern England participated in the program, and their identities and home parishes are documented, along with other personal information. Unique features of this collection include the homogenous demographic profile of the writers and the fact that all the letters were written during the first years of the migration experience, meaning that they "describe a recent journey…impressions of Upper Canada…through the eyes of a newcomer…a particular stage in emigration rather than the whole story" (Cameron et al., 2000, p. xxi). Unlike most other studies, women are well represented, with 32 out of 93 letters having a female author or co-author.

In comparing the Petworth letters to other collections, Cameron et al. (2000) observed great variance in "education, in social attitudes, in aspirations, in religion, and in temperament" but also "common themes" and "something like a family resemblance" between the writers in their collection and other 19th century British emigrant correspondents (pp. xli, xx). They found support for the claim by Ravenstein (1889) that motivations were essentially economic, in this case "low wages and competition for employment" (Cameron et al., 2000, p. xl). But they also reported that correspondence from already departed emigrants was a motivating factor, based on Socket's report that "labouring people…would not stir until they received reports from the immigrants from the year before" (Cameron et al., 2000, p. xl). Positive sentiments, such as happiness and pleasure, were associated with migrants' new-found independence and their "sense of control," whereas the negative sentiment of bitterness was associated with poor treatment as workers in England (Cameron et al., 2000, p. xlii).

In 2005, the historical theologist Donald Sinnema published an edited collection of migrant letters focused on an early 20th century Dutch settlement in Canada. True to their function of advising and attracting prospective migrants, these letters supplied information about travel, geography, climate, public infrastructure, social and religious life, the cost of living,

homesteading opportunities, farming logistics and other practical matters. While the tone was generally positive and encouraging, Sinnema (2005) reported that the mood occasionally appeared to be influenced by the weather. The letters included in the collection provide a "more intimate glimpse of life," touching on delicate subjects, such as the character of other migrants and quarrels between family members (Sinnema, 2005, p. 35). With only three letters in the collection, female migrants are underrepresented, making their pioneer experiences "rather invisible" (Sinnema, 2005, p. 36).

In 2011, historian Gur Alroey released an edited collection of 66 letters written by eastern European Jews who emigrated between 1875 and 1924. His goal was to close a gap in the scholarly study of migration experiences by focusing on "the gestation stage of the migration process...before the decision was made to emigrate and a destination was chosen" (Alroey, 2011, p. 6). While studies of this phase had been done using memoirs and oral histories, he argued that letters, produced in "real-time," offered better access to "migrants' doubts and vacillations before they set out on their way" (Alroey, 2011, p. 3). His findings indicated that the "decision making process was a rational one," not "triggered by despair," rather "a reasoned one in which the prospective emigrants tried to figure out what was best for them" (Alroey, 2011, p. 36).

Alroey (2011) asserted that understanding migrant decisions means considering omissions as well as inclusions; for example, like Kula et al. (1985), the letters in his collection made very little mention of pogroms and other forms of persecution and violence against the Jews. Instead they focused on practical matters, especially those of an economic nature. Mirroring Richards (2006), who referred to the local mentality evident in the letters of Australian immigrants, Alroey (2011) highlighted the "importance of the 'local level' in understanding the causes and characteristics of migration" (p. 3). While he identified economic opportunity as the

most important factor in determining whether Jewish migrants left eastern Europe, he suggested that the persecution they experienced there caused them to not look back.

Historian Lisa Chilton examined letters by single, middle-class, educated female immigrants in Canada, Australia, South Africa and New Zealand, who were assisted by the British Women's Emigration Association (BWEA) from the 1880s until the beginning of World War I. This focus on female migrants set Chilton (2016) apart from most other studies, which underrepresented women. Selected for the emigration scheme because of their standing as "'respectable' British women" who would promote "the growth of the 'right' sort of sociopolitical views and values" in "colonial spaces…especially…the 'rougher' frontier areas," these correspondents expressed a sense of duty to their sponsors, who expected publishable letters sharing useful information about journeys, surroundings and work experiences, of which farming, nursing and domestic service were common (Chilton, 2016, pp. 154,155).

These professional women wrote about "business ventures and charitable endeavors" as well as the more typical "struggles to…recover from major setbacks, adapt to extreme weather conditions, and overcome homesickness and feelings of social isolation" (Chilton, 2016, pp. 167, 161). Expressions of sentiment, such as "joy and comfort," were associated with the receipt of home country newspapers and other media, which served to connect the migrants with the "social world…left behind" (Chilton, 2016, pp. 161, 162). They wrote less about family because previous migration, morbidity and other factors had weakened those connections. Yet, a "desire for strong ties with people 'back home' was clearly evident" in the "genuine emotional attachment" the women expressed in their letters to BWEA staff, who along with the other assisted migrants became their "support system" in lieu of biological family (Chilton, 2016, p. 162, 166). Through their "embrace of the familial discourse" and their adoption of the roles of "surrogate daughter" and "sisters in emigration" to other women in the BWEA network, these

migrants were able to “solidify the system of support” and “squeeze more out of their relationships” (Chilton, 2016, p. 166, 167).

In 2018, Chinese history scholar Gregor Benton along with social scientist Hong Liu presented an analysis on qiaopi, a special kind of letter that contained remittances sent by Chinese migrants to the home country via special agencies designed for this purpose. These letters were functionally similar to their European equivalents in that they sustained family solidarity, carried personal news and disseminated information useful to future migrants. They exhibited similar epistolary structure, style and typology, but with less variance because of the focus on remittances. Benton and Liu (2018) reported key differences between Chinese and European letters in this important comparative study, which shed light generally on the nature of migrant letters.

Like European letters, qiaopi were disproportionately written by men, who migrated in greater numbers than women; however, the writers generally did not expect to settle permanently, as did their European counterparts. Rather, the mission of Chinese migrants was “to send back money and accumulate enough for a return in triumph,” meaning their letters were more likely to “straddle the boundary between personal and business matters” (Benton & Liu, 2018, p. 152). Low literacy rates, the availability of qiaopi agency services and the regularity of correspondence meant that these letters were more often ghost-written, dictated or prefabricated than those from European migrants.

Benton and Liu (2018) pointed out that qiaopi were less “punctuated with expressions of warmth” than European letters (p. 161). With regard to the epistolary types described by Gerber (2006), letters written by Chinese emigrants showed an “expressive deficit” but were similar to letters written by European emigrants in their regulatory and descriptive elements (Benton et al., 2018, p. 163). Unlike Gerber (2006), who claimed no correlation between gender and emotion,

Benton and Liu (2018) argued that the absence of emotion in Chinese letters was attributable to "the lesser intensity of emotional display and expression of intimacy in Chinese culture, particularly in males" (p. 163).

Chinese correspondents invoked religion and modeled its rhetoric less, and they were less likely to practice "strategic silence" than their European counterparts (Gerber, 2006a, p. 151). As absent heads of families, they issued more admonitions and directives for junior addressees. They also wrote about education more. Because qiaopi functioned as a financial record, Benton and Liu (2018) argued that they are less affected by survivorship bias than European letters. Thus, the fiscal function of the qiaopi as well as kinship dynamics in Chinese culture combined to make these letter collections "socially more inclusive" (Benton and Liu, 2018, p. 166). That said, they argued that the "practical…uniform and insubstantial" nature of qiaopi make them "less responsive to literary, textual, and content analysis" (p. 167). Finally, Chinese migrant letters were not shared and published to the extent that European letters were, suggesting that surviving collections are less subject to the biases that come with writing for an audience or editor.

The Computational Twist: Linguists at the Front

In a major step toward applying technology to the study of migrant letters, historian Patrick Fitzgerald examined emotion in machine readable letters extracted from the Irish Emigration Database. Using keyword searches, he identified letters containing an "expression of moral or emotional exchange between writers in the same family groups (Fitzgerald, 2008, pp. 269). From a quantitative perspective, Fitzgerald (2008) found support for the proposition that migrants wrote "in response to particular emotional triggers dictated by calendar dates," which can yield a "detectable pattern in correspondence" (pp. 276, 278). In a sample of 525 letters written over a 15-year period in the mid 19th century, the month of March was followed by

December and January as the highest frequency time for letter writing, likely attributable to emotional as well as economic factors. According to Fitzgerald (2008), the "early spring 'pick up' in economic activity in the northern United States and Canada interacted with the emotional triggers of St Patrick's Day (March 17) and Easter to motivate migrants to send letters with monetary enclosures to relatives (Fitzgerald, 2008, pp. 278-279). Similarly, Christmas and New Year's greetings held remittances, which were "an economic necessity for many in post-Famine" Ireland (Fitzgerald, 2008, p. 279). The slowest month for sending letters was October, possibly because of the anticipated economic slow-down in the winter months ahead and the understanding that a money-laden "Christmas letter" would be expected (Fitzgerald, 2008, p. 279). Ultimately, Fitzgerald (2008) presented more questions than answers, such as: Did letters carry more emotion after the invention of photography? Did certain calendar dates, such as a writer's birthday or the anniversary of their arrival in a new country, serve as "emotional triggers" to write home? Does the expectation of homesickness predict actual homesickness, as indicated in pre- and post-departure letters?

The movement toward the computational and quantitative study of migrant letters has been led by Emma Moreton in the field of corpus linguistics. Using word frequencies, n-grams, type/token ratios and concordances, Moreton (2012) examined 99 letters written by four sisters, who migrated to the United States from Ireland during the second half of the 19th century, to "build a picture of how, through letters, family bonds were changed and maintained over space and time" (p. 643). Using a reference corpus comprised of 42 letters written in equal parts by male and female Irish-American immigrants, she discovered the repeated pattern *I* + verb + *you* + modal / auxiliary word + verb, especially among female writers. According to Moreton (2012), these "projecting structures," such as "I hope you will write," expressed migrants' wishes and needs to their readers, along with their expectations about some form of response, thereby

"serving to maintain a psychological link" between them and their loved ones in the home country (p. 639, p. 644). She proposed extending her "transparent and replicable" procedure to include social factors, such as class and education (Moreton, 2012, p. 644).

In another study based on the same dataset, Moreton (2016) used a mixed methods approach to examine 35 letters written between 1884 and 1927 by one of the sisters, Julia Lough, who moved to the United States at the age of 13. Through a close reading of the letters, Moreton (2016) identified 24 topics, which she grouped into three categories based on function or frequency. The first category contained high-frequency topics related to letter structure (e.g., previous/future letters, greetings, weather, etc.). The second category included high-frequency, non-structural topics: Ireland/America, family/friends, religion, recollections, homesickness/separation, health/illness, work, enclosures and remittances. The third category included low-frequency topics that were "more personal and reflexive in nature" and thus showed "moments of greatest authenticity, directness, expressiveness and personal identity" (Moreton, 2016, p. 325). These included news/events, reunification, death, daily life, the writing process, identity, education, migration and transportation. Although rare, these topics assumed a primary role in the letters when they occurred, which was less often the case for higher-frequency, non-structural topics. Exceptions to this rule were the high-frequency topics of family/friends and remittances, which tended to be primary, and the low-frequency topics of identity, education and migration, which tended to be secondary.

Moreton (2016) examined the 10 most frequent part-of-speech (POS) 2-grams in narrative units falling within the recollection topic. Two of the highest ranked constructions contained a personal pronoun + verb in either the present or past tense, such as *I remember / hope / suppose / know* or *she / we / you / I used*. By qualitatively examining concordances, Moreton (2016) determined that these constructions were often used in the "act of

remembering," which served to reconnect the author with her homeland and "reinforce bonds with loved ones" (p. 328). Moreton (2016) found POS patterns to be associated with other cognitive acts, such as predicting, comparing or counting (e.g., the passage of time or the frequency of an event).

With computational linguist Rachele De Felice, Moreton (2019) proceeded to test an automated speech act tagger, developed for workplace email, on letters written by the same 19th century Irish emigrant, Julia Lough, this time including letters from her three sisters. On the small dataset comprised of 19 letters containing 621 sentences, the tagger correctly categorized almost 80% of the sentences, which is similar to its performance on the original email corpus. Sentences categorized as first-person expressive (FPF) had an 89% accuracy rate and often contained lemmas conducive to sentences projecting the letter-writer's "expectations, desires or beliefs onto the recipient" (De Felice & Moreton, 2019, p. 172). The most frequent of these lemmas, *hope*, was preceded 95% of the time by the first-person pronoun *I*. About half the time, this combination was used in greetings, inquiries about family members, or references to the exchange of letters, money or other items. The remaining instances were more "personal and reflexive in nature," exuding the "greatest authenticity, directness, expressiveness, and personal identity" and revealing the letter writer's "preoccupations and beliefs and her role within the family" (De Felice & Moreton, 2019, pp. 170, 169). Prevalent themes included education, work, womanhood and current affairs, and the most notable desire was for families to stay together. De Felice and Moreton (2019) observed that the correspondents in their study viewed their situation as "forced separation," or as Miller (1985) argued as involuntary exile rather than immigration for opportunity (p. 170).

In another 2019 study, linguist Nancy Avila-Ledesma used letters drawn from the Irish Emigration Database (IED) to compare experiences by destination: The United States versus

Australia and New Zealand. By examining collocations and concordances for the high-frequency terms *land(s)* and *situation(s)*, she found that the letters of emigrants to the southern destinations were more positive than those of emigrants to the northern one. Like Miller (1985), Avila-Ledesma (2019) reported that Irish-Americans perceived migration as "an escape from poverty and discontent," if not quite exile, whereas the Irish in Australasia cast their experiences in terms of pursuing opportunity (p. 117). Avila-Ledesma (2019) also observed that the American contingent expressed "an acute homesickness" indicating a strong ongoing attachment to and identification with the home country (p. 118). The absence of homesickness on the part of Irish-Australians conflicted with Fitzpatrick (1994a), who observed nostalgia, regret and a more profound sense of the word *home* when applied to Ireland.

With letters spanning 1840 to 1930 and the finding that "rarely did they [migrants] encourage further departures by praising America's economic opportunities," Avila-Ledesma (2019) accorded with other studies indicating a negative sentiment shift around the middle of 19th century in the United States (p. 117). This mood change supports the absorption-and-dispersion law of migration presented by Ravenstein (1889), who used census data to argue that, by 1880, the eastern United States had become saturated with newcomers to the point of dispersion, such that absorbing areas and the opportunities they offered existed only west of the Mississippi. By comparison, Irish immigrants in Australia and New Zealand were, as Fitzpatrick (1994b) noted, "nation builders, rather than late arrivals competing against an entrenched population for living space, jobs, and spouses," and thus they faced better prospects upon arrival, which is reflected in their more positive letters (p. 19). As a direction for future research, Avila-Ledesma (2019) proposed extending her comparative analysis to Irish migrants in Canada.

Weaving the Threads Together: One Scholar, Many Perspectives

The interdisciplinary unfolding of the migrant letter story might be best represented by cultural sociologist and life writing specialist Liz Stanley. She has made and continues to make major contributions to the study of historical migrant correspondence, particularly from a feminist, theoretical and methodological perspective. In 2004, Stanley conceptualized letters as being essentially dialogical (i.e., part of an exchange of ideas), perspectival (i.e., content varies over time and by addressee) and emergent (i.e., conventions are individualistic and constantly in development). Importantly, she described letters as a fractured form, full of "gaps, ellipses and mistakes" (Stanley, 2004, p. 221). As snippets of experience and parts of conversations, their interpretability as either a self-standing or intertextual narrative is limited by their missing components.

Like other scholars interested in migrant letters, Stanley (2010) reflected on the trailblazing work of Thomas and Znaniecki (1918-1920/1927), arguing that they were "on to something very important about the nature of change…and how people respond to it" and pointing out that letters and contemporary analogs like email were "absolutely central to this" (p. 149). She connected their work to the pragmatism research paradigm guiding this thesis, and more specifically to the theory of symbolic interactionism. Described by Blumer (1969), this theory emerged from ideas developed by prominent sociologists, including the authors of *The Polish Peasant in Europe and America*. According to Plummer (2004), symbolic interactionism "shuns abstract totalizing truths in favor of local, grounded, everyday observations," of which migrant letters stand as a subset (p. 1105). The key tenet is the primacy of the social self—one that is "relational, situational and sequential"—as opposed to a self that is "inner-reflecting" or fixed in any way (Stanley, 2010, p. 139).

Most of Stanley's work on migrant correspondence has been oriented around white settlers in South Africa after 1770. She argued that the features identified by Gerber (2006),

Fitzpatrick (1994b), Richards (2006) and others as definitional of a migrant letter subgenre—those being permanent absence, relationship maintenance, identity making and "parallel locations"—manifested differently within the South African context, or else they were simply true of all kinds of correspondence (Stanley, 2015, p. 401). She attributed the unique nature of the letters in her study to local economic factors and concluded that epistolary material, including letters but also email and other electronic forms, must be assessed in a wider variety of contexts before a set of subgenre defining features can be accurately established.

In 2017, Stanley said her approach to research was "to combine the very small with the very big in providing in-depth, close readings of individual texts, while also using quantitative techniques to gain purchase on overall patterns and contexts" in a "backwards-and-forwards analytical movement between the big and the little to underpin interpretation" (Stanley & Jolly, 2017, p. 231). While her approach is mostly critical, qualitative and reflexive, she has written about the potential of computational tools, such as the R programming language, to support, enhance, complement and communicate her work, for example, by enabling "large quantities of text data to be analysed quickly and efficiently" or by developing "a whole suite of new high-powered tools" to be used by her research team and visitors to her research website (Stanley, 2015, paras. 4, 5).

My Contribution of the Story of the Migrant Letter

In the century since William I. Thomas and Florian Znaniecki broke the seal on the migrant letter, social scientists and humanists have told its story from many perspectives and in various ways, eventually leading to the mixed methodologies used by Emma Moreton and Liz Stanley. Despite the range of paradigms guiding this body of research, a dominant narrative has taken shape. It tells us that the key forces at work in the migrant letter are social, economic and psychological. These are reflected in the form and function of letters but also in the topics and

emotions they encode. Social concerns are revealed in passages about family, children, friends, community and relationships, economic ones in passages about work, the cost of living and money. Psychological factors are conveyed through emotions, motivations, assessments and decisions. Meanwhile, cultural factors, such as language and religion, as well as demographic factors, like gender, education, occupation, origin, destination and time of writing seem to differentiate one kind of experience from another.

My contribution to the body of research described above will be to examine whether the story of the migrant letter collectively told by scholars is reflected when migrants' own words are examined en masse under a quantitative lens using the tools of computational text analysis and Bayesian data analysis. Specifically, I will examine the topics and emotions in the letters of North American immigrants during the long 19th century, a population already well studied by researchers using qualitative, critical and mixed methods. To more fully integrate the social science and humanistic perspectives that characterize the study of migrant letters, I will examine the degree to which topics and sentiments vary by key structural, temporal and authorial traits. To take this analysis a step further, I will explore the relationships between these variables and migration outcomes, specifically cessation of correspondence, which Erickson (1972), Fitzpatrick (1994a), Richards (2006) and others have suggested may occur when migrants feel at home in a new country.

While digital humanists Weingart (2012) and Underwood (2012) have written blog posts about the potential of Bayesian methods, the only application of them within the digital humanities that I am aware of is in the domain of machine learning and natural language processing. Whereas the Bayesian techniques to be used in this thesis will be oriented around describing and explaining the relationship between language and the social process of migration, with prediction being an application, machine learning algorithms inspired by Bayesian

probability theory are entirely oriented around prediction for specific classification tasks, such as authorship attribution, as in the ground-breaking work by Mosteller and Wallace (1964), or character identification, as in Bamman et al. (2014). This thesis will use classification algorithms to extract topics from letters, but that data will also be used in Bayesian linear models to explore the relationships between variables.

Methodology

The objective of this thesis is to learn about migration experiences by examining the topics and sentiments expressed in historical migrant letters. The dataset includes 915 letters written by 218 international migrants in North America during the long 19th century and extracted from the Alexander Street Press collection entitled *North American Immigrant Letters, Diaries and Oral Histories*.[1] The research questions are as follows:

1. What topics and sentiments are evident in the letters?
2. Do they vary by temporal, structural or biographical factors?
3. Do they predict when migrants stop writing?

The review of literature indicates that key temporal factors include when the letter was written as well as when it was written relative to the migration event (i.e., how long since the migrant left home). The structural feature of interest is the positioning of text within a letter, specifically, where is the text relative to the end of the letter. Finally, important factors on the level of the writer are gender, age, national origin, religion, occupation and country of settlement. The migration outcome to be probed in this study is cessation of correspondence, which may be a sign of social integration into the new country.

Based on the work of the scholars presented in the literature review, some findings are expected. For example, topics will likely be practical and show a local rather than a global mindset (Alroey, 2011; Richards, 2006). Topics will mostly reflect concerns about economic issues (Alroey, 2011; Erickson, 1972; Brinks, 1986; Miller et al., 2003; Cameron et al., 2000) and relationship maintenance (Gerber, 2006; Chilton, 2016; Moreton, 2012, 2016). Sentences close to letter openings will show less topical variation (Schrier, 1958; Thomas & Znaniecki,

[1] Although debated, the long 19th century is generally understood to be the period between the late 18th century and the start of World War I (Burke, 2000; Porter, 1999).

1918-1920/1927; Fitzpatrick, 1994a, 2006). Sentiment will reflect the U-shaped curve proposed by Richards (2006). The letters of Irish-Catholic migrants will be more negative than those of other anglophone migrants (Gerber, 2006; Miller & Boling, 1990). Letter sentiment will become more negative in the mid to late 19th century (Erickson, 1972; Conway, 1961; Walaszek, 1992; Avila-Ledesma, 2019). Finally, based on the idea presented by Erickson (1972), Fitzpatrick (1994a) and Richards (2006) that migrants stop writing when they feel at home in a new country, the following hypothesis will be tested: The more positive the sentiment, the more probable that correspondence will end.

The general methodological approach used in this thesis is exploratory data analysis (Tukey, 1972; Gelman, 2003). In keeping with the pragmatic research paradigm guiding this thesis, the methodology is inspired by the framework of statistical pragmatism presented by Kass (2011) and refined by Gelman (2011). Specific methods include sentiment analysis, topic modeling and Bayesian linear regression.

Data and Subset

I extracted the letters in this study from the North American Immigrant Letters, Diaries and Oral Histories database from Alexander Street Press. This database contains more than 80,000 pages of narrative material produced by people who moved to or within North America between 1800 and 1950. Although access to the online database is available through subscription service to McGill Library, I required direct access to the data files for the purpose of text mining. To permit this, library staff arranged to purchase the raw data from Alexander Street Press. I received the data in the form of a hard drive holding 8,749 text files and two Excel spreadsheets, one for text file metadata and the other for biographical data about the authors.

Using the R programming language and the comparedf function in the arsenal package, I compared the two spreadsheets to identify redundancies and inaccuracies. Although the

document level spreadsheet contained almost all the variables from the author spreadsheet, their values and attributes differed. A variable-by-variable examination revealed that the more robust biographical data was contained within the author spreadsheet. A new CSV file was created that merged the document level and author level data into one spreadsheet, in which each row was a letter and the columns contained information about either the document or the author.

The final spreadsheet contained only items meeting the following conditions: The document represents a letter, the author was a first-generation migrant, the text is in English (either originally or through translation), the letter was written in North America, and the dates of immigration and writing were between 1789 and 1914, the years widely understood to encompass the long 19th century. Text files matching the document ids from the spreadsheet were extracted from the collection of 8,749 files supplied by Alexander Street Press. These were placed into a separate folder. Further crosschecking, correcting and cleaning reduced the spreadsheet and corresponding collection of text files to 915 letters written by 218 migrants. This is the subset that served as the basis for further analysis.

Metadata: Exploration, Preparation & Summary

Temporal Variables

Time of Writing. Figure 3 shows the years the writers arrived in North America and the years they wrote their letters. The median year was 1834 for immigration and 1863 for letter production. Surges in immigration are noted between 1830 and 1834 and in the first years of the 20th century, with an isolated spike in 1856. The surges as well as the spike are mirrored by concurrent increases in letters. In addition, there are two important letter writing peaks, which occur between 1856 and 1864 and between 1880 and 1888; that is, about 20 to 25 years after the first immigration surge and the spike in 1856. The dataset does not include letters from the earliest years of the long 19th century—that is, from 1789 to 1800. Year of immigration is

missing for 146 (67%) of 218 writers whereas year of writing, usually provided as part of letter writing convention, is missing for just eight (less than 1%) of 915 letters.

Time since Immigration. The number of years between immigration and writing was calculated for each letter and this value placed into a new variable, which had a unimodal and positively skewed distribution with a median value of seven years, as shown in Figure 4. Although less frequent, letters were steadily produced up to sixty years after immigration. This variable lacks data for 29% of the letters (n = 263) as a result of year of immigration or year of writing being unknown.

Biographical Variables

Location. Fifty-two percent of the letters (n = 481) were sent from the United States compared to 46% from Canada (n = 427) and less than 1% each from Mexico (n = 3) and other destinations in North America (n = 4). There is no missing data for this variable.

Gender. As with other collections of migrant letters, women are underrepresented, with only 41 (19%) of 218 authors being female and 177 (81%) being male.[2] However, the few female writers produced an outsized proportion of the letter collection: 532 items (58%) as opposed to just 383 (42%) for men. Figure 5 shows the gender and location distributions for the group of writers and the letter collection. The disparity between the two graphs is partly attributable to three letter series, containing 100 or more items each, written by the female migrants Jette Bruns (n = 136), Susannah Moodie (n = 101) and Sarah Stretch Harris (n = 186). Together, these prolific women wrote 46% of the letters in the collection. There is no missing data for the gender variable.

Age. The letters were written by people who ranged in age from 8 to 85 years, with the mean age being 43 years (SD = 17). Female writers tended to be older than male writers with a

[2] This could stem from gender norms favoring male correspondents or lower literacy rates among females.

mean age of 47 compared to 31; however, this is again likely attributable to the three prolific women named above, who began writing letters in their 20s and stopped in their 80s. Figure 6 shows the age distribution of letters generated by these individuals. As with year of immigration, the missing data (n = 234) is more associated with men than women, in this case at a rate of 85% to 15%, respectively.

Culture. The cultural background of the writers is defined in this study using the available data: national origin and religion. About one-third of the writers (n = 73) were of Norwegian origin, with 30% (n = 64) from Great Britain and 11% (n = 25) from Holland. Religious information is missing for most of these individuals; however, the demographics and migration histories of these countries suggest that the writers were Christian and most likely Protestant. About 7% of the authors were Russian Jews (n = 16), and nearly all (13 out of 14) of the 6% of writers for whom country of origin data was missing were also Jewish. Only six individuals are shown as having Irish heritage and just one of these was Catholic. In fact, the Catholic religion is sparsely represented among the writers, with only two Czech, one Italian and one German joining the single Irish migrant in this category.

As with the female writers, Catholics were responsible for a disproportionate number of letters: They account for 2% of correspondents but 22% (n = 197) of the letters, or 39 letters per person on average. Even if all the writers categorized generally as Christian were assumed to be Protestant—a likely scenario given that most originated in England and Norway—this group represents at most 15% of writers and 45% of letters, or 13 letters per person on average. For the purpose of comparison, 16% of the authors were Jewish but produced only 4% of the letters. Figure 7 shows the distribution of national origin and religion for the writing group compared to the letter collection as a whole. It shows that Italian- and German-Catholic as well as English, probably Protestant, perspectives are more prominently represented in the data than Jewish

perspectives. The prolific writers, Jette Bruns and Susannah Moodie, were German-Catholic and English-Protestant, respectively. Sarah Harris was likewise English, but her Christian denomination is unknown.

Figure 8 integrates author age, gender and culture for the letter collection. It shows that the letters written by Catholics were written by women, with the German writers being considerably older, on average, than the Italians but also having a wider range of ages. The letters written by Protestants and Christians reflect a greater mix of genders but one that is likewise dominated by women, who tend to be older than the men. Both the English and the unspecified Christian groups include a cluster of letters written by mostly male youths. Most of the letters for which the writer's religion is unknown are male.

Occupation. Socioeconomic status (SES) is typically measured using a combination of income, education and occupation (Ellis, 2018). Only occupational data is used in this analysis because the other indicators are not included in the dataset. Writers in this study represent all of the occupational classes identified by Erickson (1972), including 39 workers in the agricultural class, 10 in the industrial class and 57 in the commercial-clerical-professional class. Two occupations (i.e., student and homemaker), associated with five individuals in the study, did not fit well with the classification system proposed by Erickson (1972). For this reason, a fourth class, "other," was added for these individuals. It must be noted that occupation data was missing for 125 people and that many writers had multiple occupations, resulting in 18 cases that mapped to more than one occupational class, as shown in Figure 9. In addition, the occupations for many women were given in terms of their husband's profession (e.g., farmer's wife, physician's wife). In these cases, the husband's job was accepted because it determined the writer's socioeconomic class, which is what this variable is intended to capture. Table 1 shows how occupations provided in the original dataset were mapped to the occupation classes used in Erickson (1972).

In order to flesh out the largest category—commercial, clerical and professional—and to better account for the jobs in the "other" category, a more nuanced classification scheme inspired by Miller et al. (2003) was developed, as shown in Table 1. In addition to agricultural, industrial, commercial and professional categories, this system included domestic, social and government categories. Counts for the agricultural and industrial classes were unchanged, but those for the commercial-clerical-professional and other classes were more widely distributed as follows: Commercial (n = 31), social (n = 25), professional (n = 15), government (n = 7) and domestic (n = 5). Individuals in the largest group, commercial, had jobs oriented around trade, while people in the next largest group, social, worked in human or community development, especially in domains like religion, education and health. The professional class was the third largest, including people whose work involved special knowledge or training and little or no manual labour (e.g., accountant, editor, engineer, etc.). The two smallest groups were associated with government work, such as politics or military, or with the domestic sphere, which included house staff as well as homemakers.

As mentioned, many writers had multiple jobs and therefore belonged to more than one occupational class. The UpSet plots in Figures 9 and 10 show cross-group membership for both classification schemes. The largest area of overlap was between the commercial-clerical-professional class and the agricultural classes, with 12 individuals pursuing work in both areas. The more nuanced scheme in Figure 10 revealed that the most common overlap was between the commercial and agricultural classes. Six writers did work that positioned them in both categories. A diversity of combinations followed, but the domestic class had the lowest rate of overlap with just 20% of its members belonging to another group. From that point on, cross-group membership rates proceeded as follows: agricultural (38%), social (40%), commercial (52%), industrial (60%), government (71%) and professional (73%).

The three female writers responsible for 46% of the letter collection were associated with multiple occupational classes. Jette Bruns was a homemaker whose husband was a physician. As such, she is a member of both the domestic and social occupational classes. Susannah Moodie was a writer and farm wife who was married to a retired army officer, positioning her in the professional, agricultural and government classes. Sarah Stretch Harris was the wife of an independent architect specializing in church design, which placed her in the commercial, professional and social classes. Thus, these prolific writers represent, to some degree, all occupational classes except for the industrial class, but only Moodie worked outside the context of home.

Migration Outcome Variable

Cessation of correspondence. Previous scholars, particularly Fitzpatrick (1994a), have suggested that migrants cease writing when they feel at home in their new country. To test whether this idea bears out in the present dataset, a new variable was constructed to indicate whether a letter was the last in a series. For the purpose of this study, a series was treated as two or more letters written by the same migrant. Single letters (n = 164) were coded as missing data for this variable. For the 751 letters included in one of the 54 series, the last letter in the sequence was coded as TRUE and all previous letters were coded as FALSE. In all but one case, the last letter in the sequence bore the latest (i.e., most recent) date. One letter was undated but accepted into the final letter subset with the understanding that editors or curators, acting on qualitative knowledge, were responsible for the letter sequencing. With 10 items each, or 38% collectively of the subset, the early 1830s and the early 1860s saw the greatest number of final letters. Most final letters were written in the first decade after immigration when writers in the subset were middle aged.

Narrative Data: Exploration, Preparation & Summary

The 915 letters in the dataset contain a total of 755,269 word tokens of which there are 25,067 unique types, producing a type-token ratio of 0.03. Lexical diversity increased to 0.09 when 19th century English language stopwords were removed, leaving 270,422 tokens and 23,849 types (Jockers & Mimno, 2013). The resulting rank-frequency distribution produces a curve that roughly follows Zipf's Law (1935), which states that the frequency of a word is inversely proportional to its rank. The top five words are time (*n* =1,648), dear (*n* =1,482), little (*n* =1,326), day (*n* =1,230) and letter (*n* =1,079). Figure 11 shows the rank-frequency distribution of the 400 most frequent content words with an inset showing the Zipf curve for terms occurring more than five times.

Structural Variable

Location of Text within Letter. In an effort to map out the internal structure of migrant letters, particularly the notion presented by Fitzpatrick (1994a, 2006) that openings are formulaic, the letters were tokenized by sentence and a new variable constructed to show the position of each sentence as a proportion of the whole letter. Thus, the range of values for this variable is 0 to 1, with smaller values nearer the beginning of the letter and larger values nearer the end. The distribution of this variable was relatively uniform (M = .51, SD = .29).

Methods for Research Questions

Measuring Sentiment

Sentiment was calculated using the Python programming language and a module called VADER (Valence Aware Dictionary and sEntiment Reasoner) developed by Hutto and Gilbert (2014). This tool calculates sentiment polarity and intensity based on scores in its dictionary, which contains 7,500 words and phrases. The items in this lexicon have been assessed by 10 human readers and assigned values ranging from negative-four (extremely negative) to four (extremely positive). Additionally, an algorithm identifies grammatical and syntactical patterns

and adjusts scores accordingly. For example, the word “nice” would be scored differently than the phrase “very nice.” Changes to polarity and intensity signaled in natural language through negation (“not nice”), contraction (“wasn’t nice”) and punctuation (“nice!”) are captured by the algorithm.

Because VADER was designed for contemporary language and its applications, such as social media, features like word shape (e.g., NICE versus nice) would also trigger scoring adjustments. Some of the texts under investigation in this thesis appear in all caps, presumably not for the purpose of emphasis but rather because that was the script style most comfortable for the migrant. For this reason, all words were converted to lowercase prior to sentiment scoring. Other preprocessing of text included the removal of ellipses, dashes, the expression “&dot” and whitespace characters.

The VADER composite sentiment score was chosen for analysis. This score is calculated by summing the value for all tokens included in the lexicon, applying the rules-based adjustments and normalizing the value to fall between negative-one (most negative) and one (most positive). This composite sentence score was placed into a new variable and then a score for the whole letter was calculated as the mean of the sentence scores. A CSV containing the 13,608 sentences in the dataset, each with its own sentence ID, sentiment score, in-letter positioning value and all letter-level data was saved for use in the topic modeling and linear regression phase of the research.

Identifying Topics

The metadata include three variables describing the subject matter, or topics covered, in each letter, but the process by which these tags were determined (i.e., human coded or machine learned) is unspecified. To standardize the topics, integrate the ideas of scholars like Moreton (2016) and probe topics at the sentence level, two forms of topic modeling were performed. The

first approach, used to extract topics for whole letters, follows the lead of Jockers and Mimno (2013), who modeled themes in 19th century novels using the Java based Mallet implementation of Latent Dirichlet allocation (LDA), developed by McCallum (2002). For the present study, the Mallet wrapper in the Python Gensim package was used (Rehurek & Sojka, 2010).

According to its developers, LDA is bag-of-words approach to topic modeling that treats documents as "random mixtures over latent topics, where each topic is characterized by a distribution over words" (Blei et al., 2003, p. 996). It assumes that a document contains a number of words, *N*, following a particular distribution (e.g., Poisson) and a number of topics, *K*, with a Dirichlet distribution. Each word (w_i) is the result of draws from these two probability distributions. With these assumptions in place, the LDA model "tries to backtrack from documents to find a set of topics that are likely to have generated the collection" (Chen, 2011). In a process inspired by Bayesian updating and multilevel modeling, the LDA algorithm does this by randomly assigning words to topics and calculating—case by case—the probability that word and topic assignments would produce the document collection.

While LDA works well with letter-length texts, it is less effective with texts of 50 words or less because of the sparsity of collocated terms. Three alternatives to Mallet LDA were used to model topics at the sentence level. The first is the author-topic model, developed by Rosen-Zvi et al. (2004) and implemented in Gensim by Mortensen (2017). In this model, authors are treated, like topics, as a Dirichlet distribution over words, which can be thought of as their personal vocabulary. A dictionary maps authors to sentences, and thereby to their words and interests. The second approach, the collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (GSDMM) introduced by Yin and Wang (2014), is built on the assumption that short texts are composed of just one topic rather than a distribution of topics, as in LDA. The third approach, the biterm algorithm developed by Cheng et al. (2014), learns topics using corpus-

wide word co-occurrences.

For all models, the key parameter—the number of topics over the whole corpus—was guided by the 24-topic system identified by Moreton (2016). At the letter level, I ran a series of LDA models with topic-number values ranging from three to 27. C_V coherence scores were compared, and the models with the highest scores were examined in detail and visualized using the pyLDAvis library for Python to check topic differentiation (Sievert & Shirley, 2014). The selected model was the one that had the highest coherence score, least overlap among topics and number of topics closest to that found by Moreton (2016).

Of the sentence model types, all performed well with 20 to 24 topics produced. The GSDMM model was selected because it produced more intelligible topics than the author-topic model and had shorter run times than the biterm model. In order to optimize GSDMM performance, additional text processing was needed on top of what was done to score sentiment and perform the letter-level topic modeling. Sentences with more than 50 words and words occurring only once in the dataset were removed along with any that contained no words after all processing was complete. This resulted in 2,592 of the 37,608 sentences (7%) having missing data for the sentence-level topic.

Modeling Relationships

Bayesian linear modeling, using the brms package in the R programming language, was used to determine if topic and sentiment varied by time, letter structure or author traits (Bürkner, 2017). Sentiment was modelled as normally distributed and topics as multinomially (i.e., categorically) distributed in single-level exploratory models. During linear regression, cases that are missing data are dropped from the analysis, which can lead to biased results if data are missing not at random (MNAR). Helbich and Kamphoefner (2006) report that the completeness of migrant letter collections—on the level of the letters themselves but also in terms of what is

known about the writers—is affected by factors such as gender, occupation, age, religion, education, wealth, mobility, personal interests and national origin of letter writer and recipient. Overall, Helbich and Kamphoefner (2006) reported that letters written by younger, female, Irish and urban migrants were less likely to survive and be supplied to researchers, resulting in their underrepresented status. This suggests that the letter-level missing data in this study is likely missing systematically rather than at random. Thus, to avoid introducing additional bias into the analysis, missing data for all predictor variables was imputed, based on the biographical factors listed above and using the MICE package in R (Van Buuren & Groothuis-Oudshoorn, 2011). New versions of the sentence- and letter-level CSV files were created to integrate the imputed values.

To account for prolific writers and to determine whether sentiment and topic predict cessation of correspondence, a multilevel structure was used in the analysis for research question three. As a preliminary step, a random-intercept model, clustered on writer ID and taking sentiment as the outcome and topic as the predictor, was performed to assess the relationship between these narrative features. Then, each of these variables was included, along with potentially confounding or mediating covariates, as the principal predictor in a multilevel model taking cessation of correspondence as the multinomially (i.e., binomially) distributed outcome variable. Based on the results of these regressions, a series of multilevel models, again clustered on author ID, were designed and the results compared using information criteria. The model with the most interpretable results and the lowest WAIC score was chosen for the final analysis. This model included sentiment and topic, with random intercepts for each writer and a random slope, conditioned on gender, for the sentiment term.

Results

The Bayesian approach to data analysis is based on probability theory, which McElreath (2017) describes as "counting the ways things can happen" (p. 20). Bayes Theorem, represented in simplified form in Eq. (1), transforms these counts into joint probability distributions, from which estimates can be made to support conjectures about the relative plausibility of outcomes.

$$\Pr(p \mid d) \propto \Pr(d \mid p) * \Pr(p) \tag{1}$$

The equation above states that the probability of a value for an unknown parameter (p) given the observed data (d) is proportional to the probability of the data given the parameter value multiplied by the prior probability of the parameter value. The equation can be further simplified as follows:

$$\text{posterior probability} \propto \text{likelihood} * \text{prior probability} \tag{2}$$

In Eq. (2), the likelihood is the linear model designed by the researcher, examples of which will be presented below. It contains the outcome and predictor variables, also known respectively as the dependent and independent variables, and the probability distributions for them and all their parameters, including their means and standard deviations. The prior is a probability distribution representing existing knowledge, which in the pragmatic style of Bayesian updating can be understood as the posterior distribution generated by the previous model run. The most plausible parameter values can be extracted from the posterior using Markov chain or Hamiltonian Monte Carlo (i.e., MCMC or HMC) algorithmic methods. These estimates can be used to determine the relative probabilities of various outcomes, conditional upon the data and model assumptions.

What follows is an introduction to how Bayesian statistical modeling can be used in the analysis of textual material. Although the merits of this approach are most abundant in the mixed-effects modeling used to answer Research Question 3, the Bayesian framework was used

throughout for the purpose of consistency and demonstration. Simple models of sentiment and topic are presented first, without any predictor variables, followed by univariate modeling showing the relationships between these narrative features and each of the temporal, structural and biographical variables. These fundamental principles will then be applied in a multivariate, hierarchical model to examine whether sentiment or topic predicts cessation of correspondence.

What sentiments and topics are evident in the letters?

Sentiments

Compound sentiment scores in the VADER system are generated by adding up the scores for each token in the lexicon, then applying rules-based adjustments, and normalizing the scores so that they fall between -1 and 1. Using a Bayesian approach to describing the distribution of values for this variable, we would use the model shown in Eq. (3).

$$\begin{gathered} sentimentLetter_i \sim Norm(\mu, \sigma) \\ \mu \sim Norm(0, 3) \\ \sigma \sim Unif(0, 10) \end{gathered} \tag{3}$$

This model states that letter sentiment scores are normally distributed around a mean, μ, and standard deviation, σ. Each of these parameters has the displayed prior distribution. For the mean, μ or mu, this is normal and centered on zero with a standard deviation of 3. For the standard deviation, σ or sigma, the distribution is uniform, meaning all values between zero and 10 are equally probable. Given the VADER scale and the possible outcomes it implies, these can be considered very uninformed priors. That is, they allow the model to explore a wide range of possible values for the parameters.

When provided with fuel (i.e., data) and a computational engine (e.g., MCMC or HMC), the model generates marginal posterior distributions for each of these parameters, including a

mean, a standard deviation and the upper and lower bounds for a credible interval that contains a pre-determined proportion of the most probable parameter values. For all models in this analysis, the lower bound of the credible interval is set at 5% and the upper bound at 95%, producing intervals that contain 90% of the probability mass. The more proximate the lower and upper bounds of the credible interval, the more certain the estimate for the mean. Taken together, the marginal distributions for the parameters comprise the joint posterior distribution.

For sentiment, the combination of marginal parameter values with the highest joint probability are .16 for the mean and .15 for the standard deviation. With 90% credible intervals of .15 and .17 for the mean and .15 and .16 for the standard deviation, the estimates generated by this model can be interpreted with confidence. Given that compound sentiment scores above .05 are considered positive in the VADER system, we can say that the letter sentiment scores in this dataset are on average mildly positive.

A similar model to that shown in Eq. (3) was used to estimate the distribution of sentiment scores on the sentence level. The mean, .13, placed the average sentiment again in the mildly positive realm, but the variance was much greater, with a standard deviation of .41. This suggests that sentiment fluctuates more from sentence to sentence within letters than between whole letters. The credible windows for both sentence-level parameters was very narrow, between .13 and .14 for the mean and .41 and .42 for the standard deviation.

Topics

A 21-topic Mallet LDA model was chosen for detailed analysis because it had the highest topic C_V coherence score, as shown in Figure 12, while also approaching the number of topics qualitatively identified by Moreton (2016) for 19th century North American immigrant correspondence. When plotted using the pyLDAvis library, this model produced reasonable intertopic distances, as measured by the Jensen-Shannon divergence method. Table 2 shows the

key terms for the topics produced by the selected model and Figure 13 shows the intertopic distances and the 30 most salient terms.[1]

The Mallet LDA model produced probability distributions across all topics for each letter in the dataset. The topic with the greatest probability value served as the dominant letter topic, the collection-wide distribution of which is expressed in Bayesian terms in Eq. (4). It states that the letter topics follow a multinomial distribution with 21 unordered categories, $topic_1$ through $topic_k$, each of which has its own probability estimate (α) normally distributed around zero with a standard deviation of two, such that the sum of these estimates equals one, or 100 percent. The prior for the standard deviation reflects that this is a logistic model that produces estimates on the log-odds scale. The prior allows the model to explore a wide range of values along this scale, on which -5 maps to near zero probability and 5 maps to almost 100% probability (i.e., $p = 1$).

$$\begin{gathered} letterTopics_i \sim Cat(softmax(topic_{1i}, topic_{2i}, \ldots, topic_{ki}) \\ topic_{1i} = 0 \\ topic_{2i}, \ldots, topic_{ki} = \alpha_{topic_2}, \ldots, \alpha_{topic_k} \\ \alpha_{topic_2}, \ldots, \alpha_{topic_k} \sim Norm(0, 2) \end{gathered} \tag{4}$$

When provided the Mallet LDA topic assignments, the model above yields the probability distribution shown in Figure 14.

Keyword lists were mapped to topic labels by examining the most salient terms in the dynamic pyLDAvis visualization and comparing these to the 24 topics identified by Moreton (2016). A word cluster reflective of the "daily life" topic in Moreton (2016) was primary in 15% of the letters ($n = 138$). This topic was followed by "farming" ($n = 88$) and "news and events" ($n = 94$), each of which were primary in about 10% of the letters. Of these, only "news and events"

[1] Dynamic version at https://nbviewer.jupyter.org/github/menyalas/ThesisPublic/blob/main/MalletLDA21.html

was a topic identified by Moreton (2016). The remaining topics identified by the topic model were primary in 2% to 7% of the letters.

In total, 18 of the 24 topics found by Moreton (2016) were evident in the results of the Mallet LDA model, but the mappings were not straightforward. In some cases, several of the qualitatively identified topics mapped to a single topic from the model. For example, "greetings," "news and events," "health," and "writing process" mapped to a new topic called "correspondence," while "illness," "weather" and "daily life" all mapped to the one carryover item "daily life." Meanwhile, "previous letters," "future letters," "salutations," "signoffs," "homeland," and "enclosures" were not evident in the topic model results.

The 22 sentence-level topics identified by the GSDMM model and shown in Table 3 follow the same distribution described in Eq. (4), but with one additional category.[2] Some topics mirror those found at the letter level, most notably "daily life" (n = 6,739), "family and friends" (n = 5,751) and "transition" (n = 4,913), which respectively account for 19%, 16% and 14% of the sentences. Other crossover topics included "correspondence" (n = 2,132), "farming" (n = 1,018) and "news and events" (n = 3). Although some of the Mallet LDA topics did not carryover directly in the GSDMM results, many are reflected in the sentence-level topics. For example, the letter-level topics "settlement," "rural places" and "pioneering" relate to "homesteading," while "religion" and "education" connect with "community." Meanwhile, some letter topics dissolve into multiple, more nuanced sentence topics: "authority" into "military" and "governance," "essentials" into "provisions" (i.e., nourishment) and "gear" (i.e., objects). Finally, some topics shift meaning—"urban places" becomes "built environment"—and new topics like "suffering" appear at the sentence scale. The letter topic of "recollection" is not clearly reflected at the sentence level, its closest cousin being "observation." Five GSDMM

[2] As with the Mallet LDA model, the single topic assigned to each textual unit, in this case sentence, was the most probable among the full set of topics discovered in the collection.

topics were associated with fewer than 1% of sentences. Among these was “news and events,” which factored much more prominently at the letter level where it was primary in 10% of letters

Do topics vary by temporal, structural or biographical factors?

For ease of interpretation and to ensure the most robust modeling, all numerical predictor variables in the univariate models to follow were standardized by subtracting the mean from each value and dividing by the standard deviation (McElreath, 2017). The result of this data transformation is that results are expressed in terms of standard deviations from the mean. Predictor variables were added to the base models for sentiment and topics using the forms that follow:

$$
\begin{array}{cc}
 & letterTopics_i \sim Cat(softmax(topic_{1i}, topic_{2i}, \ldots, topic_{ki}) \\
 & topic_{1i} = 0 \\
sentimentLetter_i \sim Norm(\mu_i, \sigma) & topic_{2i} = \alpha_2 + \beta_2 x_i \\
\mu_i = \alpha + \beta x_i & \ldots \\
\alpha \sim Norm(0, 3) & topic_{ki} = \alpha_k + \beta_k x_i \\
\beta \sim Norm(0, 3) & \alpha_2, \ldots, \alpha_k \sim Norm(0, 2) \\
\sigma \sim Unif(0, 10) & \beta_2, \ldots, \beta_k \sim Norm(0, .5)
\end{array}
\tag{5}
$$

In the above models, the coefficients for the predictors (β) are assigned normal prior distributions with a mean of zero and a standard deviation of 2 and .5, respectively, for sentiment and topic. The other terms remain the same as in the base models. The estimates extracted from the posterior distribution for each model include values for means, standard deviations and credible intervals for intercepts (α), coefficients (β) and standard errors (σ).

Temporal Variables

Year of Writing. The passage of time is associated with decreased positivity in letters, as shown in Figure 16. For each standard deviation increase in time, equivalent to 21 years, the average letter sentiment score is expected to decrease by .03 units. Given that the estimated sentiment for a letter written in 1865 was .16, it would take 63 years, or until 1928, for the

average sentiment score to dip below .05, the lower bound for positive letters in the VADER system. Therefore, the letters in this dataset remained on average positive for the duration of the long 19th century, which is generally understood to have ended in 1914.

The distribution of topics also changed with time, as shown in Figure 17, which shows the predicted, or simulated, probability of all topics by year of writing, based on the 4,000 samples in the posterior distribution and taking into account estimates for both mu and sigma. The three most common topics illustrate three different patterns: "daily life" grows toward mid-century, then tapers off to about its starting point by the early 1900s; "farming" is high for the first quarter century but then decreases rapidly; "news and events" shows moderate, steady growth for the entire period. The probability of the "settlement" topic shows an almost exponential decline, from more than 40% in the earliest letters to practically zero in the latest letters. Another topic relevant to early migration experiences, "shipboard," becomes less probable over time but in less dramatic fashion. Conversely, the probability of "nursing" and "finances" increase over time, especially after 1865, while the proximate topic of "money," oriented around a more rudimentary vocabulary, is consistently low along with all other topics not highlighted above or in Figure 17.

Time since Immigration. The passage of time relative to the year of migration is associated with increased positivity in letters, as shown in Figure 18. For each standard deviation increase in time, equivalent to 16 years, the average letter sentiment score is expected to increase by .02 units. The data become sparse at the upper limits of the predictor variable, resulting in a wider credible interval and greater uncertainty about the estimate. However, the results indicate that migrants who wrote over many decades showed increasing positivity in their letters, with the letters written within the first 13 years of migration having an estimated sentiment score of .14 compared to .22 for letters written 45 years after their arrival in the new country.

Figure 19 shows the predicted probability of letters according to the number of years elapsed since migration. Again based on 4,000 posterior samples and taking into account model uncertainty by incorporating the estimates for sigma, two topics—"authority" and "pioneers"—are expected to steadily become more likely over time. The case is the inverse, but more mildly so, for three topics that were the most probable soon after migration. In descending order, these are "daily life," "recollection" and "essentials." The model predicts that these topics are less likely to feature in letters the further the migrant is from the year of immigration. The topic of "farming" is more probable in the midrange years. "Transition" becomes slightly more probable and "finances" slightly less probable over time. All other topics showed negligible change relative to years passed since migration.

Structural Variables

Location of Text within Letter. For each additional standard deviation in position, equivalent to 28% of the letter, sentiment is expected to increase by .01 units; however, the credible window reaches zero, indicating that this positive relationship is uncertain. The position of topics within letters similarly shows very limited variation. As shown in Figure 20, only the probability estimates for "correspondence," "daily life" and "family and friends" increase or decrease over the course of a letter. Of these, just one—"family and friends"—changes by more than 5%, with an estimated probability of 13% at letter openings and 20% at closings.

Biographical Variables

Location. The estimated sentiment score for letters written by immigrants in the United States (M = .12) was .08 units lower than letters written by Canadian immigrants (M = .20), as shown in Figure 21. The credible interval on the coefficient for other North American countries spanned zero, meaning that the estimate was unreliable, a result of the sparsity of data for this group. For the same reason, the estimates for topic distribution for this group were uncertain.

Figure 22 shows the predicted probabilities of topics by author location. The highest probability topic among Canadian immigrants was “daily life” (19%) compared to “news and events” (18%) for American immigrants. The two groups differed notably on “recollection,” which had a predicted probability of 14% among Canadian immigrants but just 1% among Americans. “Farming” was relatively probable for both groups, at 11% for Canadians and 8% for Americans. “Nursing” stood out as a topic that was more important for Americans (9%) than Canadians (3%). In general, the group of other North American countries resembled the topic distributions of Canadians more than Americans.

Gender. On average, estimated mean sentiment scores did not differ for male and female writers, as shown in Figure 23, but topics did vary, as shown in Figure 24. “Recollection,” “news and events,” “nursing” and “daily life” were predicted to be more probable among female authors and “farming” much more so among male authors. Of the 21 topics, the model predicted that 12 were more likely to have been written by a man and 9 by a woman.

Age. Letter sentiment was positively associated with age, as shown in Figure 25. For each standard deviation in age, equivalent to 17 years, the average letter sentiment score increased by .02 units. Letters written by migrants aged one standard deviation below the mean, or about 23 years old, had a score of .14 compared to .18 for letters written by migrants aged one standard deviation above the mean, or about 57 years. The data become sparser around this age, meaning that the estimates for older writers are less certain.

As with calendar year, the letter topic of “daily life” shows an arched progression in probability over time, increasing as writers approached middle age and decreasing in later years. “Recollection” and “news and events” become more probable, culminating with 23% and 29% probability at the age of 75. The topic of “mind-body-spirit” becomes more likely with time while the inverse is true for “pioneers,” “nursing,” “education” and “religion,” all registering

changes of 5% or more between childhood and old age. The probability of the "farming" topic is above 10% prior to age 40 and then drops steadily to 4% at 75 years old. These trends are visualized in Figure 26.

National Origin. Figure 27 shows the coefficients and credible intervals for all groups relative to the reference group, the English, for which the average letter sentiment score was .21. The only group expected to have a higher sentiment score than the English was the Irish-Scottish, which had a coefficient of .18, yielding an average sentiment score of .39. While the directionality of the coefficient was inconclusive for a few groups, including the Irish, it was negative for most, meaning that the mean score, relative to the English, was lower. This was true for the Xhosa, Dutch, French-Prussian, German, Italian, Lithuanian, Norwegian, Russian, Scottish and Spanish. However, given the .05 positivity threshold in the VADER system, most of these groups still had overall positive expected scores. The exceptions were the Xhosa, Lithuanians and Russians, whose letters were on average neutral, and the Spanish, for whom the estimated average dipped into the negative realm below -.05.

Figure 28 shows the predicted probability of letter topics by national origin. The greatest correlation is between "nursing" and Italy (56%) as well as "news and events" and Germany (49%). The "pioneer" topic was particularly probable for French-Prussian letters (25%). Three topics—"recollection," "farming" and "daily life"—were relatively probable for all national origins, especially the Scottish (20%) and the Irish-Scottish (31%), but less so for the German and Italian writers. The Scottish were also more likely to write about "settlement" (16%). Education was a topic of interest for the Finnish (15%) and Russians (11%), with the latter also writing more about finances (12%).

Religion. As with national origin, the majority group, in this case the Protestant denomination of Christianity, was used as the base case in the model and for the interpretation of

the results. As shown in Figure 29, the estimated mean sentiment score for Protestant authored letters was .20. All other religious groups had negative coefficients and credible intervals below zero, indicating that they all had lower average sentiment scores than the Protestants. However, the estimates for these groups, including the lowest values in the credible interval, remained above the .05 threshold for positive sentiment on the VADER scale. The ranking of estimated mean scores was as follows: Protestant (.20), Christian (.17), Catholic (.13) and Jewish (.11).

As illustrated in Figure 30, the model predicts that letters from Catholic writers have a relatively high probability of being about "news and events" (34%) and "nursing" (17%). Christian, Protestant and Jewish letters are most likely to be about daily life and farming, but Protestant letters stand out as likely to be about recollection (14%). Additionally, Jewish letters have a relatively high probability of being about finances (9%), money (9%) and modernization (8%).

Occupation. Because migrants in this dataset belonged to multiple groups, occupation was treated as a series of indicator variables rather than as a single factorial variable with mutually exclusive levels. Even after using multiple imputation to predict missing values, the dataset included 10 letters for which the writer's occupational class was a zero value. The model took these "unemployed" migrants as the reference group, which has an estimated mean sentiment score of .07.

Speaking in terms of writers associated with only one occupational class, agricultural workers had the highest estimated mean sentiment score (M = .18). At .11, the professional and commercial classes were likewise expected to have mean scores higher than the unemployed class. The coefficients for all other classes spanned zero, indicating that the model was uncertain as to whether their scores were higher or lower than the base case. Only the industrial class had an estimate likely to position it below the unemployed group and a credible interval whose lower

bound potentially placed it in neutral sentiment territory. Otherwise, sentiment scores for all occupational groups were expected to be above the VADER positivity threshold of .05. Figure 31 visualizes the estimated mean sentiment scores for this variable.

The estimated scores increased with multiple group membership. For the most common combination—agricultural and commercial—the expected score was .22. The next most common combination included these two classes plus the professional and social classes, yielding an estimated average sentiment score of .28, which was the highest of all occupational combinations. Domestic workers had the lowest rate of cross-group membership, with just one combination involving the social class, producing an expected sentiment score of .13.

As with national origin, a couple of topics stand out as having strong relationships with certain occupational class: "news and events" with the domestic class (29%) and "nursing" with the social class (27%). The domestic class also showed a strong association with the "finances" topic (13%). The strongest relationships for the agricultural and commercial classes were with farming, with probability figures of 22% and 22%, respectively. "Pioneers" was central in letters by government and industrial class workers, both groups registering a probability figure of 13% for this topic. The professional class was most likely to write about daily life. Figure 32 illustrates the predicted probability of topics by occupational group.

Do sentiment or topics predict the end of correspondence?

In order to establish whether a relationship existed between the two intended predictor variables in the final model, I ran a linear regression between them, using sentiment as the dependent variable and topic as the independent variable. To address the fact that some writers produced more than one letter, random intercepts were included in this model, with the grouping variable being the ID assigned to each letter writer (n = 218). As shown in Figure 33, the estimated mean sentiment score for all topics was in the positive realm of the VADER scale.

"Recollection," "family and friends" and "farming" were most positive with estimated average sentiment scores of .29, .28 and .22, respectively. Only "nursing," "pioneers" and "finances" had estimated sentiment scores with credible intervals dipping below the -.05 positivity threshold on the VADER scale. Thus, all letter topics in this dataset are, to varying degrees, more positive than negative.

I also ran two random-intercept models, each taking cessation of correspondence as the outcome variable and one of the two narrative variables as the predictor, plus the covariates identified as potentially important in Research Question 2. These included gender, years since migration, year of writing, age at writing, religion, national origin, and membership in the agricultural and industrial classes. Because cases with missing data for the outcome variable are automatically dropped during estimation, only the 751 letters written by one of the 54 repeat authors were used for this model. Also, because sentiment was now a predictor rather than an outcome variable, it was centered around the lowest value (-.40) to facilitate interpretation. The new outcome variable, cessation of correspondence, was operationalized as follows: last letters in series were coded as 1 and all other letters coded as zero. Neither narrative feature nor any of the covariates aside from gender showed a credible fixed effect relative to the reference case, which was a Protestant man from England, average in all respects, writing about daily life or with the lowest sentiment score in the dataset.

For the final analysis, four models were compared, all taking cessation of correspondence as the outcome variable and both sentiment and topic as predictor variables. Information criteria were used to determine which of the models had the least deviance but also the least potential for overfitting the data. As shown in Table 4, the model with random intercepts and slopes as well as gender as a second-level predictor for sentiment had the best WAIC score. It was therefore selected for further analysis. Its specification is shown in Eq. (6).

$$lastLetter_{ik} \sim Bernoulli(p_{ik})$$

$$logit(p_{ik}) = \beta_{0k} + \beta_{1k}Sentiment_i + \beta_{2k}Topic_{1i} + \ldots + \beta_{21k}Topic_{20i}$$

$$\beta_{0k} = \gamma_{00} + \gamma_{01}Female + \eta_{0k}$$
$$\beta_{1k} = \gamma_{10} + \gamma_{11}Female + \eta_{1k}$$

$$\gamma_{00} \sim Norm(0, 2)$$
$$\gamma_{01}, \gamma_{10}, \gamma_{11}, \beta_{2k}, \ldots, \beta_{21k} \sim Norm(0, .5)$$

$$\eta_{0k}, \eta_{1k} = MVNorm(0, \Phi, R)$$
$$\phi_k \sim HalfCauchy(0, 2)$$
$$R \sim LKJ(2)$$

(6)

The model above specifies that unique intercepts be generated for each writer ($\beta_{0\kappa}$) and that these be influenced by gender (i.e., when the author is female rather than male). Additionally, each writer has a unique slope ($\beta_{1\kappa}$) showing the relationship between sentiment and the probability of a letter being the last, again with the female gender being the predictor for this relationship. Each of the letter topics, except for the reference category of "daily life," is included in the model as an indicator variable. These are not assigned random slopes. Neither sentiment nor topic showed a population-level effect, as shown in Table 5. Therefore, this analysis does not support the hypothesis that greater positivity in letters predicts cessation of correspondence.

Gender, however, did make a difference, with men having a higher estimated probability of ending correspondence across the sentiment spectrum, as shown in Figure 34. The same was true for topics: across the board, the estimates showed higher probabilities of men discontinuing correspondence. A similar pattern is evident when the random effects are plotted, as in Figure 35, which shows the predicted probability of correspondence ending for the male and female writers in the dataset across a range of sentiment values. The lines for the men cluster around 20% while those for women cluster below 5% and around 10% to 15%.

Discussion

The broad objective of this thesis was to examine the degree to which scholarly interpretations of migration experiences align with migrants' personal accounts. This study has focused on letters written by immigrants in North America during the long 19th century. The literature review summarized existing research focusing on letters from this time period. Produced by scholars in sociology, history, literary studies and linguistics, this body of work spans academic disciplines and research paradigms. A unifying theme, most notably reported by Thomas and Znaniecki (1918-1920/1927), Erickson (1972) and Gerber (2006), was the desire of 19th century European emigrants to avoid modernizing forces in the old country in favor of an independent, pastoral life in the New World. My contribution has been to examine this idea and the others proposed by migrant letter scholars using Bayesian data analysis, a methodology I believe capable of bridging disciplines and paradigms, in particular by making quantitative, statistical results more widely interpretable within the arts and humanities community. In this thesis, I have modelled sentiment and topics in migrant letters using factors identified as important by previous scholars. This section contextualizes the results for each of my research question and synthesizes findings across questions. It describes unmet expectations and surprising discoveries relative to previous research while also raising new questions for future work.

The results of the topic modelling provide some support for the system described by Moreton (2016). While she identified 24 topics, the Mallet LDA and GSDMM tools used here found the ideal number to be 21 for the whole-letter analysis and 22 for the sentence-level analysis. Six topics from Moreton (2016) were not clearly represented in the results of the whole-letter topic model. Four of these—"previous letters," "future letters," "salutations" and "signoffs"—are related to letter-writing processes and conventions, making it appropriate to

group them with the “correspondence” topic. This leaves two outlying topics from Moreton (2016). To a degree, these fit into the categories produced by the topic model; for example, “homeland” into “recollection” and “enclosures” into “money” or “finances.” Thus, the number of topics as well as many of the labels produced by the letter-level model are in good alignment with the system proposed by Moreton (2016).

However, some differences exist. Subjects like farming, pioneering and rural life are absent from the classification system in Moreton (2016). This probably results from her focus on 35 letters written by one person: Julia Lough, who moved from Ireland to the United States in 1884, at the age of 13 years old, eventually becoming a successful independent dressmaker in Connecticut. The topics identified by Moreton (2016) are specific to a professional person rather than a farmer or pioneer. Also, like many of the Irish who arrived in the United States in the latter half of the 19th century, Julia’s lifestyle was more urban than rural. In contrast, the letters modelled in the present study were written by a diversity of people over a longer period of time, resulting in greater topic variance.

The sentence-level topics varied from the letter-level topics in this study, but they followed a similar pattern relative to Moreton (2016). Again, the topic model produced fewer categories specific to letter-writing practice, but overall a similar number of topics, in this case 22 topics instead of 24. The fact that neither the Mallet LDA nor the GSDMM approach registered the range of correspondence related topics found by Moreton (2016) shows that human and computerized processes are different but potentially complementary. Human readers may better detect nuance within prominent themes, leading to multiple subcategories, whereas computational tools may better detect sparse topics that go unnoticed by human readers. Interestingly, neither of the topic models produced the “reunification” theme, which Richards

(2006) and Gerber (2006) join Moreton (2016) in reporting. It would seem that the automatic processes, or my interpretation of their results, missed this topic.

While topics were mostly locally oriented, as described by Alroey (2011) and Richards (2006), the term "world" drove broader subjects, such as "modernization" and "authority," indicating that migrants wrote about "larger issues of the day," as reported by Chávaz-Garcia (2018) for 20th century migrants (p. 16). Gerber (2006), Brinks (1986), Erickson (1972) and others argued that topics reflected economic concerns and this was born out, with multiple letter- and sentence-level topics defined by key words relating to monetary and financial notions. Finally, the kind of relational concerns, oriented toward social networks in both the old and new countries, which were described by scholars like Gerber (2006), Richards (2006) and Fitzpatrick (1994a), were evident in the topic model results, especially at the sentence scale.

Overall the topics varied by temporal and biographical factors in unsurprising ways. For example, "farming" is prominent up to 40 years of age and in the midrange years after migration, while showing an overall decline during the course of the entire study period. These results make sense given the natural course of life and the industrializing processes in play during the long 19th century. However, the relationships between other topics and the covariates are less intuitive. The decreasing probability of the topic "recollection" over time when measured from the moment of emigration is perplexing because the topic becomes more probable when time is measured by migrant age. This seems to indicate that older people are generally more interested in remembering than younger people but that for all migrants, whatever their age and base level of interest, the activity is more important in the years immediately after leaving home and becomes less so the longer they spend in the new country.

In terms of sentiment, the narratives in this study were positive, both when measured as whole letters and when broken into sentences, but only mildly so, with average scores of .16 and

.13, respectively, on a scale ranging from -1 to 1, where .05 is considered positive. These findings align with previous research describing the tone of migrant correspondence as more positive than negative (Sinnema, 2005; Fitzpatrick, 1994a; Blegen, 1955; Houston and Smyth, 1990).

The univariate model examining the relationship between calendar year and sentiment shows decreased positivity over time, providing partial support for previous research indicating a negative shift in tone around the middle of the 19th century (Conway, 1961; Walaszek, 1992; Miller & Boling, 1990). However, on average, the letters in this study did not become negative, according to the VADER sentiment scale. They merely became less positive.[1] The less severe downturn in sentiment might be attributable to the diversity and character of the letter collection under investigation. Many of the previous scholars reporting the shift focused on immigrants in the United States, specifically Walaszek (1992) for the Polish, Conway (1961) for the Welsh, and Miller and Boling (1990) for the Irish. In a comparison of post-1840 letters written by Irish emigrants in Australia/New Zealand and the United States, Avila-Ledesma (2019) found more positivity in letters sent from the British dominions. Fitzpatrick (1994a) similarly found that Irish-Australian immigrants wrote with consistently positive sentiment. Immigrants in Canada may have likewise maintained a more upbeat tone throughout the 19th century, resulting in a tempered downward shift in tone in the North American dataset used for this study. Furthermore, many of the previous reports about the negative sentiment shift were oriented around Irish-Americans, a group that accounts for a small proportion of the letters used here. Although sentiment became less positive during the 19th century, this trend is not as pronounced as expected, given the attention given to it by previous scholars. It might be that the mid-century

[1] Details about the VADER scoring system are available in Hutto and Gilbert (2014) and at https://github.com/cjhutto/vaderSentiment#about-the-scoring

shift would be more apparent if the relationship between sentiment and calendar year were modelled as nonlinear using quadratic or cubic parameters.

While letters in the dataset become less positive during the 19th century, they become more positive as years passed from the migration event; that is, the longer a migrant was in the destination country, the more positive the sentiment conveyed in their letters. This is an interesting finding, given the background trend toward neutral sentiment. It may provide partial support for the suggestion by Richards (2006) that migration experiences, as described through letters, follow a U-shaped curve defined by more positive emotion in later years. In his analysis of a letter series representing a "classic" experience, the curve begins at a high point in the pre-emigration years and dips after emigration, with the "upturn…achieved only after a long transition" (p. 67). The collection used for this thesis do not include pre-departure letters, meaning that only the long post-departure trough and the final upswing are potentially evident. The steep slopes expected of a U-shape cannot be captured by a linear model like the one used here, but the plotted data points seem to suggest a gentle rise rather than a deep trough. This may be related to a drop-off in correspondence over time, which Fitzpatrick (1994a) proposed may occur when immigrants begin to feel at home in their new setting. That is, the overwhelmingly positive letters do not exist because the migrants stopped writing. It may also be due in part to the overall decreasing sentiment during the 19th century, found in the present study and by Walaszek (1992), Conway (1961), and Miller and Boling (1990). An improved model of letter sentiment would assume a nonlinear shape and include both temporal variables—calendar year and years since immigration—as covariates to control for this potentially cofounding relationship.

The structural variable under study in this thesis is the location of text within letters. This variable was included to examine whether latent features, such as sentiment and topics, reflected

the regular composition described by Thomas and Znaniecki (1918-1920/1927), Schrier (1958) and Fitzpatrick (1994a). Specifically, the question being asked is: Are letter openings and closings different from middle sections, which Fitzpatrick (1994a) described as conversational and intimate. In a related vein, within-letter sentiment patterns might reflect the categories described by Gerber (2006) as regulative, descriptive and expressive, the purpose of the last one being to communicate emotion. This analysis showed no discernible pattern in sentiment at the sentence level. This might be due to the use of the compound VADER sentiment score instead of the finer grained option of extracting the proportion of positive, negative and neutral vocabulary in each sentence. It is also possible that the scores assigned to the words in the VADER lexicon are not appropriate to 19th century English, as suggested by Hamilton et al. (2016), who found that more than 5% of positive or negative words changed polarity between 1850 and 2000. A direction for future research would be to repeat the study for the period 1850 to 1914 and modify the VADER lexicon to reflect the sentiment scores found by Hamilton et al. (2016) for this time period.[2] Finally, a tool that measures emotional states—such as joy, sadness, anger, fear, disgust and surprise—rather than just the polarity and valence of terms might be better suited to examining how sentiment varies by the location of text within letters (Mohammad, 2016).

Topics showed some variation by in-letter positioning, but again the patterns were not as distinct as expected, given the emphasis placed on this subject by migrant letter scholars. The findings here do not justify the omission of openings, closings or any other letter parts prior to quantitative analysis. The weak association between topic and letter structure might be attributable to methodological limitations, especially those related to short text topic modeling (STTM). As described by Wu et al. (2020), various computational tools falling into three classes are currently under development to help overcome the problem of sparsity in short texts. The

[2] Decade-by-decade lexicons for adjectives and high-frequency words are available at https://nlp.stanford.edu/projects/socialsent/

approach used in this thesis was from the Dirichlet multinomial mixture (DMM) class, but it did not include word embeddings, which Wu et al. (2020) argue can improve performance in certain domains. Kozlowski et al. (2019) recently proposed word embeddings as a solution for historical and cultural contexts, where the semantic relationships between words and ideas vary widely over time. Further investigation using a combination of DMM and word embeddings may show that topics vary enough over the course of letters to focus attention on particular sections.

Turning now to biographical variables, letters written by Canadian immigrants were more positive (M = .20) than those written by American immigrants (M = .12). It is necessary here to consider the three prolific writers who collectively produced almost half the letter collection. Two of them—Susannah Moodie and Sarah Stretch Harris—were English emigrants in Canada, whose letters contained above average sentiment of .21, compared to .16 for the entire writing group. Their influence was offset to a degree by the third prolific writer, the German-American Jette Bruns, whose letters carried an average sentiment score of .13. When all three of these writers are removed from the dataset, Canadian letters (M = .19) remain more positive than American letters (M = .12). This result echoes the observation by Avila-Ledesma (2019), who found that letter sentiment varied by country of settlement, with letters by American immigrants being less positive than those by immigrants in Australia and New Zealand. Importantly, the average scores for all countries in the current study, both with and without the prominent writers, was above the .05 threshold marking the boundary between neutral and positive sentiment; thus, the American letters remain on average positive, just less positive than the Canadian letters.

Initial results of the sentiment-gender model did not find a difference between letters written by men and women. This finding would partially support both Gerber (2006), who argued that there was no correlation between gender and emotion, and Benton and Liu (2018), who speculated that qiaopi letters showed an “expressive deficit” because the male writers

usually responsible for them were less likely than women to express intense emotion (p. 164).

Given that the prolific writers in this dataset were all female, it is possible that their patterns of emotional expression biased the results. Indeed, when they are removed from the dataset, a gender difference appears, with men (M = .16) having a mean score twice that of women (M = .08) but also a greater diversity of scores, as measured by interquartile range (IQR) wherein higher values indicate more spread in the middle of the distribution. For men, the IQR was .20 as compared to .12 for women. The letters by the two prolific female writers in Canada have IQR values of .16 and .20, making them resemble or approach the letters written by men. However, these are balanced by the German-American's letters, which had an IQR of just .09.

Adding random intercepts to the sentiment-gender model, such that letters are clustered by author, produces revised estimates similar to those above without excluding data points from the analysis. As shown in Figure 36, this multilevel model produces estimates based on the full letter collection while also showing the scale of variation among the letter series. Expressed as a standard deviation of the group-level (i.e., writer-level) intercepts, the value in this case is .09, meaning that 95% of letter-series sentiment scores fall between -.09 and .27, or a window of four standard deviations equivalent to .36 units on the VADER scale. Furthermore, the estimates for the prolific writers (i.e., the three who produced more than 100 letters each) can be isolated, as shown by the red dots in Figure 36. In these ways, the multilevel model provides more information that supports better interpretation of the results. Future work could extend multilevel modeling to all temporal, structural and biographical independent variables used in this study.

This closer examination of gender shows that the results do not in fact support the idea that men express less intense emotion than women nor that gender and sentiment are unrelated. Rather, the results show that gender is complex and surprising. As for the idea that men are more stoic, the inverse appears to be the case for the letters in this dataset. The initial results showing

no difference between male- and female-authored letters appears to be related to the fact that two of the three most prolific writers, who together produced more than a third of the collection, were women expressing sentiment in ways more typical of the men in the dataset. As for the second point about gender and emotion being uncorrelated, the average female writer in the dataset expressed a very different pattern than her male counterpart. Her sentiments were less positive (but not negative) while also being more centered.

Related to the U-shaped curve theorized by Richards (2008) is the notion that sentiment changes with age, in his view, positively in later years when the migrant emerges from the long period of transition. Conversely, Gerber (2006) suggests that the trend is negative, the result of "emotional and physical decline" and the "shrinking of horizons" that comes with aging (p. 209; p. 114). The present analysis shows a positive correlation between migrant age and sentiment, thus favoring the Richardsonian perspective. This is especially remarkable given the background decline in positive sentiment during the course of the century reported in this study and by historians Walaszek (1992), Conway (1961), and Miller and Boling (1990). While migrants in general became less positive, aging migrants became more positive. When the prolific writers are removed from the analysis, this effect becomes slightly more pronounced ($\beta = .03$).

Seventeen letters written by a migrant whose national origin is given as "Irish; Scottish" carry the most positive sentiment in this study. Joseph Carrothers was an Ulster Protestant who migrated from Northern Ireland to Canada in 1847 to work as a farmer and a tradesperson. His letters date from one to 23 years after migration, spanning his 55th to 77th year of life. In his last letter, written to his brother Willy, he described working steadily through the winter on his horticultural projects: "no day goes better with me than the day I am at work at the bench and I work steady. The people wonder that I work as I do being the age I am" (Carrothers, 1870). Aside from this aged but busy and upbeat Canadian immigrant, and two Czech-American

immigrants, the people who wrote with the most positive sentiment belonged to the largest of the national origin groups: the English. The term "invisible immigrant," as used by Erickson (1972), describes people who are settling into a country where they are culturally similar to the dominant resident group. This was the case for the English arriving in North America, where native populations had been decimated, displaced or subsumed. As explained by Gerber (2006), the integration of culturally similar people, such as the English in Canada and the United States, is expected to be easier, their experiences unrepresentative of most migrants. However, he challenges this "commonsense notion," pointing out that the Scottish, Irish and English letter-writers in his study expressed distrust and negative opinions of each other as well as resident whites (Gerber, 2006, p. 27). The high rank of English emigrant letters on the sentiment scale lends support to the idea that culturally similar groups integrate more smoothly. But how far does "invisibility" extend?

Looking at national origin alone does not reveal a clear pattern. As mentioned, the estimate for Czech letters was higher than for English letters, and the Finnish higher than the Scottish. However, the intervals for many of the estimates limit the potential for definitive statements. Of note, however, is the relatively positive sentiment in the Irish letters, which is somewhat surprising, given the negativity reported by Miller & Boling (1990) and Avila-Ledesma (2019). This may be explained by the fact that only four of the 27 letters produced by Irish emigrants in this dataset were written by individuals in the United States. Furthermore, most letters were written before 1880, potentially before sentiment became less positive in mid- to late-century (Benton and Liu, 2018).

Another component of culture, religion, was key to sentiment. The four national groups with the lowest sentiment scores were the Spanish, Russian, Lithuanian and Xhosa. On close

examination, this last group included just one man who was of European descent.[3] Of the letters produced by individuals in these groups, 22 out of 24 were written by Jews. All the other religions represented in the dataset were Christian, meaning that the Jewish migrants would have been the most different, or the least "invisible." In the topic distribution for this group, "money" and "finance" figured more prominently than for other religions. This result appears to be related to occupation. Most Jewish writers were members of the commercial or agricultural classes, but they were also the most likely to have no assigned occupation, accounting for eight of the 10 individuals treated as "unemployed" in this study. It might be that, as Cancian and Wegge (2016) report, financial instability leads to a greater focus on monetary issues. But this did not manifest, as Cancian and Wegge (2016) suggest, in less attention to other topics. Jewish letters cover a mix of topics comparable to other groups, if not more balanced, and perhaps even more contemplative, as evidenced by a particular focus on "modernization." As expected, the most invisible religious group—Protestants—produced letters with the most positive sentiment. These results accord with Kula et al. (1973/1986), who found that Protestant migrants in America had better experiences than Catholics or Jews, thus supporting the idea that the more culturally similar a migrant is to the resident population, especially in terms of religion, the more easily they adapt to their new setting.

Several migrant letter scholars have treated occupation as an organizing concept in their analyses (Erickson, 1972; Miller et al., 2003; Walaszek, 1992). Indeed, it was a central feature in the ground-breaking work on migrant correspondence by Thomas and Znaniecki (1918-1920/1927) and it may help explain the decline in positive sentiment reported by many previous studies as well as by the current one. According to scholars, 19th century migration to the

[3] Although his cultural assignment is "Dushane; Xhosa," this individual's race is given as white and his point of emigration is a town in the Netherlands. It is possible he is a Boer of Dutch ancestry who fled English rule and/or the Anglo-Boer wars in the Cape Colony in southern Africa, migrating to the United States via the Netherlands.

Americas was driven by crowding, industrialization and a lack of opportunity in Europe. Many migrants dreamed of a self-sufficient pastoral life, something that was widely achievable in North America during the first half of the 19th century. The negative emotion that arose in the latter part of the century was associated with a shift from rural to urban living, from agricultural to industrial work. This pattern can be seen in the results of this study, with agricultural work being associated with the most positive sentiment, by a factor of almost three over the next closest occupational classes: professional and commercial.

Unsurprisingly, industrial work was associated with the least positive sentiment, one of the few cases in this study where the interval for the estimate dips below the VADER positivity threshold. However, a temporal shift from agricultural to industrial work is not apparent. In other words, the occupational classes of the letter writers do not change from agricultural to industrial with the passage of time. The median calendar years for these classes are 1861 and 1864, respectively, with the spread being relatively comparable. But another distinction is clear: More of the letters by industrial workers were written by immigrants in the United States, as opposed to Canada, whereas the inverse is true for letters by agricultural workers. This may reflect the later shift toward industrial work in Canada relative to the United States. According to Conrad (2012), Canadian industrialization occurred between 1885 and 1914, with the "boom years" being 1901 to 1911 (p. 178). By comparison, the United States was undergoing its transformation as early as 1866 and had become "an urban-industrial nation" by 1900 (Boyer, 2012, p. 63). What this might suggest is that the independent, pastoral lifestyle sought by so many Europeans—the "yeoman vision" described by Gerber (2006)—remained within reach in Canada longer than in the United States, resulting in letters expressing more positive sentiment from immigrants north of the border (p. 17).

The interests of migrants, pastoral or otherwise, are evident in the occupational

distribution of topics. As expected, “farming” is most likely among the agricultural and commercial classes, which have the highest degree of cross-membership, as shown in Figure 10. Some topic-occupation associations are less intuitive and upon further investigation trace to uneven representation of writers in the dataset. Specifically, the high probability of the topic “news and events” for the domestic class is attributable to the prolific American immigrant Jette Bruns, who is responsible for 90% of the instances of this topic. This type of effect occurs for lower frequency writers, such as the Catholic frontier missionary Sister Blandina Segale, who wrote 94% of the letters about “nursing,” resulting in its high probability for the social work class. These results highlight the problem of uneven representation in cultural datasets and raise the question of how much the topics produced by the Mallet LDA model, in this context, reflect the interests of individual authors as opposed to corpus-wide topics. It might be worth revisiting the author-topic model for future studies because it explicitly models both author interests and topics. Also, because it is optimized for cases of multiple authorship, it might be helpful in disambiguating between the voices of husband and wife in co-written letters or exploring how sentiment and topics vary with the involvement of an amanuensis.

Neither sentiment nor topic was found to predict the end of correspondence. However, gender did have a notable effect, with women being less likely to stop writing. The hypothesis that positive sentiment is associated with the end of correspondence is not supported; in fact, these findings suggest that the inverse is closer to the truth. That is, as sentiment becomes more positive so too does the probability that correspondence will continue, at least for women, as shown in Figure 34. When examined at the group-level (i.e., by writer), as in Figure 36, this pattern appears to be more relevant to individuals who have a higher base probability of ending correspondence. However, the uncertainty around sentiment and the gender-sentiment interaction is too great to make definitive statements about this potential pattern. This study therefore

accords more with Gerber (2006), who claimed no clear pattern for the discontinuation of correspondence, than with Fitzpatrick (1994a) and others, who have argued that migrants stopped writing when they felt at home in the new country. A time-series model that sequences letters by date might reveal clearer patterns in sentiment or topics leading to cessation of correspondence. Adding random slopes for the topic variable, using gender as a second-level predictor, or including an interaction term between sentiment and topic might also yield more robust findings. Finally, selection bias, as described by Helbich and Kamphoefner (2006), may impact upon the analysis for this question. Specifically, the higher probability that women would continue writing may be tied to the fact that female authored letters were less likely to survive. Thinking in evolutionary terms, being numerous and enduring contributed to the survival of female authored letters, whereas these features mattered less for the survival of male authored letters. Bayesian imputation could potentially be used to model missing female letters, which might make the correlation between sentiment and cessation of correspondence more distinct.

To synthesize, this study found that migrant letters were mildly positive, with sentiment that varied over time and according to the biographical features of the author, but not by the position of text within letters. Topics were mostly practical, locally oriented and geared around relationship maintenance, but some reflected abstract contemplation. Topics varied according to temporal and biographical factors in ways that mostly make intuitive sense. Neither sentiment nor topic predicted cessation of correspondence. Three key themes ran throughout the analysis. First, although many variables showed relationships with sentiment, the key covariate appears to be occupation: Writers in this dataset who were engaged in agricultural work expressed more positive sentiment. Second, cultural invisibility is associated with positive sentiment, with religion mattering more than national origin. Third, dedicated female correspondents, such as Susannah Moodie, Jette Bruns and Sarah Stretch Harris, blur gender distinctions. Although their

voluminous works can introduce bias into cultural datasets, this can be controlled through multilevel modeling, thereby allowing the voices of a prolific few to potentially help speak for an unrepresented many.

Conclusion

This thesis was motivated by a call by the International Migration Organization to listen to migrant stories for the purpose of improving the policy framework within which they move. The stories at the center of this work are told through letters, mostly by people who relocated from Europe to Canada and the United States during the long 19th century, the assumption being that migrants separated by more than 100 years and an array of cultural, social and economic differences share common ground. Scholars from a wide range of academic disciplines have turned to the letters of European settlers to explore, from many angles and at many scales, how humans experience migration.

My contribution has been to approach this narrative material with a quantitative mindset and a computational toolkit featuring Bayesian data analysis. The research questions included: What sentiments and topics are evident in the correspondence? How do they vary by time, letter structure and author traits? Do they predict cessation of correspondence? I hypothesized that positivity in letters would correlate with an increased probability that a letter would be the last. In addition to highlighting the key findings, interpretations, limitations and opportunities, this concluding chapter will revisit some of the overarching issues raised in the introduction: namely, how well are migrants' own words reflected in scholarly research and how might the methods demonstrated in this thesis be used to achieve greater equity in quantitative cultural research.

Summary of Findings

Sentiment was found to be mildly positive, with a subtle decline over the 19th century but increases with age and years elapsed since migration. Positivity in letters was associated with male, Protestant and English authorship and membership in the agricultural class. Letters from immigrants in the United States and in the industrial class were less positive. Topics were mostly oriented around practical matters, relationship maintenance and local affairs, but some suggested

contemplation about broader, more abstract subjects. The distribution of letter topics varied most notably by gender and country: Men wrote more about farming and women about recollection, with American immigrants writing less about both topics, which were among the three highest scoring for sentiment. Neither topic nor sentiment predicted cessation of correspondence, offering no support for the hypothesis. In fact, a hint of the opposite pattern was detected, at least for women, who were more likely to continue writing when letter sentiment was more positive. However, women were more likely, in general, to maintain correspondence, irrespective of sentiment and topics.

Industrialization, Invisibility and the Female Experience

While gender was the variable showing the strongest effects, it may obscure the role of other factors, particularly occupation, which scholars have identified as central to 19th century migrant experiences and letters. This study found the topic of "farming" to vary with gender, occupation and sentiment as well as with the passage of time and the country of settlement in ways that fit nicely with the one of the most prominent themes in migrant letter scholarship. That is, 19th century migrants were motivated by a desire to lead an independent, pastoral lifestyle, leading to a negative shift in sentiment when population growth and industrializing processes made this dream unattainable.

Indeed, this study showed sentiment was highest when farming was the topic or agriculture the occupation, conditions that were more prevalent among the male writers in the dataset as well as those living outside of the United States, where the proportion of writers in the agricultural class was just half that in Canada. Individuals involved in industrial work, who were more likely to be men in the United States, had the least positive sentiment. Thus, it seems that occupation, and the changes that came with industrialization, may explain sentiment in migrant letters better than the gender or the physical location of the writer.

While the more positive tone of the Canadian immigrants may be attributable to their greater association with agricultural work, it might also stem from the fact that they were overall less likely to be part of a visible minority group. By comparison, the letters from immigrants in the United States represent a broader spectrum of cultural backgrounds than those from Canada, whose writers were mostly Protestants originating in England or its European dominions. In particular, the proportion of letters by Jewish migrants is six times greater in the U.S. collection than in the Canadian collection. Regardless of the writers' national origins, which varied widely, the Jewish letters were the least positive. This seems to suggest that of the variables used in this study to operationalize culture—that is, religion and national origin—the former is more important in determining sentiment.

The traditional occupation of 19th century women—that is, work oriented around home and family—may have contributed to the more centered sentiment scores detected in their letters. That is, external occupations may be related to a wider range of sentiments, such as those observed in the male authored letters. The less positive tone expressed in letters by typical (i.e., nonprolific) female writers raises more questions. Does it stem from occupational dissatisfaction, a lack of agency or some other factor? The fact that two of the three prolific female writers—Susannah Moodie and Sarah Stretch Harris—had sentiment profiles similar to the average man is intriguing. Nothing in the religion or national origin of these women sets them apart from their female peers, but occupation could be a factor. Moodie was the prolific female writer with the highest average sentiment score and, as a published author of novels, children's books, poetry and memoirs, the only one with a profession of outside of home and family. A critical reading of the female-written letters in the dataset could help explain how the relationships, work, interests, perspectives, social lives or simply the practice of correspondence may have contributed to the unusual sentiment profiles of Moodie and Harris while illuminating why the third prolific writer,

Jette Bruns, as well as the typical females in the dataset expressed less positivity and less variability in sentiment.

Limitations and Opportunities

While the findings of this study are generally in alignment with the migrant letter story told by qualitative and critical scholars, a few expectations based on their work were unmet. This might be because variables were not effectively operationalized. For example, occupation may have been confounded by gender, and national origin and religion may not have adequately captured the "invisibility" effect, which is complex and location specific. Future research might use a nonbinary measure of gender or a qualitatively constructed variable for culture that better captures race and ethnicity, including religion, language and social network factors. These might take into account whether the migrant traveled alone or in a group, such as under an assisted migration scheme, and whether they joined an established expat or diasporic community. The relationship between narrative features and other migration outcomes, such as permanent settlement, occupational class change or out-of-group marriage (i.e., exogamy), could also be examined in future studies. Finally, sample imbalance, associated with the prolific writers or the underrepresentation of infrequent female writers, could be addressed by linking with other datasets (e.g., the Irish Emigration Database) or a more extensive application of multilevel modeling, which could incorporate Bayesian imputation of missing data rather than the multiple imputation approach used in this study.

Migrants, Scholars and Equity in Cultural Datasets

Returning to the word frequency clouds presented in the introduction, which were generated from the titles of migration related articles in academic journals, it would seem that the most central themes in the scholarship reflect concerns voiced by historical migrants in their letters. In particular, family, community and work are subjects reflective of the letters in this

study. Humanities journals might overemphasize religion. Only a couple of topics in this study included religious terms, which were nested in keyword lists oriented around community rather than spirituality, doctrine or the practice of faith. However, religion, like gender, was found to be a powerful social construct potentially related to sentiment, meaning that it is a valuable subject of scholarly inquiry even if migrants did not expound on it explicitly in their letters.

To examine whether scholarly interpretations reflect migrants' experiences, as communicated through personal correspondence, this study took a quantitative approach, grounded in the pragmatic research paradigm, operating within the framework of exploratory data analysis and employing the tools of natural language processing and Bayesian linear modeling. This study attempted to demonstrate how such a conceptual and methodological framework might be used in a way that is transparent and proactive about the problems of representativeness in cultural datasets. It also produced quantitative results framed in straightforward mathematical terms that I hope will be useful and entirely accessible for criticism within an interdisciplinary community of scholars of all methodological persuasions. In the same way that "all models are wrong," the results of this study are an imperfect approximation of a complex subject based on incomplete data (Box, 1976). They are an effort to listen not only to migrants whose letters persisted through the years but also, through them, to discern the rough outlines of the many more unlettered experiences, past and present.

A Final Act of Researcher Reflexivity

In the interest of transparency and for the purpose of identifying potential researcher-side bias, it should be noted that the migrant-author of this thesis entertains her own pastoral dreams and bears a striking resemblance, in name but also occupational pursuits, to one of the prolific writers mentioned in this study. No known shared lineage exists with Susannah Moodie, but she does have a new fan.

References

Alroey, G. (2011). *Bread to eat and clothes to wear: Letters from Jewish migrants in the early twentieth century*. Wayne State University Press.

Alvstad, C. (2013). The Transatlantic Voyage as a Translational Process: What Migrant Letters Can Tell Us. In M. Boyden, H. Krabbendam, & L. Vandenbussche (Eds.), *Tales of transit: Narrative migrant spaces in Atlantic perspective, 1850-1950* (pp. 103-120). Amsterdam University Press. http://hdl.handle.net/1854/LU-4193170

Avila-Ledesma, N. E. (2019). "Believe My Word Dear Father that You Can't Pick Up Money Here as Quick as the People at Home Thinks It": Exploring Migration Experiences in Irish Emigrants' Letters. *Corpus Pragmatics, 3*, 101-121. https://doi.org/10.1007/s41701-018-00051-8

Bamman, D., Underwood, T., & Smith, N. A. (2014). A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 370-379). Association for Computational Linguistics. 10.3115/v1/P14-1035

Barton, H. A. (1975). *Letters from the promised land: Swedes in America, 1840–1914*. University of Minnesota Press.

Benton, G., & Liu, H. (2018). *Dear China: Emigrant letters and remittances, 1820-1980*. University of California Press. https://doi.org/10.1525/california/9780520298415.001.0001

Blegen, T. (1931). *Norwegian migration to America, 1825-1860.* The Norwegian-American Historical Association. https://hdl.handle.net/2027/mdp.39015046452465

Blegen, T. (Ed.). (1955). *Land of their choice: The immigrants write home*. University of Minnesota Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(4/5), 993–1022. https://dl.acm.org/doi/10.5555/944919.944937

Blumer, H. (1969). *Symbolic interactionism: Perspective and method.* Prentice-Hall.

Blumer, H., & Bain, R. (1939). *An appraisal of Thomas and Znaniecki's The Polish Peasant in Europe and America*. Social Science Research Council. https://hdl.handle.net/2027/mdp.39015026698004

Borges, M., & Cancian, S. (2016). Reconsidering the migrant letter: From the experience of migrants to the language of migrants. *The History of the Family, 21*(3), 281-290. https://doi.org/10.1080/1081602X.2016.1222502

Boyer, P. S. (2012). *American history: A very short introduction*. Oxford University Press. https://doi.org/10.1093/actrade/9780195389142.001.0001

Box, George, E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Brinks, H. (1986). *Write back soon: Letters from immigrants in America.* CRC Publications.

Burke, E. (2000, May 19-21). *Modernity's Histories: Rethinking the Long Nineteenth Century, 1750-1950* [Conference presentation]. University of California World History Workshop, Davis, CA, United States. https://escholarship.org/uc/item/2k62f464

Bürkner P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software. 80*(1), 1-28. https://doi.org/10.18637/jss.v080.i01

Cameron, W., Haines, S., & Maude, M. (2000). *English immigrant voices: Labourers' letters from upper Canada in the 1830s*. McGill-Queen's University Press.

Cancian, S., & Wegge, S. A. (2016). 'If it is not too expensive, then you can send me sugar': money matters among migrants and their families. *The History of the Family, 21*(3), 350–367. https://doi.org/10.1080/1081602X.2016.1147372

Carrothers, J. (1870, March 8). [Letter to Willy]. North American Immigrant Letters, Diaries and Oral Histories (Document ID S9635-D032), ProQuest, Alexandria, VA, USA.

Chávez-García, M. (2018). Introduction: An archive of intimacy. In *Migrant longing: Letter writing across the U.S.-Mexico borderlands*. University of North Carolina Press. https://doi.org/10.5149/northcarolina/9781469641034.003.0001

Chen, E. (2011, November 22). Introduction to Latent Dirichlet Allocation. *Edwin Chen*. http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering, 26*(12), 2928–2941. https://doi.org/10.1109/TKDE.2014.2313872

Chilton, L. (2016). Letters "home" from Canada: British female emigrants and the imperial family of women. In F. Iacovetta, & M. Epp (Eds.), *Sisters or strangers? Immigrant, ethnic, and racialized women in Canadian history* [Second Edition] (pp. 153-171). University of Toronto Press.

Conrad, M. (2012). Making Progress, 1885–1914. In *A Concise History of Canada* (pp. 164-193). Cambridge University Press. https://doi.org/10.1017/CBO9781139032407.008

Conway, A. (1961). *The Welsh in America: Letters from the immigrants*. University of Minnesota Press.

D'Ignazio, C., & Klein, L. (2018). Chapter seven: The power chapter. In *Data feminism*. MIT Press Open. https://bookbook.pubpub.org/pub/7ruegkt6

De Felice, R., & Moreton, E. (2019). Identifying speech acts in a corpus of historical migrant correspondence. *Studia Neophilologica, 91*(2), 154-174. https://doi.org/10.1080/00393274.2019.1616216

Ellis, L., Hoskin, A. W., Ratnasingam, M. (2018). Conceptualizing and Measuring Social Status. In L. Ellis, A. W. Hoskin & M. Ratnasingam (Eds.), *Handbook of Social Status Correlates* (pp. 1-14). Academic Press. https://doi.org/10.1016/B978-0-12-805371-3.00001-7.

Erickson, C. (1972). *Invisible immigrants: The adaptation of English and Scottish immigrants in nineteenth-century America.* University of Miami Press.

Felman, S., & Laub, D. (1992). *Testimony: Crises of witnessing in literature, psychoanalysis, and history*. Routledge. https://doi.org/10.4324/9780203700327

Fender, S. (1992). *Sea changes: British emigration and American literature*. Cambridge University Press.

First Peoples' House (n.d.). *Traditional Territory*. McGill University. Retrieved April 14, 2020, from https://www.mcgill.ca/fph/welcome/traditional-territory

Fisher, M. (2013). *Migration: A world history*. Oxford University Press.

Fitzgerald, P. (2008). Exploring transnational and diasporic families through the Irish Emigration Database. *Journal of Intercultural Studies*, *29*(3), 267-281. https://doi.org/10.1080/07256860802169204

Fitzpatrick, D. (1994a). Emigrant letters: I take up my pen to write these few lines. *History Ireland, 2*(4), 15-19.

Fitzpatrick, D. (1994b). *Oceans of consolation: Personal accounts of Irish migration to Australia*. Cornell University Press.

Fitzpatrick, D. (2006). Irish emigration and the art of letter-writing. In B. S. Elliott, D. A. Gerber, & S. M. Sinke (Eds.), *Letters across borders* (pp. 97-106). Palgrave Macmillan.

Ganzevoort, H. (Ed.). (1999). *The last illusion: Letters from Dutch immigrants in the land of opportunity, 1924-1930*. University of Calgary Press.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review, 71*(2), 369–382. https://doi.org/10.1111/j.1751-5823.2003.tb00203.x

Gelman, A. (2011). Bayesian statistical pragmatism. *Statistical Science, 26*(1), 10–11.

Gerber, D. A. (1997). The immigrant letter between positivism and populism: The uses of immigrant personal correspondence in twentieth-century American scholarship. *Journal of American Ethnic History, 16*(4), 3-34. https://www.jstor.org/stable/27502216

Gerber, D. A. (2006). *Authors of their lives: The personal correspondence of British immigrants to North America in the nineteenth century*. NYU Press.

Gerber, D. A. (2006a). Epistolary masquerades: Acts of deceiving and withholding in immigrant letters. In B. S. Elliott, D. A. Gerber, & S. M. Sinke (Eds.), *Letters across borders* (pp. 141-157). Palgrave Macmillan.

Greenwood, M. (2019). The migration legacy of E. G. Ravenstein. *Migration Studies, 7*(2), 269-278. https://doi.org/10.1093/migration/mny043

Hale, F. (1986). *Their own saga: Letters from the Norwegian global migration*. Minnesota Press.

Hamilton, W. L, Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 595–605). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1057

Harte, N. (2008, July 16). Professor Charlotte Erickson: Meticulous historian of migration. *The Independent*. https://www.independent.co.uk/news/obituaries/professor-charlotte-erickson-meticulous-historian-of-migration-868657.html

Helbich W., & Kamphoefner W. D. (2006) How Representative are emigrant letters? An exploration of the German case. In B. S. Elliott, D. A. Gerber, & S. M. Sinke (Eds.), *Letters across borders* (pp. 29-55). Palgrave Macmillan.

Held, M. B. E. (2019). Decolonizing Research Paradigms in the Context of Settler Colonialism: An Unsettling, Mutual, and Collaborative Effort. *International Journal of Qualitative Methods, 18*, 1-16. https://doi.org/10.1177/1609406918821574

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

Hoerder, D. (1992). German immigrant workers' views of "America" in the 1880s. In M. Debouzy (Ed.), *In the shadow of the Statue of Liberty: Immigrants, workers, and citizens in the American republic, 1880-1920* (pp. 5-22). University of Illinois Press.

Houston, C., & Smyth, W. (1990). Irish emigration and Canadian settlement: Patterns, links, and letters. University of Toronto Press.

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14)* (pp. 216-225). AAAI Press https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109

International Organization for Migration (2015). *IOM outlook on migration, environment and climate change*. https://doi.org/10.18356/9ba951ac-en.

International Organization for Migration (2017). *World Migration Report 2018*. https://doi.org/10.18356/f45862f3-en.

International Organization for Migration (2019). *World Migration Report 2020*. https://doi.org/10.18356/b1710e30-en.

Irick, R. L. (1982). *Ch'ing policy toward the coolie trade, 1847-1878* (Asian library series, no. 18). Chinese Materials Center.

Irish Emigration Database [Online archive]. Mellon Centre for Migration Studies. http://www.dippam.ac.uk/ied/

Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. Poetics, 41(6), 750–769. https://doi.org/10.1016/j.poetic.2013.08.005

Kass, R. E. (2011). Statistical inference: The big picture. *Statistical Science, 26*(1), 1–9. https://doi.org/10.1214/10-STS337

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: analyzing the meanings of class through word embeddings. *American Sociological Review, 84*(5), 905–949. https://doi.org/10.1177/0003122419877135

Kula, W., Assorodobraj-Kula, N., Kula, M., & Wtulich, J. (1986). *Writing home: Immigrants in Brazil and the United States, 1890-1891*. East European Monographs. (Original work published 1973)

Lucassen, J., Lucassen, L., & Manning, P. (2010). *Migration history in world history: Multidisciplinary approaches*. Brill.

Manning, P., & Trimmer, T. (2013). *Migration in world history* (2nd edition). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781351256681

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press/Taylor & Francis Group.

Mertens, D. (2010). Transformative mixed methods research. *Qualitative Inquiry, 16*(6), 469–474.

Miller, K. (1985). *Emigrants and exiles: Ireland and the Irish exodus to North America*. Oxford University Press.

Miller, K. A., & Boling, B. D. (1990). Golden streets, bitter tears: The Irish image of America during the era of mass migration. *Journal of American Ethnic History, 10*(1-2), 16–35. https://www.jstor.org/stable/27500798

Miller, K., Schrier, A., Boling, B. D., & Doyle, D. N. (2003). *Irish immigrants in the land of Canaan: Letters and memoirs from colonial and revolutionary America, 1675-1815*. Oxford University Press.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion Measurement* (pp. 201-237). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100508-8.00009-6

Moreton, E. (2012). Profiling the female emigrant: A method of linguistic inquiry for examining correspondence collections. *Gender & History, 24*(3), 617-646. https://doi.org/10.1111/j.1468-0424.2012.01699.x

Moreton, E. (2016) 'I never could forget my darling mother': The language of recollection in a corpus of female Irish emigrant correspondence. *The History of the Family, 21*(3), 315-336. https://doi.org/10.1080/1081602X.2016.1155469

Morgan, D. L. (2014). Pragmatism as a paradigm for social research. *Qualitative Inquiry, 20*(8), 1045–1053. https://doi.org/10.1177/1077800413513733

Mortari, L. (2015). Reflectivity in research practice: An overview of different perspectives. *International Journal of Qualitative Methods, 14*(5). https://doi.org/10.1177/1609406915618045

Mortensen, O. (2017). The Author-Topic Model [White Paper]. Technical University of Denmark. http://www2.imm.dtu.dk/pubdb/edoc/imm6971.pdf

Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. Addison-Wesley.

Mulder, W. (1954). Through immigrant eyes: Utah history at the grass roots. *Utah Historical Quarterly, 22*(1), 41-55.

Perelman, L. (1991). The medieval art of letter writing: Rhetoric as institutional expression. In C. Bazerman and J. Paradis (Eds.), *Textual Dynamics of the Professions: Historical and Contemporary Studies of Writing in Professional Communities* (pp. 97-119). University of Wisconsin Press. https://wac.colostate.edu/docs/books/textual_dynamics/chapter4.pdf

Plummer, K. (1983). *Documents of life: An introduction to the problems and literature of a humanistic method*. G. Allen & Unwin.

Plummer, K. (2004). Symbolic interactionism. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods* (Vol. 1, pp. 1105-1105). SAGE Publications, Inc. https://doi.org/10.4135/9781412950589.n992

Porter, A. (Ed.). (1999). *The Oxford history of the British empire, volume III: The nineteenth century*. Oxford University Press.

Ravenstein, E. (1889). The laws of migration. *Journal of the Royal Statistical Society, 52*(2), 241-305. https://doi.org/10.2307/2979333

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (pp. 46-50). University of Malta. http://is.muni.cz/publication/884893/en

Richards E. (2006) The Limits of the Australian Emigrant Letter. In B. S. Elliott, D. A. Gerber, & S. M. Sinke (Eds.), *Letters across borders* (pp. 56-74). Palgrave Macmillan. https://doi.org/10.1057/9780230601079_3

Risam, R. (2018). *New digital worlds: Postcolonial digital humanities in theory, praxis, and pedagogy*. Northwestern University Press.

Rockwell, G., & Sinclair Stéfan. (2016). *Hermeneutica: computer-assisted interpretation in the humanities.* MIT Press. https://doi.org/10.7551/mitpress/ 9780262034357.001.0001

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 487–494). AUAI Press. https://doi.org/10.5555/1036843

Schrier, A. (1958). *Ireland and the American emigration, 1850-1900*. University of Minnesota Press.

Seganti, F. R. (2010). Practicing reflexivity in the study of Italian migrants in London. *The Qualitative Report, 15*(4), 966-987. http://www.nova.edu/ssss/QR/QR15-4/seganti.pdf

Sievert, C., & Shirley, K. E. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, (pp. 63–70). Association for Computational Linguistics. https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf

Sinnema, D. (Ed.). (2005). *The first Dutch settlement in Alberta: Letters from the pioneer years, 1903-14*. University of Calgary Press.

Smith, S., & Watson, J. (2010). *Reading autobiography: A guide for interpreting life narratives.* University of Minnesota Press.

Stanley, L. (2004). The epistolarium: On theorizing letters and correspondences. *Auto/Biography, 12*(3), 201–235. https://doi.org/10.1191/0967550704ab014oa

Stanley, L. (2010). To the letter: Thomas and Znaniecki's *The Polish Peasant* and writing a life, sociologically. *Life Writing, 7*(2), 139–151. https://doi.org/10.1080/14484520903445271

Stanley, L. (2015, June 9). Digital humanities institute, NHC. *Whites Writing Whiteness: Letters, Domestic Figurations and Representations of Whiteness in South Africa 1770s-1970s*. https://www.whiteswritingwhiteness.ed.ac.uk/blog/digital-humanities-institute-nhc/

Stanley, L., & Jolly, M. (2017). Epistolarity: Life after death of the letter? *Auto/Biography Studies: A/b, 32*(2), 229–233. https://doi.org/10.1080/08989575.2016.1187040

Thomas, E. (1978). Herbert Blumer's critique of *The Polish Peasant*: A post mortem on the life history approach in sociology. *Journal of the History of the Behavioral Sciences*, *14*(2), 124-131. https://doi.org/10.1002/1520-6696(197804)14:2<124::AID-JHBS2300140205>3.0.CO;2-2

Thomas, W. I., & Znaniecki, F. (1927). *The Polish peasant in Europe and America* (2nd ed., Vols. 1–2). Alfred A. Knopf. https://hdl.handle.net/2027/uc1.b3115547 and https://hdl.handle.net/2027/uc1.b3427821 (Original work published 1920)

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Underwood, T. (2012, January 3). A brief outburst about numbers. *The Stone and the Shell*. https://tedunderwood.com/2012/01/03/a-brief-outburst-about-numbers/

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 3*(45), 1-67. https://dx.doi.org/10.18637/jss.v045.i03

Walaszek, A. (1992). In America Poles work like cattle: Polish Peasant Immigrants and work in America, 1880-1921. In M. Debouzy (Ed.), *In the shadow of the Statue of Liberty: Immigrants, workers, and citizens in the American republic, 1880-1920* (pp. 83-94). University of Illinois Press.

Weingart, S. (2012, January 10). Doing Bayesian data analysis. the scottbott irregular. https://scottbot.net/doing-bayesian-data-analysis/

Wiesel, E. (1977). The Holocaust as a literary inspiration. In *Dimensions of the Holocaust: Lectures at Northwestern University* (pp. 5-19). Northwestern University Press.

Wu, X., Yuan, Y., Li, Y., Qian, Z., & Qiang, J. (2020). Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge & Data Engineering, 1*(1), 1–1. https://doi.org/10.1109/TKDE.2020.2992485

Yin, J., & Wang, J. (2014). A Dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 233-242). Association for Computing Machinery. https://doi.org/10.1145/2623330.2623715

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin.

Table 1: Occupation Class Mappings

Job	Erickson (1972)	Miller (2003)
Accountant	Commercial, Clerical, Professional	Professional
Architect	Commercial, Clerical, Professional	Professional
Artist[a]	Commercial, Clerical, Professional	Commercial
Banker	Commercial, Clerical, Professional	Professional
Businessman	Commercial, Clerical, Professional	Commercial
Clergy	Commercial, Clerical, Professional	Social
Cook	Commercial, Clerical, Professional	Domestic
Diplomat	Commercial, Clerical, Professional	Government
Editor	Commercial, Clerical, Professional	Professional
Educator	Commercial, Clerical, Professional	Social
Engineer	Commercial, Clerical, Professional	Professional
Explorer[b]	Industrial	Commercial
Factory worker	Industrial	Industrial
Farmer	Agricultural	Agricultural
Government appointee	Commercial, Clerical, Professional	Government
Homemaker	Other	Domestic
Housekeeper	Commercial, Clerical, Professional	Domestic
Jeweler	Commercial, Clerical, Professional	Commercial
Laborer[c]	Agricultural	Agricultural
Manufacturer	Industrial	Industrial
Merchant	Commercial, Clerical, Professional	Commercial
Military	Commercial, Clerical, Professional	Government
Miner	Industrial	Industrial
Missionary	Commercial, Clerical, Professional	Social
Nun	Commercial, Clerical, Professional	Social
Nurse	Commercial, Clerical, Professional	Social
Physician	Commercial, Clerical, Professional	Social
Plantation manager	Agricultural	Agricultural
Politician	Commercial, Clerical, Professional	Government
Rancher	Agricultural	Agricultural
Religious leader	Commercial, Clerical, Professional	Social
Religious worker	Commercial, Clerical, Professional	Social
Retail worker	Commercial, Clerical, Professional	Commercial
Royal governor	Commercial, Clerical, Professional	Government
Secretary	Commercial, Clerical, Professional	Professional
Servant	Commercial, Clerical, Professional	Domestic
Social worker	Commercial, Clerical, Professional	Social
Student	Other	Social
Surveyor	Commercial, Clerical, Professional	Professional
Tailor	Commercial, Clerical, Professional	Commercial
Teacher	Commercial, Clerical, Professional	Social
Tradesman	Commercial, Clerical, Professional	Commercial
Writer	Commercial, Clerical, Professional	Professional

[a] *Six out of the seven artists cross-listed as architects or business people, thus classified as commercial.*
[b] *These individuals were cross-listed as miners, thus likely prospectors who would sell their discoveries. This job matches the peddler occupation included in the commercial class proposed by Miller et al. (2003).*
[c] *Five out of six individuals categorized only as laborers wrote from Hawaii, where the main 19th century industries were whaling, sandlewood and sugar, which is why they are categorized as agricultural rather than industrial workers.*

Table 2: Letter Topics from Mallet LDA Model

No.	Topic	Key Words	N	%
1	Shipboard	day, morning, ship, hour, sea, night, time, wind, water, evening	38	0.04
2	Urban Places	city, house, town, building, street, place, shop, railroad, work, business	30	0.03
3	Nursing	sister, man, room, hospital, mother, place, patient, door, train, number	52	0.06
4	Correspondence	letter, brother, time, wife, health, news, mail, thing, child, today	54	0.06
5	Education	child, school, mother, boy, time, teacher, parent, music, picture, father	43	0.05
6	Recollection	book, paper, work, dear, copy, friend, world, memory, kind, term	66	0.07
7	Transition	time, account, month, manner, situation, nature, difficulty, matter, interest, change	21	0.02
8	Settlement	place, land, country, settlement, town, year, settler, river, family, emigrant	31	0.03
9	Authority	man, government, day, soldier, war, member, house, company, officer, expense	18	0.02
10	Family/friends	family, friend, girl, woman, daughter, death, husband, life, year, hand	25	0.03
11	Religion	church, people, place, person, year, immigrant, case, congregation, time, number	30	0.03
12	Daily Life	week, day, weather, boy, baby, morning, doctor, night, today, town	138	0.15
13	Rural Places	foot, gold, time, snow, road, water, place, man, side, tree	14	0.02
14	Essentials	body, world, food, man, business, water, ship, air, blood, life	14	0.02
15	Farming	land, country, acre, farm, year, farmer, wheat, crop, price, kind	88	0.10
16	Mind-body-spirit	day, eye, hand, heart, room, life, mind, head, night, hour	30	0.03
17	Pioneers	company, fire, day, horse, wagon, animal, water, camp, wood, night	30	0.03
18	News/events	thing, time, people, year, lot, house, uncle, trip, week, brother	94	0.10
19	Finances	year, wife, money, time, care, interest, husband, answer, debt, business	37	0.04
20	Modernization	people, state, country, law, year, hand, world, property, idea, opinion	29	0.03
21	Money	work, dollar, month, day, money, cent, time, week, pay, cost	33	0.04
			915	1.00

Table 3: Sentence Topics from GSDMM

No.	Topic	Key Words	N	%
1	Homesteading	land, year, country, settle, acre, farm, state, time, family, pay	1346	0.04
2	Appearances	wear, dress, look, shirt, send, pair, color, head, skin, hair	256	0.01
3	Military	company, move, throw, machine, cannon, hand, fire, supply, arm, cartridge	77	0.00
4	Correspondence	letter, write, send, receive, time, arrive, day, hear, week, answer	2132	0.06
5	Governance	take, state, law, people, election, man, return, member, day, officer	859	0.02
6	Transactions	money, pay, dollar, work, take, want, year, give, sell, time	2662	0.08
7	Gear	bring, table, tool, knife, clothe, dress, chair, chest, work, bed	140	0.00
8	Transition	country, give, people, year, work, time, write, life, letter, place	4913	0.14
9	Employment	work, dollar, day, pay, month, wage, week, man, gold, find	846	0.02
10	Living Body	body, blood, food, air, lung, tube, call, nerve, matter, pass	105	0.00
11	Currency	dollar, piece, coin, dime, people, franc, nation, cent, size, quarter	20	0.00
12	Daily Life	day, take, come, time, sister, go, week, room, school, place	6739	0.19
13	Built Environment	house, build, room, foot, place, tree, brick, roof, building, street	493	0.01
14	Community	church, school, people, preach, place, attend, congregation, day, take, child	1101	0.03
15	Observation	eye, look, see, hand, take, man, face, horse, turn, walk	916	0.03
16	News/events	draw, spirit, destroy, wealth, health, chat, do, treaty, lumber, trade	3	0.00
17	Provisions	pound, cent, bread, tea, water, day, meat, bring, dollar, flour	669	0.02
18	Farming	acre, land, year, crop, wheat, tree, grow, plant, farm, potato	1018	0.03
19	Suffering	feel, time, heart, day, suffer, child, pain, year, fever, life	1966	0.06
20	Family/friends	child, write, sister, time, give, brother, send, letter, year, family	5751	0.16
21	Arrival	boat, steam, place, ship, take, town, building, city, people, inhabitant	324	0.01
22	Transit	day, water, foot, snow, see, ship, fall, time, night, come	2680	.08
			35016	1.0

Table 4: Cessation of Correspondence Model Comparison

ID	Model Elements	WAIC
oneLevel	Sentiment + Topic	361.88
randomIntecept	Sentiment + Topic + (1 | Writer)	335.60
randomSlope	Sentiment + Topic + (1 + Sentiment | Writer)	335.54
randomSlopePredictor	**Sentiment*Gender + Topic + (1 + Sentiment | Writer)**	**331.31**

Bold indicates the model selected for interpretation.

Table 5: Fixed Effects from the Multilevel Regression Model for Cessation of Correspondence

	Estimate	Est.Error	Q5	Q95
Intercept	-1.627	0.382	-2.257	-0.999
Sentiment	-0.020	0.467	-0.803	0.751
Female	-0.721	0.405	-1.386	-0.048
Urban Places	-0.207	0.428	-0.917	0.493
Nursing	-0.358	0.468	-1.134	0.410
Correspondence	0.145	0.386	-0.483	0.778
Education	0.293	0.418	-0.409	0.975
Recollection	-0.116	0.453	-0.856	0.623
Transition	0.177	0.455	-0.577	0.902
Settlement	-0.325	0.435	-1.058	0.377
Authority	0.195	0.452	-0.562	0.930
Family/Friends	-0.132	0.435	-0.856	0.578
Religion	0.614	0.421	-0.087	1.284
Shipboard	-0.213	0.425	-0.919	0.492
Rural Places	-0.103	0.475	-0.889	0.665
Essentials	0.379	0.444	-0.352	1.100
Farming	0.425	0.344	-0.152	0.985
Mind-Body-Spirit	0.028	0.458	-0.733	0.768
Pioneers	-0.073	0.436	-0.797	0.640
News/Events	-0.412	0.443	-1.155	0.310
Finances	0.017	0.464	-0.744	0.781
Modernization	0.204	0.434	-0.523	0.925
Money	0.029	0.431	-0.680	0.728
Sentiment:Female	-0.189	0.462	-0.948	0.582

The intercept represents men who are writing about daily life with the least positive sentiment (-.40).

Figure 1

Word Frequencies: Titles of Migration Articles in Social Science Journals (2015-2016)

Note. Reprinted from *World Migration Report 2018*, p. 106, Copyright 2017 International Organization for Migration (IOM).

Figure 2

Word Frequencies: Titles of Migration Articles in Arts and Humanities Journals (1999 – 2020)

Note. The code that generated this graphic is available at https://nbviewer.jupyter.org/github/menyalas/ThesisPublic/blob/main/20200504_AM_HumanitiesReview.ipynb

Figure 3

Distribution of Years When Writers Immigrated and Produced Letters

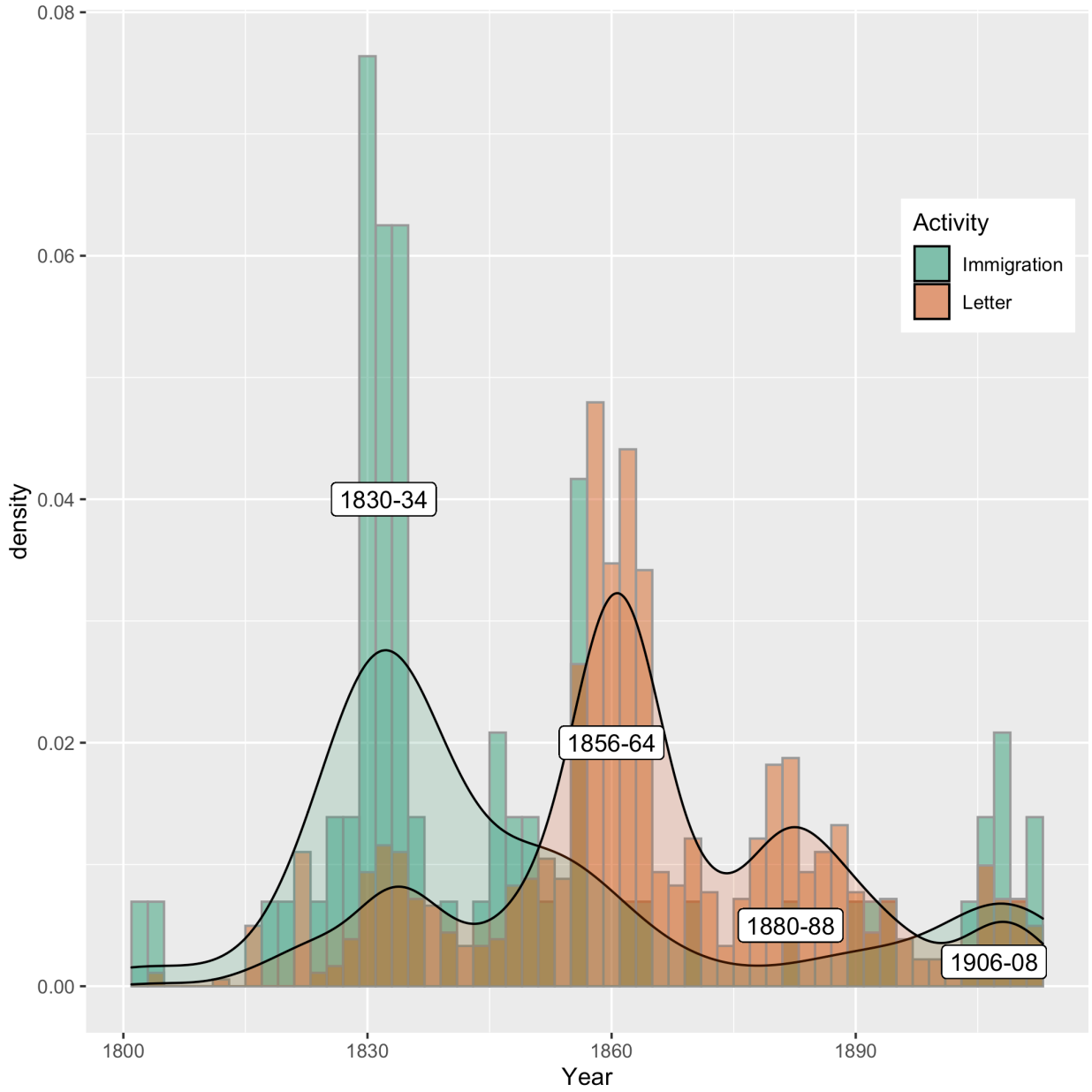


Figure 4

Distribution of the Number of Years between Immigration and Letter Writing

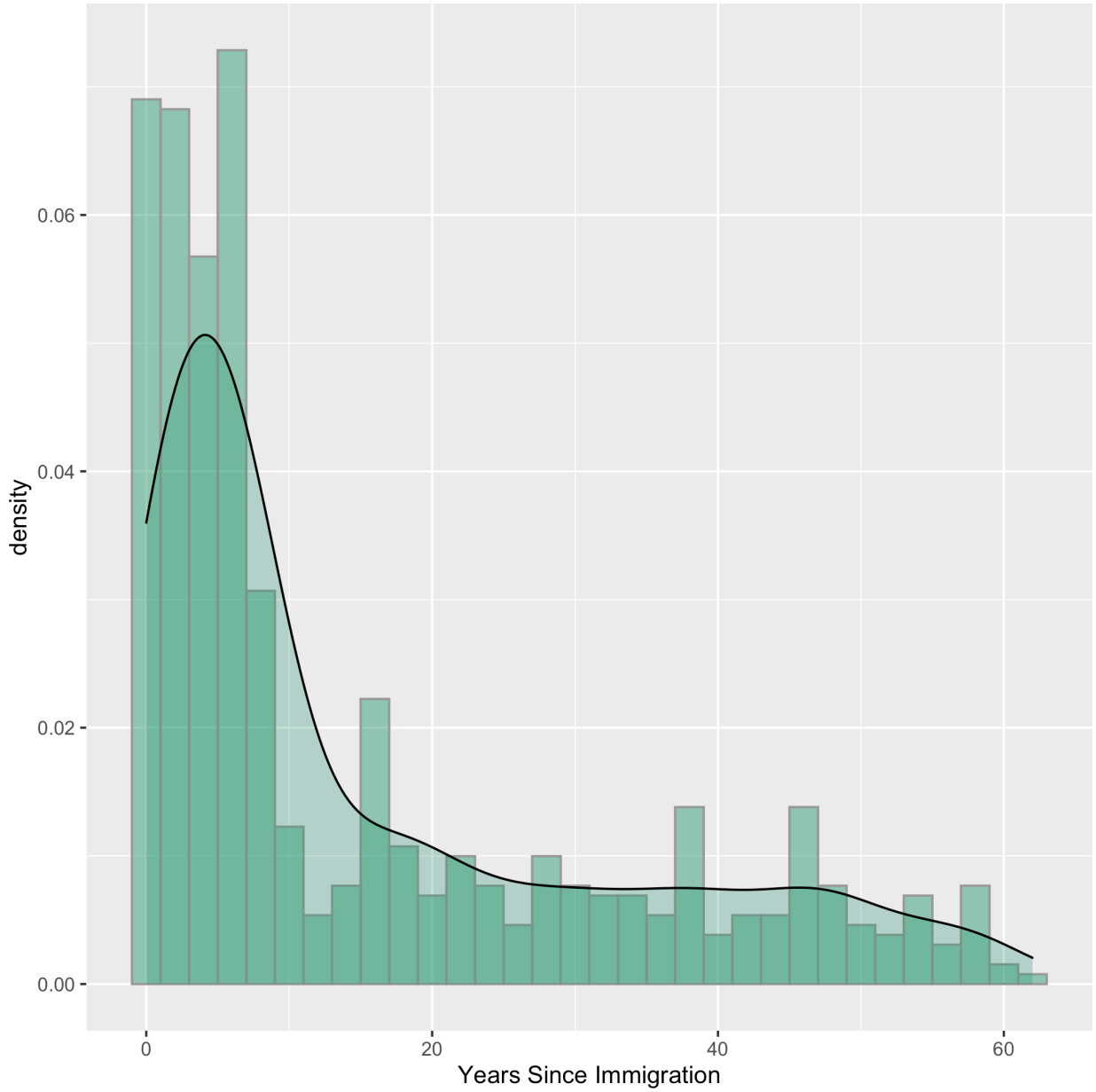


Figure 5

Writer Location and Gender for Writing Group and Letter Collection

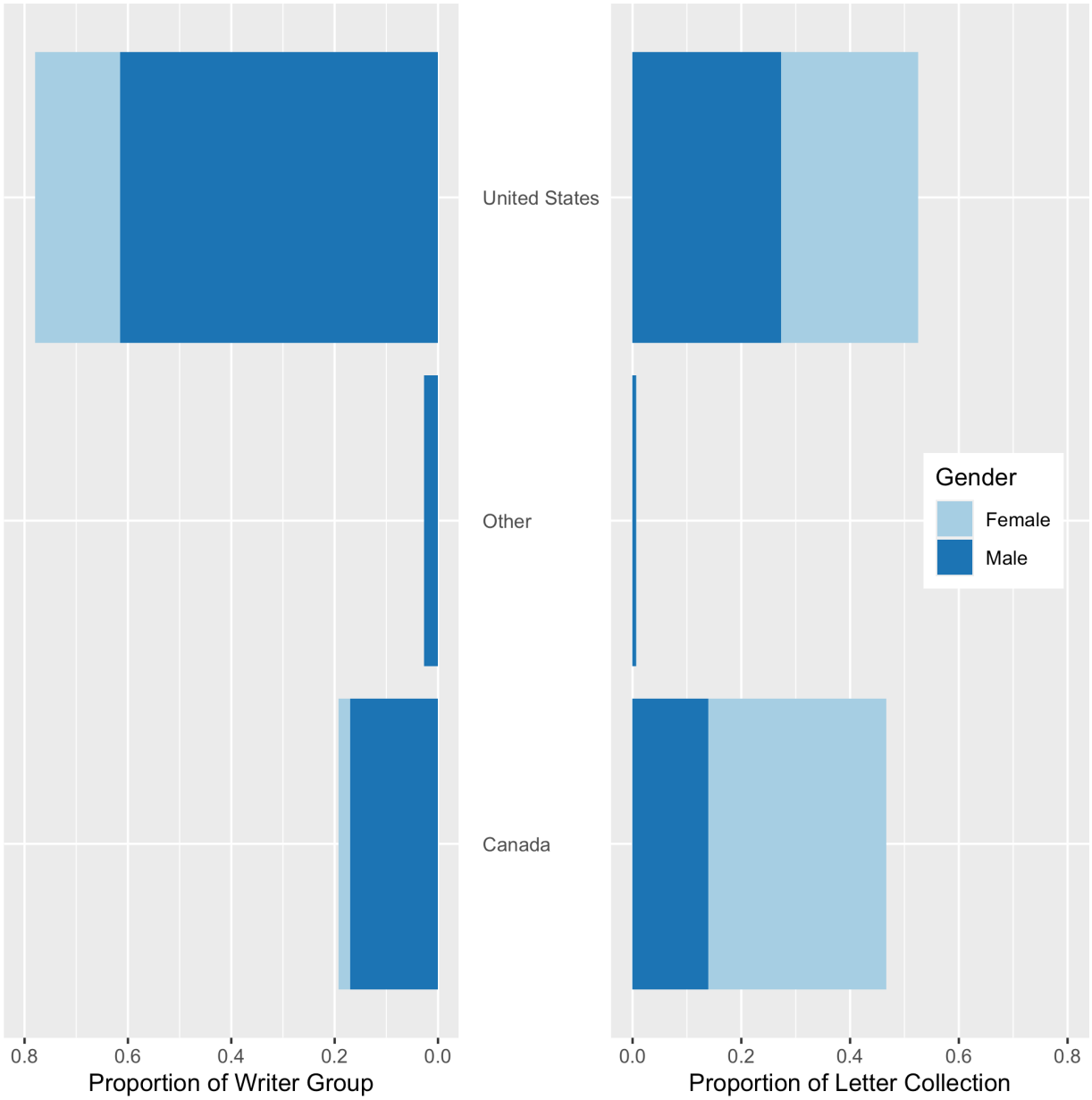
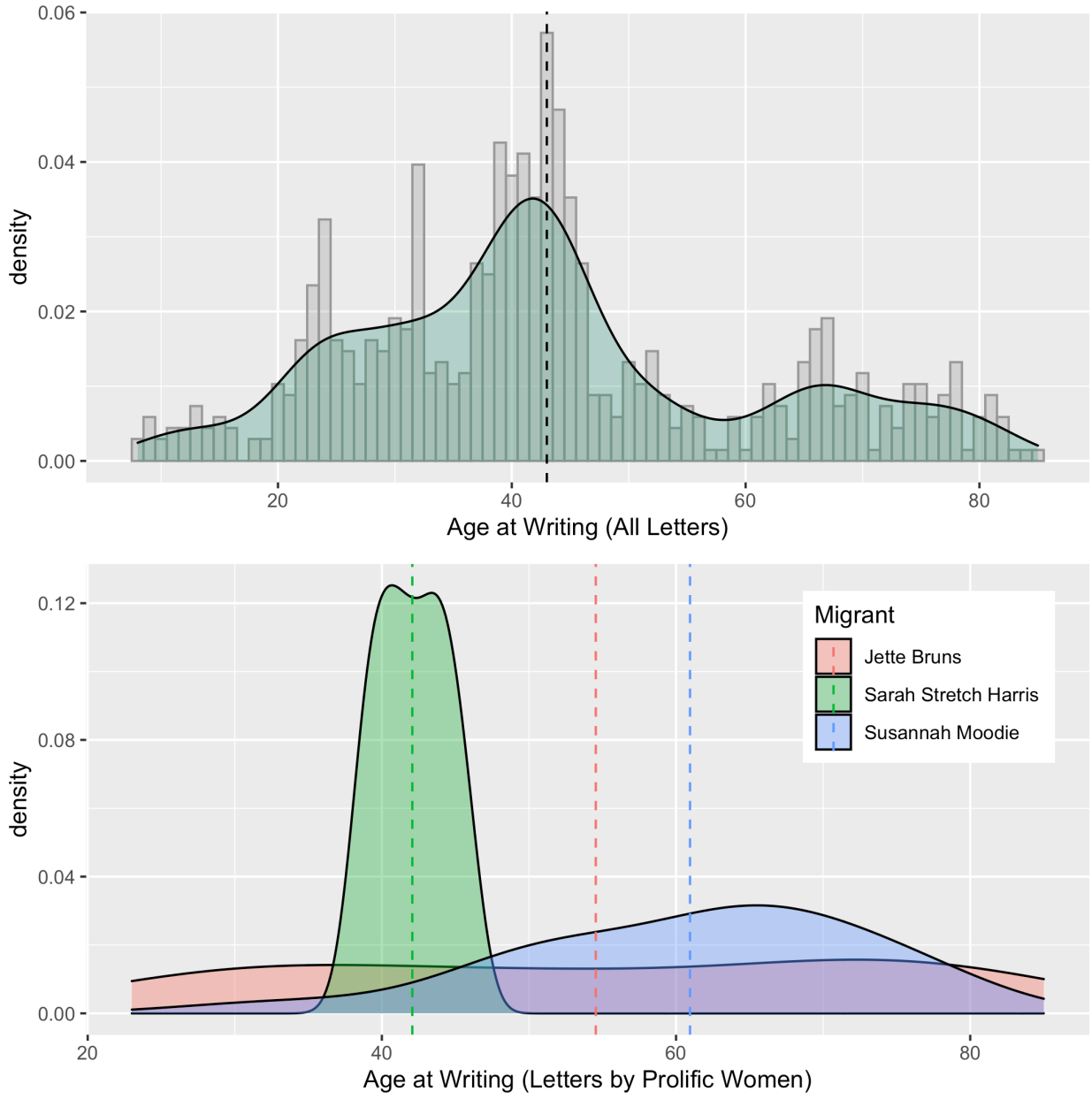


Figure 6

Distribution of Letter Writer Ages, Overall and by Prolific Individuals



Note. The three individuals represented in this plot produced 423 of the 915 letters in the dataset. Their mean ages at the time of writing are 55 years for the American immigrant Jette Bruns (*SD* = 20) and 42 and 61 years, respectively, for the Canadian immigrants, Sarah Stretch Harris (*SD* = 2) and Susannah Moodie (*SD* = 11).

Figure 7

National Origins and Religions for Group of Writers and Letter Collection as a Whole

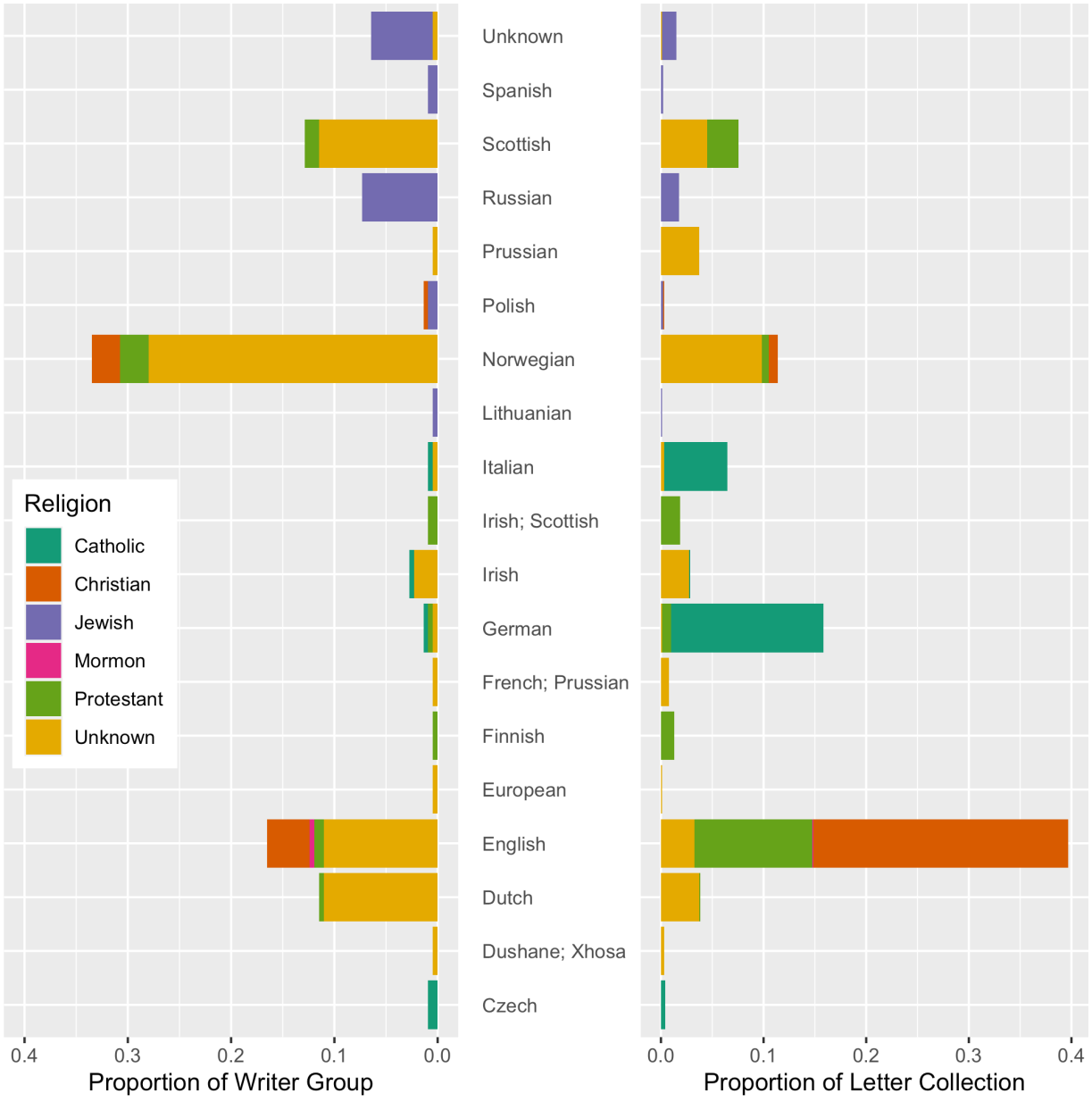
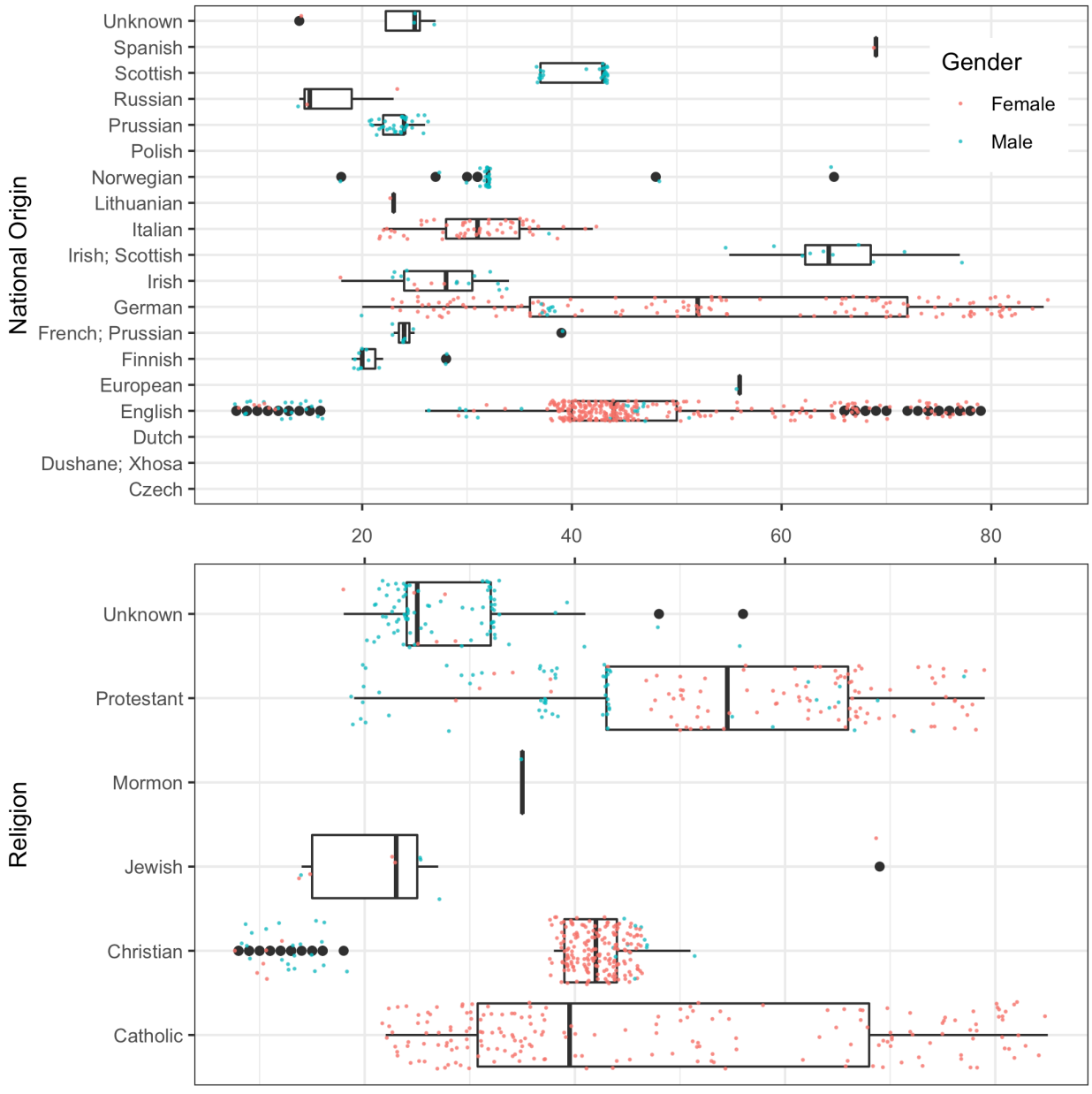


Figure 8

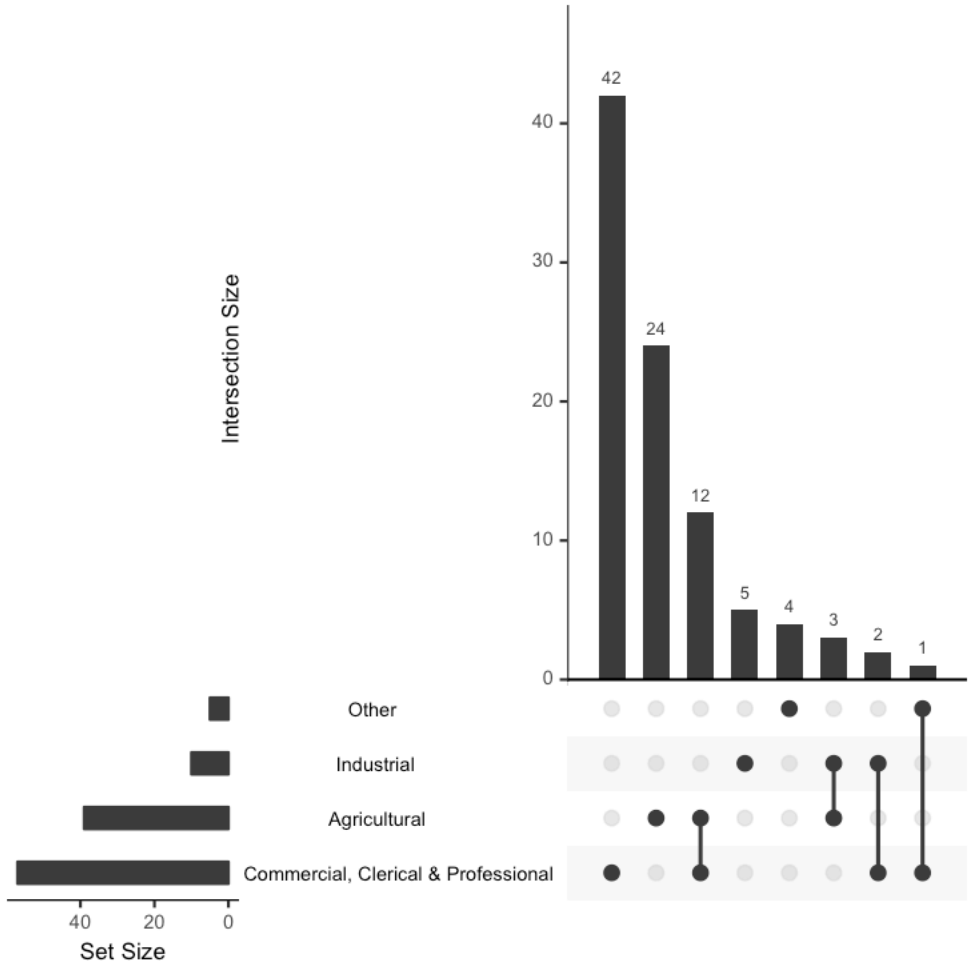
Letter Writer Age by National Origin and Religion with Gender



Note. Boxes indicate the interquartile range with the lower bound being the 25th percentile and the upper being the 75th percentile for each group. The horizontal lines are the median ages. Dots indicate the gender of the letter writer.

Figure 9

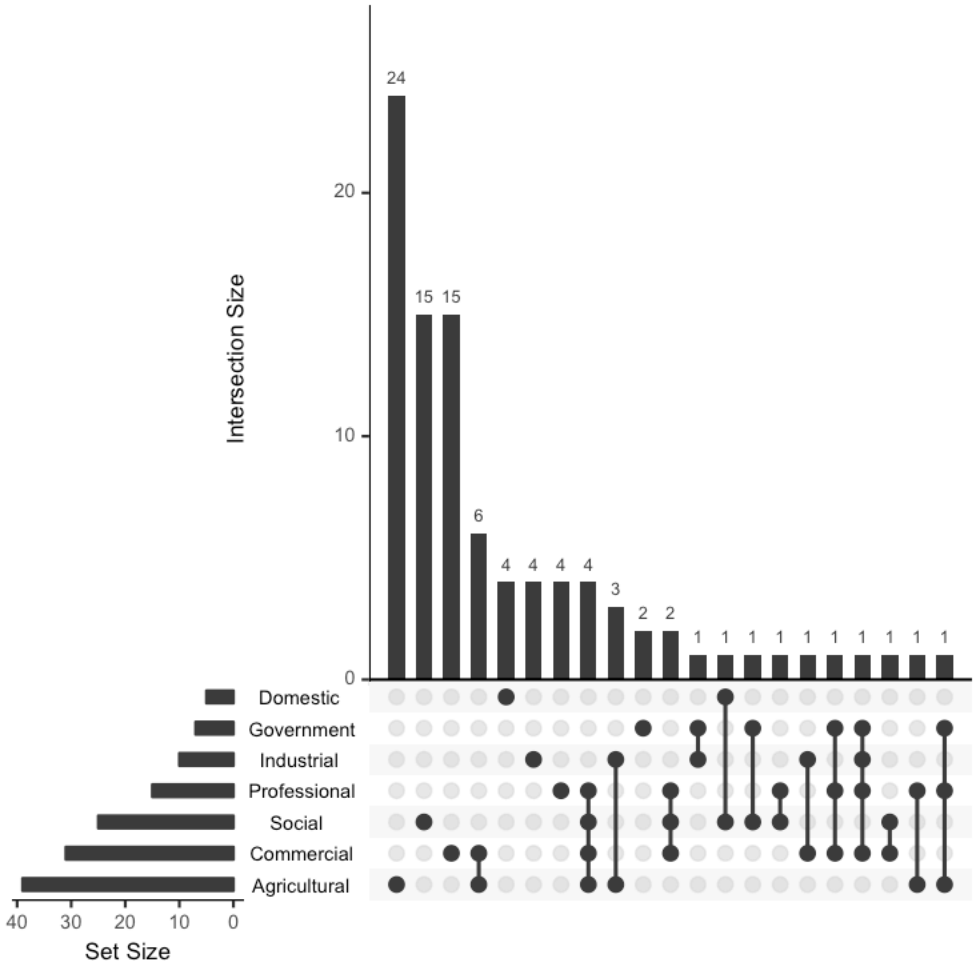
North American Occupation Classes of Writers: 3-Group Scheme



Note: This classification system uses the categories presented by Erickson (1972).

Figure 10

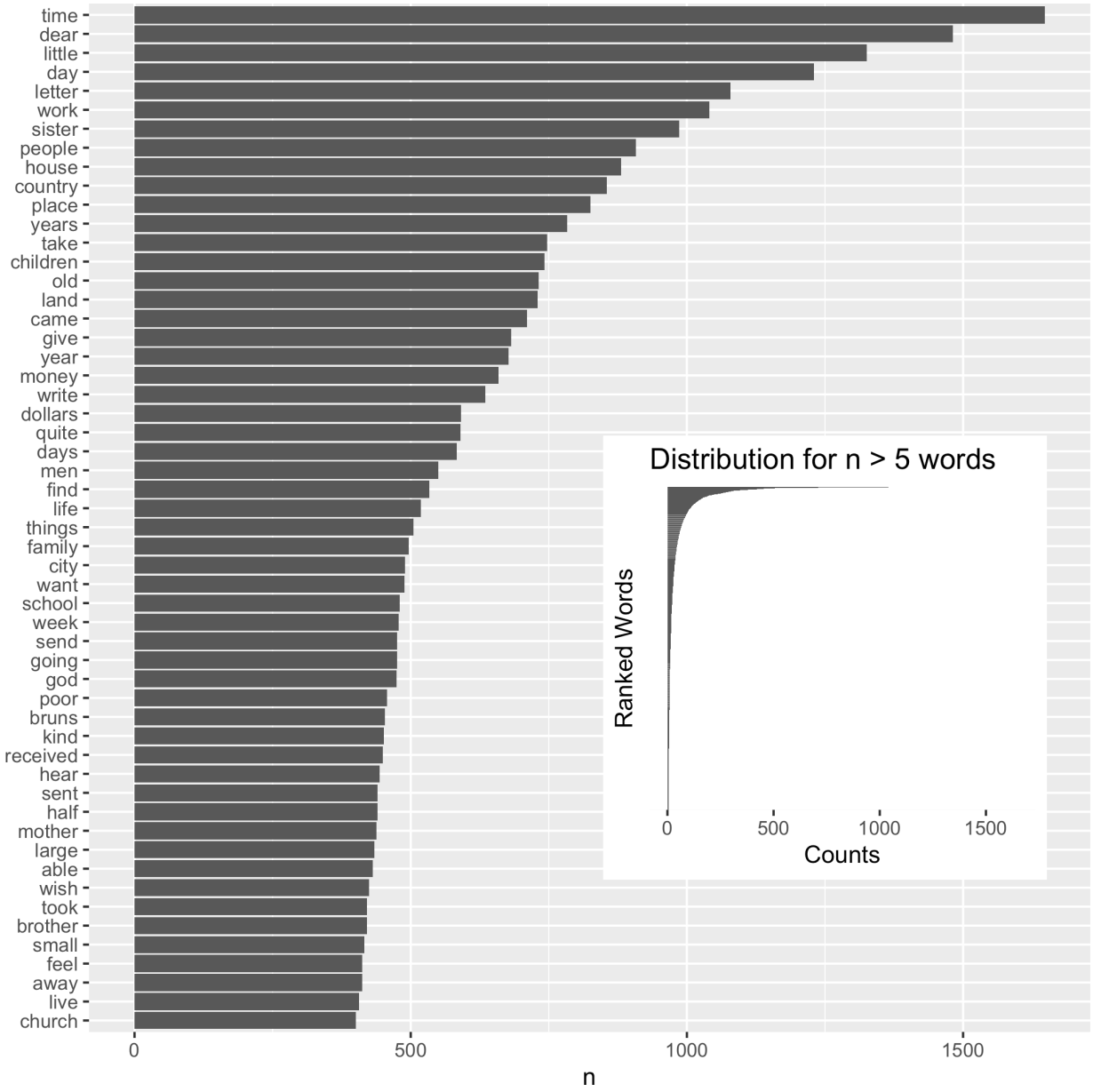
North American Occupation Classes of Writers: 7-Group Scheme



Note: This classification system uses the categories presented by Erickson (1972) with the exception that the commercial, clerical and professional group is subdivided using a system inspired by Miller et al. (2003)

Figure 11

Rank-Frequency Distribution of Content Words



Note. The words listed in this graph appeared more than 400 times. The inset shows a greater portion of the overall distribution to illustrate its adherence to Zipf's Law. Stopwords include 5,631 items identified by Jockers and Mimno (2012) for topic modeling of 19th century English literary texts (https://www.matthewjockers.net/macroanalysisbook/expanded-stopwords-list/)

Figure 12

Coherence Scores (C_V) for Mallet LDA Models

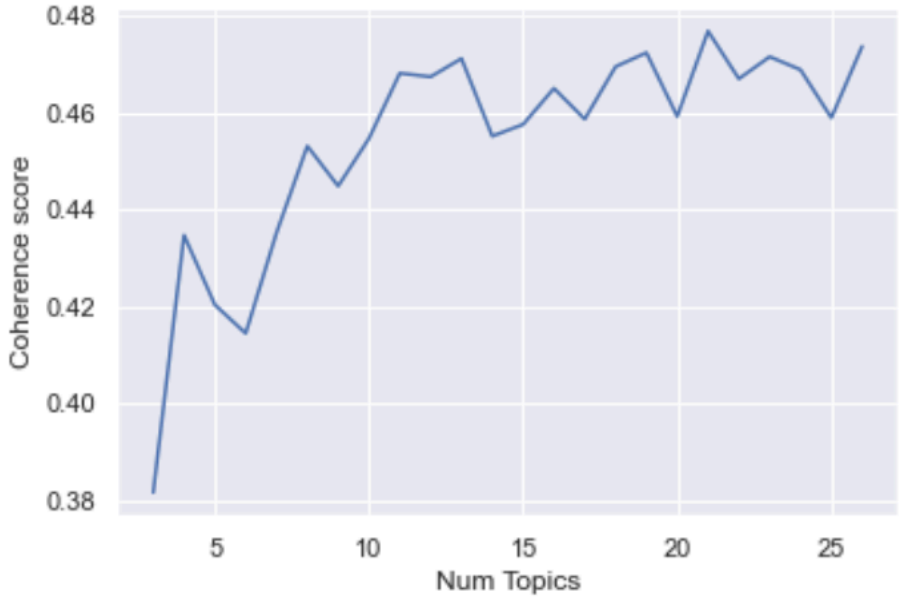
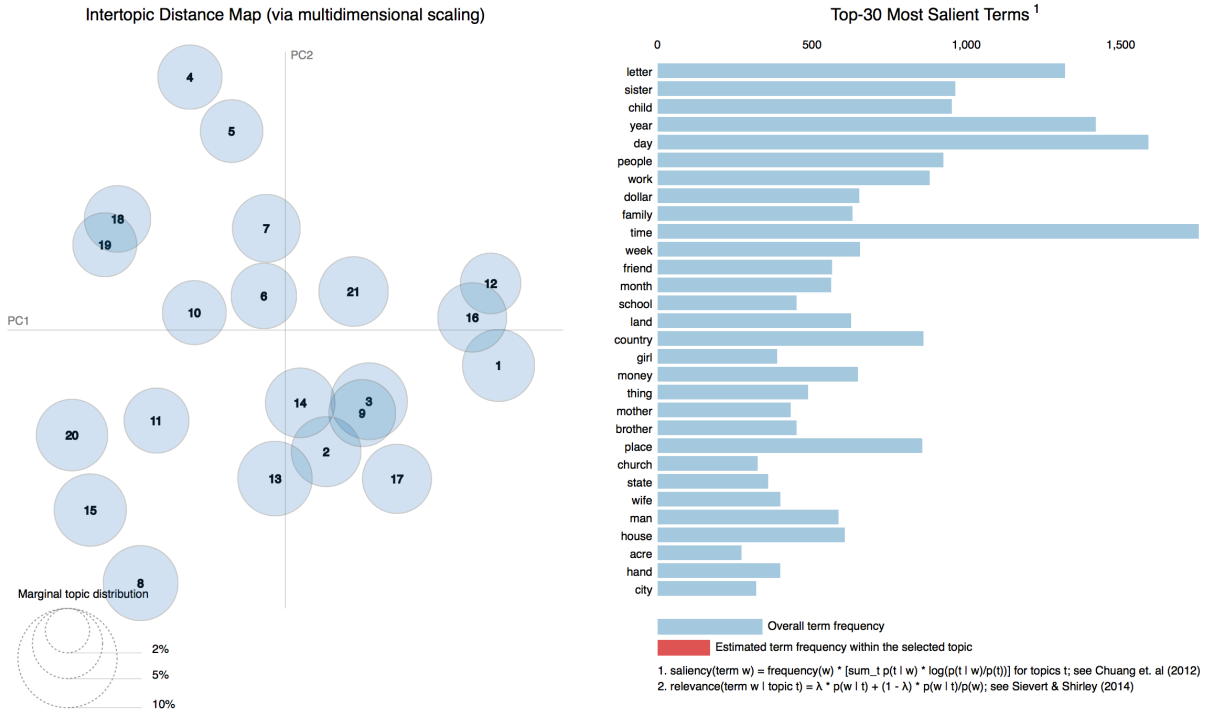


Figure 13

Intertopic Distance for Letters and Most Salient Terms



Note. Topic labels are shown in Table X. A dynamic version of this plot, showing the most salient terms for each topic, is available at https://nbviewer.jupyter.org/github/menyalas/ThesisPublic/blob/main/MalletLDA21.html

Figure 14

Distribution of Letter Topics

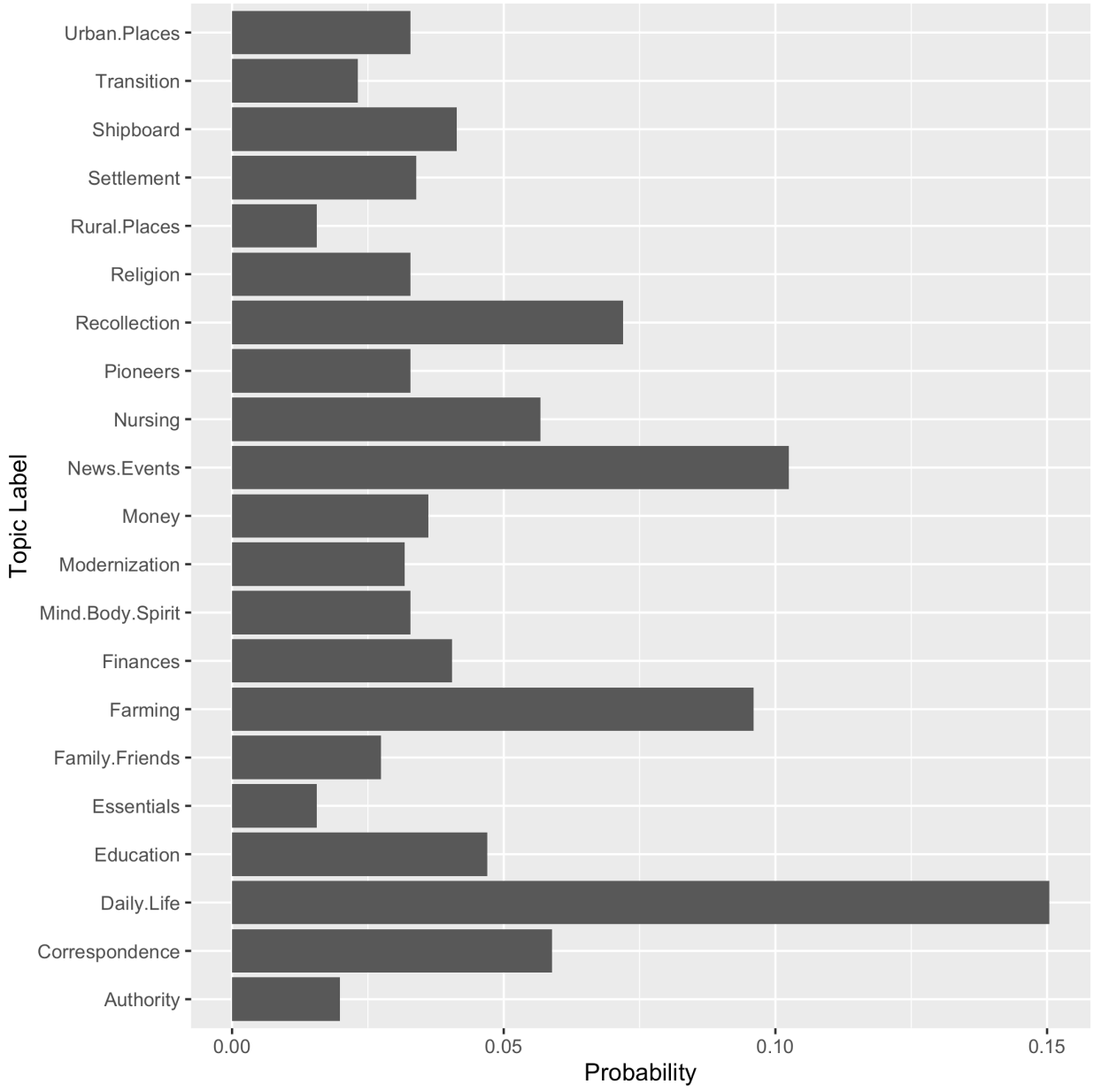


Figure 15

Distribution of Sentence Topics

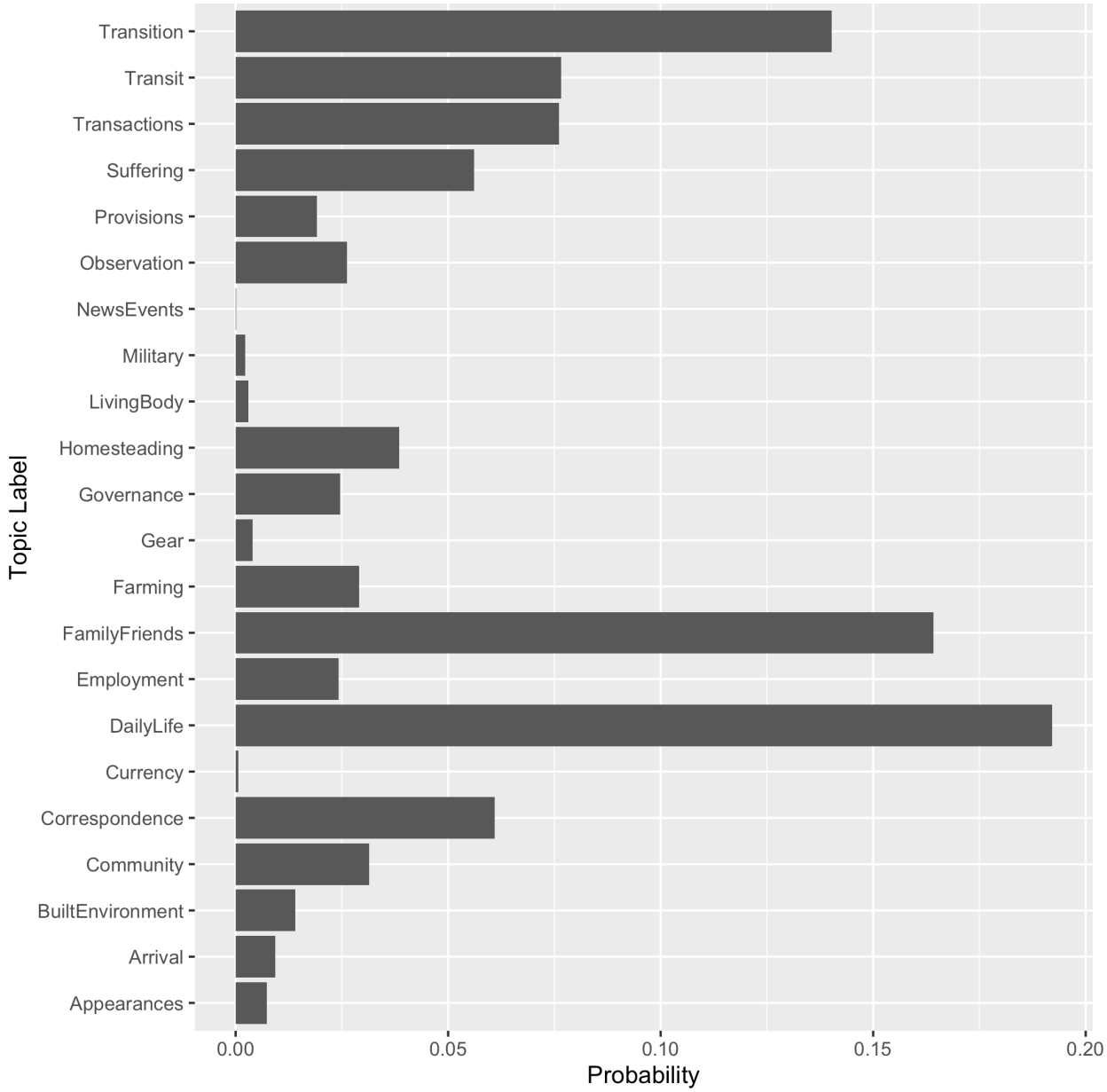


Figure 16

Estimated Sentiment Score by Calendar Year

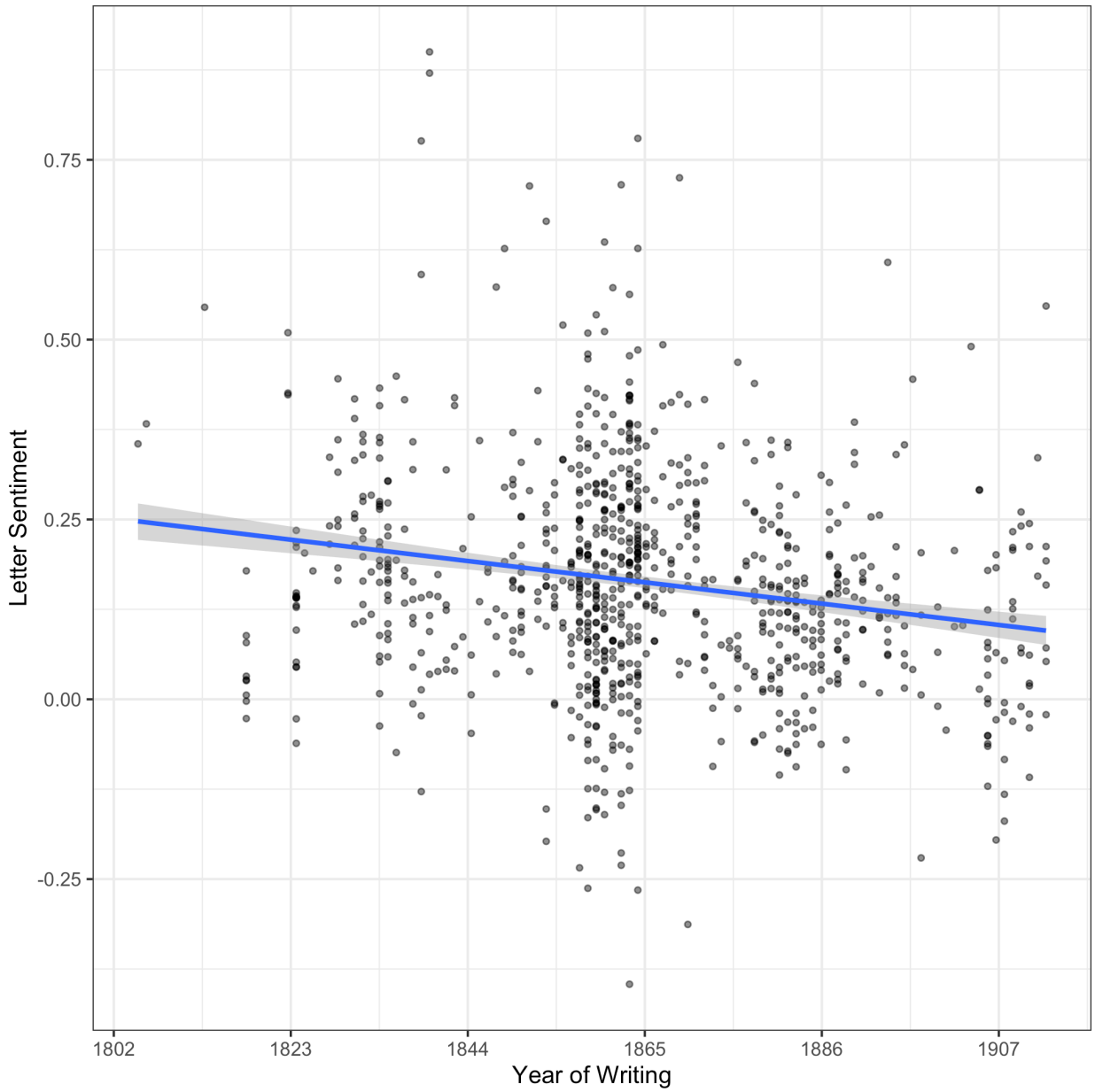


Figure 17

Predicted Probability of Topics by Year of Writing (1800 to 1914)

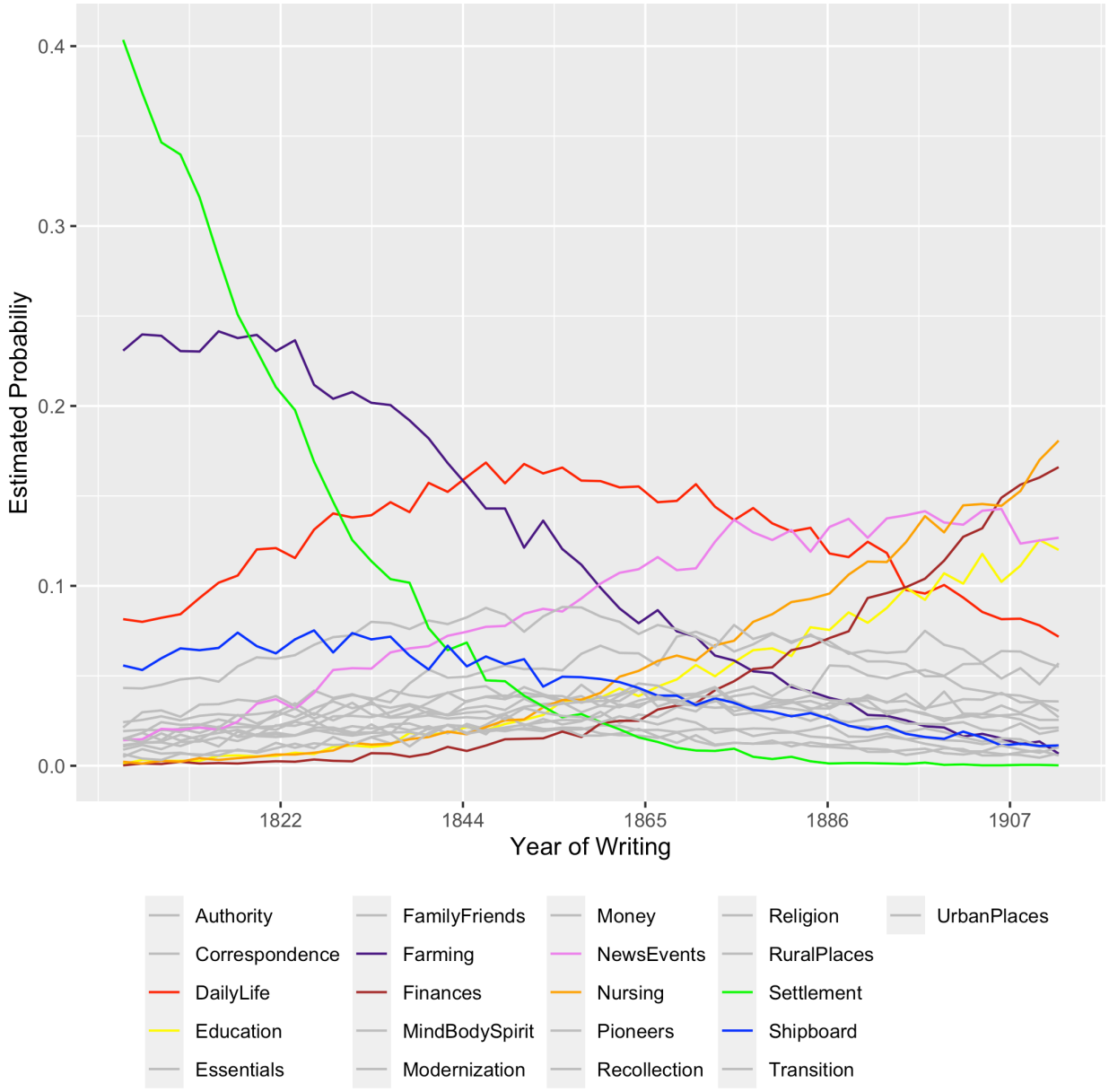


Figure 18

Estimated Sentiment Score by Number of Years Elapsed Since Migration

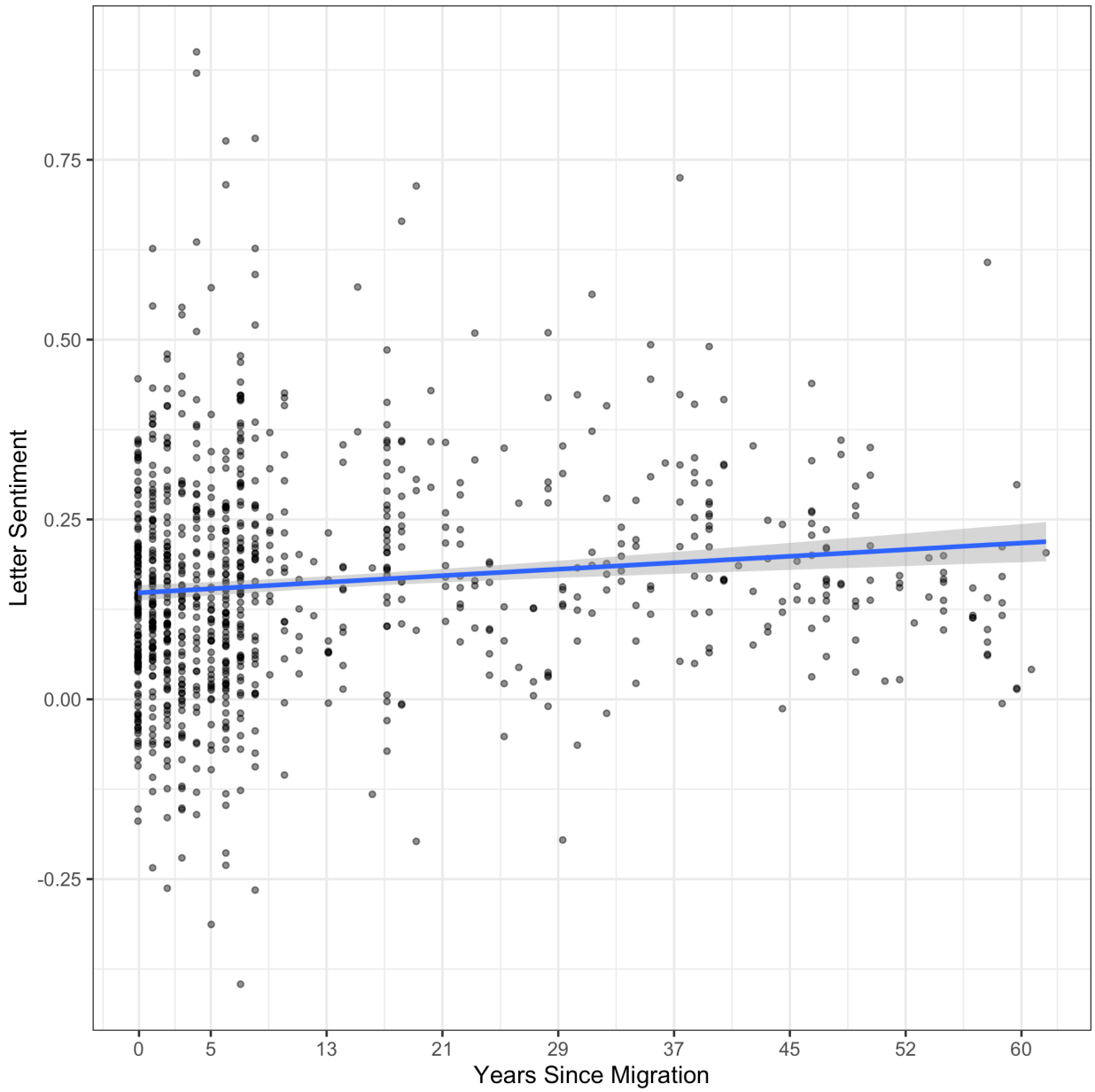


Figure 19

Predicted Probability of Topics by Number of Years Since Migration

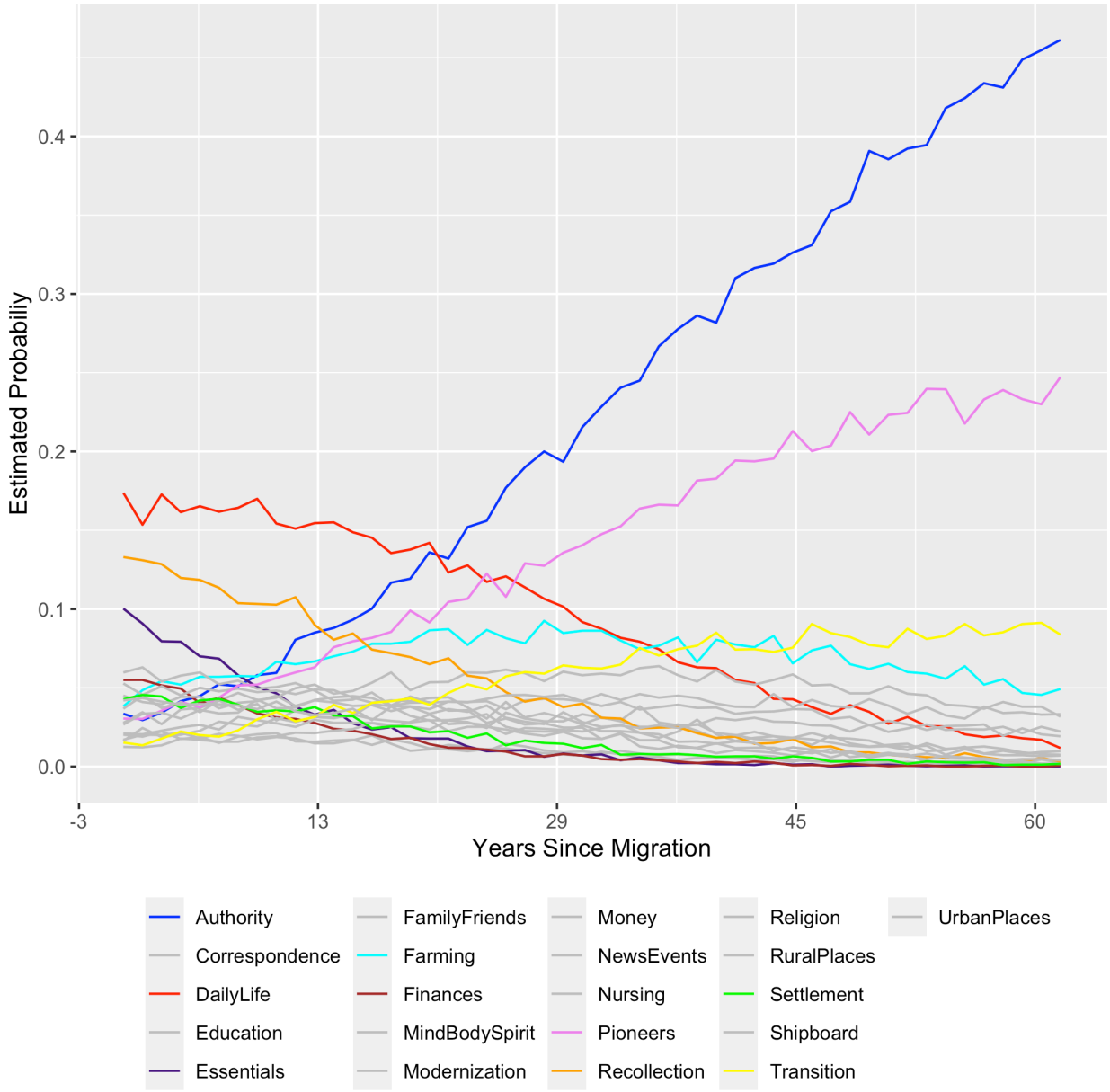


Figure 20

Predicted Probability of Topic by Position of Sentence in Letter

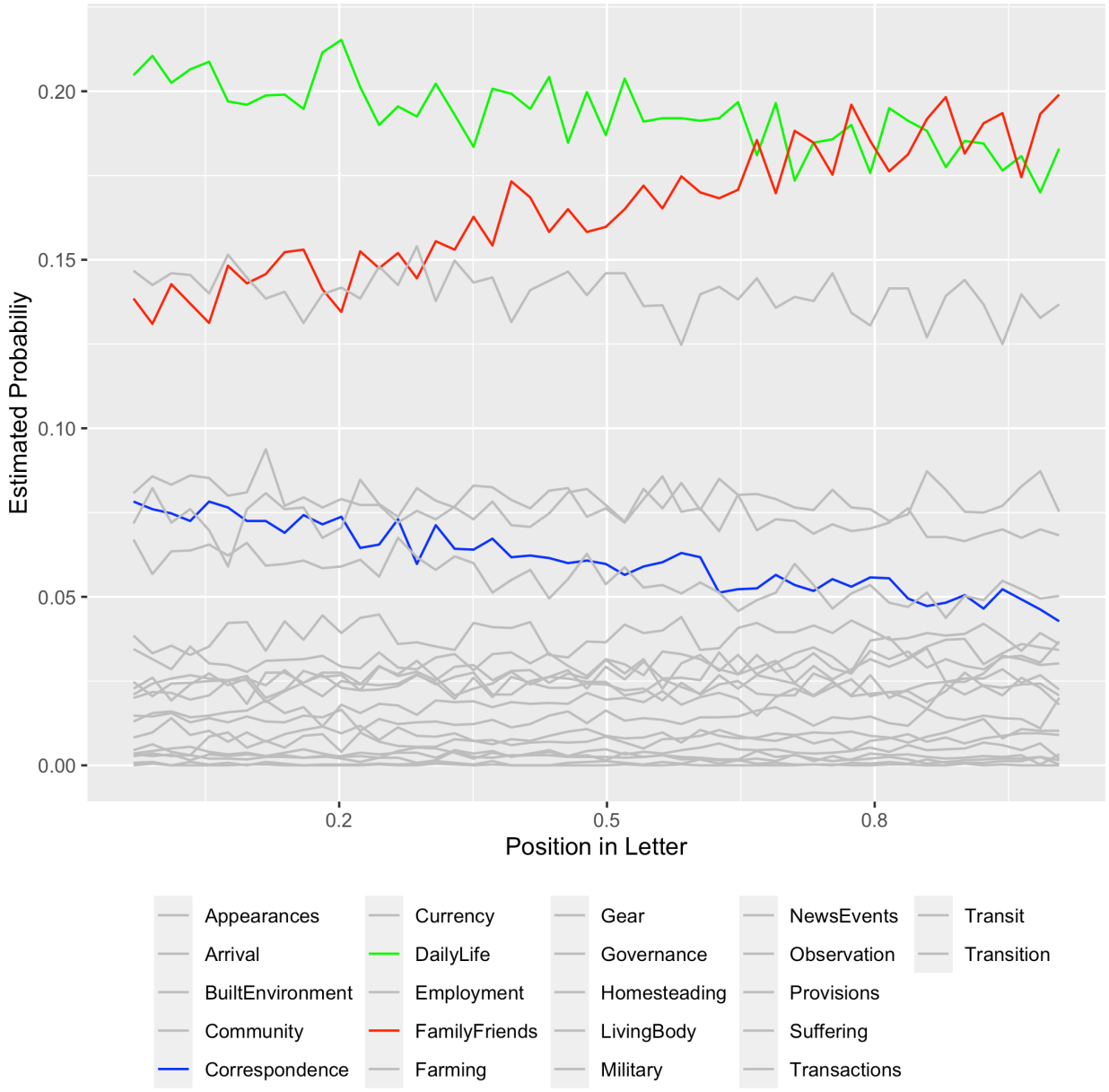


Figure 21

Estimated Sentiment by Author Location

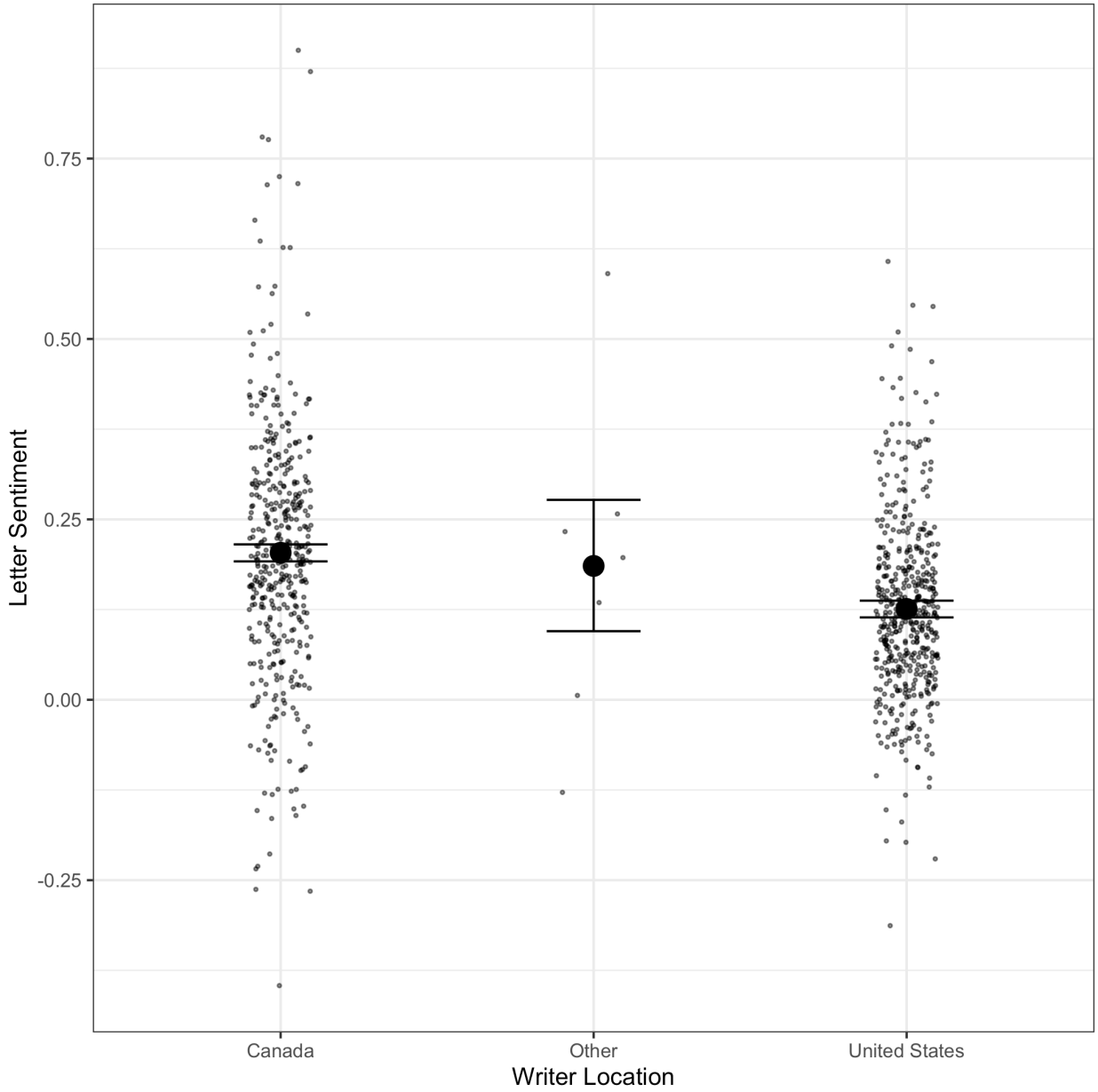


Figure 22

Predicted Probability of Topic by Author Location

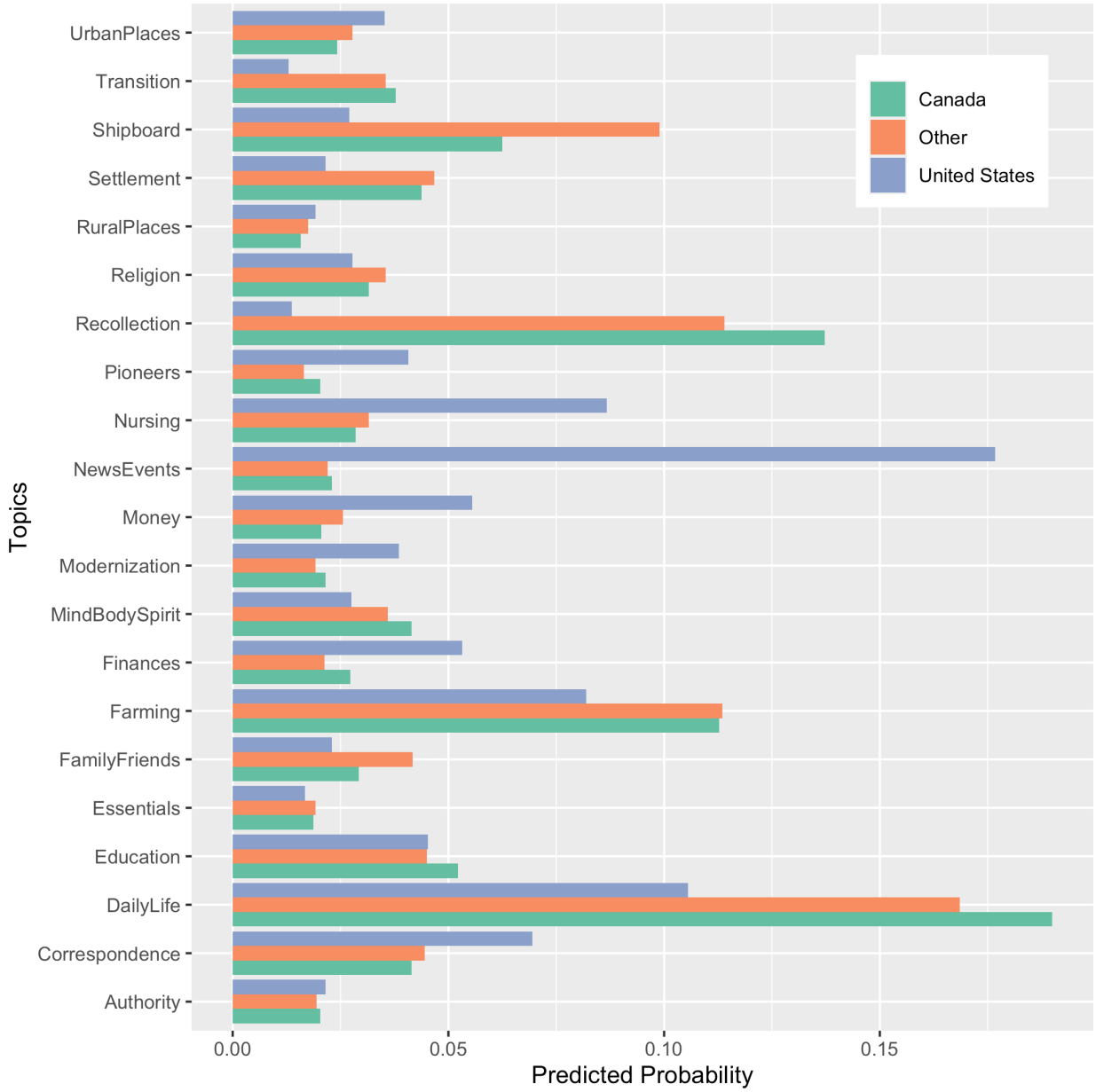


Figure 23

Estimated Sentiment Score by Author Gender

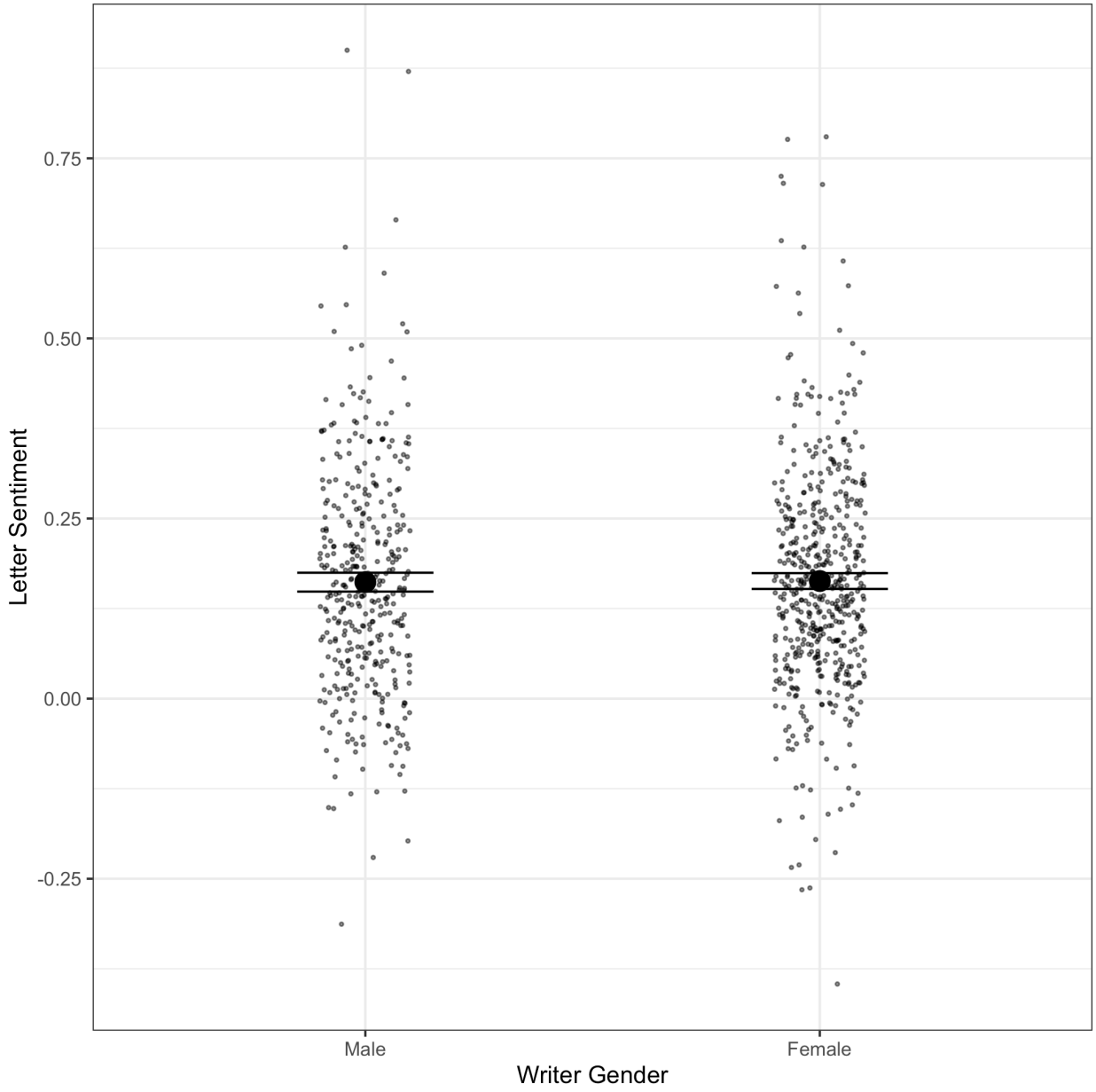


Figure 24

Predicted Probability of Letter Topics by Author Gender

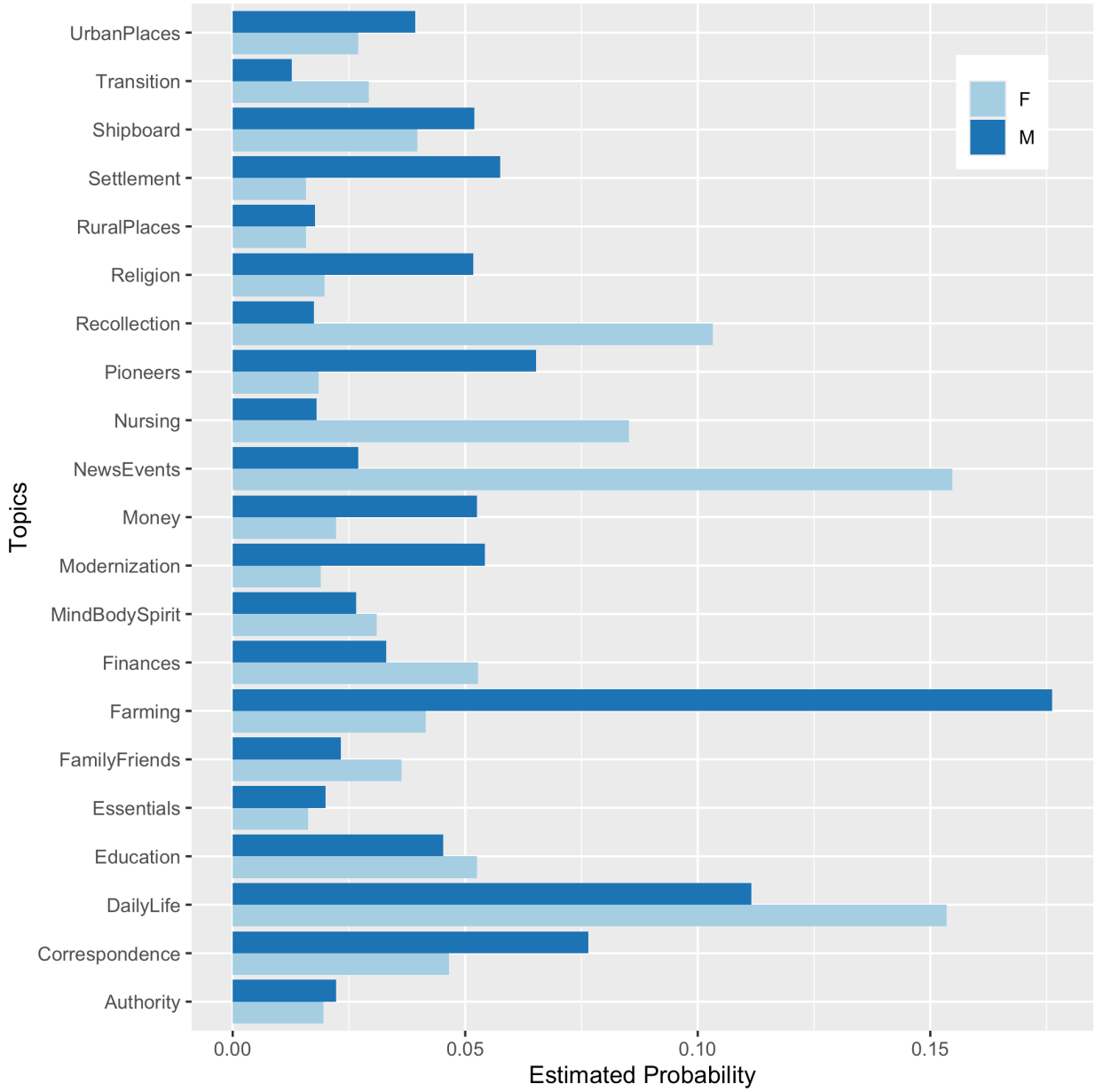


Figure 25

Estimated Sentiment Scores by Author Age

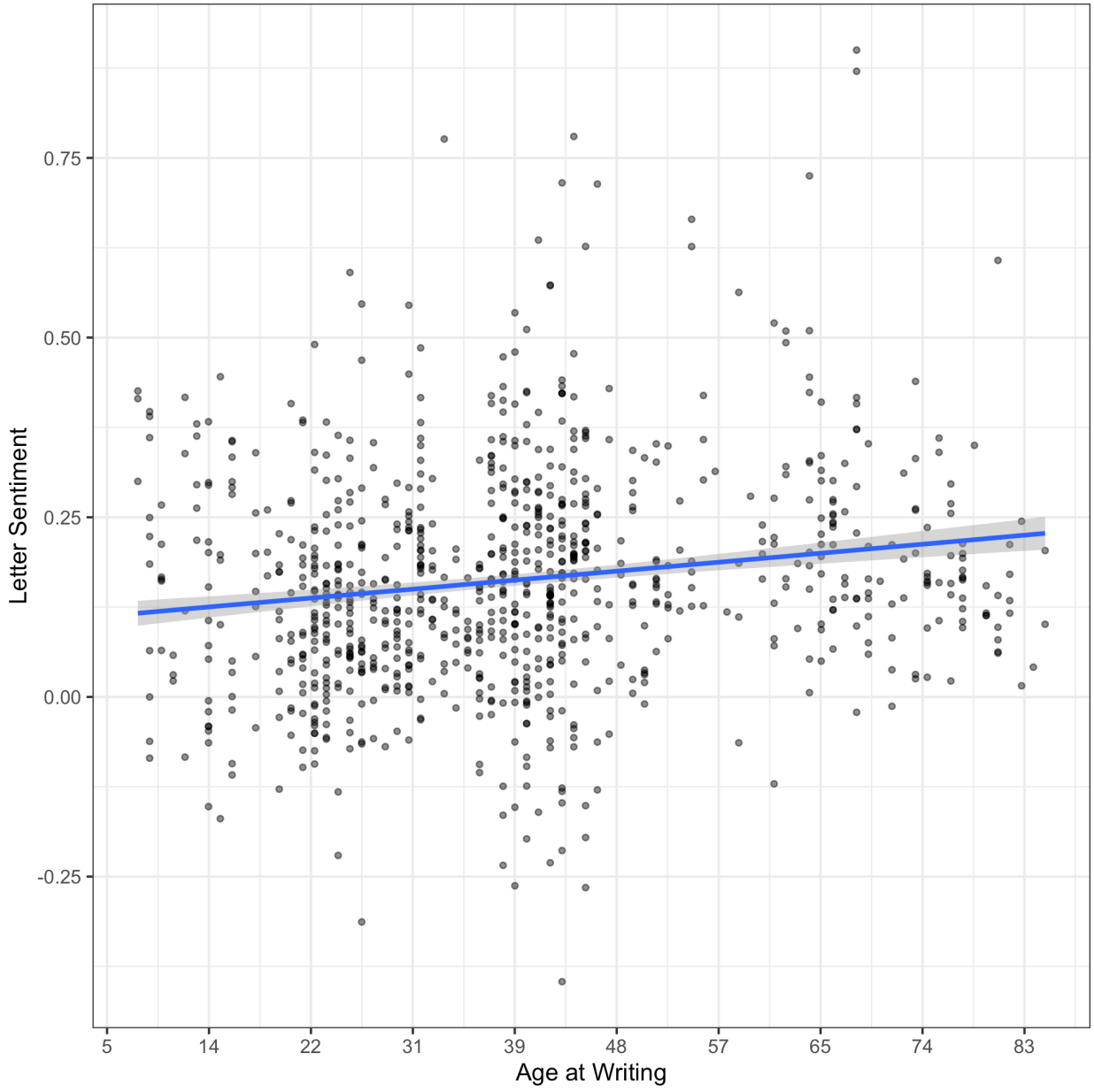


Figure 26

Predicted Probability of Letter Topics by Author Age

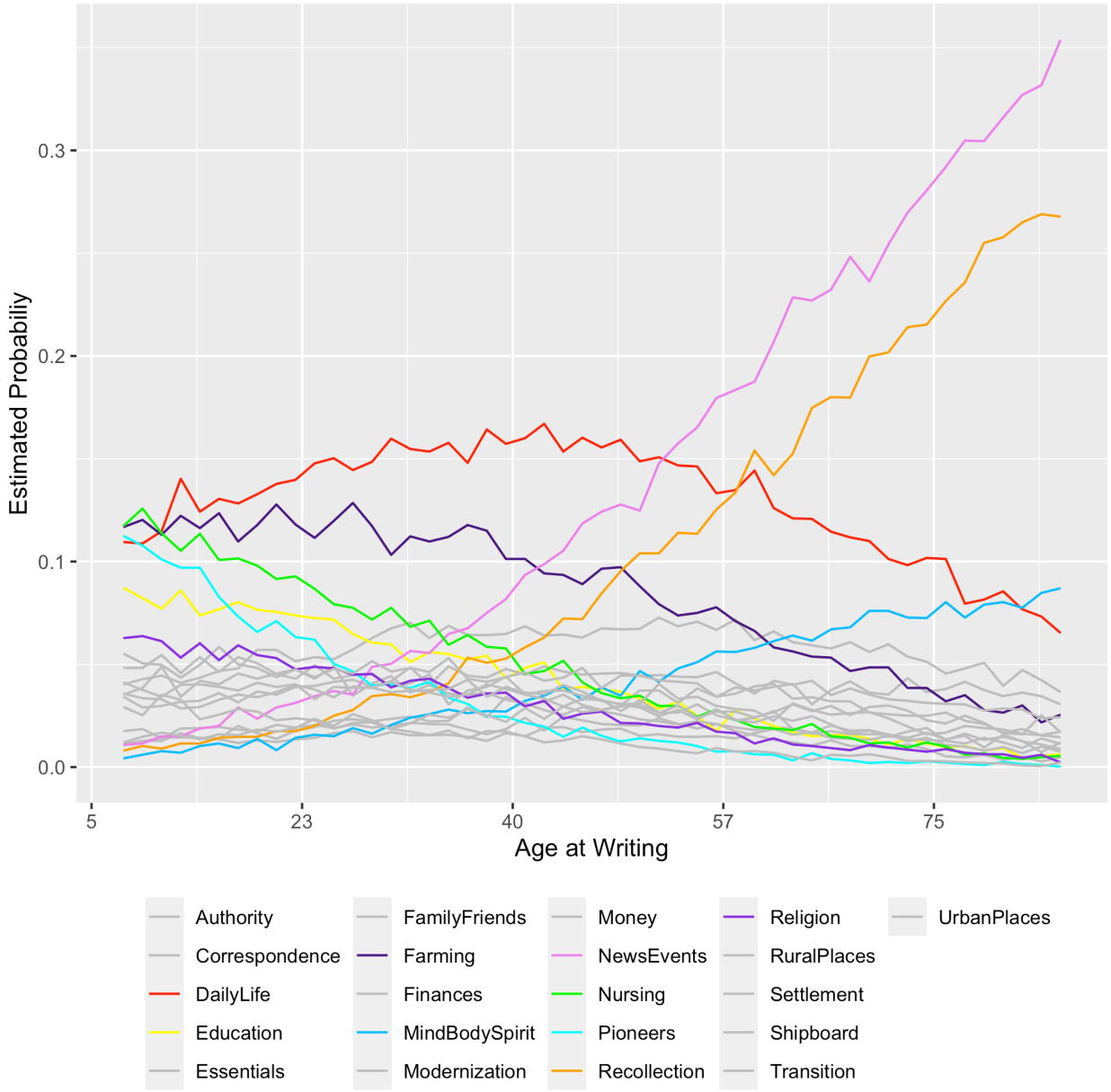


Figure 27

Estimated Sentiment Scores by National Origin

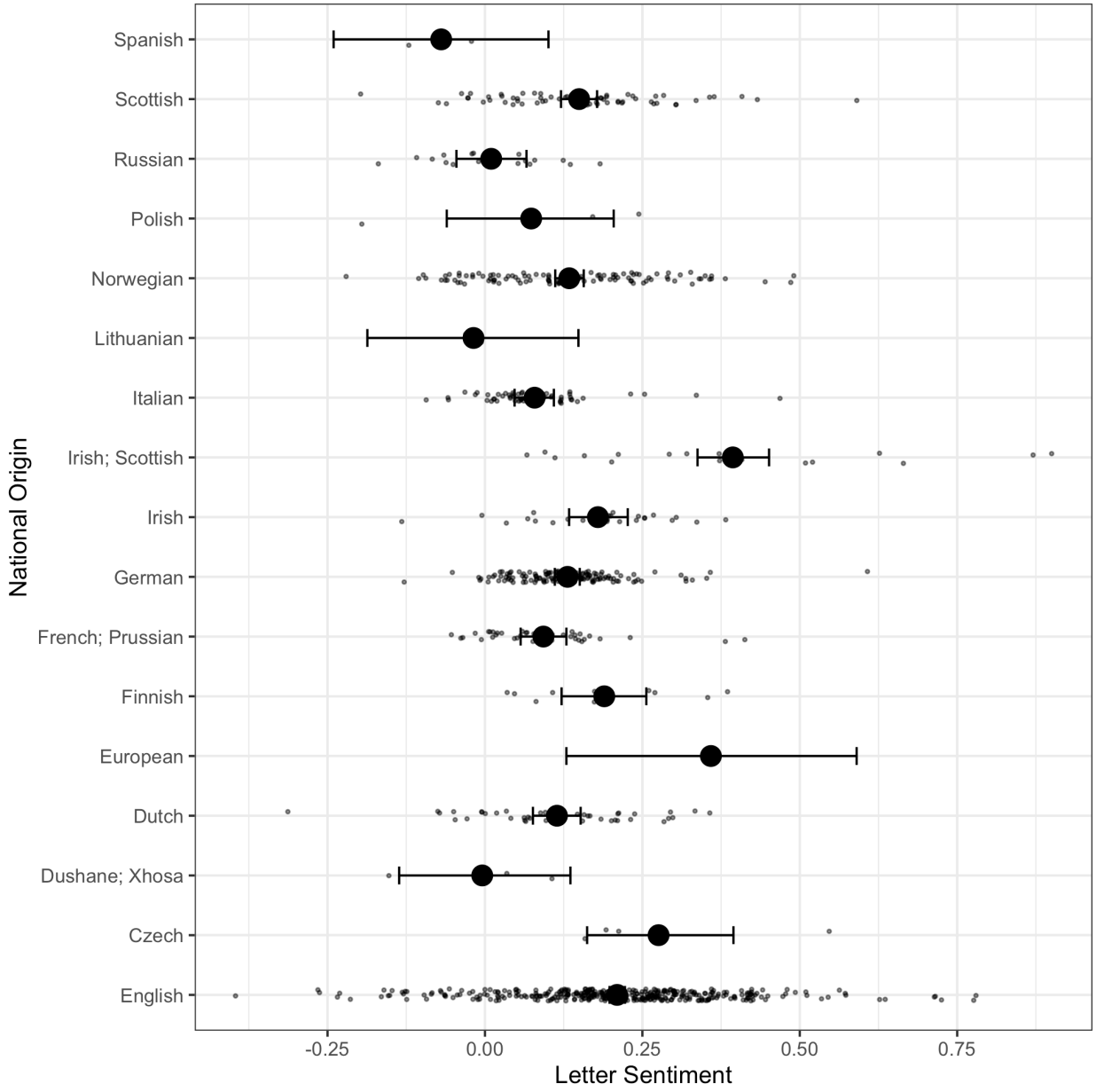


Figure 28

Predicted Probability of Letter Topics by National Origin

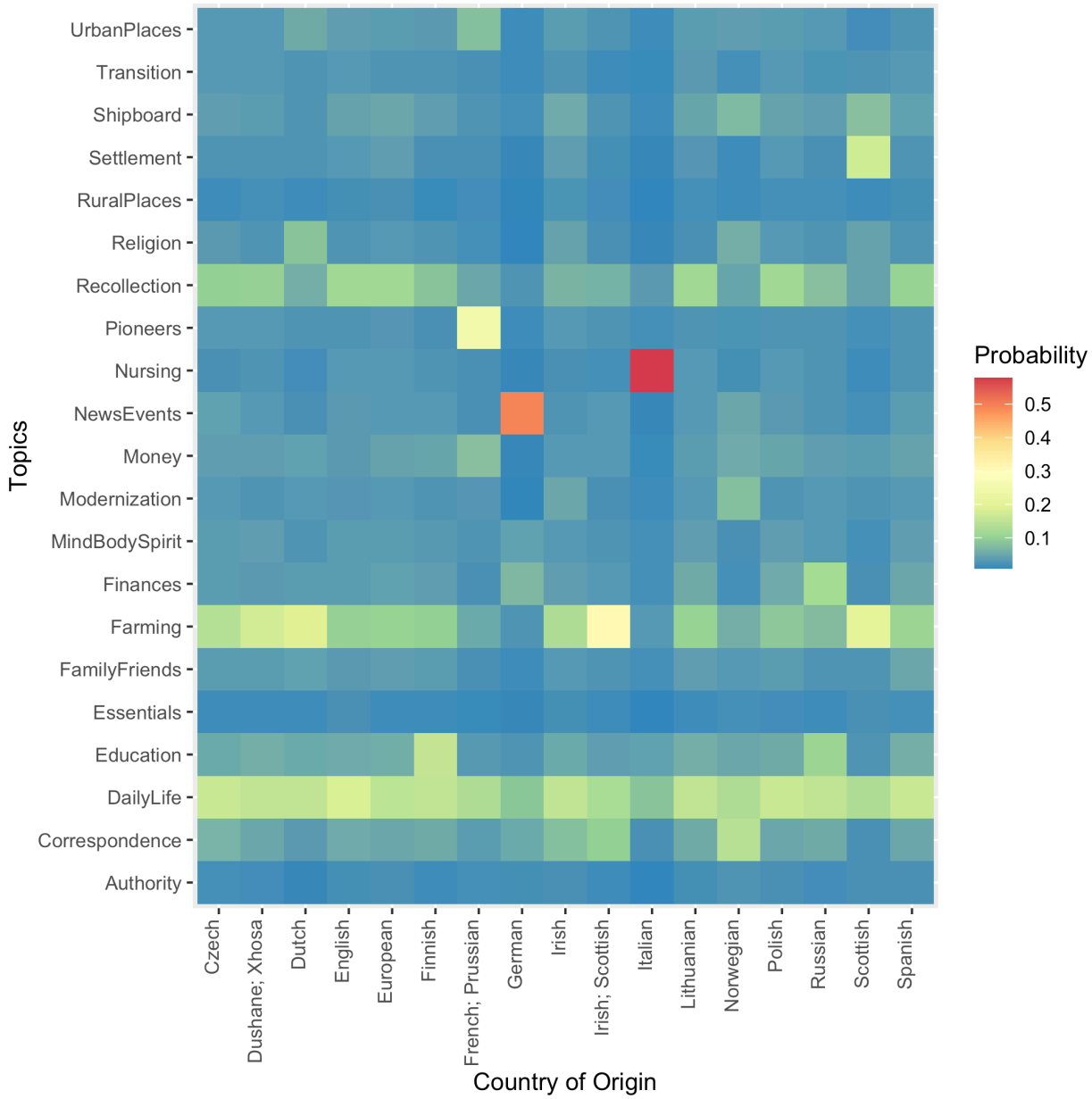


Figure 29

Estimated Sentiment Scores by Religion

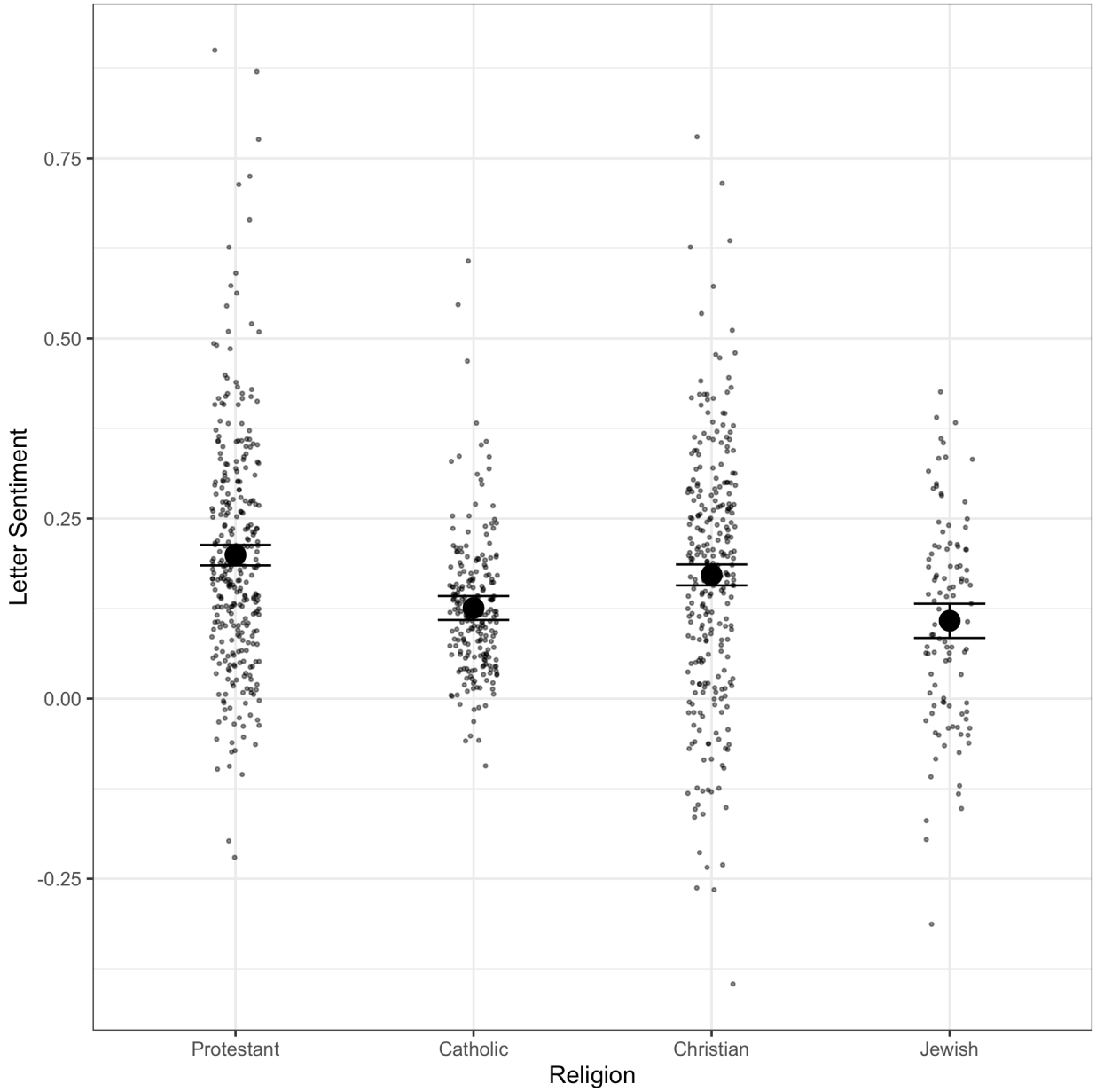


Figure 30

Predicted Probability of Letter Topics by Religion

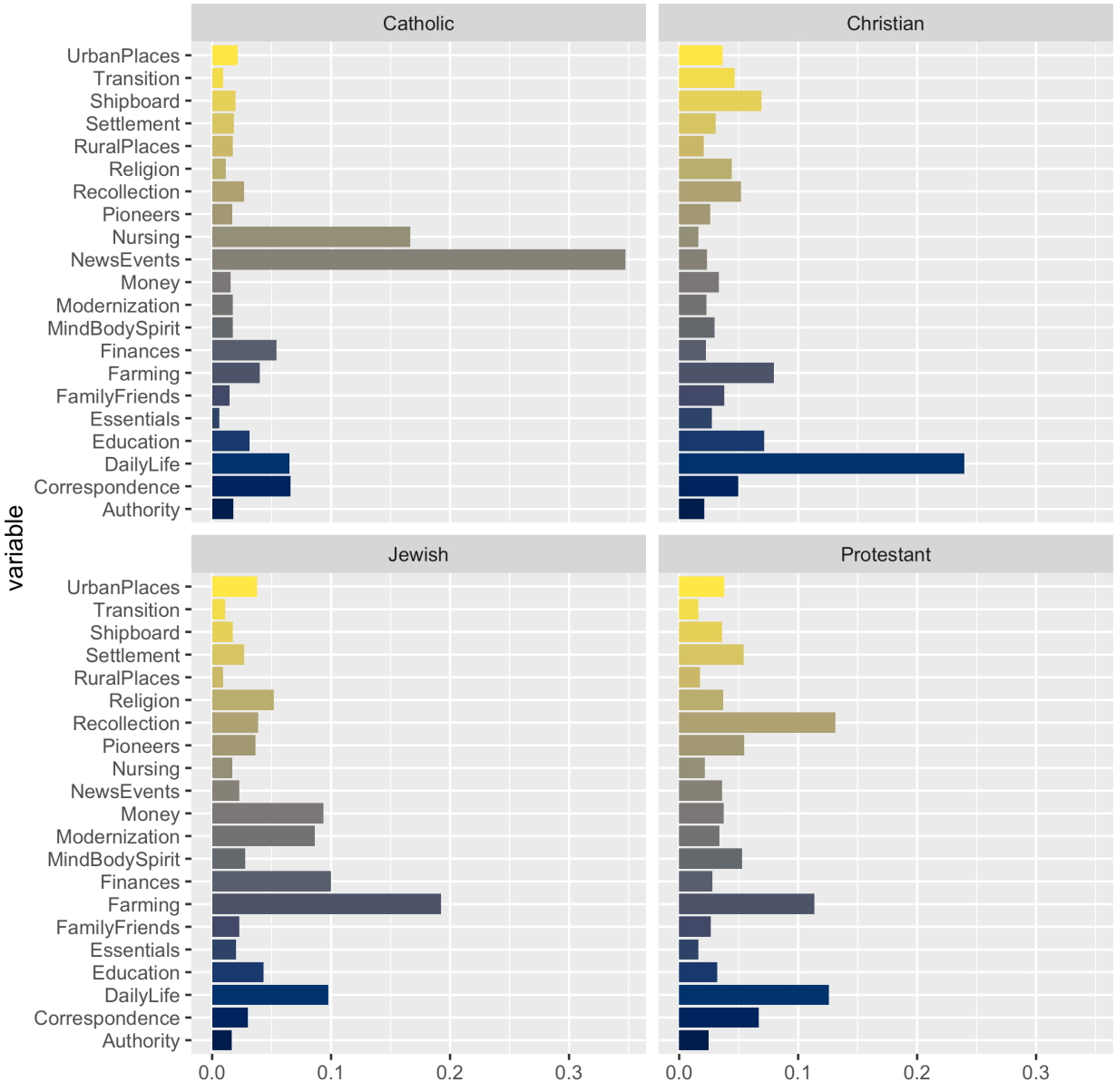


Figure 31

Estimated Sentiment Scores by Occupational Class

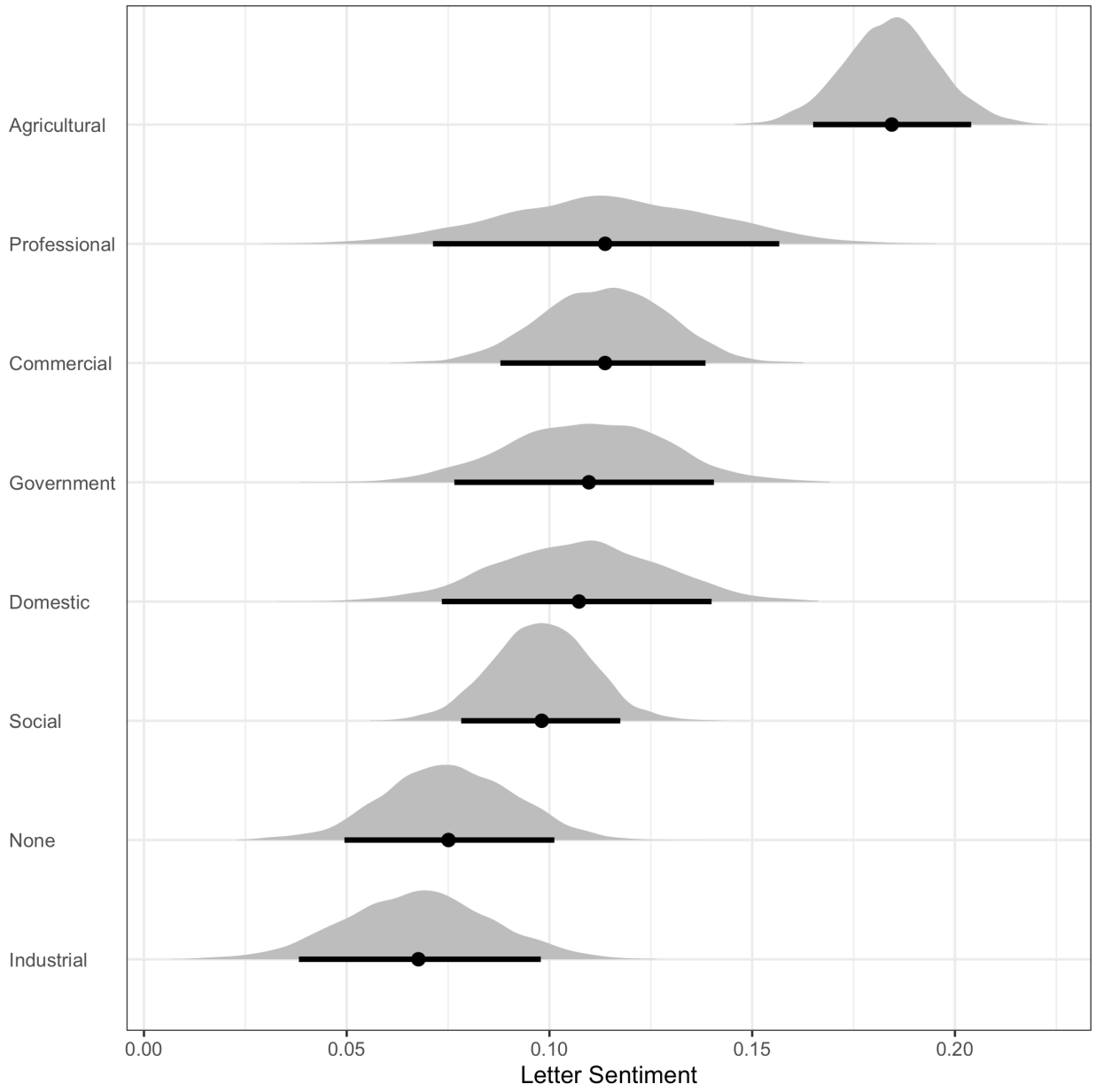


Figure 32

Predicted Probability of Letter Topics by Occupational Class

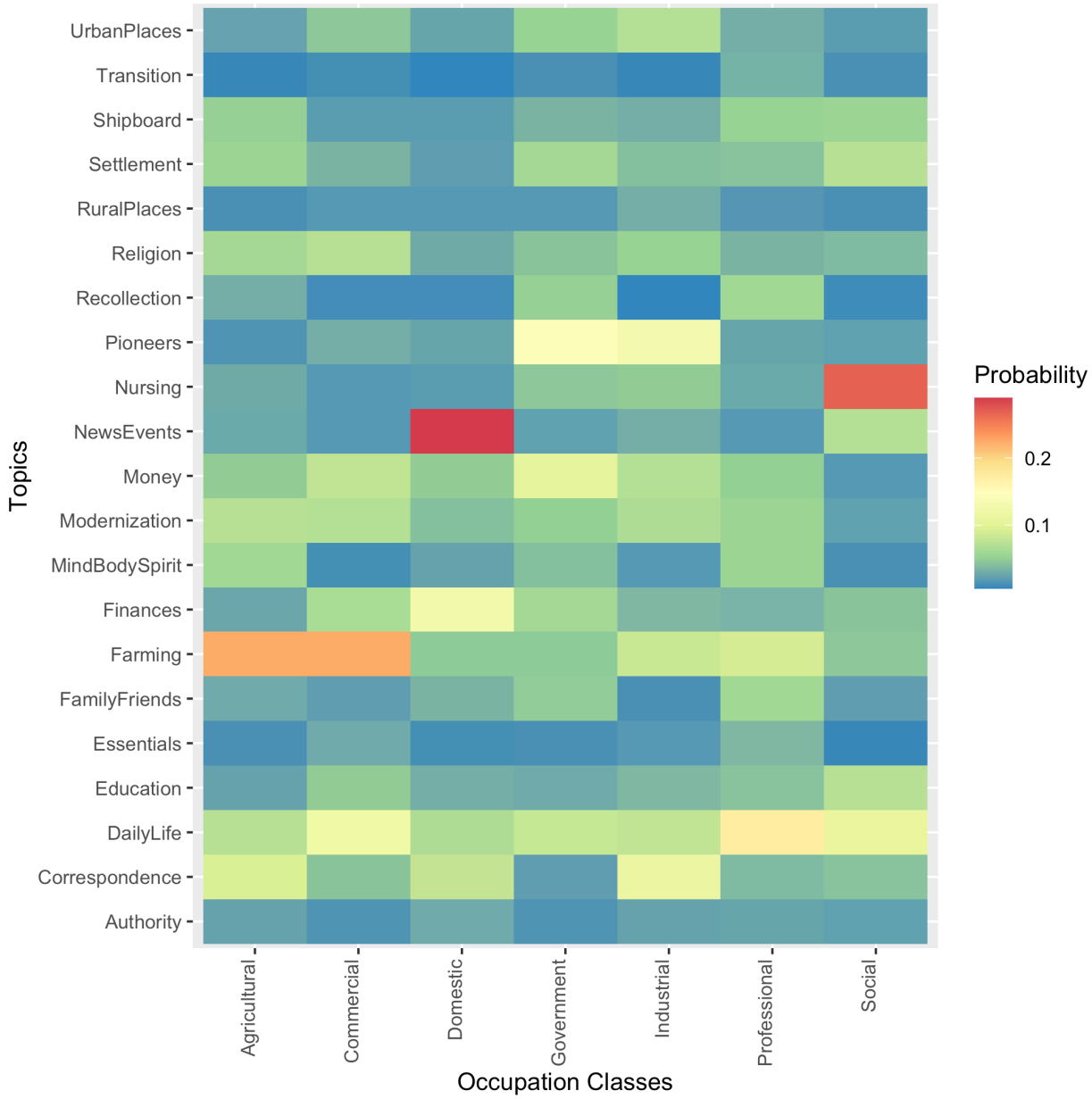


Figure 33

Estimated Sentence Scores by Letter Topic

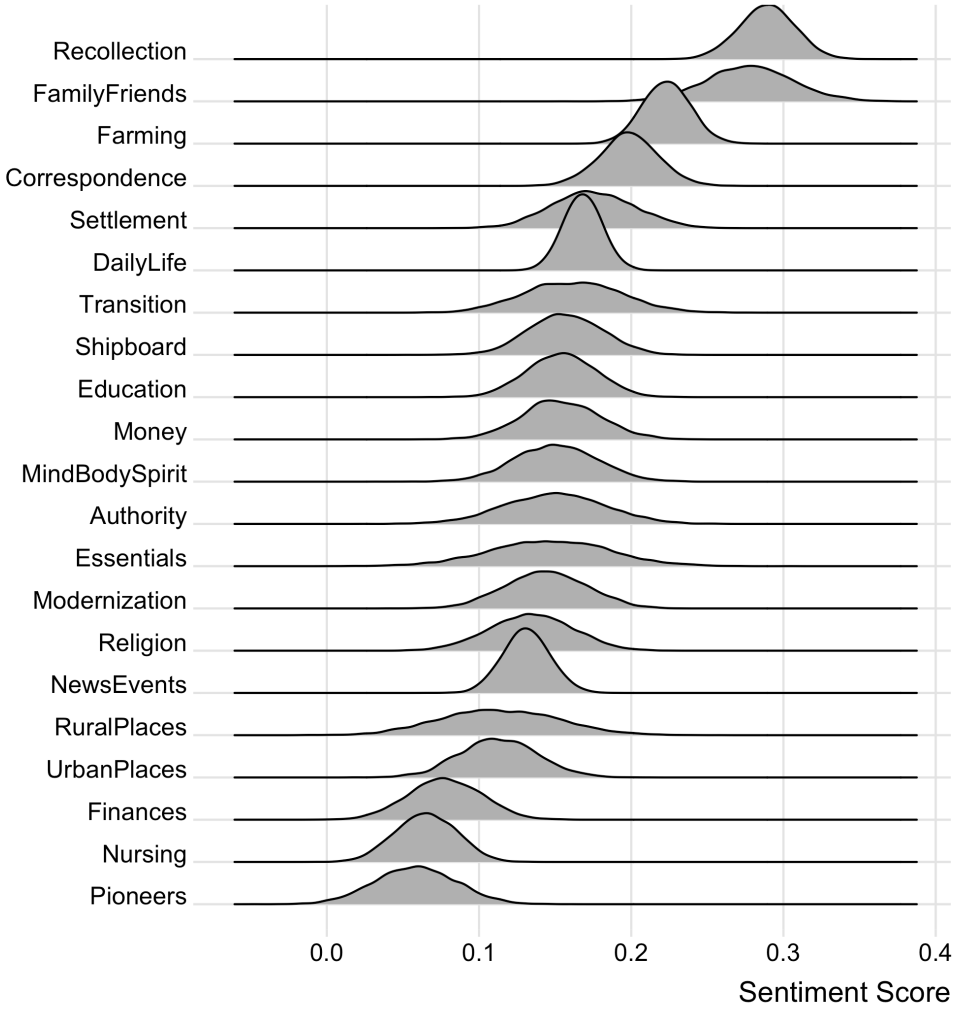


Figure 34

Gender Differences in Estimated Probability of Correspondence Cessation Relative to Sentiment for the Topic "Daily Life"

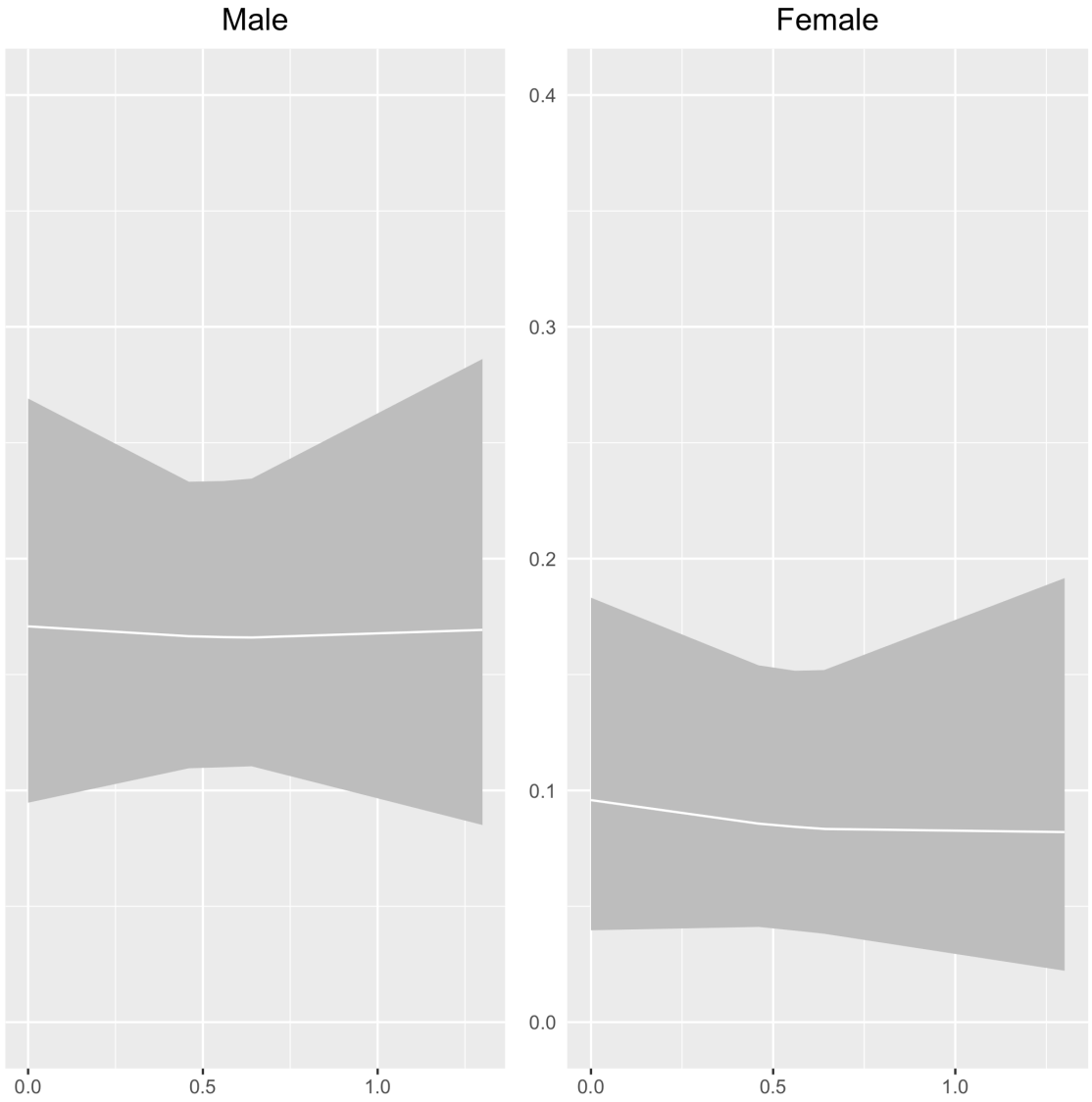


Figure 35

Predicted Probability of Correspondence Cessation for Each Writer Conditioned on Sentiment and Gender

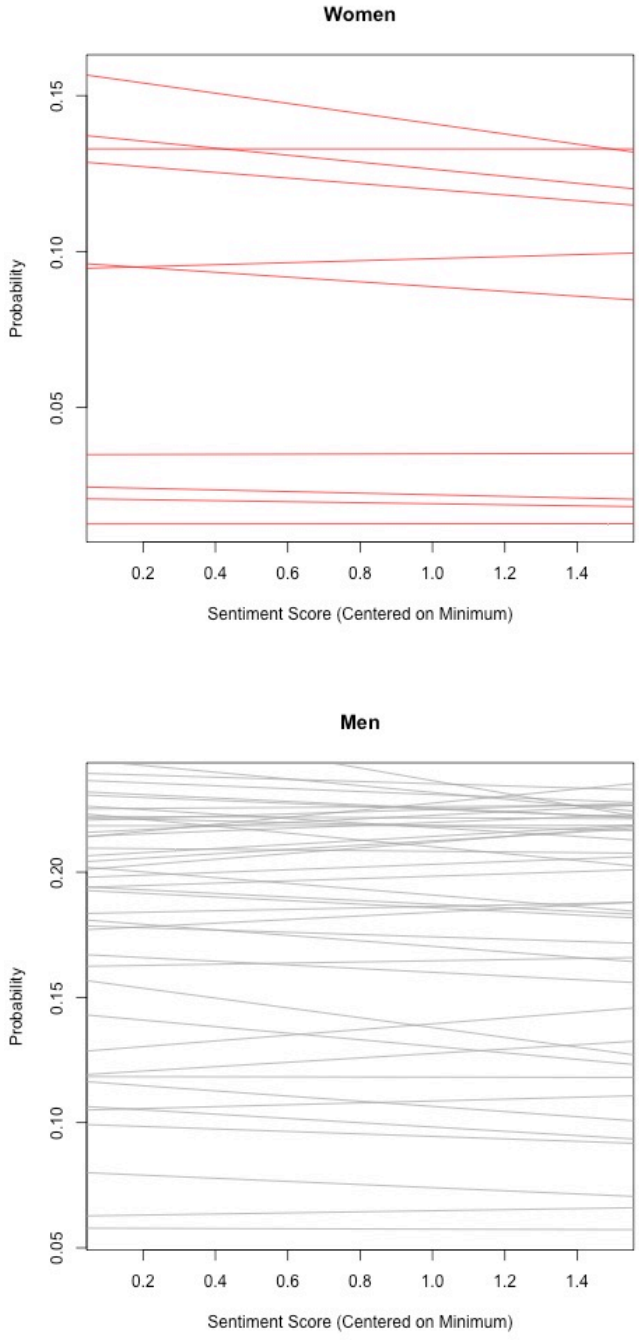


Figure 36

Fixed and Random Effects in Multilevel Model of Sentiment and Gender